Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses

Editors:

MANUEL CÁNOVAS

Dept. Biochemistry and Molecular Biology
Faculty of Chemistry
University of Murcia, Spain

JOSÉ L. IBORRA

Dept. Biochemistry and Molecular Biology
Faculty of Chemistry
University of Murcia, Spain

ARTURO MANJÓN

Dept. Biochemistry and Molecular Biology
Faculty of Chemistry
University of Murcia, Spain

Fundación CajaMurcia
Murcia, Spain, 2006

Understanding and Exploiting Systems Biology in Biomedicine and Bioprocesses.

ISBN 84-611-1135-4

D. L. MU-1036-2006

PREFACE

This volume contains contributions presented at the 1st International Symposium on Systems Biology, called "From genomes to *In silico* and back" that took place at Murcia (Spain) during the first two days of June, 2006.

The major objective of this Symposium was to point out the importance of Systems Biology in describing a biological system. Biological systems consist of a large number of heterogeneous components interacting selectively with other components in the system. These components must be connected in a proper way, so that an entire system can be functional. Precise molecular models are required to represent and understand biological systems, opening a broad field of applications. Thus, the exchange of information and experience in research and close communication between international participants will help identify future needs and new aspects in the use of this new scientific field. An intense feedback between fundamentals and applications of bioprocess and biomedicine research was stimulated.

Topics

- Fundamentals and tools of Systems Biology.
- Systems Biology applications in bioprocesses.
- Systems Biology applications in biomedicine

Special acknowledge is to be paid to Fundación CajaMurcia, for their logistic and economic support, both to the Symposium and to the edition of this volume. Additional acknowledgements are given to the Spanish Ministry of Science and Education, to the Fundación Genoma España and to the Fundación Séneca for their economic support to the Symposium, as well as to the European Federation of Biotechnology and the Spanish Society of Biotechnology (SEBIOT) for their scientific support.

Finally, we would like to thank the authors for their cooperation in the prompt and carefull preparations of their manuscripts.

The Editors:

Manuel Cánovas
José L. Iborra
Arturo Manjón

CONTRIBUTORS

Ahsan, Zaid - Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110067.

Alonso, Antonio A. - Process Engineering Group. IIM-CSIC (Spanish Council for Scientific Research), Vigo, Spain.

Arense, P. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Argüelles, J.C. - Area de Microbiología. Facultad de Biología. Universidad de Murcia. E-30071 Murcia. Spain.

Bachmann, J. - Systems Biology of Signal Transduction. German Cancer Research Center (DKFZ) Heidelberg, Germany.

Balsa-Canto, E. - Process Engineering Group. IIM-CSIC (Spanish Council for Scientific Research), Vigo, Spain.

Banga, Julio R. - Process Engineering Group. IIM-CSIC (Spanish Council for Scientific Research), Vigo, Spain.

Bernal, C. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Bernal, V. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Bond, D.R - Department of Microbiology, University of Massachusetts, Amherst, MA, US.

Buceta, J. - Parc Científic de Barcelona, Centre de Recerca en Química Teòrica (CeRQT), Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain.

Butler, J. E. - Department of Microbiology, University of Massachusetts, Amherst, MA, US.

Canela-Xandri, O. - Parc Científic de Barcelona, Centre de Recerca en Química Teòrica (CeRQT), Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain.

Cánovas, M. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Carbajosa, G. - Centro Nacional de Biotecnología (CNB – CSIC) Madrid, Spain.

Cascante, Marta - Department of Biochemistry and Molecular Biology, Faculty of Chemistry, and CERQT-Parc Cientific de Barcelona (PCB), University of Barcelona, Marti i Franques 1, 08028 Barcelona, Spain.

Cases, I. - Centro Nacional de Biotecnología (CNB – CSIC) Madrid, Spain
Structural Bioinformatics Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain.

Ceron, Julian - Massachusetts General Hospital Cancer Center and Harvard Medical School, Building 149, 13th Street, Charlestown, 02129 MA, USA.

Cocaign-Bousquet, Muriel - Laboratoire Biotechnologie-Bioprocédés, UMR 5504 INSA/CNRS & UMR 792 INSA/INRA, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France.

Coppi, M. V. - Department of Microbiology, University of Massachusetts, Amherst, MA, US.

Curto, Raul - University School of Experimental Sciences and Technology (EUCET), Universitat Internacional de Catalunya, Nova Estacio s/n, 43500 Tortosa, Spain.

De Lorenzo V. - Microbial Biotechnology Department, National Center for Biotechnology-CSIC, Cantoblanco, Spain.

Edosa, Oseghale Lucky - Department Biostatistical Science, Faculty of Science Ambrose Alli, University, EKPOMA, Nigeria.

Egea, José A. - Process Engineering Group, IIM-CSIC, Vigo, Spain.

Esteve-Núñez, A. - Department of Microbiology, University of Massachusetts, Amherst, MA, US.
Centro de Astrobiología, INTA, Madrid, Spain.

García, Míriam R. - Process Engineering Group, IIM-CSIC, Vigo, Spain.

Garg, Lalit C. - Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110067.

Gayen, Kalyan - Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai-400076, India.

Golebiewski, Martin - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Gonzalez-Alcón C. - Dpto. Estadistica, Investigacion Operativa y Computacion, Universidad de La Laguna, Spain.

González-Párraga, P. - Area de Microbiología. Facultad de Biología. Universidad de Murcia. E-30071 Murcia. Spain.

Herranz, H. - ICREA and Institut de Recerca Biomèdica (IRB), Parc Científic de Barcelona, Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain.

Hormiga, J.A. - Grupo Tecnología Bioquímica y Control Metabólico, Departamento de Bioquímica y Biología Molecular. La Laguna, Spain.

Iborra, J. L. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Jaktaji, Pourahmad R. - Biology Department, The University of Shahrekord, Shahrekord 88186/34141, IRAN.

Jensen, Peter Ruhdal - Center for Microbial Biotechnology, Biocentrum-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

Kania, Renate - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Koebmann, Brian - Center for Microbial Biotechnology, Biocentrum-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

Klingmüller, U. - Systems Biology of Signal Transduction. German Cancer Research Center (DKFZ) Heidelberg, Germany.

Klipp, Edda - Max Planck Institute for Molecular Genetics, Berlin, Germany.

Krebs, Olga - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Llaneras, F. - Dept. of Systems Engineering and Control, Technical University of Valencia, Valencia, Spain.

Lloyd R. G. - Institute of Genetics, Queen's Medical Centre, The University of Nottingham, Nottingham NG7 2UH, UK.

Loubiere, Pascal - Laboratoire Biotechnologie-Bioprocédés, UMR 5504 INSA/CNRS & UMR 792 INSA/INRA, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France.

Lovley, D.R. - Department of Microbiology, University of Massachusetts, Amherst, MA, US.

Mahadevan, R. - Department of Microbiology, University of Massachusetts, Amherst, MA, US.
Genomatica, San Diego, California, US.

Manrique, Marina - Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain.

Marin-Sanguino A. - Dpto. Bioquimica y Biologia Molecular, Universidad de La Laguna, Spain.

Martínez-Esparza, M. - Area de Inmunología. Facultad de Medicina. Universidad de Murcia. E-30071 Murcia. Spain.

Martínez-Vicente, E. - Area de Microbiología. Facultad de Biología. Universidad de Murcia. E-30071 Murcia. Spain.

Masdemont, B. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Mathur, Divya - Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110067.

Milán, M. - ICREA and Institut de Recerca Biomèdica (IRB), Parc Científic de Barcelona, Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain.

Mir, Saqib - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Momodu, Omoike Maliki - Department Bioinformatics and Molecular Biostatistics, Faculty of Science, Lagos State University.

Okwudili, Okoye U. - Department Microbiology, Facultyof Science Ambrose Alli, University, EKPOMA, Nigeria.

Otero Muras, Irene - Process Engineering Group, Spanish Council for Scientific Research, IIM-CSIC. Spain.

Pareja, Eduardo - Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain.

Pareja-Tobes, Pablo - Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain.

Pareja-Tobes Tobes, Eduardo - Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain.

Pazos F. - Protein Design Group, National Center for Biotechnology-CSIC, Cantoblanco, Spain.

Pedreño, Y. - Area de Microbiología. Facultad de Biología. Universidad de Murcia. E-30071 Murcia. Spain.

Pfeifer, A. C. - Systems Biology of Signal Transduction. German Cancer Research Center (DKFZ) Heidelberg, Germany.

Picó, J. - Dept. of Systems Engineering and Control, Technical University of Valencia, Valencia, Spain.

Poyatos, Juan F. - Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain.

Raynaud, Sandy - Laboratoire Biotechnologie-Bioprocédés, UMR 5504 INSA/CNRS & UMR 792 INSA/INRA, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France.

Redon, Emma - Laboratoire Biotechnologie-Bioprocédés, UMR 5504 INSA/CNRS & UMR 792 INSA/INRA, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France.

Reigada, R. - Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.

Rodríguez-Fernández, María - Process Engineering Group. IIM-CSIC (Spanish Council for Scientific Research), Vigo, Spain.

Rojas, Isabel - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Ros, J.M. - Departamento de Tecnología de los Alimentos, Nutrición y Bromatología. Facultad de Veterinaria. Universidad de Murcia. E-30071 Murcia. Spain.

Sagués, F. - Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.

Schaber, Jörg - Max Planck Institute for Molecular Genetics, Berlin, Germany.

Sevilla, A. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Solem, Christian - Center for Microbial Biotechnology, Biocentrum-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

Surovtsova, Irina - EML Research gGmbH, Bioinformatics and Computational Biochemistry, Heidelberg, Germany.

Teruel, R. - Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Tiwari, Madhulika - Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110067.

Tobes, Raquel - Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain.

Torres Darias, Néstor.V. - Grupo Tecnología Bioquímica y Control Metabólico, Departamento de Bioquímica y Biología Molecular. Facultad de Biología, Universidad de La Laguna, 38206 La Laguna, Spain.
Instituto Canario de Investigación del Cáncer (ICIC), Islas Canarias, Spain.

Trigo, A. - Centro Nacional de Biotecnología (CNB – CSIC) Madrid, Spain
Structural Bioinformatics Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain.

Valencia, A. - Centro Nacional de Biotecnología (CNB – CSIC) Madrid, Spain
Structural Bioinformatics Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain.

Venkatesh, K. V. - Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai-400076, India.

School of BioSciences & Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai-400076, India.

Vera-González, Julio - Systems Biology and Bioinformatics Group, Department of Computer Science, University of Rostock, Albert Einstein Str. 21, 18051 Rostock, Germany.

Grupo Tecnología Bioquímica, Departamento de Bioquímica y Biología Molecular, Facultad de Biología, Universidad de La Laguna, 38206 La Laguna, Tenerife, Islas Canarias, Spain.

Vilas, Carlos - Process Engineering Group, IIM-CSIC, Vigo, Spain.

Voit E.O. - Dept. of Biomedical Engineering, Georgia Institute of Technology, USA.

Weidemann, Andreas - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Wittig, Ulrike - Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany.

Wolkenhauer, O. - Systems Biology and Bioinformatics Group, Department of Computer Science. University of Rostock. Rostock, Germany.

Zobeley, Jürgen - EML Research gGmbH, Bioinformatics and Computational Biochemistry, Heidelberg, Germany.

CONTENTS

SYSTEMS BIOLOGY: FUNDAMENTALS AND TOOLS

A Curated Database for Reaction Kinetics

Renate Kania, Ulrike Wittig, Martin Golebiewski, Olga Krebs, Andreas Weidemann, Saqib Mir and Isabel Rojas

Scientific Databases and Visualization Group, EML Research gGmbH, Heidelberg, Germany. e-mail: Renate.Kania@eml-r.villa-bosch.de

Keywords: Reaction kinetics, Database, Systems Biology, Enzyme kinetics.

1. Abstract

Simulations of complex biochemical reaction networks require reliable kinetic data. In order to facilitate the search and retrieval of kinetic data we have developed SABIO-RK (**S**ystem for the **A**nalysis of **Bio**chemical Pathways - **R**eaction **K**inetics), a database with information about biochemical reactions and their kinetics. The data is manually extracted from literature and verified by curators concerning standards, formats and controlled vocabularies.
SABIO-RK does not only contain and merge information about reactions such as reactants and effectors (activators or inhibitors), details about the catalyzing enzyme, organism, tissue and cellular location, but also the reaction kinetics are included. The type of the kinetics, modes of inhibition or activation and corresponding equations are shown with their parameters, measured values and experimental conditions. Links to other databases like Swiss-Prot and PubMed enable the user to gather further information about proteins corresponding to the enzymes, and to refer to the original publication, respectively.
Users can query the database by specifying reactions, enzymes, organisms, locations and experimental conditions. Kinetic data of the selected reactions can be exported in SBML (Systems Biology Mark-up Language) format, allowing the use of the data as the basis for the definition of biochemical network models.
Availability: http://sabio.villa-bosch.de/SABIORK/

2. Introduction

In order to understand the networks of biochemical reactions in a living cell, scientists are trying to combine experimental data with theoretical methods. Mathematical models for the simulation of biochemical networks are being developed. These models require information about the kinetics of each of the reactions participating in the network, such as the kinetic laws describing the dynamics of the reactions with their respective parameters determined under certain experimental conditions.
Access to reliable data about reaction kinetics is thus crucial for the development of computer-based models of biochemical pathways. There are a couple of

databases containing relevant information as summarized in the following. BRENDA (Schomburg et al, 2004) is a comprehensive database on information about enzymes extracted from literature. The enzyme entries also contain information about the reactions catalysed by the enzyme and in some cases information about the mechanism associated with the reaction's kinetics. In most cases the parameters associated to the combination enzyme-reaction kinetics are also given. Swiss-Prot (Boeckmann et al, 2003) started to include experimental data like pH- and temperature dependence and kinetic parameters as comments related to biophysicochemical properties. The BioModels database (Le Novere et al, 2006) stores published mathematical models of biological interest that are annotated and linked to relevant data resources (e.g. publications or databases). The models include kinetic law equations and their parameters represented in SBML (Systems Biology Mark-up Language) format (Hucka et al. 2003) and can be used for simulations of biochemical reactions or networks. However, none of these databases links experimental kinetic parameter data for single reactions to complete sets of information comprising the corresponding rate equations, environmental conditions and concentrations of reactants and modifiers used for the determination, independent of a simulation model.

In order to assure the comparability, kinetic parameters need to be standardised and related to kinetic mechanisms, rate equations and environmental conditions. As kinetic constants highly depend on environmental conditions, they only can be specified completely by describing these conditions used for determination. Data sets based on experiments that are assayed under similar experimental conditions should be associated to each other to facilitate the comparison.

SABIO-RK (**S**ystem for the **A**nalysis of **Bio**chemical Pathways - **R**eaction **K**inetics) is designed to merge and structure all these data to support researchers interested in information about biochemical reactions and their kinetics. The system allows the creation of complex queries for reactions, kinetic laws or kinetic parameters, based on their characteristics, e.g.: reaction participants (substrate, product, enzyme), environmental conditions (pH, temperature), locations, biological sources (cell type, tissue, organism) or pathways to which the reaction can be related to. Beside the search facilities, SABIO-RK enables the user to export the chosen kinetic data in a standard format such as SBML (Hucka et al, 2003).

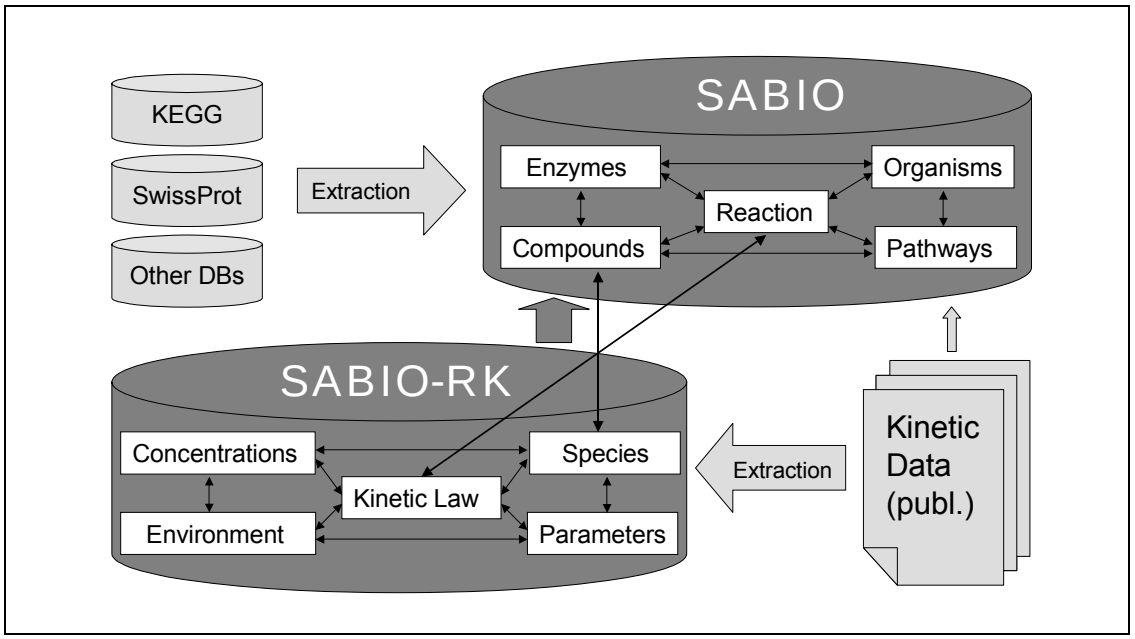


Figure 1. Population, content and schematic relation of SABIO and SABIO-RK. SABIO contains general information about biochemical pathways and reactions in different organisms, including details about corresponding enzymes and reactants. Most of these data are collected from other databases like KEGG or Swiss-Prot. SABIO-RK extends SABIO by storing information about the reaction's kinetic properties, such as the kinetic laws with their corresponding parameters and environmental conditions under which they were determined.

3. The Database

SABIO-RK represents an extension of the SABIO (**S**ystem for the **A**nalysis of **Bio**chemical Pathways) biochemical pathway database, also developed at EML Research (Rojas et al, 2002). SABIO stores all fundamental information about biochemical pathways, like reactions and their participants (enzymes, compounds, etc.). SABIO-RK combines the general data about biochemical reactions stored in SABIO with information about their kinetic properties (Fig.1). A kinetic law is associated with a relation between a biochemical reaction (defined by its substrates, products and effectors) and a catalysing enzyme (typically defined by an Enzyme Classification number and a description of the enzyme variant, e.g. isoenzyme or mutant). This relation is originally defined within a publication based on experimental conditions under which the reaction kinetics where determined (e.g. pH, organism, temperature, etc.). Thus, a reaction can have multiple kinetic laws in the database, dependent on the environmental and experimental conditions, enzyme variants, and the absence or presence of effectors. The highly structured and strongly linked nature of the data stored in the database, facilitates the definition of complex queries, consisting of different search criteria.

3.1. Database population

For the population of SABIO-RK, we combine information about biochemical reactions automatically collected from other database resources with manually entered and curated data. In order to establish a broad information basis for the database, compounds, reactions, their associations with biochemical pathways and their enzymatic classifications are regularly downloaded from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa et al, 2006). Also extracted from KEGG is the information on which reactions occur in which organisms, based on the annotation of the enzyme proteins. This is being done in a progressive manner, determined by the organisms for which we have kinetic data.

The information about the kinetics of the biochemical reactions is extracted manually from literature. It is often the case that a reaction, reported in a paper revised, is not part of KEGG, or sometimes not even the participating compounds are present. This in turn requires the definition of new compounds and reactions within the SABIO database. Determining whether a reaction or a compound is already included in SABIO, is not a trivial issue, given that the search by name may not suffice to determine synonymic expressions. To support the curators, linguistic methods are developed to obtain compound structures from names and compare compounds at the level of their chemical structure (see chapter 3.2).

Students with biochemical background carry out the extraction of the information from text. We developed a web-based input interface to support them in entering the data in a consistently organized structure. The interface is also used by the curators to check, supplement and revise the entries and supports them in the administrational work (assignment of papers etc.). The publications to be revised have been obtained from PubMed (PubMed), by using several queries leading to papers, which very likely will contain information about biochemical reaction kinetics. Ideally students extract the following information for each reaction reported within a publication:

- Reaction defined by substrates and products
- Modifiers of the reaction (activators, inhibitors, catalysts, cofactors)
- Cellular location of compounds
- Enzyme classification number
- SwissProt accession number (of the enzyme)
- Variant of the enzyme (wild type or a certain isoenzyme or mutant)
- Kinetic law type (e.g. Michaelis-Menten, Ping Pong Bi Bi)
- Kinetic law formula
- Kinetic parameters (e.g. Km, kcat, Vmax)
- Concentrations used for reactants, enzymes and modifiers
- Experimental conditions (e.g. temperature, pH, buffer composition)
- Biological source (e.g. cell type, tissue, organism, strain)
- Information source (reference)

For most of this information, comment lines are available to add information, for example about synthetic or labelled derivatives of physiological compounds. The described parameter values are entered with their standard deviations.
The students' work is supported by offering search facilities to look for reactions and lists for compound names, locations, organisms, tissues/cell types, kinetic law types and parameter units already existing in the SABIO or SABIO-RK database. This helps to avoid redundancies just because of aberrant notations or typing errors. Beside that, it is possible to include new entries, if the reported data, for example, reaction or organism is not yet in the database. All entries completed by the students undergo a curation process before they are loaded into the final database SABIO-RK. The entries are checked, complemented and verified by a team of biological experts to eliminate possible errors and inconsistencies.

3.2. Curation

As standards for publishing data of biochemical reactions and reaction kinetics are still missing, the curators are faced with problems like synonymic or aberrant notations of compounds and enzymes, multiplicity of parameter units and missing information about assay procedures and experimental conditions. Missing information about the organism or experimental conditions sometimes can be found by reference searches on a limited scale. Very often alternative substrates are tested in the experiments, but the products or the stoichiometry is not given in the paper. In such cases the expert knowledge is needed.
The curation process includes the unification and standardisation of the data. Already existing standards for data formats are applied as well as new standards are defined if necessary. For example, the unification of parameter units or chemical compound names involves existing standards as the International System of Units (SI) for unit notation or the nomenclature recommendations for chemical compounds of the International Union of Pure and Applied Chemistry (IUPAC). In contrast to, for enzyme specifications (mutants, isoforms, etc.) database-internal norms are assigned additionally to the enzyme classification system of the International Union of Biochemistry and Molecular Biology (IUBMB). Already existing controlled vocabularies are used for the representation of organisms, tissues, cellular locations etc. There are cases were information is still stored as free text. For example a buffer description can be very complex containing information about coupled enzyme assays. Therefore, information about the buffer composition currently is stored as a free text. A comment line belonging to the entire data set, contain information about host organisms in which proteins are expressed (e.g. recombinant enzymes expressed in Escherichia coli) and additional information about proteins e.g. their composition of subunits.
As mentioned above, many compounds are known by a variety of names and scientists very often use trivial names instead of the IUPAC nomenclature, which's make the curators work cumbersome. If a compound name written in the paper can not be found in the provided compound list, the curators have to check, if the "new" compound is already stored in the database with another name, or if

it is necessary to insert a new compound into the database. The following workflow includes exhaustive searches in other databases storing information about chemical compounds like ChEBI (**Ch**emical **E**ntities of **B**iological **I**nterest) (ChEBI) or PubChem (PubChem). A new compound is included in the database by using the names and characteristics according to the paper and supplemented by information from these databases. A tool for the linguistic analysis of the names of organic compounds has been developed, named CHEMorph (Anstein et al, to appear). CHEMorph analyses systematic and semi-systematic names, class terms, and also otherwise underspecified names, by using a morpho-syntactic grammar developed in accordance with IUPAC nomenclature. It yields an intermediate semantic representation of a compound, which describes the information encoded in a name. The tool provides SMILES strings (Weininger, 1988) for the mapping of names to their molecular structure and also classifies the terms analysed.
The curators' work is also supported by some automatic routines to check the consistency of the entered data. For example, when a kinetic law formula is entered, it is verified for the correct mathematical format. Moreover, the list of parameters is checked, if each of the parameters contained in the kinetic law formula is defined. The parameters are assigned to a parameter type (e.g. Km, kcat, concentration), which allows an extension of database searches on parameter types. All parameter values are specified as scalable SI units compatible with the unit requirements of the SBML specification and making the data comparable. This procedure includes the standardisation of different notations for one and the same unit, as mM and mmol/l, as well as the conversion of comparable units, like µmol/min (International Standard Unit) and katal (mol/s) for enzymatic activities.

4. Current Search Facilities

The current version of the SABIO-RK web interface allows users to perform searches for reactions by specifying characteristics (one or many) of the reactions of interest (Fig2). For example the user can specify the pathway to which the reactions searched should belong to, e.g. Glycolysis; or can specify more characteristics to obtain, for example:

- all reactions of the Glycolysis pathway in Yeast.
- all reactions in human liver which use D-Glucose as substrate at pH >7.0

Figure 2. SABIO-RK web interface to specify search criteria. Searches for reactions can be defined by specifying one or more reaction participants (reactants or enzymes), pathways, biological sources (organisms, tissues, cell types), cellular locations, environmental conditions, or publications.

The system will return all reactions and related enzymes satisfying the given search criteria (Fig3). The user can specify whether all reactions should be shown or only those for which kinetic data are available. By clicking on a reaction more information about it will be shown. For each reaction the result screen gives information whether or not there is kinetic data available in the database, using a three color-code to indicate this. Green means that for the associated reaction there are kinetic data available matching all search criteria, in the second example above this would mean that there is kinetic data reported on the respective reaction in human liver and measured at pH>7.0. Yellow means there are kinetic data available, but not matching all search criteria, for the same example this would mean that there is kinetics data available but for example not in liver but in heart, or not in human but in mouse. Red indicates that there are no kinetic data available for the reaction reported. Considering the cases where there are no kinetic data, or at least no kinetic data for the exact matches (no green entries for a given reaction) we found important to offer information about the availability of kinetic data for other reactions catalyzed by the same enzyme. The

availability of kinetic data for the enzyme is reported by another clickable box (using the same three color-code) beside the enzyme's classification number.

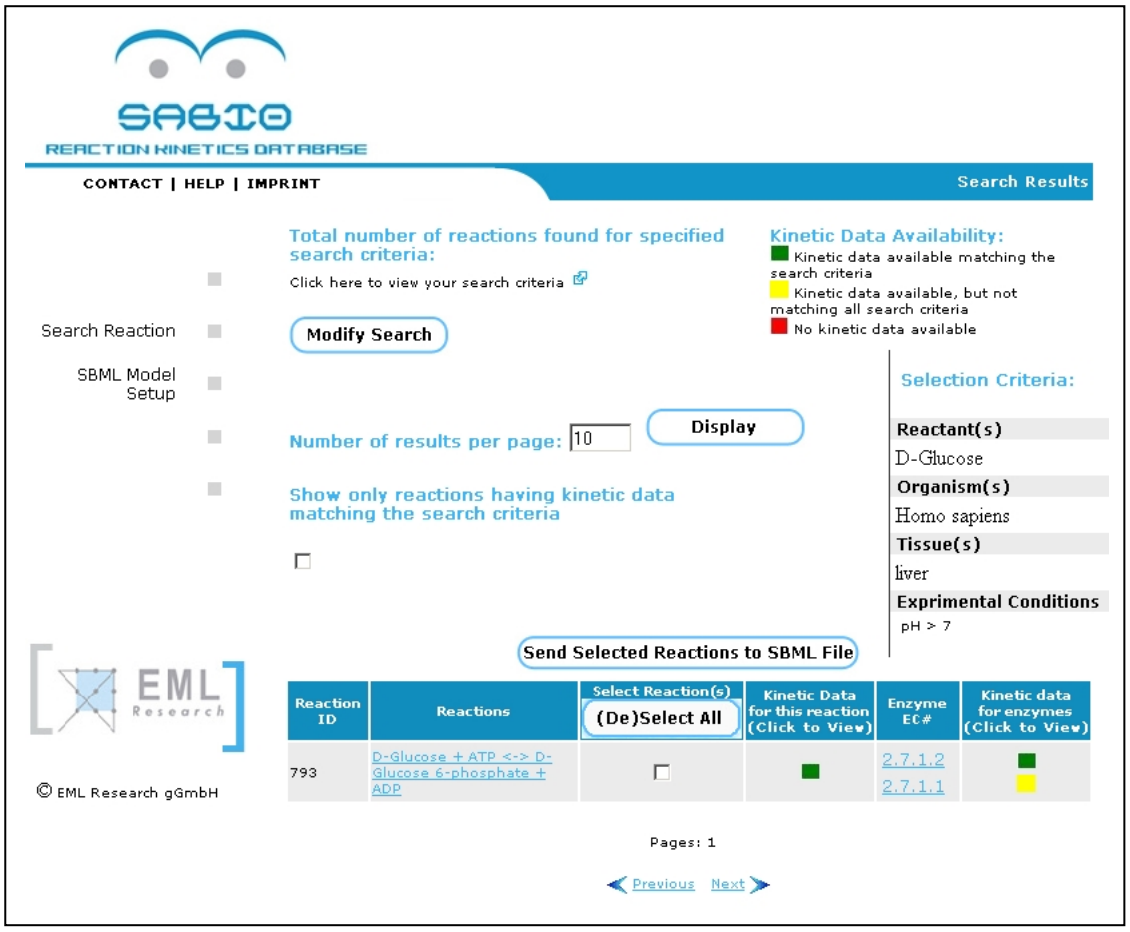


Figure 3. SABIO-RK query result page. The query result is represented as a list of reactions and related enzymes satisfying the search criteria. Information about the availability of kinetic data is shown using a three color code.

The user can then either view the kinetic data belonging to the specified reaction, or all kinetic data available for the enzymes catalyzing this reaction. In a new window the entries containing kinetic data are listed according to the selection done. To get a general idea, in the overview only organism, tissue, enzyme classification and the variant of the enzyme are shown. The expanded version then shows all the kinetic data and additional information extracted from a publication. Also the information source of each database entry is clearly shown and linked to the PubMed database in order to allow the user to refer to the original paper to obtain additional information about the experiment described (Fig 4).

Entry Nr. 2580 [⊞][⊟] Select

Organism:	Homo sapiens
Tissue:	liver
EC Class: 2.7.1.2	**Variant:** wildtype

Reversability: reversible

Substrates

name	location	comment
ATP	unknown	-
D-Glucose	unknown	-

Products

name	location	comment
ADP	unknown	-
D-Glucose 6-phosphate	unknown	-

Modifiers

name	location	effect	comment
Glucokinase	unknown	Modifier-Catalyst	-

Kinetic Law

(Vmax*S^n) / (S_h^n+S^n)

Kinetic Law Type: Hill Cooperativity

Parameters

name	species	type	St_value	Deviation	End_value	unit	comment
A	ATP	concentration	1.0	-		mM	-
S	D-Glucose	concentration	0.0	-	100.0	mM	-
Vmax		Vmax	12.3	0.45		mU/ml	-
n		Hill	1.6	0.03		-	-
S_h	D-Glucose	Km	5.7	0.08		mM	-
E	Enzyme	concentration	60.0	-		nM	-

Experimental conditions

	St_value	end_value	unit
pH	7.1	-	
Temperature	22	-	°C

Buffer: 25 mM Hepes, 25 mM KCl, 5 mM mercaptoethanol, 1 mM NADP, 2.5 mM MgCl2, 2 units/ml glucose 6-phosphate dehydrogenase

Comment: expressed in E. coli

PUBMEDID: 14988235

Figure 4. SABIO-RK database entry. An example data set represents a specified reaction including kinetic data, experimental conditions and additional information extracted from a publication.

The user can create a SBML file using the selected data. A reaction can be included into the SBML file without any kinetic data, i.e. only its stoichiometry or the user can select data from an entry (displaying kinetic data) for its inclusion into the SBML file. Due to the limitations of the SBML file format, the data exported is limited and in some cases simplifications are needed. For example no information about the experimental conditions, under which the parameters were determined, can be exported, although we plan to incorporate this information as annotations. Because parameter values can only be single values, no ranges, we include as parameter value the mean of the parameter range (if given). The SBML file lists all the compounds (in SBML named species) belonging to the

reactions. If a compound is present in more than one reaction, it will only be defined once in the file and will be referred to in the corresponding reactions. The SBML file is generated using the LibSBML library (Hucka et al, 2003).

5. Summary

SABIO-RK is a web-accessible database containing curated data about biochemical reactions and their kinetics. The system merges general information about reactions, mainly retrieved from other databases, with kinetic data extracted from publications. The kinetic data comprises information such as the type of kinetics, modes of inhibition or activation and kinetic laws with their parameters, providing, whenever available, information about the experimental conditions under which this information was determined. The user can search for reactions (with their kinetic data) based on their characteristics, such as pathways in which they participate, catalysing enzyme, reactants, etc., plus the characteristics of the experimental conditions from the assays used to determine the reactions' kinetics. To support modellers, selected kinetic data can be exported in SBML format to build models for the simulation of complex biochemical processes.

6. Future Perspectives

There are still many features that we plan to add to the database, both at the level of the search capabilities as well as at the level of the database content. We are working on a database schema that enables the insertion of detailed descriptions of the kinetic reaction mechanism. This should allow us to represent kinetic properties of elementary reaction steps or binding events in the database.
Additional database search functions will be less reaction oriented, offering, (for example) capabilities to search for kinetic parameters and law types, e.g. search for all enzymes of the pathway glycolysis for which Km values are known. Users will also be able to search for reaction networks or paths of reactions between two defined compounds or enzymes.
Although we already have annotated our data to some other database resources (like PubMed or Swiss-Prot), we will extend the annotation of the database content. This includes the usage of controlled vocabularies and annotation to the corresponding ontologies, like the Open Biomedical Ontologies (OBO) and the Systems Biology Ontology (SBO).
Data export functions will be expanded, since a lot of the information stored in SABIO-RK cannot yet be formally described in SBML, for example the environmental conditions like pH or temperature. At the level of the SBML file we will also add more annotations to the different entries, such as KEGG and ChEBI identifiers to the species, as well as SBO identifiers to the kinetic laws.
A very important aim for us is to convince wet-lab scientists to use the input interface to enter data directly into the database. Thus, all the needed information

can be given by the experimenters and no information is lost. In doing so, users would be able to directly compare their own experimental results in SABIO-RK with kinetic data extracted from literature.
Further development and implementation of tools for information extraction and retrieval, and supporting data curation is also planned in order to accelerate the database population process. However, due to the form how the data extracted is presented in the literature, i.e. scattered in the text, in tables, in figures, and in formulas, we will still require much manual processing.

Acknowledgements. *The project is funded by the Klaus Tschira Foundation and partially by the German Research Council (BMBF). We would also like to thank the members of the Bioinformatics and Computational Biochemistry and the Molecular and Cellular Modelling Groups of EML Research for their helpful discussions and comments. Last but not least, we thank all the student helpers, who have contributed to the population of the database.*

References

Anstein S, Kremer G, Reyle U (2006) Identifying and Classifying Terms in the Life Sciences: The Case of Chemical Terminology. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. To appear

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003) The Swiss-Prot Protein Knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-70.

ChEBI (Chemical Entities of Biological Interest); Available from: http://www.ebi.ac.uk/chebi/

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524-31.

SI (International System of Units); Available from: http://www.bipm.fr/en/si/

IUBMB (International Union of Biochemistry and Molecular Biology); Available from: http://www.chem.qmul.ac.uk/iubmb/enzyme/

IUPAC (International Union of Pure and Applied Chemistry); Available from: http://www.chem.qmul.ac.uk/iupac/

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354-7.

Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*, **34**, D689-91.

OBO (Open Biomedical Ontologies); Available from: http://obo.sourceforge.net/

PubChem; Available from: http://pubchem.ncbi.nlm.nih.gov/

PubMed; Available from: http://www.pubmed.gov

Rojas I, Bernardi L, Ratsch E, Kania R, Wittig U, Saric J (2002) A database system for the analysis of biochemical pathways. *In Silico Biol* **2**,0007.

Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, **32**, D431-3.

SBO (Systems Biology Ontology); Available from: http://www.ebi.ac.uk/compneur-srv/sbo/

Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, **28**, 31-36.

Modelling of Signal Transduction in Yeast – Sensitivity and Model Analysis

Edda Klipp and Jörg Schaber

Max Planck Institute for Molecular Genetics, Berlin, Germany
{klipp,schaber}@molgen.mpg.de

Keywords: Signal transduction, mathematical modelling, sensitivity analysis, parameter dependence.

1. Abstract

Experimental research has revealed components and mechanisms of cellular stress sensing and adaptation. In addition, mathematical modelling has proven to foster the understanding of some basic principles of signal transduction and signal processing as well as of sensitivity and robustness of information perception and cellular response. Here we review some modelling principles, results and open questions exemplified for a model organism, the yeast *Saccharomyces cerevisiae*.

2. Introduction

During their life span, cells face a multitude of stresses and changes in the environment. Most of those changes are normal processes that can happen more or less frequently, like temperature changes, variation in nutrient supply or appearance of a mating partner. Therefore, species had to adapt to such types of stress during evolution and to develop appropriate, specific and efficient mechanisms to cope with such typical demands.
In the last few years, a series of modelling approaches has been used and adopted to support the understanding of the complex behaviour of signalling networks. The concepts range from very abstract models that elucidate some key properties of signalling pathways (e.g. Heinrich et al., 2002, Papin and Palsson, 2004) to very detailed models that precisely monitor the dynamics of specific regulatory events (e.g. Vaseghi et al., 2001, Schoeberl et al., 2002, Yi et al., 2003, Swameye et al., 2003). Systematic overview on structural properties and dynamic features of signalling pathway models are given in (Papin et al., 2005, Tyson et al., 2003).
The complexity of biochemical networks is far from being resolved experimentally. Nevertheless there is need to understand their behaviour in a rational way, which is often hard to achieve by intuition. Establishing models of such networks supports the integration of experimental knowledge into a consistent picture, the formulation of hypotheses and cognitions in a precise language. It serves to test, support, or falsify hypotheses about the underlying

Understanding and Exploiting Systems Biology

biological mechanism. Modelling may integrate different parts of the whole and thereby allow analysis of properties that only emerge upon the interaction of elements in a comprehensive network. A sound model can produce predictions that can be experimentally tested and it can simulate processes that are experimentally hidden.

3. Modelling: Mathematical Techniques and Tools

3.1. Purpose of Modelling

The development of a model serves the abstract and condensed representation of facts in order to allow for the analysis of their relations and to gain understanding about their internal organization and their communication with the environment.
Although the number of data in biological research currently explodes, such data is useless without sufficient interpretation. A computational model can on hand serve the data interpretation; on the other hand it can point to biological aspects that are still not sufficiently experimentally resolved. Within the field of Systems Biology, the view has been established that experimental research and model development should go hand in hand in an iterative manner including formulation of an initial model, hypothesis generation, experimental testing of hypotheses, model-based experimental design, model refinement upon new data, and so on.
The iterative modeling and experimentation process is hard to follow in publications, since they often only represent the final results. Model improvement with time and with accumulating experimental information is documented e.g. for yeast cell cycle (Novak et al., 1999, Chen et al., 2000, Chen et al., 2004 and others) and for signaling pathways (Bhalla, 2004, Bhalla, 2002, Bhalla and Iyengar, 2001, Bhalla and Iyengar, 1999).

3.2. Model Development

Usually, an experimental observation inspires the formulation of a hypothesis as a first step. In the second step we define what questions the model is supposed to answer, i.e. the *scope* of the model. The scope determines what components and processes the model will take into account or omit and it defines the system's boundaries. Omitting certain processes from the models even though they might play a role is based on the assumption that they have only a minor influence on the event under study, that their values remain constant in the experimental setup, or that they simply cannot be described with the currently available means. For example, the effect of regulated gene expression is usually neglected in the modelling of metabolic networks although modellers are certainly aware of production and degradation of enzymes. But the different time scales of protein turnover and metabolic reactions justify this simplification in many cases. The initial model is usually formulated as a word model. The word model itself is also subjected to a process of refinement and sophistication in the course of model development. A graphical representation of the model structure, e.g. a diagram, is also helpful.

Subsequently, the word model is translated into a mathematical model (for an overview on mathematical techniques see below). To assure that our model is in principle able to answer our initial question we must *verify* whether our model can achieve this independently of choice of specific parameter values, i.e. in a qualitative way. For example, when we want to explain an observed temporal oscillation of a cellular compound, we must test whether our model is structured in a way that it is able to produce oscillations. This might not be as trivial as it sounds in some cases. For example, until now there exists no general theorem for the existence of oscillations in chemical systems with more than three compounds (Heinrich & Schuster 1996). When no mathematical theorem is available that tells us something about the general properties of our system, verification of the proposed model behaviour is generally obtained by playing around with the model structure and its parameters, checking whether it behaves in the way we want. Verification of the model structure is an important step in the process of model development because it can save much time and effort later on. When the model is not able to fit observed data, this might be a general problem of the model structure. Having checked this in advance we can avoid validating a model in vain.
Generally, it is also desirable to learn more about general properties of the model, like e.g. steady states and bifurcation points. When we analyse metabolic systems, we can apply mathematical tools like Metabolic Control Theory to analyse the system.

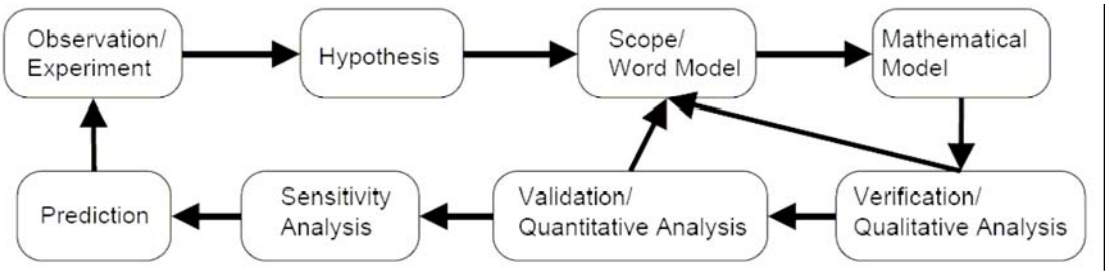


Figure 1: Model development flow chart.

Having verified that the model can principally reproduce our expectations we can now *validate* that the model can also reproduce our observations in a quantitative manner. This is generally achieved by adjusting the model parameters such that the components of the model match observational data. It is important to gain further support for our model by testing whether it is also able to reproduce independent data without changing the fitted parameters. Independent in this sense means that the data was neither used to fit the parameters nor to develop our model. We need a training data set and test data set. The test data generally describe the same phenomena but under slightly different conditions. It is a prerequisite for a sound model validation that the model is able to reproduce observed data under different conditions but with the same parameters that were used to reproduce the training data set. This is supposed to reflect the fact that our model accurately describes the intrinsic structure of the studied system and,

like nature, is able to adequately adjust its reaction to a changing environment/input without changing internal structure and interactions.
It is important to know the limits of applicability of a model. They determine to what extent possible predictions and conclusion hold. Moreover, it is important to know what parameters are sensitive, i.e. whose changes have a substantial impact on the systems behaviour, and thus have to be determined with great accuracy. To this end we must conduct a *sensitivity* analysis. Usually, this is achieved by changing one parameter value at a time and looking at the resulting change of a specific output variable. A classical measure of sensitivity is the relative sensitivity S that is defined as

$$S = \frac{\Delta O}{O} \cdot \frac{p}{\Delta p} \tag{1}$$

where $\Delta O/O$ is the relative change of some output of interest and $\Delta p/p$ is the relative parameter change, compared to the initial state of parameter, respectively. S is easy to interpret, as $S = 1$ means that a certain percentage change of a parameter yields the same percentage change of the considered output. Usually, when $|S| \leq 1$, p is considered as non-sensitive. When $|S| >> 1$, p is considered as sensitive. The range in which p is changed depends on the uncertainty with which p is determined. This can be the measurement error or some other knowledge about the range in which p can vary. With no such knowledge, it is usually a good start to change p by 50%.
Classical sensitivity analysis studies the reaction of one or more output variable to the change of one parameter at a time. Generally, it cannot be assumed that parameters have an independent influence on the considered output. In most cases the sensitivity of one parameter depends on the state of one or more other parameters. However, manipulating individual parameters can be viewed as unusual perturbation of the system by, e.g. a mutation or other kind of damage. It is reasonable to assume that under the conditions we are mostly interested in it is unlikely that many parameters change or are perturbed at the same time.
Having determined sensitive parameters gives us important information about our system. It not only tells us where small measurements errors can have drastic consequences for the system behaviour but also where additional research or measurements might be adequate. Sensitive parameters can also be interesting targets for drug developers as it makes sense to manipulate a system where it is most sensitive. Sensitivity analysis tells us something about the robustness and resilience of the system.
It is not only important to explore the sensitivity of the system to parameter changes but also to changes in the input stimuli. Biological systems are always subjected to varying environmental conditions and we must check whether our system is as flexible as we expect it to be. Moreover, a structural sensitivity analysis, i.e. not only changing parameters but also model formulas, can give valuable information what features of the model are necessary to exhibit a certain behaviour and what parts can be omitted or simplified.
The sensitivity analysis relates to and complements the two preceding steps

verification and validation. Verification tells us something about the theoretical properties of our model system, how the model could behave, i.e. the qualitative structure of the state space. Validation determines a concrete state of the system that reflects observed biological phenomena, i.e. tells us where our system is quantitatively located in the theoretical state space. Finally, sensitivity analysis provides us with a quantitative picture of the state space around our system.
We can then use the model to explore more systematically regions of the state space that are of particular interest, i.e. make predictions. The model ideally should be able to predict future experiments. When the model correctly predicts the experiments we gain confidence in the model and also in the original hypothesis. Moreover, the model can be used to design future experiments. In combination with the sensitivity analysis we can determine where additional measurements give us most information about the system.
In case, the model does not correctly predict the experiments it has to be checked whether the experiments still comply with the original hypothesis. If it does we have to modify the model, otherwise we have to modify the hypothesis. Both ways, we close the cycle.

3.3. Mathematical Description of Dynamic Processes

Depending on the available experimental information, the purpose of modelling, the experience and preference of the modeller, signalling pathways can be described with different techniques. In general, all approaches rely on a description of the network structure with a graph representing as edges the interaction (activation, inhibition, complex formation) between the nodes, i.e. the different signal molecules. Boolean networks or Petri nets describe the states of individual nodes in a discrete fashion and these states are updated along a discretised time axis according to the rules assigned to the edges. In their basic version, Boolean networks allow only for two states (1 or 0, i.e. active or not active). Petri nets assign individual tokens to the places (i.e. nodes). More sophisticated approaches tend to consider more different states and update rules.
The dynamics on a continuous time scale can be simulated in a stochastic manner, e.g. with one of Gillespie's methods (e.g. Gillespie, 1977) by assuming discrete state values, e.g molecule numbers. A frequent approach is the description with ordinary differential equations (ODEs), where the state space is continuous (concentrations or activities) and the time is continuous. In the following we will focus on the ODE model approach.
The dynamics of the biochemical reaction network is expressed by the balance equations

$$\frac{d\boldsymbol{S}(t)}{dt} = \boldsymbol{N}\boldsymbol{v}(\boldsymbol{S}(t), \boldsymbol{p}) \tag{2}$$

where $\boldsymbol{S}$, $\boldsymbol{v}$, and $\boldsymbol{p}$ denote the vectors of concentrations, reaction rates, and parameters of the system, respectively, and t is the time. The matrix $\boldsymbol{N}$ contains the stoichiometric coefficients. Typical expressions for the reaction rates are the so-called mass action rate law

$$v_i(S_j) = k_i \cdot S_j \tag{3}$$

or the Michaelis-Menten rate law

$$v_i(S_j) = \frac{V_{max} S_j}{K_M + S_j} \tag{4}$$

or the Hill kinetic

$$v_i(S_j) = \frac{V_{max} S_j^n}{K_{0,5}^n + S_j^n} \tag{5}$$

The mass action law implies a linear dependence of rate on substrate concentration, while hyperbolic Michaelis-Menten kinetics and sigmoid Hill kinetics show saturation. Note that more elaborated kinetic mechanisms are described, especially for more substrates and for reversible reactions (Cornish-Bowden, 2004).

In the cell, signalling pathways have to cross several boundaries: the cell membrane, the nuclear envelope, the mitochondrial membranes or others. This may make it necessary to include different compartments into the model. Moving between compartments has different effects in discrete or continuous settings: if one molecule leaves a compartment, then one molecule will arrive in the neighbouring compartment. If one µm of a substance leaves a compartment, the concentration change in the neighbouring department depends on their relative volumes.

3.4. Analysis of Models

The model can be analyzed in various ways, first to test whether its behaviour really reflects the aspects that we wanted to represent, second to deduce predictions based on a presumably appropriate description.

Purely based on the stoichiometry, i.e. on the wiring, is the analysis of the stoichiometric matrix ***N***. The linear dependence of rows of the stoichiometric matrix points to moiety conservation in the system, i.e. it reveals which compounds or moieties are neither produced nor degraded by the network in total, such as the sum of differently modified forms of a protein. In mathematical terms, one has to find a regular matrix ***G*** such that $G \cdot N = 0$. Then $G \cdot S = const.$ expresses the conservation relations. The linear dependence of columns of ***N*** ($N \cdot K = 0$ with regular matrix ***K***) reveals the dependence of fluxes in steady state, i.e. steady state fluxes are linear combinations of the columns of matrix ***K***. For example, in an unbranched pathway, all fluxes must be the same in case of steady state.

Flux balance analysis (FBA) is based on the relations revealed for fluxes in steady state. To elucidate operation modes of the cell under different environmental conditions or to suggest such modes for biotechnological

processes, it calculates from all possible steady state fluxes that set of fluxes that maximizes or minimizes a certain function of these fluxes, e.g. by linear programming.

Metabolic control analysis (MCA) seeks to quantify the impact of individual rates or parameters on the steady state values of variables by calculating the respective derivative. In MCA a version of the above-defined sensitivity S is often applied, the response coefficient R, that is actually nothing else than the sensitivity S of the linearised system

$$R = \frac{p}{O} \cdot \frac{\partial O}{\partial p} \quad (6)$$

The theorems of MCA (Reder, 1988) establish a relation between R, which is a property of the whole system, and the local sensitivities of the individual rates with respect to the compound concentrations and the network stoichiometry $\boldsymbol{N}$. Especially interesting for signalling pathways is the analysis of time-dependent response coefficients $^{v_i}S_j$

$$R(t) = \frac{p}{O(t)} \cdot \frac{\partial O(t)}{\partial p} \quad (7)$$

which show the impact of a parameter value on the dynamics of a compound, not only on its steady state value (Ingalls and Sauro, 2003).

4. Modelling Cell Signalling: Concept and Examples

4.1. Components of Signalling Pathways

Despite their diversity in function and design, many signalling pathways use the same essential components, which are often highly conserved through evolution and between species. For example, proteins in yeast pathways have homologs in human pathways and G proteins or MAP kinases are conserved throughout kingdoms. Here, we will introduce the most prevalent signalling pathway modules that are frequently connected in series.

Receptors receive extracellular stimuli by ligand binding and transmit a signal to intracellular signalling molecules. Many receptors are transmembrane proteins. Upon signal sensing, they change their conformation and become active (**Figure 2A**), now being able to initiate downstream processes. Cells can regulate the number and the activity of specific receptors, e.g. in order to shut off the signal transmission during sustained stimulation. An interplay of production and degradation regulates the number of receptors (for a model involving receptor internalization in the yeast pheromone pathway see (Yi et al., 2003)). Phosphorylation of serine/threonine or tyrosine residues in the cytosolic domain by protein kinases can regulate the activity and thereby adapt the signalling system to input signals of different intensity.

A possible way of signal transmission from the receptor is the binding to and the activation of G proteins. The heterotrimeric G protein consists of the subunits α,

β, and γ (**Figure 2B**). Upon activation, a GDP bound to the α-subunit is exchanged with a GTP, and the G protein dissociates into different subunits, which transmit the signal to downstream processes. As soon as the GTP is hydrolyzed to GDP, the subunits can re-associate to form the initial heterotrimeric G protein.
The change between GTP- or GDP-bound states is also characteristic for so-called small G proteins like Ras, Rho, Rab, Ran, or Arf. They have different activities in both forms (**Figure 2C**). Transformation from the GDP state to the GTP state is catalyzed by the Guanine Exchange Factor (GEF), while the reverse process is facilitated by a GTPase-activating protein (GAP), which induces hydrolysis of the bound GTP (Schmidt and Hall, 2002).

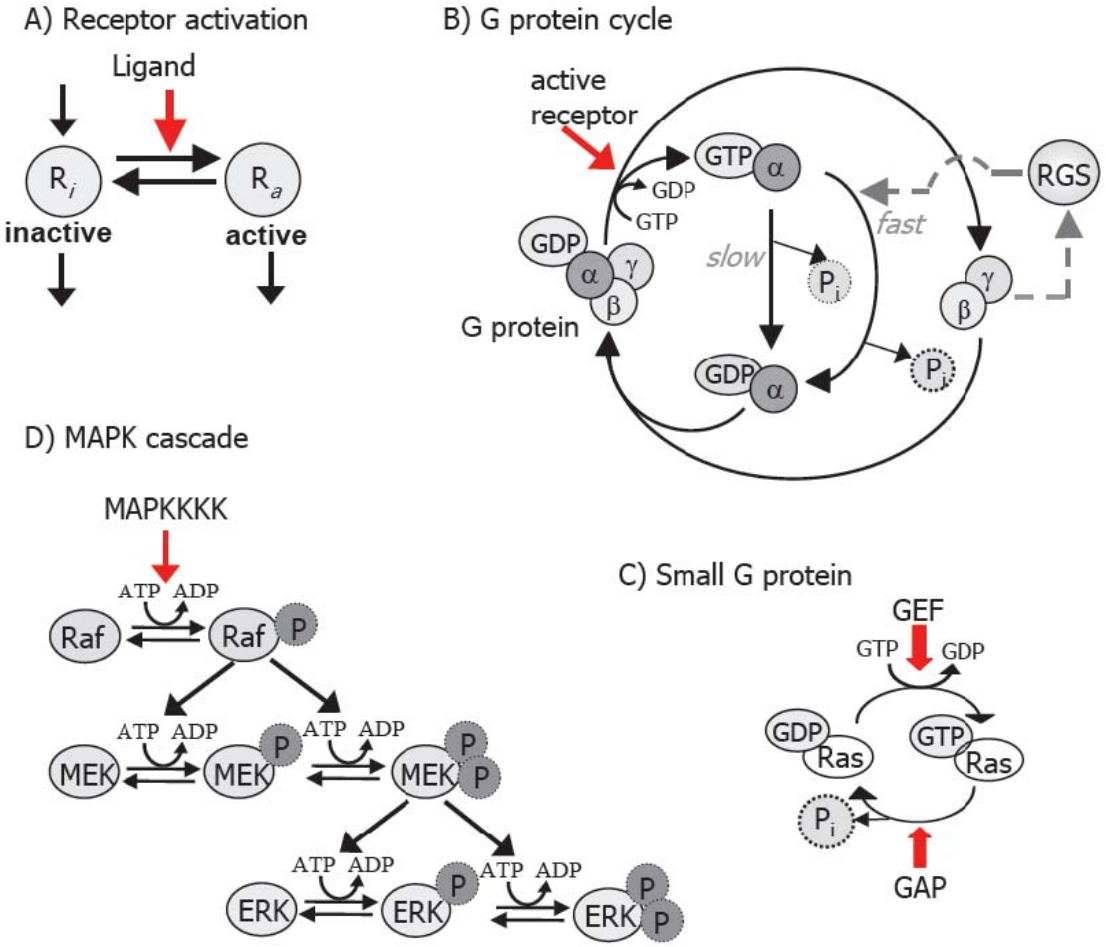


Figure 2: Building blocks of signalling pathways. A) Activation of the receptor by a ligand, B) G protein cycle including slow and fast mode; the fast mode is activated by feedback loop involving a protein (RGS), C) Small G protein switch between two states, GDP-bound and GTP-bound, D) the MAP kinase cascade involves several successive phosphorylation events.

Extracellular signal-regulated kinase (ERK) or mitogen-activated protein kinase (MAPK) cascades consist of three or four different proteins that specifically

catalyse the phosphorylation of the subsequent proteins (**Figure 2D**). According to their roles, these kinases are called MAP kinase (MAPK), MAP kinase kinase (MAPKK), and so on. The dephosphorylation is ensured by phosphatases that are often less specific, but can also be very specific to certain targets. In some cases, the MAP kinases bind to a scaffold protein forming a complex.
Several functions for such scaffold formation are discussed, such as to ensure the physical vicinity of components or their correct molecular orientation or an increase in signal amplification. Scaffolding can account for the fact that signalling pathways often appear to be decoupled although they contain common components.

4.2. Stress Response Pathways in Yeast

The response of yeast cells to external stimuli, environmental changes, nutrient supply or availability of a mating partner is ensured by a variety of signaling pathways that partly overlap by the use of common proteins (**Figure 3**).

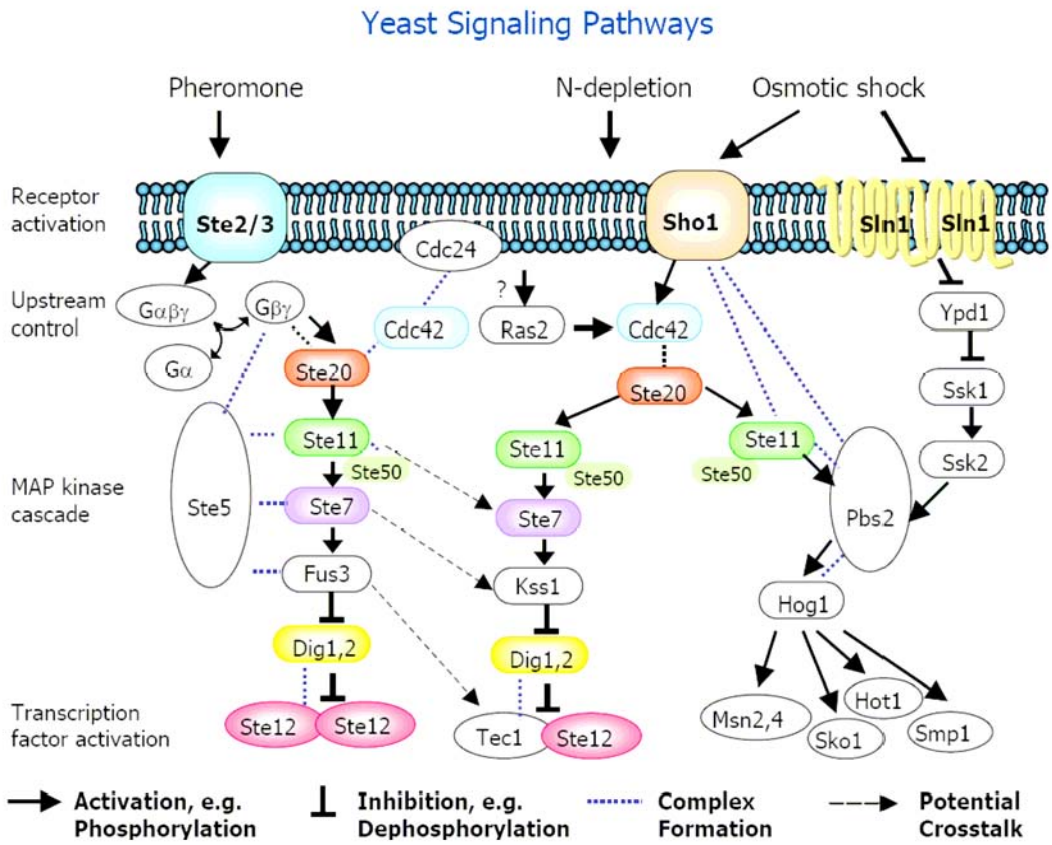


Figure 3: Selected signalling pathways of the yeast Saccharomyces cerevisiae. Shown are the pheromone pathway, the filamentous growth pathway (responding to starvation signals) and high osmolarity glycerol (HOG) pathway. These pathways share several components, and mechanisms for ensuring signal specificity and appropriate signal integration are still under investigation.

Signal transduction in yeast has been studied thoroughly; an overview is given for example in (Hohmann, 2002). Several quantitative models have been published so far and some of them are collected in databases like JWSonline (Snoep and Olivier, 2003). Yi and colleagues (Yi et al., 2003) presented a first model of the G protein activation within the pheromone pathway. This model takes into account G protein activities that have been measured using fluorescence resonance energy transfer (FRET). It comprises the production, degradation and activation of the G protein coupled α-receptor (Ste2), the activity cycle of the G protein and its regulation by the regulator of G protein (RGS) Sst2 (compare **Figure 3**).

This model has been adapted and incorporated into a more comprehensive model of the pheromone pathway (Kofahl and Klipp, 2004), which includes downstream processes of the activation of Gβγ. As shown in **Figure 3**, the components of the MAP kinase cascade bind to the scaffold protein Ste5. Binding of Ste5 to Gβγ and the MAP KKKK Ste20 brings Ste20 into the vicinity of Ste11, the MAP KKK, permitting its activation. Furthermore, a cycle of binding, phosphorylation and release of the MAPK Fus3 is considered. Phosphorylated Fus3 triggers the following events including the activation of the transcription factor Ste12, the activation of the cell cycle regulator Far1 and the activation of the RGS Sst2.

The pheromone pathway model includes several feedback loops that help to downregulate the pathway after successful signal transduction. First, the activation of Fus3 leads to a repeated phosphorylation of more Fus3 molecules. Secondly, the activation of Sst2 itself depends on the activation of Fus3. It accelerates the closing of the G protein cycle by enhancing the rate of hydrolysis of Gα-bound GTP. Yi et al. (Yi et al., 2003) studied strains with either constitutively active or inactive Sst2. Third, the transcription factor Ste12 enhances the expression of the protease Bar1, which is exported, and cleaves the α-factor, and thereby counteracts the input signal. Hence, the pathway design ensures the long-term downregulation of the pathway after successful activation of target processes.

The parameters of this model have been estimated from literature values. The impact of individual values has been tested by sensitivity analysis. Although this model is not based on data specifically measured to support it, its predictions for graded response to increasing concentration of α-factor or for the behaviour of mutant cells match very well with experimental observations.

The response of yeast to osmotic stress has been described by a model (Klipp et al., 2005) that comprises the high osmolarity glycerol (HOG) pathway, transcriptional regulation, the effect on metabolism and the change in the production of glycerol and an additional model describing regulation of volume and osmotic pressure. The HOG pathway consists of two input branches, the Sln1 branch and the Sho1 branch (which is not considered in the model). The receptor Sln1 is a membrane protein that regulates a phosphorelay system. Under normal conditions, it is continuously phosphorylated and transmits its phosphate group to Ypd1, which in turn passes it on to Ssk1. In this way, Ssk1 is kept

phosphorylated and inactive. Upon osmotic stress, phosphorylation of Sln1 is interrupted and Ssk1 switches to a non-phosphorylated, active state. In this form, it triggers the HOG MAP kinase cascade, which involves the redundant proteins Ssk2 and Ssk22 as well as Pbs2 and Hog1. Phosphorylated Hog1 can enter the nucleus and regulate the transcription of a series of genes.
An interesting feature of this pathway is that it is downregulated despite sustained activation by external osmolarity. This cellular response could not be explained by modelling the signalling pathway in isolation. It was argued that the cells sense turgor pressure instead of the external salt concentration. The turgor pressure is partially regulated by glycerol. Active Hog1 activates the expression of genes coding for enzymes that are involved in the production of glycerol.
The parameters for this model have been determined on the basis of a standard experiment applying 0.5M NaCl to wild type cells and have been tested for various experimental scenarios with mutant cells and different salt concentrations.
Model simulations have revealed details of the signalling process, enlightening the role of the glycerol channel Fps1 in glycerol accumulation, and the feedback control exerted by protein phosphatases in the MAP kinase pathway. It turns out that Fps1 is responsible for the immediate control on the internal glycerol concentration, while the stimulated expression of GPD1/2 and GPP1/2 and the resulting increased glycerol production preserves a high level of glycerol during growth in high osmolarity. The model implies that the HOG pathway is shut off by to glycerol accumulation, cell re-swelling, and turgor increase rather than by enhanced expression of phosphatases. This result has been confirmed by the experimental fact that the pathway can be fully reactivated by a second osmotic stress.

4.3. Studied Phenomena

4.3.1. Relative Importance of Kinases and Phosphatases

MAP kinase cascades are regulated by the activity of kinases that phosphorylate the proteins, and by phosphatases that in turn ensure the dephosphorylation. While kinases activate and phosphatases deactivate, both partners are necessary to determine the basic level of activation in absence of external stimuli, but also strength and duration of activation in its presence. It has been discussed that kinases are responsible for the amplitude of the signal, while phosphatases determine its duration [Hornberg, 2005]. Interestingly, this holds only for weakly activated cascades [Heinrich, 2002], while strongly activated cascades show the tendency of prolonged activation upon increase of stimulus. This is based on conservation of MAP kinase proteins on each level, which limits the increase of the active form upon strong activation (**Figure 4**).

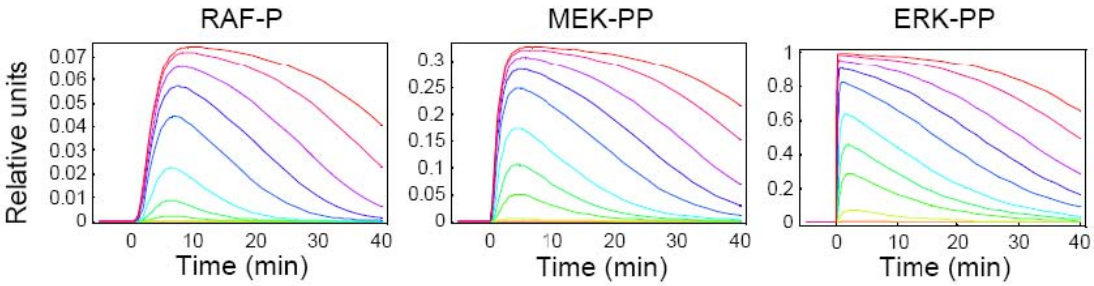


Figure 4: Time courses of the concentration of the phosphorylated forms of three kinases (Raf, MEK, ERK) in the MAP kinase cascade as in Figure 2D, i.e. their activation profiles over time: low activation of the receptor leads to an increase of the amplitude, stronger receptor activation causes longer activation. All rate laws are mass action kinetics with rate constants of kinases and phosphatases equal to 1 and the initial concentration values of the phosphorylated proteins were 0 and of the non-phosphorylated proteins were 1.

4.3.2. Dynamic Behaviour and Parameters

The specific behaviour of a biochemical network is determined by (i) its wiring, expressed by the stoichiometric matrix ***N***, (ii) by the kinetic laws of the individual reactions including the involvement of modifiers that are not substrate or product of this reaction, (iii) by the values of the kinetic parameters and (iv) by the concentrations involved, like initial concentrations and conserved moieties.

In order to obtain a satisfactory picture of the studied object, all four aspects must be appropriate. The wiring scheme is frequently (but not always!) sufficiently well known from experimental information. For some metabolic reactions, the kinetic mechanism is also determined together with the respective parameters. However, kinetic laws and parameters are often not well-defined by experimental information, whereas concentration or number of molecules involved are often known to a satisfactory extent.

To develop models with predictive value, high-quality data is necessary. Time series data must cover the regions, in which the dynamics of the pathways take place. Moreover, for sound model validation and parameterization it is necessary to have a measure of uncertainty for the measured data, as standard deviations, for instance. This requires measurement repetitions to be done that are unfortunately often not available.

4.3.3. Signalling: Network Versus Pathway

The original perception of signalling pathways stems from the experimental analysis that could connect a stimulus of the cell with a measurable effect and could trace the path connecting both. Nowadays, it becomes obvious that cells possess a comprehensive arsenal of signalling molecules that may interact in various combinations giving rise to the transmission of various signals, but also to the integration and separation of diverse types of information.

It is now a matter of taste whether modelling starts immediately with the

complete signalling network, or whether one starts with the individual traditional pathways that are sometimes well understood and then tries to integrate them. Coupling of pathways may be performed in the same way as modelling individual pathways: pathway structure is merged and individual reaction rates are adopted using a mixture of handcrafted rules and intuition. Approaches for systematic model integration are rare. A starting point is SBMLmerge, which combines models implemented in SBML.

4.3.4. Crosstalk Between Pathways

There are many different ways in which signalling pathways can interact with each other, a phenomenon often called crosstalk. For example, different pathways can be triggered by the same receptor or they can share components that, once activated by one pathway, leak into another pathway and thereby activate it. For an overview of different ways of pathway crosstalk see (Schwartz & Baron 1999, Schwartz & Madhani 2004, Cowan & Storey 2003). In modelling crosstalk there has been the issue of quantifying the amount of crosstalk. Some studies analysed the topological and structural properties of signalling networks by, e.g., classifying modes of interaction (Papin & Palsson 2004) or by counting the theoretically possible interactions between pathways (Binder & Heinrich 2004).

As signalling is a transient process one can argue that it is the dynamic behaviour of interacting pathways that is important rather than the static features. Two recent studies address the dynamic features of pathway crosstalk. By analysing the activation of pathways by a so-called intrinsic and an extrinsic stimulus, respectively, one study defined measures for pathway specificity and fidelity (Komarova et al., 2005). These measures give useful insights how pathways interact with each other. However, it is important to note that these measures refer to responses to one stimulus at a time. However, it can be assumed that cells usually process multiple information in parallel and these measures give no clue how signals interact while being transmitted concomitantly, It can be expected that signals amplify or inhibit each other, when transmitted at the same time. Thus, it does not suffice to study each signal in isolation but also to study the cell's response to multiple stimuli at the same time. Schaber et al. (under review) proposed crosstalk measures that include parallel multiple pathway activation called the intrinsic and extrinsic specificity that yield a better understanding of how the pathways dynamically interact.

4.3.5. Modelling and Standards

The purpose of modelling is to provide an abstract description of an instance that fosters the understanding/representation of specific aspects of this instance. Such a model must neglect other aspects for the sake of simplicity, and these neglected aspects will change with a change of the specific question to be answered by the model. Therefore, one cannot establish fixed rules for a model that are valid once and forever. On the other hand, the growing modelling community and the need to communicate with experimental researchers make it necessary to establish some rules how specific aspects should be expressed in a model of a certain type.

A prominent approach for the development of such a standard is the Systems Biology Markup Language (SBML) (Hucka et al., 2003), which serves as a unified exchange language for the description of biochemical network models. Another standard is the Minimal Requirements in the Annotation of Models (MIRIAM) (Novere et al., 2005), a standard for the description and documentation of models in a publication.

Acknowledgements. *This work was supported by the German Ministry for Education and Research via the Berlin Center for Genome based Bioinformatics (Grant 031U109C) and by the European Commission via the QUASI project (Grant FP6-2002-LSH-503 230).*

References

BHALLA, U. S. (2002) Biochemical signaling networks decode temporal patterns of synaptic input. *J Comput Neurosci,* **13,** 49-62.

BHALLA, U. S. (2004) Models of cell signaling pathways. *Curr Opin Genet Dev,* **14,** 375-81.

BHALLA, U. S. & IYENGAR, R. (1999) Emergent properties of networks of biological signaling pathways. *Science,* **283,** 381-7.

BHALLA, U. S. & IYENGAR, R. (2001) Robustness of the bistable behavior of a biological signaling feedback loop. *Chaos,* **11,** 221-226.

CHEN, K. C., CALZONE, L., CSIKASZ-NAGY, A., CROSS, F. R., NOVAK, B. & TYSON, J. J. (2004) Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell,* **15,** 3841-62.

CHEN, K. C., CSIKASZ-NAGY, A., GYORFFY, B., VAL, J., NOVAK, B. & TYSON, J. J. (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol Biol Cell,* **11,** 369-91.

CORNISH-BOWDEN, A. (2004) *Fundamentals of Enzyme Kinetics,* London, Portland Press.

GILLESPIE, D. T. (1977) Exact Stochastic Simulation of coupled chemical-reactions. *J Phys Chem,* **81,** 2340-2361.

HEINRICH, R., NEEL, B. G. & RAPOPORT, T. A. (2002) Mathematical models of protein kinase signal transduction. *Mol Cell,* **9,** 957-70.

HOHMANN, S. (2002) Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol Mol Biol Rev,* **66,** 300-72.

HUCKA, M., FINNEY, A., SAURO, H. M., BOLOURI, H., DOYLE, J. C., KITANO, H., ARKIN, A. P., BORNSTEIN, B. J., BRAY, D., CORNISH-BOWDEN, A., CUELLAR, A. A., DRONOV, S., GILLES, E. D., GINKEL, M., GOR, V., GORYANIN, II, HEDLEY, W. J., HODGMAN, T. C., HOFMEYR, J. H., HUNTER, P. J., JUTY, N. S., KASBERGER, J. L., KREMLING, A., KUMMER, U., LE NOVERE, N., LOEW, L. M., LUCIO, D., MENDES, P., MINCH, E., MJOLSNESS, E. D., NAKAYAMA, Y., NELSON, M. R., NIELSEN, P. F., SAKURADA, T.,

SCHAFF, J. C., SHAPIRO, B. E., SHIMIZU, T. S., SPENCE, H. D., STELLING, J., TAKAHASHI, K., TOMITA, M., WAGNER, J. & WANG, J. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics,* **19,** 524-31.

INGALLS, B. P. & SAURO, H. M. (2003) Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. *J Theor Biol,* **222,** 23-36.

KLIPP, E., NORDLANDER, B., KRUGER, R., GENNEMARK, P. & HOHMANN, S. (2005) Integrative model of the response of yeast to osmotic shock. *Nat Biotechnol,* **23,** 975-82.

KOFAHL, B. & KLIPP, E. (2004) Modelling the dynamics of the yeast pheromone pathway. *Yeast,* **21,** 831-50.

KOMAROVA, N. L., ZOU, X., NIE, Q. & BARDWELL, L. (2005) A theoretical framework for specificity in cell signaling. *Mol Sys Biol*.

NOVAK, B., TOTH, A., CSIKASZ-NAGY, A., GYORFFY, B., TYSON, J. J. & NASMYTH, K. (1999) Finishing the cell cycle. *J Theor Biol,* **199,** 223-33.

NOVERE, N. L., FINNEY, A., HUCKA, M., BHALLA, U. S., CAMPAGNE, F., COLLADO-VIDES, J., CRAMPIN, E. J., HALSTEAD, M., KLIPP, E., MENDES, P., NIELSEN, P., SAURO, H., SHAPIRO, B., SNOEP, J. L., SPENCE, H. D. & WANNER, B. L. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol,* **23,** 1509-15.

PAPIN, J. A., HUNTER, T., PALSSON, B. O. & SUBRAMANIAM, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol,* **6,** 99-111.

PAPIN, J. A. & PALSSON, B. O. (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J Theor Biol,* **227,** 283-97.

REDER, C. (1988) Metabolic control theory: a structural approach. *J Theor Biol,* **135,** 175-201.

SCHMIDT, A. & HALL, A. (2002) Guanine nucleotide exchange factors for Rho GTPases: turning on the switch. *Genes Dev,* **16,** 1587-609.

SCHOEBERL, B., EICHLER-JONSSON, C., GILLES, E. D. & MULLER, G. (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol,* **20,** 370-5.

SNOEP, J. L. & OLIVIER, B. G. (2003) JWS online cellular systems modelling and microbiology. *Microbiology,* **149,** 3045-7.

SWAMEYE, I., MULLER, T. G., TIMMER, J., SANDRA, O. & KLINGMULLER, U. (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc Natl Acad Sci U S A,* **100,** 1028-33.

TYSON, J. J., CHEN, K. C. & NOVAK, B. (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol,* **15,** 221-31.

VASEGHI, S., MACHERHAMMER, F., ZIBEK, S. & REUSS, M. (2001) Signal transduction dynamics of the protein kinase-A/phosphofructokinase-2 system in Saccharomyces cerevisiae. *Metab Eng,* **3,** 163-72.

YI, T. M., KITANO, H. & SIMON, M. I. (2003) A quantitative characterization of the yeast heterotrimeric G protein cycle. *Proc Natl Acad Sci U S A,* **100,** 10764-9.

Focusing on Dynamic Dimension Reduction for Biochemical Reaction Systems.

Irina Surovtsova∗[a], and Jürgen Zobeley[a]

[a]*EML Research gGmbH, Bioinformatics and Computational Biochemistry, Heidelberg, Germany. e-mail: irina.surovtsova@eml-r.villa-bosch.de*

Keywords: Complexity reduction, dimension monitoring, intrinsic low–dimensional manifold (ILDM), quasi–steady state assumption.

1. Abstract

Due to the complexity of biochemical reaction networks the so–called complexity reduction algorithms play a crucial role for making simulations realizable "in silico". Our first approach (Zobeley et al., 2005) of dynamic dimension reduction is based on different time scales of biokinetics and seizes the distinction between "fast" and "slow" modes detected adaptively. This modified ILDM method ("intrinsic low–dimensional manifold") is suited not only for steady states, but for all possible dynamics and provides a systematic tool for an automated complexity reduction of arbitrary biochemical reaction networks.
Continuing this dynamic modification of ILDM, the present study focuses on a numerical question that we believe to be still open: the period of "adequate" approximation, i.e. how long the differential–algebraic equations of ILDM provide acceptable approximations of the (biochemical) ODEs.
Numerical simulations are to give a first answer here — considering the example of glycolysis in yeast presented in (Wolf and Heinrich, 2000).

2. Introduction

In recent years, biochemistry has made huge steps towards quantified results. This is basically a consequence of improved experimental techniques providing more and more data about molecules in living cells ("in vivo") — instead of former experiments in separated test tubes ("in vitro"). In fact, the amount of experimental data and its precision has also revealed a new challenge in scientific computing. Today standard computing facilities come up against limiting factors when simulating biochemical models in their full complexity available. This experience leads to the important question how to reduce the "complexity" of both models and calculations so that the relevant effects are still simulated correctly.

∗ Corresponding author

Understanding and Exploiting Systems Biology

Former concepts of complexity reduction in biochemistry are based on the restriction to the steady state behaviour of the biological system (such as Kauffman et al., 2002, and Price et al., 2003). Assuming steady states, however, is only suitable for few examples such as simple microorganisms in a fermenter whereas the majority of biochemical systems is characterized by highly nonlinear dynamics. Very popular examples are calcium oscillations (in plants and animals) realizing information processing in cells (see e.g. Berridge at al., 1998), metabolic oscillations in neutrophils (Petty et al., 2000), and glycolytic oscillations (Duysens et al., 1957, and Frenkel, 1968).

Furthermore, environmental conditions exert an essential influence on organisms and are usually changing permanently. Thus, any satisfactory concept of complexity reduction has to take dynamic aspects into account.

There are different approaches to reduce models describing complex chemical and physical processes. The procedures involve application of conservation relations, lumping of species and sensitivity analysis. One of the most popular approaches is based on the presence of a wide range of characteristic time scales in a chemical system. At present, there are several methods in chemistry which are built on the concept of time scale decomposition. An incomplete list includes the computational singular perturbation (CSP) method (as, for instance, Lam and Goussis, 1994) and method of intrinsic low–dimensional manifolds introduced by Maas and Pope in 1992.

Time scale decomposition should be a quite promising approach to dynamic complexity reduction of biochemical systems since biological processes proceed on a wide range of time scales (from fraction of seconds by signal transduction to several hours by events like gene expression). Nevertheless, there is an essential distinction, though, preventing us from applying immediately reduction methods used in chemistry. Indeed, biochemical reactions can hardly ever be described by the law of mass action and thus, ordinary differential equations of concentrations usually contain reaction terms that are more difficult to compute than polynomials and more sensitive with respect to calculation errors.

As a consequence, existing numerical methods have to be reviewed whether their results are still reliable for biochemical reaction networks.

In our previous paper **(**Zobeley et al., 2005), we have introduced an adapted ILDM method (based on Deuflhard and Heroth, 1996) for the use of time–dependent complexity reduction in the context of systems biology. This approach starts from the notion of different time scales and seizes the distinction between “fast” and “slow” modes detected adaptively**.**

Continuing this dynamic modification of ILDM, the present study focuses on a numerical question that we believe to be still open: the period of “adequate” approximation, i.e. how long the differential–algebraic equations of ILDM provide acceptable approximations of the (biochemical) ordinary differential equations. This is directly related to the question how long the distinction between “fast” and “slow” modes is appropriate.

3. Mathematical Methodology

3.1. The quantitative description of a biochemical reaction network

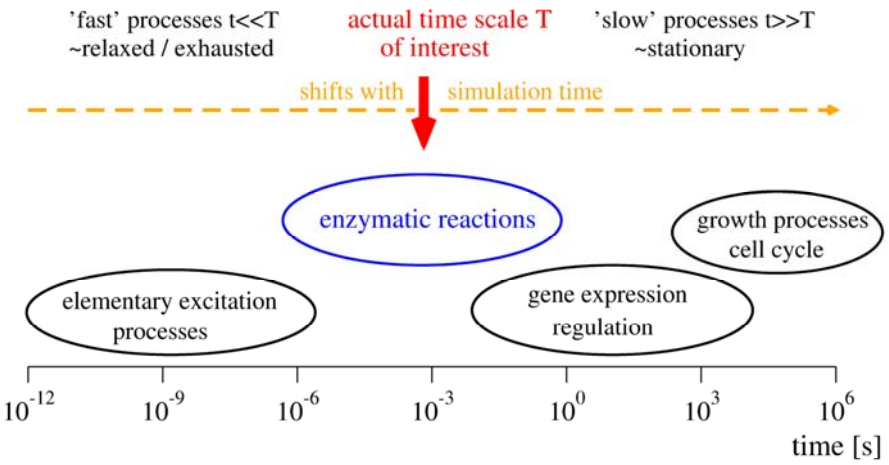
Considering an arbitrary biochemical reaction network, each of the n species is described by its time–dependent concentration $c_j = c_j(t)$. So in particular, spatial dependencies are not taken into consideration here. These scalar concentrations are united into a time–dependent vector $\boldsymbol{c} = (c_1 \dots c_n)^T$ and, its dynamics is described by an autonomous system of ordinary differential equations (shortly, ODEs) in combination with the initial state $\boldsymbol{c_0}$, i.e.

$$\frac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{f}(\boldsymbol{c}(t)) \quad \text{for all } t \geq 0, \qquad \boldsymbol{c}(0) = \boldsymbol{c_0}$$

Due to existing relations of mass conservation, such a system of ODEs often exhibits linear dependencies that can be detected and removed in a systematic and automated way. Indeed, applying the tools of stoichiometric network analysis (to the stoichiometric matrix) lays the basis for reducing the ODE system successively so that finally, we can assume the ODE system to be in its reduced and linearly independent form. Further details of this network analysis can be found in (Heinrich and Schuster, 1996) and (Reder, 1988), for example.

3.2. Time scale decomposition

Whenever several processes come together, they usually differ from each other in regard to their characteristic duration. To be slightly more precise, (only) the direct comparison reveals which process is rather fast in comparison with others or, in contrast, which process can be regarded as (almost) stationary in comparison with others. Of course, such a distinction between "quasi–stationary" and (highly) nonstationary need not be fixed, but has to be adapted to the current state $\boldsymbol{c} = (c_1 \dots c_n)^T$.



As a mathematical criterion for the several time scales occurring in a reaction network, we seize the eigenvalues of the Jacobian matrix

$$D\boldsymbol{f}(\boldsymbol{c}) = \left(\frac{\partial f_i}{\partial c_j}\right)_{1 \le i,j \le n}.$$

This is because, for short times, the solutions of the ordinary differential equation $\frac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{f}(\boldsymbol{c}(t))$ and of its linearization $\frac{d}{dt}\boldsymbol{c}(t) = D\boldsymbol{f}(\boldsymbol{c}(t))$ have key properties in common. If, in addition, the Jacobian is a constant diagonal matrix, i.e.

$$D\boldsymbol{f}(\boldsymbol{c}(t)) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \qquad \text{with real coefficients } \lambda_1 \dots \lambda_n,$$

then the solution is very easy to formulate explicitly, i.e. $c_j(t) = c_{0,j}\, e^{\lambda_j t}$, and $\tau_j = \frac{1}{|\lambda_j|}$ indicates the "characteristic time scale" of component j (if $\lambda_j \neq 0$). Even in the situation of complex coefficients $\lambda_1 \dots \lambda_n$ of the constant diagonal matrix, the explicit solutions are the same although they might be difficult to interpret physically. The corresponding characteristic time scale and the qualitative behaviour of the solution now depend on the real part of λ_j. Indeed, for $\mathrm{Re}\,\lambda_j < 0$, the solution $c_j(t)$ is always decaying and tends to 0 (while time is increasing) and, for $\mathrm{Re}\,\lambda_j > 0,$ the absolute value of $c_j(t)$ is getting arbitrarily large. A nonvanishing imaginary part of λ_j indicates oscillatory features of $c_j(t)$.

Now we seize this simple example of calculus and try to exploit it when searching "fast" and "slow" modes of the ODE system $\frac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{f}(\boldsymbol{c}(t))$ at a given initial point $\boldsymbol{c_0}$. In general, the Jacobian matrix $D\boldsymbol{f}(\boldsymbol{c_0})$ need not have diagonal form. So we transform the physical variables (describing the concentration $\boldsymbol{c} = (c_1 \dots c_n)$ of species) to a new coordinate system that is easier to handle mathematically. The components of the transformed state vector are called *modes*.

On the one hand, this transformation ought to be simple and not to complicate further calculations and so, we want to choose it linear and constant. On the other hand, the Jacobian of the transformed right–hand side $\boldsymbol{f}$ should be similar to handle as the diagonal form before, but we cannot assume $D\boldsymbol{f}(\boldsymbol{c_0})$ to be always diagonalizable. So the so–called *Schur decomposition* provides a useful tool of linear algebra:

A well–known algorithm provides an orthogonal matrix $\boldsymbol{Q} \in \mathrm{IR}^{n \times n}$ such that the transformation of the Jacobian matrix $D\boldsymbol{f}(\boldsymbol{c_0})$ has the form

$$\boldsymbol{Q}\ D\boldsymbol{f}(\boldsymbol{c_0})\ \boldsymbol{Q}^{-1} \;=\; \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \cdots & \boldsymbol{S}_{1m} \\ 0 & \boldsymbol{S}_{22} & & \boldsymbol{S}_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{S}_{mm} \end{pmatrix}$$

with each diagonal submatrix $\boldsymbol{S}_{jj}$ being either a (1×1)–block (with a real eigenvalue of $D\boldsymbol{f}(\boldsymbol{c_0})$) or a (2×2)–block (corresponding to a complex conjugate pair of eigenvalues). An adequate composition of Givens rotations facilitates in addition that the corresponding eigenvalues $\lambda_1 \ldots \lambda_n$ are sorted such that $|\mathrm{Re}\,\lambda_1| \le |\mathrm{Re}\,\lambda_2| \le \ldots \le |\mathrm{Re}\,\lambda_n|$, as shown in Golub and van Loan, 1996, for example. Using orthogonal matrices for all these transformations implies the advantage that computational errors usually do not increase significantly.
Replacing now the physical quantity $\boldsymbol{c}(t) = (c_1(t) \ldots c_n(t))$ (of concentrations) by its transformation

$$\boldsymbol{z}(t) := \boldsymbol{Q}\,\boldsymbol{c}(t),$$

this time–dependent vector of modes satisfies the ordinary differential equation

$$\tfrac{d}{dt}\boldsymbol{z}(t) = \boldsymbol{Q}\cdot\tfrac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{Q}\cdot\boldsymbol{f}(\boldsymbol{c}(t)) = \boldsymbol{Q}\cdot\boldsymbol{f}(\boldsymbol{Q}^{-1}\,\boldsymbol{z}(t))$$

and thus, the Jacobian of the right–hand side at the initial state $\boldsymbol{z}(0) = \boldsymbol{Q}\,\boldsymbol{c_0}$ reveals the wanted eigenvalues $\lambda_1 \ldots \lambda_n$ by construction. Furthermore the "characteristic time scales" of the first components of $\boldsymbol{z}(t)$ are (possibly) much larger than their counterparts of the last components because $\frac{1}{|\mathrm{Re}\,\lambda_1|} \ge \frac{1}{|\mathrm{Re}\,\lambda_2|} \ge \ldots \ge \frac{1}{|\mathrm{Re}\,\lambda_n|}$. So Schur decomposition (in combination with Givens rotations) has laid the basis for a time scale decomposition – at least close to the initial points $\boldsymbol{c_0}$ and $\boldsymbol{z}(0) = \boldsymbol{Q}\,\boldsymbol{c_0}$, respectively.

3.3. The distinction between "slow" and "fast" modes: ILDM method (of Maas and Pope).

The key question now is to exploit these multiple time scales for numerical calculations.
Many examples of biochemical reaction kinetics have in common that maximum and minimum of the characteristic time scales differ from each other tremendously, i.e. $\frac{1}{|\mathrm{Re}\,\lambda_1|} >> \frac{1}{|\mathrm{Re}\,\lambda_n|}$. As a consequence, the last mode $z_n(t)$ seems to respond "instantaneously" to changes of some other modes. In other words, $z_n(t)$ appears to be in a steady state — after a very short time that we are willing to neglect.
Mathematically speaking, this step of approximation is based on a substitution. Indeed, the (exact) ordinary differential equation for $z_n(t)$, i.e. $\frac{d}{dt}z_n(t) = \left(\boldsymbol{Q}\cdot\boldsymbol{f}(\boldsymbol{Q}^{-1}\,\boldsymbol{z}(t))\right)_n$, is replaced by the algebraic equation $0 = \left(\boldsymbol{Q}\cdot\boldsymbol{f}(\boldsymbol{Q}^{-1}\,\boldsymbol{z}(t))\right)_n$.
This idea is called *quasi–steady state assumption* (QSSA) for $z_n(t)$.
Of course, we are free to make QSSA for more than one component of $\boldsymbol{z}(t)$.
Motivated by the difference in time scales, the modes with QSSA are called "fast" whereas the others are called "slow".

This notion leads to the method of *low–dimensional intrinsic manifolds*. It was introduced for the example of combustion by Maas and Pope in 1992. Roughly speaking, it consists of three basic steps:

1) Make a suitable linear transformation revealing the eigenvalues of Jacobian and thus the characteristic time scales.
2) Decide how many modes (after the transformation) are regarded as "slow". The rest of the modes are considered "fast".
3) Use the ordinary differential equations for the "slow" modes and determine the corresponding "fast" modes alternately by means of their algebraic equations (according to their quasi–steady state assumptions).

For implementing this algorithm, many details have to be specified. From now on, r abbreviates the number of slow modes and, $n - r$ is the corresponding number of fast modes.

Firstly, we require a "suitable" linear transformation defined by an invertible matrix. Starting in a given initial point $\boldsymbol{c_0}$, the orthogonal matrix $\boldsymbol{Q}$ resulting from Schur decomposition is an obvious suggestion. The linearization of the transformed ODE reveals a potential weakness, though. Indeed, uniting slow and fast modes into separated vectors leads to the form

$$\boldsymbol{Q}\ D\boldsymbol{f}(\boldsymbol{c_0})\ \boldsymbol{Q}^{-1} \;=\; \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \cdots & \boldsymbol{S}_{1m} \\ 0 & \boldsymbol{S}_{22} & & \boldsymbol{S}_{2m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{S}_{mm} \end{pmatrix} \;=\; \begin{pmatrix} \boldsymbol{S}_{\text{slow}} & \boldsymbol{S}_{\text{coup}} \\ 0 & \boldsymbol{S}_{\text{fast}} \end{pmatrix}$$

with the submatrix $\boldsymbol{S}_{\text{coup}} \in \mathrm{IR}^{r\times(n-r)}$ reflecting the sensitivity of slow modes with respect to their fast counterparts. So in general, we cannot exclude that the slow modes depend very much on the exact values of fast modes. For moderating the necessity of precision, a further linear transformation is to provide the "decoupling" effect of $\boldsymbol{S}_{\text{coup}} = 0.$ Choosing

$$\boldsymbol{Id}_n - \begin{pmatrix} 0 & \boldsymbol{Z} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{Id}_r & -\boldsymbol{Z} \\ 0 & \boldsymbol{Id}_{n-r} \end{pmatrix} \qquad \text{with some } \boldsymbol{Z} \in \mathrm{IR}^{r\times(n-r)} \text{ and identity matrix } \boldsymbol{Id}_n \in \mathrm{IR}^{n\times n}$$

as an ansatz for the second linear transformation, $\boldsymbol{Z}$ has to solve the Sylvester equation

$$\boldsymbol{S}_{\text{slow}}\,\boldsymbol{Z} - \boldsymbol{Z}\,\boldsymbol{S}_{\text{fast}} \;=\; -\boldsymbol{S}_{\text{coup}}$$

and thus can be calculated by a classical algorithm given in Golub and van Loan, 1996, for example. So the composition of these two linear transformation is described by the matrices

$$\tilde{\boldsymbol{Q}} := \left(\boldsymbol{Id}_n - \begin{pmatrix} 0 & \boldsymbol{Z} \\ 0 & 0 \end{pmatrix}\right) \cdot \boldsymbol{Q}, \quad \boldsymbol{T} := \tilde{\boldsymbol{Q}}^{-1} = \boldsymbol{Q}^T \cdot \left(\boldsymbol{Id}_n + \begin{pmatrix} 0 & \boldsymbol{Z} \\ 0 & 0 \end{pmatrix}\right)$$

leading to the transformed Jacobian

$$\boldsymbol{T}^{-1}\ D\boldsymbol{f}(\boldsymbol{c_0})\ \boldsymbol{T}\ =\ \widetilde{\boldsymbol{Q}}\ D\boldsymbol{f}(\boldsymbol{c_0})\ \widetilde{\boldsymbol{Q}}^{-1}\ =\ \begin{pmatrix} \widetilde{\boldsymbol{S}}_{\text{slow}} & 0 \\ 0 & \widetilde{\boldsymbol{S}}_{\text{fast}} \end{pmatrix}.$$

Obviously, $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ is invertible, but it need not be orthogonal (as $\boldsymbol{Q}$). So we have paid a possibly high price for the advantage of decoupling slow and fast modes linearly, i.e. computational errors might increase significantly.
Applying now the transformation to all quantities involved, the state vector $\boldsymbol{c} = (c_1 \dots c_n)^T$ is substituted by

$$\boldsymbol{x} := \begin{pmatrix} \boldsymbol{x}_{\text{slow}} \\ \boldsymbol{x}_{\text{fast}} \end{pmatrix} := \widetilde{\boldsymbol{Q}}\,\boldsymbol{c}\ =\ \boldsymbol{T}^{-1}\,\boldsymbol{c}$$

and, the corresponding ordinary differential equation is

$$\tfrac{d}{dt}\,\boldsymbol{x}(t)\ =\ \boldsymbol{g}(\boldsymbol{x}(t)) \qquad \text{with}\ \ \boldsymbol{g} := \begin{pmatrix} \boldsymbol{g}_{\text{slow}} \\ \boldsymbol{g}_{\text{fast}} \end{pmatrix} := \boldsymbol{T}^{-1} \cdot \boldsymbol{f}(\boldsymbol{T}\,\cdot).$$

So the quasi–steady state assumption about the fast modes means that this ODE system is replaced by the differential–algebraic equation (DAE)

$$\begin{aligned} \tfrac{d}{dt}\ \boldsymbol{x}_{\text{slow}}(t) &= \boldsymbol{g}_{\text{slow}}(\boldsymbol{x}_{\text{slow}}(t),\ \boldsymbol{x}_{\text{fast}}(t)), \quad \boldsymbol{x}_{\text{slow}}(0) = \boldsymbol{x}_{0,\ \text{slow}} := \left(\boldsymbol{T}^{-1}\ \ \boldsymbol{c_0}\right)_{j\ =\ 1\dots r} \in \mathbb{R}^r \\ 0 &= \boldsymbol{g}_{\text{fast}}(\boldsymbol{x}_{\text{slow}}(t),\ \boldsymbol{x}_{\text{fast}}(t)), \end{aligned}$$

Here the initial value of $\boldsymbol{x}_{\text{fast}} = \boldsymbol{x}_{\text{fast}}(t)$ has still to be specified. Transforming the given (physical) state $\boldsymbol{c_0} \in \mathbb{R}^n$ leads to the candidate $\boldsymbol{x}_{0,\ \text{fast}} := \left(\boldsymbol{T}^{-1}\ \ \boldsymbol{c_0}\right)_{j\ =\ r+1\dots n} \in \mathbb{R}^{n-r}$, but $\boldsymbol{x}_{0,\ \text{fast}}$ might be "inconsistent" with the DAE, i.e. it does not satisfy the algebraic equation $0 = \boldsymbol{g}_{\text{fast}}(\boldsymbol{x}_{0,\text{slow}},\ \boldsymbol{x}_{0,\text{fast}})$. Then a solution $\hat{\boldsymbol{x}}_{0,\ \text{fast}} \in \mathbb{R}^{n-r}$ of $0 = \boldsymbol{g}_{\text{fast}}(\boldsymbol{x}_{0,\text{slow}},\ \hat{\boldsymbol{x}}_{0,\text{fast}})$ is required as initial value for solving the DAE and, due to the aim of approximation, it should be close to $\boldsymbol{x}_{0,\ \text{fast}}$. Several methods are available for solving such a nonlinear equation numerically. Among the very popular examples are Newton method and its simplified modifications.
The next important question is how many modes are regarded as "fast" and "slow", respectively. This question is sometimes called *dimension monitoring* (as e.g. in Deuflhard and Bornemann, 2002).
Obviously, candidates for fast modes are merely those components $x_{r+1} \dots x_n$ of $\boldsymbol{x}$ whose (maybe complex) eigenvalues $\lambda_{r+1} \dots \lambda_n$ of $D\boldsymbol{g}(x_{0,slow}, x_{0,\,fast}) = \boldsymbol{T}^{-1}\,D\boldsymbol{f}(\boldsymbol{c_0})\,\boldsymbol{T}$ have negative real parts since this feature reflects decaying. Roughly speaking, the larger the absolute value $\left|\operatorname{Re}\lambda_j\right| = -\operatorname{Re}\lambda_j > 0$ is, the faster this mode x_j converges to its steady state (as the linear example mentioned in the beginning suggests). So we rely on the characteristic time scales $\tau_j = \frac{1}{\left|\operatorname{Re}\lambda_j\right|} < \infty\ \ (j = 1 \dots n)$ for

specifying r and exploit $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n$ due to the choice of transformation in section 3.2.

Deuflhard and Heroth, 1996, suggest a criterion justified by the analytical methods of singular perturbation theory. For a given error tolerance $tol > 0$, the number r of slow modes is chosen such that the corresponding decomposition $\boldsymbol{x} = (\boldsymbol{x}_{\text{slow}}, \boldsymbol{x}_{\text{fast}})^T \in \mathrm{IR}^r \times \mathrm{IR}^{n-r}$ still satisfies

$$\tau_{r+1} \cdot \left| \boldsymbol{g}_{\text{slow}}(\boldsymbol{x}_{0,\text{slow}}, \boldsymbol{x}_{0,\text{fast}}) - \boldsymbol{g}_{\text{slow}}(\boldsymbol{x}_{0,\text{slow}}, \hat{\boldsymbol{x}}_{0,\text{fast}}) \right| \leq tol$$

with the vectors $\boldsymbol{x}_{0,\text{slow}} \in \mathrm{IR}^r$ and $\boldsymbol{x}_{0,\text{fast}}, \hat{\boldsymbol{x}}_{0,\text{fast}} \in \mathrm{IR}^{n-r}$ denoting the initial values mentioned before.

An important advantage of this criterion is that it is very cheap to calculate. Moreover, Deuflhard, Heroth and Maas, 1996, extend earlier reports about highly satisfactory results when simulating

- Hydrogen–oxygen combustion (due to Hoppensteadt, Alfeld and Aiken, 1981)
- Thermal decomposition of n–hexane (due to Isbarn, Ederer and Ebert, 1981)
- Oregantor (due to Field and Noyes, 1974)

Their numerical implementation is based on a smart combination of methods: The DAE system resulting from ILDM is discretized according to the linearly implicit Euler method and, few steps for extrapolation (with respect to the step size $\tau > 0$ of time) provide sufficiently accurate results. This implementation and further simplifying modifications of the Euler method are justified by the asymptotic error expansion presented by Deuflhard, Hairer and Zugck in 1987.

3.4. Open question: How long is the distinction between "fast" and "slow" modes appropriate?

From our point of view, an important question has not been investigated sufficiently. It concerns the period of time in which the preceding distinction between "fast" and "slow" modes can be preserved.

For extending its role of an efficient reduction method to biochemical reaction networks, ILDM has to fulfil an essential condition. Indeed, the transition from the original ODE system to the approximating DAE system provides sufficient accuracy merely up to some time $\hat{\tau} > 0$ and, this additional time scale $\hat{\tau}$ ought to be large in comparison with the (shortest) characteristic time scale $\tau_{r+1} = \frac{1}{|\mathrm{Re}\,\lambda_{r+1}|}$ of the "fast" modes: $\hat{\tau} >> \tau_{r+1}$. Otherwise all arguments of singular perturbation theory and asymptotic expansion (used by Deuflhard et al.) fail definitely. In fact, we have found an implicit remark (about this necessary feature) in Deuflhard and Bornemann, 2004, but there were no details available to us whether this condition has been verified numerically so far.

For investigating the period $\hat{\tau} > 0$, we would like to avoid all further systematic errors due to approximating hypotheses. So first, the original ODE system describing the physical quantities of concentrations

$$\frac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{f}(\boldsymbol{c}(t)), \qquad \boldsymbol{c}(0) = \boldsymbol{c_0}$$

is solved numerically up to a fixed time $t = t_{\max}$. Then we select several points $t_0 \in [0, t_{\max}]$ of time subsequently and execute the following steps:

1) Calculate the time scale decomposition at time t_0 according to section 3.2, i.e. using Schur decomposition and Givens rotations for determining the eigenvalues of the Jacobian $D\boldsymbol{f}(\boldsymbol{c}(t_0))$.
2) Determine the maximal number $r \in \{1 \ldots n\}$ of slow modes according to the criterion of Deuflhard and Heroth (for given and fixed error threshold $tol > 0$).
3) The transformed initial value problem

$$\tfrac{d}{dt}\,\boldsymbol{x}(t) \;=\; \boldsymbol{g}(\boldsymbol{x}(t)) \;=\; \begin{pmatrix} \boldsymbol{g}_{\text{slow}}(\boldsymbol{x}(t)) \\ \boldsymbol{g}_{\text{fast}}(\boldsymbol{x}(t)) \end{pmatrix} \quad \text{for } t \geq t_0, \qquad \boldsymbol{x}(t_0) = \boldsymbol{T}^{-1}\,\boldsymbol{c}(t_0) \in \mathrm{I\!R}^n$$

is directly related to the approximating DAE system for $\boldsymbol{y}(t) = (\boldsymbol{y}_{\text{slow}}(t), \boldsymbol{y}_{\text{fast}}(t)) \in \mathrm{I\!R}^r \times \mathrm{I\!R}^{n-r}$

$$\begin{aligned} \tfrac{d}{dt}\;\boldsymbol{y}_{\text{slow}}(t) \;&=\; \boldsymbol{g}_{\text{slow}}(\boldsymbol{y}_{\text{slow}}(t),\; \boldsymbol{y}_{\text{fast}}(t)), \quad \boldsymbol{y}_{\text{slow}}(t_0) = \left(\boldsymbol{T}^{-1}\;\boldsymbol{c}(t_0)\right)_{j=1\ldots r} \in \mathrm{I\!R}^r \\ 0 \;&=\; \boldsymbol{g}_{\text{fast}}(\boldsymbol{y}_{\text{slow}}(t),\; \boldsymbol{y}_{\text{fast}}(t)), \quad \boldsymbol{y}_{\text{fast}}(t_0) \in \mathrm{I\!R}^{n-r} \text{ as solution of} \\ &\qquad\qquad\qquad\qquad 0 = \boldsymbol{g}_{\text{fast}}(\boldsymbol{y}_{\text{slow}}(t_0),\; \boldsymbol{y}_{\text{fast}}(t_0)) \text{ close to } (\boldsymbol{T}^{-1}\;\boldsymbol{c}(t_0))_{j=r+1\ldots n} \end{aligned}$$

that is also solved numerically.
4) Approximate the maximal period $\hat{\tau}_{\text{rel}} > 0$ such that the relative error comparing the DAE solution $\boldsymbol{T}\,\boldsymbol{y}(t_0 + \hat{\tau}_{\text{rel}}) \in \mathrm{I\!R}^n$ (after the inverse transformation) with the original solution $\boldsymbol{c}(t_0 + \hat{\tau}_{\text{rel}})$ (of the biochemical ODEs) is below a given threshold.
5) Approximate the maximal period $\hat{\tau}_{\text{matrix}} > 0$ such that the submatrices $\hat{\boldsymbol{S}}_{11} \in \mathrm{I\!R}^{r\times r}$, $\hat{\boldsymbol{S}}_{21} \in \mathrm{I\!R}^{(n-r)\times r}$, $\hat{\boldsymbol{S}}_{22} \in \mathrm{I\!R}^{(n-r)\times(n-r)}$ of the Jacobian $D\boldsymbol{g}(t_0 + \hat{\tau}_{\text{matrix}}) = \begin{pmatrix} \hat{\boldsymbol{S}}_{11} & \hat{\boldsymbol{S}}_{12} \\ \hat{\boldsymbol{S}}_{21} & \hat{\boldsymbol{S}}_{22} \end{pmatrix}$ satisfy upper threshold conditions of the following quantities:

- absolute quotient of the smallest eigenvalue of $\hat{\boldsymbol{S}}_{11} \in \mathrm{I\!R}^{r\times r}$ and λ_r (i.e. the smallest eigenvalue of corresponding submatrix $\tilde{\boldsymbol{S}}_{\text{slow}}$ of $D\boldsymbol{g}(t_0)$)
- absolute quotient of the largest eigenvalue of $\hat{\boldsymbol{S}}_{22} \in \mathrm{I\!R}^{(n-r)\times(n-r)}$ and $\lambda_{r+1} < 0$ (i.e. largest eigenvalue of corresponding submatrix $\tilde{\boldsymbol{S}}_{\text{fast}}$ of $D\boldsymbol{g}(t_0)$)
- quotient of the Euclidean norms $\left\|\hat{\boldsymbol{S}}_{21}\right\|_{\mathrm{I\!R}^{(n-r)\times r}} \Big/ \left\|\hat{\boldsymbol{S}}_{22}\right\|_{\mathrm{I\!R}^{(n-r)\times(n-r)}}$

Indeed, the relative error mentioned in step (4) gives the best impression how accurate the approximation $\boldsymbol{y}(t)$ is *after* its transformation back to physical quantities and thus, $\boldsymbol{T}\,\boldsymbol{y}(t)$ is compared with the "exact" solution $\boldsymbol{c}(t)$.
In our opinion, the matrix indicators mentioned in step (5) also play a crucial role for ILDM. The period $\hat{\tau}_{\text{matrix}} > 0$ indicates how long the time scale decomposition is appropriate. Here both the fastest component among the "slow" modes and the

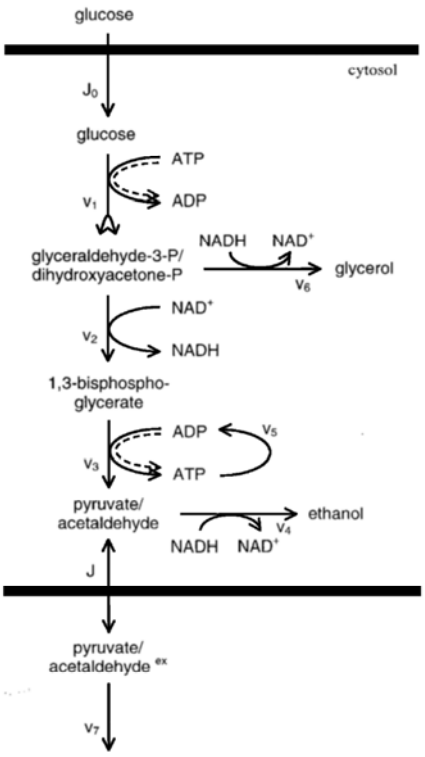
slowest component among of the "fast" modes are taken into account. For drawing any conclusions about time scales merely from $\hat{\boldsymbol{S}}_{11} \in \mathbb{R}^{r\times r}$ and $\hat{\boldsymbol{S}}_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$, though, the submatrix $\hat{\boldsymbol{S}}_{21} \in \mathbb{R}^{(n-r)\times r}$ of "coupling" should be close to 0 — as it was at time t_0 by construction. This leads to the third quantity investigated in step (5).

So finally, we would like to verify if $\hat{\tau}_{\text{rel}} > 0$ and $\hat{\tau}_{\text{matrix}} > 0$ can be regarded as "large" in comparison with $\tau_{r+1} = \frac{1}{|\text{Re}\,\lambda_{r+1}|}$.

4. Numerical Simulation

In order to illustrate our theoretical considerations, we computed and analyzed the time scale decomposition in the glycolysis reaction system in yeast as described in Wolf and Heinrich, 2000. The model includes the main steps of anaerobic glycolysis, and the production of ethanol and glycerol, as well as the effect of intercellular coupling. Depending on the kinetic parameters, the model shows both stationary and oscillatory behaviour.

The reaction network of a single cell is described in the scheme on the right–hand side (see Wolf and Heinrich, 2000), where J_0 is the input of glucose via the cellular membrane;



reaction 1	lumped reactions of hexokinase, phosphoglucoisomerase and PFK
reaction 2	glyceraldehydes-3-phosphate dehydrogenase reaction;
reaction 3	lumped reactions of phosphoglycerate kinase, phosphoglycerate mutase, enolase and pyruvate kinase;
reaction 4	alcohol dehydrogenase reaction;
reaction 5	non–glycolytic ATP consumption;
reaction 6	formation of glycerol from triose phosphates;
reaction 7	degradation of the coupling substance in the extracellular medium.

The model includes also the membrane transport of the coupling substance, characterized by the flux J.

The reaction rates are described by linear and bilinear functions of the concentrations, except for the first reaction (lumped reaction of hexokinase and PFK), where inhibition by ATP, according to the substrate inhibition of PFK, is additionally taken into account.

We refer to Wolf and Heinrich, 2000 for justifying the model. The differential equation system of the model for single cell reads:

$$\frac{dc_1}{dt} = J_0 - v_1 = J_0 - k_1 c_1 c_6 f(c_6)$$
$$\frac{dc_2}{dt} = 2v_1 - v_2 - v_6 = 2k_1 c_1 c_6 f(c_6) - k_2 c_2 (N - c_5) - k_6 c_2 c_5$$
$$\frac{dc_3}{dt} = v_2 - v_3 = k_2 c_2 (N - c_5) - k_3 c_3 (A - c_6)$$
$$\frac{dc_4}{dt} = v_3 - v_4 - J = k_3 c_3 (A - c_6) - k_4 c_4 c_5 - J$$
$$\frac{dc_5}{dt} = v_2 - v_4 - v_6 = k_2 c_2 (N - c_5) - k_4 c_4 c_5 - k_6 c_2 c_5$$
$$\frac{dc_6}{dt} = -2v_1 + 2v_3 - v_5 = -2k_1 c_1 c_6 f(c_6) + 2k_3 c_3 (A - c_6) - k_5 c_3$$
$$\frac{dc_7}{dt} = \varphi J - v_7 = \varphi J - k_7 c_7$$

With $f(c_6) = \left[1 + \left(\frac{c_6}{K_1}\right)^q\right]^{-1}$ and $J = \kappa(c_4 - c_7)$ (the kinetic constant κ is related to the permeability of the membrane for the coupling substance). The concentration of glucose is represented by the variable c_1, and that of 1,3–biphosphoglycerate by c_3.

c_5, c_6 and c_7 correspond to the ATP, NADH and coupling substance in the external solution, respectively, Owing to the fact that several glycolytic reactions are omitted and that other reactions are lumped, the model variables c_2 and c_4 denote the concentrations of pools of intermediates.

The kinetic parameters and initial values of concentrations used in this study are listed in Tables 1, 2.

Table 1. *Kinetic parameters for modelling glycolysis in yeast*

Parameter	*Value*	
J_0	3.0	$mM \cdot min^{-1}$
k_1	varied	$mM^{-1} \cdot min^{-1}$
k_2	6.0	$mM^{-1} \cdot min^{-1}$
k_3	16.0	$mM^{-1} \cdot min^{-1}$
k_4	100.0	$mM^{-1} \cdot min^{-1}$
k_5	1.28	min^{-1}
k_6	12.0	$mM^{-1} \cdot min^{-1}$
k_7	1.5	min^{-1}
κ	3.0	min^{-1}
q	4.0	
K_I	0.52	mM
N	1.0	mM
A	4.0	mM
φ	0.1	

Table 2. *Initial values of concentrations*

Metabolite	*Concentration*
c_1	6.2
c_2	0.8
c_3	0.18
c_4	0.32
c_5	0.1
c_6	2.5
c_7	0.01

Varying the parameter k_1 leads to different kinds of nonlinear behaviour which mimics the experimental observation. We explicitly studied the features for $k_1 = 10\ \mathrm{mM \cdot min^{-1}}$ and $k_1 = 50\ \mathrm{mM \cdot min^{-1}}$ corresponding to small amplitude oscillations

reaching the steady state (Figure 1) and high–amplitude oscillations (Figure 2), respectively.

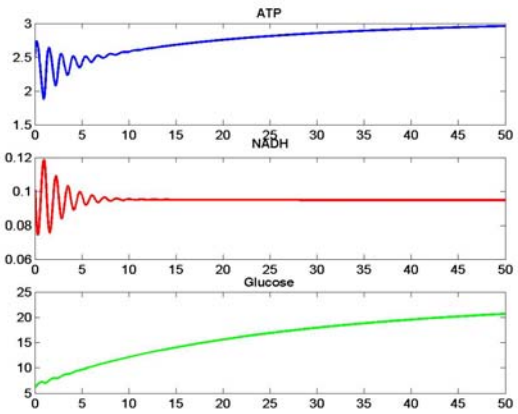


Figure 1. *Simulation of the full model for k_1= 50 mM min^{-1}*

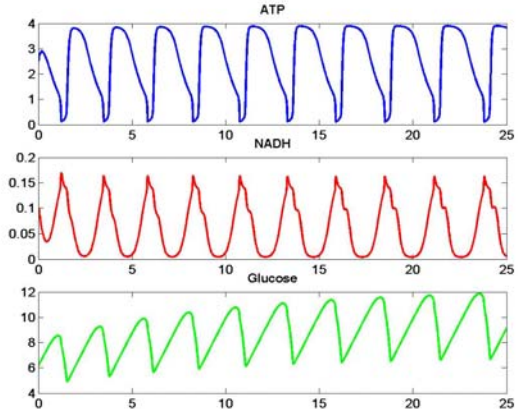


Figure 2. *Simulation of the full model for k_1 = 10 mM min^{-1}*

Table 3. Numerical analysis of glycolysis model for k_1= 50 mM min^{-1}: Number of "slow" modes *r*, matrix indicators (as in section 3.4) and contribution (in %) of metabolites to "slow" space.

time t_0	*r*	$\hat{\tau}_{\text{matrix}}/\tau_{r+1}$	$\hat{\tau}_{\text{rel}}/\tau_{r+1}$	c_1	c_2	c_3	c_4	c_5	c_6	c_7
0.7500	4	5.1200	1.0737	14.3623	19.5467	22.5031	5.9761	11.6605	15.3361	10.6152
1.2500	4	12.0000	1.0737	17.9475	18.5497	23.6323	4.2620	11.7120	10.7574	13.1390
1.7500	4	8.0000	1.0737	16.9367	18.9644	24.6826	3.9002	12.8088	13.7319	8.9754
2.7500	3	4.0960	1.0737	19.1841	23.3106	26.8022	2.9766	4.1741	16.7119	6.8406
13.7500	3	154.070	953.962	20.4577	17.8148	29.1398	3.6469	4.1409	22.3181	2.4818

Table 4. Numerical analysis of glycolysis model for k_1= 10 mM min^{-1}: Number of "slow" modes *r*, matrix indicators (as in section 3.4) and contribution (in %) of metabolites to "slow" space.

time t_0	*r*	$\hat{\tau}_{\text{matrix}}/\tau_{r+1}$	$\hat{\tau}_{\text{rel}}/\tau_{r+1}$	c_1	c_2	c_3	c_4	c_5	c_6	c_7
0.5000	4	1.0737	1.0737	34.4562	32.1218	0.9824	3.9576	10.7611	3.9561	13.7648
0.7500	3	1.0737	1.0737	33.7617	32.8314	1.1462	4.0381	7.9193	16.4806	3.8226
1.0000	3	1.0737	1.0737	13.0690	11.8107	18.6353	11.6513	16.7766	14.1726	13.9746
1.5000	5	1.0737	1.0737	13.3830	8.0924	11.9151	18.3115	19.2211	14.5981	14.4788
4.5000	5	1.0737	1.0737	13.7484	12.2284	10.7754	19.7374	17.9116	10.6344	14.9643

All simulations are performed using MATLAB library routines.

In Figure 1, the numerical simulations of the full ODE system are presented for $k_1 = 50\,\text{mM}\cdot\text{min}^{-1}$. First the system displays transient relaxation oscillations and

then, it settles into steady states. This specific situation offers the excellent possibility of observing qualitatively differing behaviour in complexity reduction on a single run. The results of analysis are shown in Table 3. The first column refers to the selected points of time, the second one indicates the numbers of "slow" modes according (only) to the criterion of Deuflhard and Heroth, 1996. The comparisons between the maximal periods $\hat{\tau}_{\text{matrix}}, \hat{\tau}_{\text{rel}}$ and the characteristic time scale τ_{r+1} (of the slowest "fast" mode) are listed in the third and fourth column, respectively. At the first 4 points of time, they are of the same order of magnitude. So there is no obvious reason why to neglect the "fast" modes. In contrast, the oscillations have almost decayed at the last point of time and thus, the indicators support applying the ILDM approximation.
In order to complete our analysis of "slow" space, Table 3 summarizes the contributions (in %) of each metabolite to slow modes. These conclusions are drawn from the components of the current transformation matrix $\boldsymbol{T}^{-1}$ as in Zobeley et al., 2005.
Figure 2 and Table 4 show the corresponding results for the kinetic parameter $k_1 = 10\,\text{mM}\cdot\text{min}^{-1}$. The situation is now dominated by the oscillating features of metabolites. In fact, the indicating periods $\hat{\tau}_{\text{matrix}}, \hat{\tau}_{\text{rel}}$ preserve the same order of magnitude as the characteristic time scale τ_{r+1} and thus, we doubt that reducing to dimension 3–5 can be justified by asymptotic analysis.

5. Conclusion

The ILDM approach to the decomposition of the original ODE system is based on the assumption that the main part of dynamics, being of real interest for the researchers, belongs to the intrinsic "slow" manifold. The fast period of the motion is of minor importance.
Nevertheless, many ODE systems of biochemical models do not meet this basic principle of ILDM. So an effective complexity reduction method should take both fast and slow parts of a trajectory into account. On the fast part, the fast variable changes only, whereas the slow one is constant. On the slow part of the trajectory, rates of changes of both variables are balanced (described by algebraic equations).
The presented study is a starting point for developing such an adaptive "combined" method. In this context, the time step providing "acceptable" precision will play a crucial role for deciding which modes to focus on. It ought to be chosen not only in terms of the initial linearization, but depends on the dynamic features of the transformation matrix. Our simulation of glycolysis might give useful hints for further improvements in this direction.

Acknowledgements. *The authors would like to thank the Federal Ministry of Education and Research (BMBF) and the Klaus Tschira Foundation (KTS) for financial support. I. Surovtsova thanks Th. Lorenz (Univ. Heidelberg, Germany)*

for continued help and very fruitful mathematical discussion, as well as for his patience.

References

Berridge, M.J., Bootman, M.D., and Lipp, P. (1998): Calcium – a life and death signal. *Nature,* **395**, 645–648.

Correa, C., Niemann, H., Schramm, B. and Warnatz, J. (2001): Reaction mechanisms reduction for higher hydrocarbons by the ILDM method. *Proc. Comb. Inst*., **28**, 1607–1614.

Deuflhard, P. and Bornemann, F. (2002): *Numerische Mathematik 2. Integration gewöhnlicher Differentialgleichungen*, 2nd edition, Springer.

Deuflhard, P., Hairer, E., and Zugck, J. (1987): One–step and extrapolation methods for differential–algebraic systems, *Numer. Math.,* **51**, 501–516.

Deuflhard, P. and Heroth, J. (1996): Dynamic dimension reduction in ODE models, in: Keil, F. et al. (Eds.), *Scientific Computing in Chemical Engineering*, Springer, 29–43.

Duysens, L.N.M., and Amesz, J. (1957): Fluorescence sprectrophotometry of reduced phosphopyridine nucleotide in intact cells in the near-ultraviolet and visible region. *Biochim. Biophys. Acta,* **24**, 19–26.

Field, J., and Noyes, R.M. (1974): Oscillations in chemical systems IV. Limit cycle behaviour in a model of a real chemical reaction. *J. Chem. Phys.,* **60**, 1877–1884.

Frenkel, R. (1968): Control of reduced diphosphopyridine nucleotide oscillations in beef heart extracts.I. Effect of modifiers of phosphofructokinase activity, *Arch. Biochem. Biophys.,* **125**, 151–156.

Golub, G.H. and van Loan, C.F. (1996): *Matrix computations*, Johns Hopkins University Press, Baltimore.

Heinrich, R. and Schuster, S. (1996): *The regulation of cellular systems*, Chapman and Hall, New York.

Hoppensteadt, F.C., Alfeld, P., and Aiken, R. (1981): Numerical treatment of rapid chemical kinetics by perturbation and projection methods. in: Ebert, K.H. et al. (Eds.), *Modelling of Chemical Reaction Systems.* Springer Series in Chem. Phys., vol. 18, 31–37.

Isbarn, G., Ederer, H.J., and Ebert, E.H. (1981): *The thermal decomposition of n-hexane: kinetics, mechanism and simulation.* in: Ebert, K.H. et al. (Eds.), *Modelling of Chemical Reaction Systems.* Springer Series in Chem. Phys., vol. 18, 235–248.

Kauffman, K.J., Pajerowski, J.D., Jamshidi, N., Palsson, B.O., and Edwards, J.E. (2002): Description and analysis of metabolic connectivity and dynamics in the human red blood cell. *Biophys. J.,* **83**, 646–662.

Lam, S.H. and Goussis, D.M. (1994): The CSP method for simplifying kinetics. *International Journal of Chemical Kinetics*, **26**, 461–486.

Maas, U. and Pope, S.B. (1992): Simplifying chemical reaction kinetics: Intrinsic low–dimensional manifolds in composition space. *Combustion and Flame,* **88**, 239–264.

Petty, H.R., Worth, R.G., and Kindzelskii, A.L. (2000): Imaging sustained dissipative patterns in the metabolism of individual cells. *Phys. Rev. Lett.,* **84**, 2754–2757.

Price, N.D., Reed, J.L., Papin, J.A., Famili, I., and Palsson, B.O. (2003): Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys. J.,* **84**, 794–804.

Reder, C. (1988): Metabolic control theory: a structural approach. *J. Theor. Biol,.* **135**, 175–201.

Schmidt, D., Blasenbrey, T. and Maas, U. (1998): Intrinsic low–dimensional manifolds of strained and unstrained flames. *Combust. Theory Modelling,* **2**, 135–152.

Wolf, J., Heinrich, R,, (2000): Effect of cellular interaction on glycolytic oscillations in yeast: a theoretical investigation. *Biochem. J.,* **345,** 321–334**.**

Zobeley, J., Lebiedz, D., Kammerer, J., Ishmurzin, A. and Kummer, U. (2005): A new time–dependent complexity reduction method for biochemical systems, in: Prami, C. et al. (Eds.), *Transactions on Computational Systems Biology*, Springer LNCS 3380, 90–110.

Transcriptional Regulation integrated into a Biodegradation database

Carbajosa, G.* Trigo, A., Valencia, A. and Cases, I.[§]

Centro Nacional de Biotecnología(CNB – CSIC) Madrid, Spain
[§]corresponding author icases@cnb.uam.es

Keywords: Biodegradation, Transcription Regalation, Databases, Environmental

1. Abstract

The increasing amount of information on the strains, compounds, enzymes, reactions and, what we are specially interested in, regulators implicated in microbial biodegradation of toxic pollutants provides us with the building blocks for formulating a "Global Biodegradation Network". We have created a relational database containing information both about the metabolism of several biodegradative pathways and their transcriptional regulation, covering transcriptional factors and their actions on promoters and operons implicated, thus integrating information on both metabolism and regulation features of the network. The information on regulation is extracted from the bibliography and it covers a range of 141 different species. By now, we have found data about 113 regulatory proteins, 113 promoters and 130 regulatory binding sites. Also, we have information on around 200 transcriptional complexes formed by the proteins, binding sites and effectors. All these complexes perform 353 registered different actions on promoters: inducing, inhibiting and repressing them. All this information is stored including the DNA and protein sequences and the genomic context, when available. Our data model supports very detailed molecular information, as well as more undefined regulatory mechanisms.

2. Introduction

Natural microbial communities have acquired the ability to degrade external chemical compounds that are beyond their Standard Metabolism, such as chemical pollutants (xenobiotics) that appear as side effects of industrial activity. Such communities are composed of a complex mixture of species and strains working co-ordinately. The final chain of reactions leading to chemical mineralization is frequently a puzzle of reactions carried out by enzymes from several species. From a biological perspective this biodegradation network presents very interesting properties that differentiate it from the standard metabolic pathways, i.e., it has an inter-species composition, and it is the result of a fast adaptation to new environmental conditions. Interestingly, this can also be seen as a new scenario for analysis with a Systems Biology viewpoint,

offering possibilities different and complementary to the study of the organization and evolution of classical metabolic networks. The first steps in order to understand this complexity have been the study of the general properties of the known biodegradation network (Pazos et al., 2004). However the analysis of this metabolic network is not sufficient to understand its behaviour in the natural context. The presence of a gene that codifies for a specific enzymatic activity does not guarantee the presence of that enzymatic activity, that gene needs to be expressed in enough quantity, this being determined by the environmental conditions, and this expression is often integrated in layers of iterative regulatory networks that ensure the performance not only of the whole cell, but also of the bacterial population, and even the microbial community, in a changing environment (Cases and de Lorenzo, 2005). Therefore, the understanding of how specific regulation, in response to a given substrate, and superimposed levels of regulation, determined by the physiology of the cell or the presence of alternative carbon sources, is a must if we aim to describe the behaviour of the biodegradation network in the presence of different stimuli as it happens in the real microbial communities in the environment. So, we present here a database of regulatory elements of biodegradation pathways. This database includes specific and general transcription factors, their binding sites in the DNA, the organization of biodegradative operons and their promoters, and the conditions under these two elements interact giving as result transcription regulation. Some preliminary analysis on the collected data is also presented and discussed here. All this information is linked to the metabolic data already available, and together offer a extraordinary resource for the understanding on how biodegradation of xenobiotic compounds occurs in the environment.

3. Results and Discussion

As a first approach for data collection, regulatory information was extracted from the bibliography, taking as a starting point two reviews on the subject (Tropel et al., 2000, Diaz and Prieto, 2000) and searches for each regulatory protein, its regulated genes and the references available in the Genbank and Geneprotein files through Pubmed. All this data has been integrated with that extracted from the Minnesota Biodegradation database (Ellis et al., 2003) and the protein and gene sequences of the enzymes that were available from Genbank and that were searched in the literature. At this moment The transcriptional regulation database covers more than 600 genes where199 regulatory complexes, including the protein and gene sequences (up to 110) of their components and also of their target genes; 132 effectors, 130 binding sites and 113 mapped promoters for 196 operons (all of them traceable to coordinates in a Genbank entry) and the genomic context when it was available. All the structural information that was disposable in PDB and SCOP was included too as well as the ontology terms found in GO. Nevertheless, all this numbers represent an estimation as the regulatory framework is still under development, but still covers a significant part of the about 1025 reactions in 144 pathways described among 147 micro

organisms All the information about metabolism and regulation has been stored in a database that will be available through a web server called BioNeMo (Biodegradation Network Modelling).

We have performed preliminary analysis of some properties of the biodegradation transcription regulation mechanisms. We first analysed the Transcription Factors Binding sites (TFBS). The average length of the binding sites for transcriptional regulators is somewhere between 16 and 20 nucleotides as is typical of bacterial transcription factors (Fig. 1a). Regarding the position of TFBS relative to the transcription start site, we observed a wide distribution with TFBS, which some as far as 250 bp upstream and some interesting ones sited 250 bp downstream, with a peak between -25 and -50, and a second smaller one between -150 and -175.

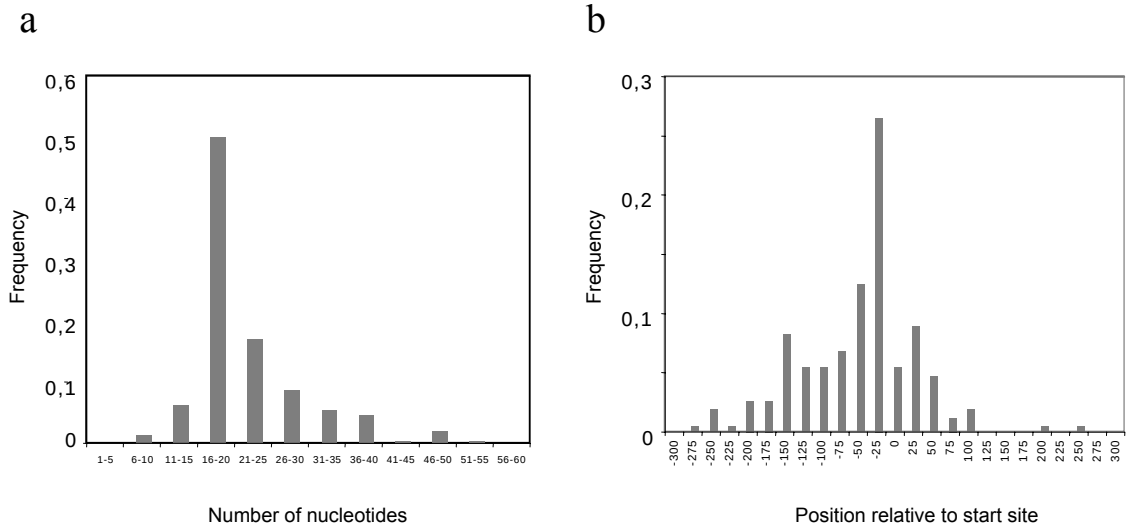


Figure 1. Distribution of length and position of Transcription Factors Binding sites. a) nucleotide length of the binding sites versus the frequency (number of sites/total number of sites). b) position of the binding sites relative to the transcription start site versus the frequency (number of sites/total number of sites).

The distribution of locations can be interpreted in terms of regulatory interaction: previous research into what is the position of a regulator binding site on the DNA relative to the transcription start site found that it is one of the factors that determine the protein regulatory function (Collado-Vides et al., 1991). The distance to the transcriptional start site shows that the regulators that act as repressors bind upstream close to the start site (from –45 to 40), maybe because most of them are able to disturb the transcriptional initiation by blocking the access of the polymerase. Most of the activators bind upstream related to the repressors (specially from –40 to –80, but also further), probably because they can interact better with the polymerase around those positions or maybe to drive changes in the DNA conformation that allow a better interaction with the RNA polymerase. Around –150 we can find the factors interacting with sigma54 that need to perform a loop in the DNA to interact with the polymerase and thus they need to be this far from the transcriptional initiation site to do so. The frequencies

observed fit well with the distribution of regulators and repressors present in the biodegradation regulatory network (Table 1). We found roughly that two thirds of biodegradation regulators are activators, and one third repressors. Interestingly this is significantly different from the *E. coli* transcription network, where repressors and activators are more or less equally represented. This difference between *E. coli* and the biodegradation network is less pronounced if we consider regulatory actions (single event of regulation between a transcription factor and a promoter). The prevalence of activation in the Biodegradation Network could be related with the biochemistry of the biodegradation metabolism. Biodegradative pathways normally involve large number of enzymes and many reactions that are expensive energetically, it is tempting to hypothesise that activation can be favoured as a regulatory mechanism, since mutation in the regulator or the binding site would not lead to an spurious expression of the pathway, as it would happen in the case of repression.

Table 1. Transcription Factors and their actions in the Biodegradation Regulatory Network.

Regulators	Activators	Repressors
Biodegradation	71 (**65%**)	38 (**35%**)
E. coli	101 (**50%**)	103 (**50%**)
Actions	Activation	Repression
Biodegradation	216 (**61%**)	137 (**39%**)
E. coli	1563 (**56%**)	1206(**43%**)

*E. coli data have been obtained from RegulonDB (Salgado et al., 2005)

We have also look at the interaction between transcription factors and effectors molecules (Fig. 2). This so called effectors are molecules that interact with the regulators driving changes that perform actions on promoters, such as inducing or repressing their expression Most of regulators interact with only a few molecules, 75% of them with only 3 or less. Interestingly a few regulators are able to interact with up to 9 or 10 different molecules (Fig. 2a).

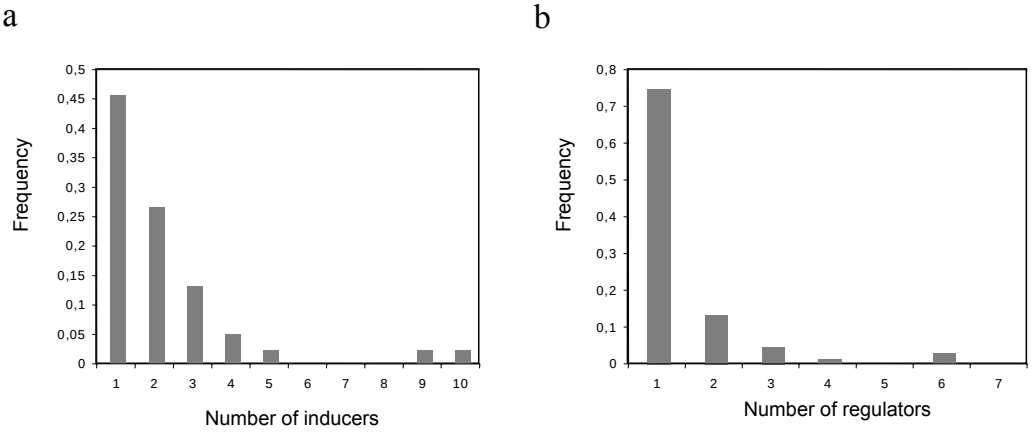


Figure 2. Interaction between Regulators and inducers. a) number of inducers versus the frequency (number of regulators that interact with n inducers/total number of regulators). b) number of regulators versus the frequency (number of inducers that interact with n regulators/total number of inducers).

This fact could be explained in the context of the 'regulatory noise hypothesis' which proposes that "transcriptional control systems develop responsiveness to new signals due to the leakiness and lack of specificity of pre-existing promoters and regulators. When needed, these may become more specific through suppression of undesirable signals and further fine-tuning of the recruited proteins to interact with distinct chemicals" (de Lorenzo V. et al 1996). Similarly most of the molecules are able to interact only with one or few regulators. However, there are a few molecules are able to interact with many transcription factor, and thus are predicted to strongly effect the expression of the biodegradation enzymes when introduced in the environment. These molecules could thus used in the remediation process of polluted sites, since they will increase the biodegradation potential of the local microbial community.

4. Conclusion

The complex transcriptional regulation behind the biodegradation network is still far from being totally understood. As the first analysis show, it seems that there are similarities in its behaviour compared to single organisms systems (as *E coli*), but the same analysis show differences, like the higher ratio of activators in the biodegradation set. We have also shown that this resource also allows to ask question which answer can help in the better understanding of the properties and evolution of the regulatory mechanism involved in the biodegradation, and also open novel approaches for the application of bioremediation technology. Our effort at this moment is focused on the integration between the metabolic network and all the information we have been collecting about regulation to try to understand the mechanism of the integration between regulation and metabolism, and the forces that are driving the biodegradation processes in microbial communities.

Acknowledgements.
This work have been supported by the Fundación Banco Bilbao Vizcaya Argentaria (BIOCON-3) and The Spanish Ministry of Education and Science (BIO2004-03512). The authors want to thank Prof. de Lorenzo for his support and stimulating discussion.

References

Cases, I. and de Lorenzo, V. (2005) Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol*. **3**,105-18.

Collado-Vides, J., Magasanik, B. and Gralla, J.D.(1991) Control site location and transcriptional regulation in Escherichia coli. *Microbiol Rev*. **55**, 371-94.

Diaz, E. and Prieto, M.A. Bacterial promoters triggering biodegradation of aromatic pollutants. (2000). *Curr Opin Biotechnol*. **11**, 467-75.

Ellis, L.B., Hou, B.K., Kang, W. and Wackett, L.P. (2003). The University of Minnesota Biocatalysis/Biodegradation Database: Post-Genomic Datamining. *Nucl. Acids Res*. **31**, 262 -265.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A. and Collado-Vides, J. (2005) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*. **34,** D394-D397.

de Lorenzo, V. and Perez-Martin, J. (1996) Regulatory noise in prokaryotic promoters: how bacteria learn to respond to novel environmental signals. *Mol Microbiol*. **19**, 1177-84.

Tropel, D. and van der Meer, J.R.(2004) Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol Mol Biol Rev.* **68**, 474-500.

The Linkage between Flux Distributions and Elementary Modes Activity Patterns: An Interval Approach

F. Llaneras∗[a], J. Picó [a]

[a]*Dept. of Systems Engineering and Control, Technical University of Valencia, Valencia, Spain, e-mail: {frallaes@doctor.upv.es, jpico@ai2.upv.es}*

Keywords: Elementary modes, Extreme pathways, α-spectrum, Metabolic flux analysis.

1. Abstract

In this paper a new approach to determine the α-spectrum is presented. The approach is based on the use of an interval representation of fluxes, making it possible to compute α-spectrum from an uncertain or even partially unknown flux distribution. In addition, as a complement of metabolic flux analysis, a new method is proposed that allows the calculation of the ranges of possible values for each non-calculable flux. The presented methods are illustrated with the example of CHO cells.

2. Introduction

This work is focused on mathematical methods for translating a metabolic flux distribution into an elementary modes or extreme pathways activity pattern. These methods determine how much flux is being carried by each e. mode or e. pathway under some particular set of circumstances. Hence, the pooly informative flux distribution can be translated into a simpler and more meaningful representation. Unfortunately, this translation has not a unique solution but a range of solutions. Thus, two options are possible: choosing a particular solution (Poolman et al., 2004; Schwarz et al., 2005), or dealing with the whole solutions region. When choosing one solution, the validity of the obtained activity pattern depends on the validation of the underlying assumptions. Following the second option, the α-spectrum, the range of possible values for each e. mode or e. pathway activity, can be determined (Wiback et al. 2003).

Herein, a new approach that allows determining the α-spectrum when fluxes are represented with an interval is presented. This representation is useful when a) flux measurements are uncertain, and b) when some non-measured fluxes cannot be uniquely determined (Klamt et al., 2002). In addition, a method to flux

∗ Corresponding author

calculation is presented as a complement of metabolic flux analysis (MFA). In many cases, when using MFA, the resulting system is undetermined and the complete flux distribution cannot be computed. In these cases, by using a similar procedure to the one used to determine the α-spectrum, it is possible to calculate the ranges of possible values for each non-calculable flux.

3. Theoretical

A biological network can be represented with a stoichiometric matrix *N*, where rows correspond to the *m* metabolites and columns to the *n* reactions. Including irreversible reactions as v_i, the mass balance of the network at steady state (Stephanopoulos et al., 1998) can be formulated as:

$$N \cdot v = 0 \qquad v_i \geq 0 \tag{1}$$

In general, as *n* is bigger than *m*, the system is undetermined. Nevertheless, the solution region can be spanned by convex combination with e. modes or e. pathways:

$$v_m = E \cdot \alpha \qquad \alpha_i \geq 0 \tag{2}$$

Where v_m is a flux distribution, *E* denotes the matrix formed with each e. mode or e. pathway as a column and α is a vector representing the non-negative activity for each e. mode or e. pathways. Despite differences between e. modes and e. pathways (Papin et al., 2004), the proposed methods can be applied in both cases, and therefore from this point only the term e. mode will be used.

3.1. TRANSLATING A FLUX DISTRIBUTION INTO A E. MODES ACTIVITY PATTERN

System (2) can be analyzed using the procedure proposed in (Klamt et al., 2002). The number of e. modes *e* is always bigger or equal than *n-m,* the number of linear independent vectors needed to span the solution region. Therefore the rank of *E* is equal to *n-m*. When *e=n-m* the system is exactly determined, and the unique solution can be calculated by using E^{-1}. But in general *e>n-m*, and the system is undetermined with *e-(n-m)* degrees of freedom. Then, the general solution of (2) can be considered:

$$\alpha_G = \alpha_p + K(E) \cdot \lambda \qquad \alpha_i \geq 0 \tag{3}$$

Where α_p denotes a particular solution, *K(E)* the null space of *E* and λ an arbitrary vector representing the indeterminacy of equation. Thus, only such elements α_{Gi} of α_G whose corresponding row in *K* is a null row, are determined (its value can be taken from the non-negative least squares solution).

3.1.1. α-spectrum: The interval Approach

In (Wiback et al., 2003) the concept of α-spectrum is defined to work with the solution region. Basically, 2**e* linear programming problems are solved to compute the range of possible values for each e. mode activity. Here, a slight modification of the method makes it possible to compute the α-spectrum when the fluxes are represented as an interval:

$$\begin{aligned} &\forall \alpha_j, \min/\max\{\alpha_j\} \qquad j \in [1,e] \\ &subject: \qquad E\cdot\alpha \le v^+ \quad E\cdot\alpha \ge v^- \quad \alpha_i \ge 0 \end{aligned} \tag{4}$$

where v^+ and v^- are vectors with extreme values for each flux. The interval representation implies reducing the restrictions of the problem, and therefore the solution ranges will be bigger. Nevertheless, if the interval representation is well justified, the obtained solution will be less precise, but more realistic.

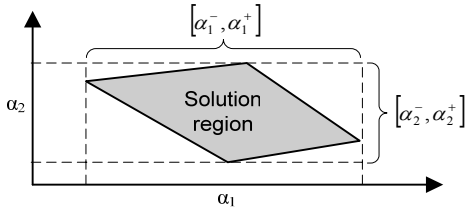


Figure 1. The α-spectrum.

This method makes it possible to compute the α-spectrum in two common situations: a) when the flux distribution is uncertain and b) when it is partially unknown. Additionally, it provides a straight method for dealing with inconsistency: Only if the flux region, defined with an interval, contains one consistent flux distribution, the linear programming problem has a solution. In figure 2, the different representations of fluxes based on certainty, consistency and completeness are summarized.

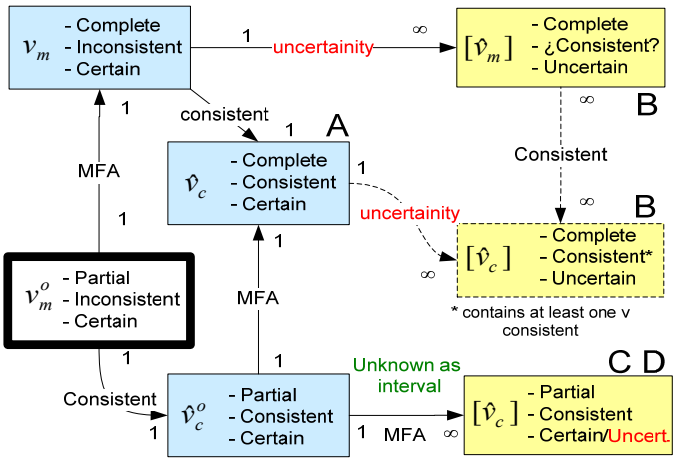


Figure 2. Fluxes as intervals.

3.2. METABOLIC FLUX ANALYSIS AND THE FLUX SPECTRUM

Although intracellular fluxes can be measured *in vivo* with tracer experiments (Sauer, 2004), there are several situations where these techniques are not suitable. In these cases, MFA can be used to calculate intracellular fluxes by using a set of measured fluxes and applying mass balances around metabolites (Stephanopoulos et al., 1998). Basically, making a partition between measured (subindex *m*) and unknown fluxes (subindex *u*), equation (1) can be transformed into:

$$N_u \cdot v_u = -N_u \cdot v_u \tag{5}$$

Following (Klamt et al., 2002), the determinacy and the redundancy of (1) can be analyzed. If the system is determined, a unique solution can be computed; nevertheless, very often it is necessary to deal with underdetermined systems, where some fluxes cannot be uniquely computed (Klamt et al., 2002).

3.2.1. Flux-spectrum

To deal with these undetermined systems, a new approach is proposed that allows the calculation of the ranges of possible values for each non-calculable flux, resulting in a region that could be termed flux-spectrum. Again, these ranges can be obtained by solving a set of linear programming problems:

$$\begin{aligned} &\forall v_{uj}, \min/\max\{v_{uj}\} \qquad j \in [1, nu] \\ &subject: \qquad N_u \cdot v_u = -N_m \cdot v_m \qquad v_i \geq 0 \end{aligned} \tag{6}$$

Thus, when some fluxes cannot be calculated, the flux-spectrum provides a method to compute its ranges of values. Obviously, it is also possible to compute the flux-spectrum when the know fluxes are represented with an interval (as a previous step the extreme values of $-N_m \cdot v_m$ need to be calculated).

4. Results

In (Llaneras et al., 2006), the presented methods have been applied to the central metabolism of CHO cells (Provost and Bastin, 2004). Including a *6x18* matrix *P* linking extracellular fluxes with intracellular ones, the extended system has 16 metabolites (*me*) and 22 reactions (*ne*).

4.1. A-SPECTRUM AND. PARTIAL KNOWLEDGE

For example, when only v_1 *(G),* v_{21} *(CO_2)* and v_{20} *(Q)* are measured, the system is undetermined: the rank of *Nu* (16) is less than the number of unknown (18). Therefore the complete flux distribution cannot be determined by using MFA. Nevertheless, even when the flux distribution is partially unknown, the α-spectrum can be computed by using the method presented in 3.1.1 (interval fluxes are given in Table 9).

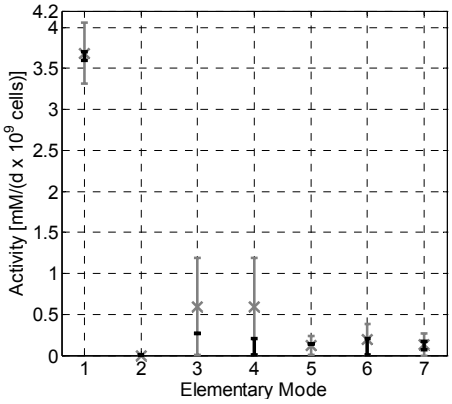


Figure 3. α-spectrum computed from the complete flux distribution (●) and from a incomplete one (x).

Table 1. Partially unknown flux distribution represented as a set of intervals (nM/(d x 10^9 cells)).

v_1 *(G)*	v_2- v_{19}	v_{20} *(Q)*	v_{21}*(CO_2)*	v_{22}
4.4305	[0,∞*]	1.186	2.5574	0

4.2. FLUX-SPECTRUM

As the system is undetermined, at least one flux cannot be uniquely determined. Moreover, there is not any calculable flux (matrix *K*, the kernel of *N*, has no null rows). Nevertheless, by using the concept of flux-spectrum it is possible to calculate the range of possible values for each non-calculable flux.

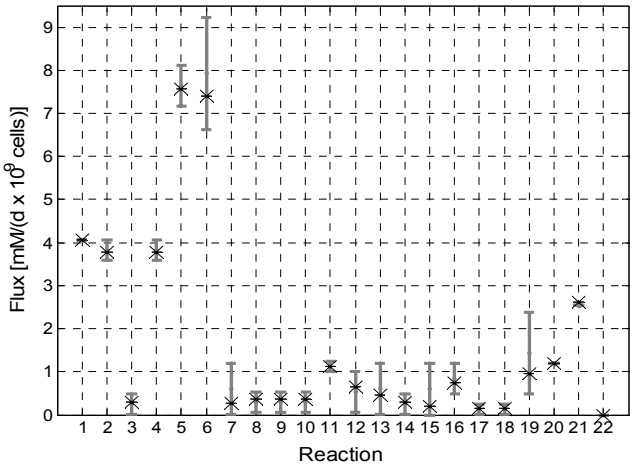


Figure 4. Exact flux distribution (x) and flux-spectrum computed from an partial flux distribution (●).

Moreover, if a unique and exact value is needed and depending on the size of the ranges, the use of each range middle point as an estimation is a sensible approach.

5. Conclusion

The translation of a metabolic flux distribution into an e. modes or e. pathways activity pattern has been investigated. A new approach to determine the α-spectrum was presented. Additionally, a method to calculate the ranges of possible values for non-calculable fluxes was proposed, as a complement to MFA.

Acknowledgements. *This research has been partially supported by the Spanish Government (CICYT-FEDER DPI2005-01180). First author is recipient of a fellowship from the Spanish Ministry of Education and Science (FPU AP2005-1442).*

References

Klamt S, Schuster S, Gilles ED. (2002) Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnol. Bioeng*. **77**(7):734-751.

Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO. (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol*. **22**(8):400-405.

Poolman MG, Venkatesh KV, Pidcock MK, Fell DA. (2004) A method for the determination of flux in e. modes, and its application to *Lactobacillus rhamnosus*. *Biotechnol. Bioeng*. **88**(5):601-612.

Provost A and Bastin G. (2004). Dynamic metabolic modelling under the balanced growth condition. *J. Process Control* **14**(7):717-728.

Llaneras F, Picó J. (2006) The Linkage between flux distributions and elementary modes activity patterns: an interval approach. Internal Report. DISA, Technical University of Valencia.

Sauer U. (2004) High-throughput phenomics: Experimental methods for mapping fluxomes. *Curr. Opin. Biotechnol*. **15**(1):58-63.

Schwarz R, Musch P, von Kamp A, Engels B, Schirmer H, Schuster S, Dandekar T. (2005) YANA - a software tool for analyzing flux modes. *BMC Bioinformatics* **6**(1):135.

Stephanopoulos GN, Aristidou AA, Nielsen J. (1998) *Metabolic engineering: Principles and methodologies*. San Diego: Academic Press.

Wiback SJ, Mahadevan R, Palsson BO. (2003) Reconstructing metabolic flux vectors from e. pathways: Defining the alpha-spectrum. *J. Theor. Biol*. **224**(3):313-324.

Towards a rational approach to metabolic engineering: Indirect Optimization Methods

Marin-Sanguino A. [a], Gonzalez-Alcón C. [b], Voit E.O.[c], and Torres N.V.[dr]

[a]Dpto. Bioquimica y Biologia Molecular, Universidad de La Laguna, Spain, e-mail: amarin@ull.es
[b]Dpto. Estadistica, Investigacion Operativa y Computacion, Universidad de La Laguna, Spain, e-mail: cgalcon@ull.es
[c]Dept. of Biomedical Engineering, Georgia Institute of Technology, USA, e-mail: eberhard.voit@bme.gatech.edu
[d]Dpto. Bioquimica y Biologia Molecular, Universidad de La Laguna, Spain, e-mail: ntorres@ull.es

Keywords: Biochemical Systems Theory, Optimization, IOM, Geometric Programming.

1. Abstract

In this work we will present the family of Indirect Optimization Methods (IOM). Among the advantages of such methods are its computational efficiency and versatility. IOM are not limited to biochemical models and has also been applied to systems where also genetics or industrial modelling are involved (Marin-Sanguino and Torres 2000). The main drawback of these methods are the need to approximate some of it's constraints as single term power laws (monomials) in order to cope with them, which sometimes leads to violations of some constraints. As a possible answer to this problem, Geometric IOM is presented. This variant uses geometric programming instead of linear programming as the core solver in order to add flexibility to the standard linear IOM methods. The new method is applied to a simple theoretical model in order to illustrate the concepts involved and show some of its possibilities.

2. Introduction

One of the driving forces of systems biology is the need to understand biosystems as deeply as possible in order to obtain technological applications. Among the technologies that might arise from a better understanding of biological complexity, we find biotechnology and biomedicine. In these and many other applications, the aim is to modify the behaviour of part or a whole biological system in order to comply with certain specifications such as

r Corresponding author

metabolite overproduction, resistance to environmental conditions or operation within some limits that are considered to be healthy.
In this endeavour, mathematical modelling based on high throughput data is a promising field that has received a great deal of attention. But biological models are often big and complex, which difficults their analysis. The optimal way of influencing their behaviour tends to be counter intuitive and far from trivial. The development of frameworks such as Biochemical Systems Theory (BST) (Savageau, 1969a-b, Voit 2000) and Metabolic Control Analysis has always paid a special attention to this problem, leading to different techniques for model driven improvement.
The steadily increasing availability of good models of biotechnologically relevant systems opens the possibility of rational redesign of industrial strains to improve their performance. In order to achieve this, mathematical methods for the analysis of big non-linear systems are needed. Any method devised to deal with biological systems should have some features:

1) It should scale well as the number of variables grows.
2) It must be general enough to cope with the wide range of different components (metabolites, RNA, proteins,...)
3) Quality tests should be provided to evaluate the viability and reliability of solutions.

Among such are the Indirect Optimization Methods (IOM) (Voit 1992, Torres and Voit 2002). The aim of IOM is to find a steady state for the system that optimizes it's performance -usually maximizing a flux or flux ratio- while keeping the steady state and guaranteeing cell viability.
IOM methods search for a solution of model which rely on the regular structure of s-systems, a formalism within Biochemical Systems Theory that

$$\dot{x}_i = \alpha_i \prod x^{g_{i,j}} - \beta_i \prod x^{h_{i,j}}$$

in the steady state can be rearranged as

$$\frac{\alpha}{\beta} \prod x^{g_{i,j}} - x^{h_{i,j}} = 1$$

Through a logarithmic transformation, these equations become linear which enables them to be used as constraints for linear programming. Together with the steady state constraints, IOM approaches allow setting further limits to the solutions such as establishing a range of variation to prevent some variables or fluxes to move too far away from their basal values, which can jeopardize cell viability. It is metabolic burden. Such burden can be considered of two sorts. On one hand, the total amount of protein present in the cell can be growth limiting. The cost of producing too much protein might be the cause for the instability of some mutations and the trend to lose some plasmids. On the other hand, the pool of metabolites must be added to the cell content in protein and other macromolecules in order to consider osmolarity issues and the activity of water.

For these reasons upper limits are often set to these or any other pools. As they are not linear in logarithmic coordinates, these constraints have to be approximated through power laws. The main problems when using IOM arises precisely due to these approximated constraints, as the obtained steady state is that of the approximated problem and not the original one. Therefore, IOM solutions are not global when applied with approximated constraints and sometimes lead to violations of the original constraints when the IOM solution is implemented in the original system. These difficulties have been addressed by dividing the problem in two linear phases (Marin-Sanguino and Torres 2003), iterative methods (Marin-Sanguino and Torres 2000) and using additional constraints to estimate error (Xu, Shao and Xiu, 2005). These issues can be more effectively overcome due to the application of geometric programming.

3. Theoretical

Geometric programming (GP) problems have the following structure:

Min $P_0(x)$
Subject to:
$P_i(x) < 1$ i=1 ... n
$Q_i(x) = 1$ i= 1...m (1)
$Q_i(x)$ must be monomials , also known as power law terms:
$Q(x)= a\, x_1^{b1}\, x_2^{b2} \ldots x_n^{bn}$ (2)

where all coefficient and exponents are positive real numbers (Zener, 1971).
$P_i(x)$ are posynomials, consisting in sums of several of the above mentioned monomials. Geometric programming shares many advantages with linear programming as they are both convex programming problems. Geometric programs can be solved quite efficiently and solvers are starting to become available (Koh et al 2006). For example, a geometric program of 1000 variables and 10000 constraints can be solved in less than a minute on a desktop computer (Boyd et al 2006). and it gets even better with sparse problems such as those found in metabolic engineering. Besides, optimal solutions found in GP are global (Boyd and Vanderberghe 2004).

3.1. Case study

We will illustrate the working of Geometric programming with a simple model of some genes coding for a linear pathway with feedback regulation on the reactions and gene expression. We will use a purely theoretical model with arbitrary units in which every variable is normalized such that its steady state value is one.

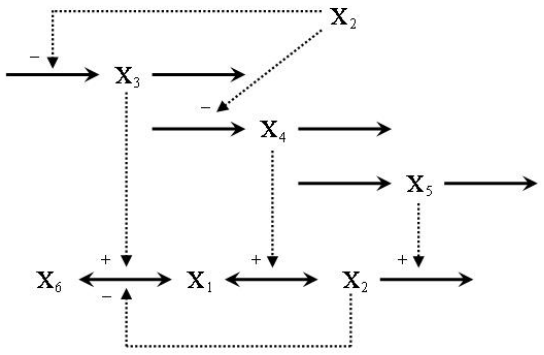


Figure 1. A linear pathway with the expression and regulation of its corresponding genes.

$$\dot{x}_1 = 60\, x_1^{-0.1}\, x_2^{-0.6}\, x_3\, x_6^{0.5} - 60\, x_1^{0.3}\, x_2^{-0.1} x_4$$
$$\dot{x}_2 = 60\, x_1^{0.3}\, x_2^{-0.1} x_4 - 60\, x_2^{0.7}\, x_5$$
$$\dot{x}_3 = 10\, x_7\, x_2^{-0.3} - 10\, x_3$$
$$\dot{x}_4 = 10\, x_8\, x_2^{-0.3} - 10\, x_4$$
$$\dot{x}_5 = 10\, x_9 - 10\, x_5$$

4. Results and Discussion

The optimization of our test system, under conditions of a maximum total enzyme five times the basal level, yields the solutions presented in table 1. Applying the traditional IOM method, an approximate solution is obtained that, not being the global optimum, yields the same flux with a negligible violation of the total enzyme constraint. Geometric IOM solution, not being too distant has been obtained in the same computation time and is guaranteed to be the global maximum flux attainable. Furthermore, the swiftness of IOM methods allows the possibility of solving the problem repeatedly for different constraint limits. Fig 2 shows the result of varying the maximum value of total metabolites pool. This curve can be used as a guide when the total pool is suspected to be critical and yields the exact cost of reducing its level . It also shows clearly the futility of relaxing this limit beyond it's basal level of two as no improvement is expected from such relaxation.

Table 1. Results of optimizing the model with traditional and geometric IOM.

	IOM	GIOM
x1	1.2	1.2
x2	0.8833	0.9934
x3	4.7844	5.1315
x4	4.7326	4.7863
x5	5.5205	5.0822
x6	1	1
x7	4.6095	5.1213
x8	4.5597	4.7769
x9	5.5205	5.0822
total enzyme	15.0376	15
Flux	303.6687	303.5257

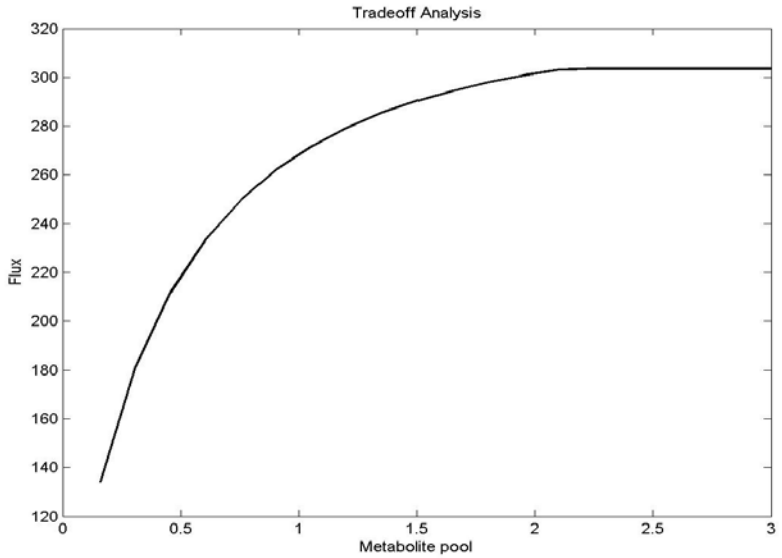


Figure 2. Substrate Optimal flux for different limits on the metabolites pool.

5. Conclusion

The IOM provide the versatility and speed needed for systems level analysis of biological systems. Their geometric variant keeps the advantages of the old versions while adding a more flexible structure and the posibility of obtaining global solutions in cases in which it was previously impossible. Further extensions of this methods such as iterative application of the geometric method or the use of recasting as a tool for optimizing a wider set of non-linear systems point to GIOM as a strong candidate for optimization in metabolic engineering.

References

Marin-Sanguino, A and Torres NV. (2003) Optimization of biochemical systems by linear programming and general mass action model representations. *Math Biosci*. **184**(2):187-200.

Marin-Sanguino, A. and Torres NV. (2000) Optimization of tryptophan production in bacteria. Design of a strategy for genetic manipulation of the tryptophan operon for tryptophan flux maximization. *Biotechnol Prog*. **16**(2):133-45.

Boyd, S., Kim, SJ., Vandenberghe, L. and Hassibi. (2006) A tutorial on geometric programming. To be published in *Optimization and Engineering*.

Boyd, S. and Vandenberghe, L.(2004) *Convex Optmization*. Cambridge University Press.

Koh, K., Kim, S., Mutapic, A. and Boyd, S. (2006) GGPLAB: A simple Matlab toolbox for Geometric Programming. version 0.95.

Savageau, M.A. (1969a) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol*. **25**(3):365–9.

Savageau, M.A. (1969b) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J Theor Biol*, **25**(3):370–9.

Torres, NV. and Voit, EO. (2002) *Pathway Analysis and Optimization in Metabolic Engineering*. Cambridge University Press.

Voit, E.O. (1992) Optimization in integrated biochemical systems. *Biotechnol. Bioeng*. **40**:572–582.

Voit EO. (2000) *Computational analysis of biochemical systems*. Cambridge University Press.

Xu, G. Shao, C. and Xiu Z. (2005). A Modified Iterative IOM Approach for Optimization of Biochemical Systems. eprint arXiv:q-bio/0508038.

Zener, C.M. (1971) *Engineering Design by Geometric Programming*. John Wiley and Sons, Inc.

Bioconductor project an open software development for computational biology and bioinformatics in Africa

Okoye U. Okwudili, Oseghale Lucky Edosa, Omoike Maliki Momodu

Department Microbiology, Faculty of Science Ambrose Alli, University, EKPOMA, Nigeria.
Department Biostatistical Science, Faculty of Science Ambrose Alli, University, EKPOMA, Nigeria.
Department Bioinformatics and Molecular Biostatistics, Faculty of Science, Lagos State University.

1. Abstract

The Bioconductor project in Africais is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics (CBB). Biology, molecular biology in particular, is undergoing two related transformations. First, there is a growing awareness of the computational nature of many biological processes and that computational and statistical models can be used to great benefit. Second, developments in high-throughput data acquisition produce requirements for computational and statistical sophistication at each stage of the biological research pipeline. The main goal of the Bioconductor project is creation of a durable and flexible software development and deployment environment that meets these new conceptual, computational and inferential challenges. We strive to reduce barriers to entry to research in CBB. A key aim is simplification of the processes by which statistical researchers can explore and interact fruitfully with data resources and algorithms of CBB, and by which working biologists obtain access to and use of state-of-the-art statistical methods for accurate inference in CBB, This paper describe details of our aims and methods, identify current challenges, compare Bioconductor to other open bioinformatics projects, and provide working examples.

2. Introduction

The Bioconductor project in Africais IS an initiative for the collaborative creation of extensible software for computational biology and bioinformatics (CBB). Biology, molecular biology in particular, is undergoing two related transformations. First, there is a growing awareness of the computational nature of many biological processes and that computational and statistical models can be used to great benefit. Second, developments in high-throughput data acquisition produce requirements for computational and statistical sophistication

Understanding and Exploiting Systems Biology

at each stage of the biological research pipeline.The main goal of the Bioconductor project in Africa is creation of a durable and flexible software development and deployment environment that meets these new conceptual, computational and inferential challenges, We strive to reduce barriers to entry to research in CBB. A key aim is simplification of the processes by which statistical researchers can explore and interact fruitfully with data resources and algorithms of CBB, and by which working biologists obtain access to and use of state-of-the-art statistical methods for accurate inference in CBB.
Among the many challenges that arise for both statisticians and biologists in Africa are tasks of data acquisition, data management, data transformation, data modeling, combining different data sources, making use of evolving machine learning methods, and developing new modeling strategies suitable to CBB. We have emphasized transparency, reproducibility, and efficiency of development in our response to these challenges. Fundamental to all these tasks is the need for software; ideas alone cannot solve the substantial problems that arise.

Transparency. High-throughput methodologies in CBB are extremely complex, and many steps are involved in the conversion of information from low-level information structures to statistical databases of expression measures coupled with design and covariate data, Credible work in this domain requires exposure of the entire process.

Pursuit of reproducibility. Experimental protocols in molecular biology are fully published lists of ingredients and algorithms for creating specific substances or processes. Accuracy of an experimental claim can be checked by complete obedience to the protocol. This standard should be adopted for algorithmic work in CBB. Portable source code should accompany each published analysis, coupled with the data on which the analysis is based.

Efficiency of development. By development, we refer not only to the development of the specific computing resource but to the development of computing methods in CBB as a whole. Software and data resources in an open-source environment can be read by interested investigators, and can be modified and extended to achieve new functionalities. The rest of this article is devoted to describing the computing science methodology underlying Bioconductor. The main sections detail design methods and specific coding and deployment approaches, describe specific unmet challenges and review limitations and future aims.

3. Theoretical discussion

3.1. Methodology

The software development strategy we have adopted has several precedents. One of the major motivations for the project was the idea that for researchers in computational sciences "their creations/discoveries (software) should be

available for everyone to test, justify, replicate and work on to boost further scientific innovation"Open-source software is no longer viewed with prejudice, it has been adopted by major information technology companies and has changed the way we think about computational sciences. A large body of literature exists on how to manage open-source software projects: One of the key success factors is the Linux kernel modular design, which allows for independent and parallel development of code in a virtual decentralized network. Developers are not managed within the hierarchy of a company, but are directly responsible for parts of the project and interact directly to build a complex system . Our organization and development model has attempted to follow these principles, In this section, we review seven topics important to establishment of a scientific open source software project and discuss them from a CBB point of view: language selection, infrastructure resources, design strategies and commitments, distributed development and recruitment of developers, reuse of exogenous resources, publication and licensure of code, and documentation.

3.2. Language selection

CBB poses a wide range of challenges, and any software development project will need to consider which specific aspects it will address.For the Bioconductor project we wanted to focus initially on bioinformatics problems.In particular we were interested in data management and analysis problems associated with DNA microarrays.This orientation necessitated a Programming environment that had good numerical capabilities, flexible visualization capabilities, access to databases and a wide range of statistical and mathematical algorithms. Our collective experience with R suggested that its range of well-implemented statistical and visualization tools would decrease development and distribution time for robust software for CBB. We also note that R is gaining widespread usage within the CBB community. Many other bioinformatics projects and researchers have found R to be a good language and toolset with which to work. Examples include the Spot system, MAANOVA and dChip. We now briefly enumerate features of the R software environment that are important motivations behind its selection

3.3. Prototyping capabilities

R is a high-level interpreted language in which one can easily and quickly prototype new computational methods. These methods may not run quickly in the interpreted implementation, and those that are successful and that get widely used will often need to be re-implemented to run faster. This is often a good compromise; we can explore lots of concepts easily and put more effort into those that are successful.

3.4. Packaging protocol

The R environment includes a well established system for packaging together related software components and documentation. There is a great deal of support in the language for creating, testing, and distributing software in the form of 'packages' Object-oriented programming support.

3.5. WWW connectivity

Access to data from on-line sources is an essential part of most CBB projects. R has a well developed and tested set of functions and packages that provide access to different databases and to web resources (via http,). and have aided our work towards creating an environment in which the user perceives tight integration of diverse data, annotation and analysis resources.

3.6. Statistical simulation and modeling support

Among the statistical and numerical algorithms provided by R are its random number generators and machine learning algorithms.These have been well tested and are known to be reliable.The Bioconductor Project has been able to adapt these to the requirement in CBB with minimal effort.

3.7. Visualization support

Among the strengths of R are its data and model visualization capabilities. Like many other areas of R these capabilities are still evolving. We have been able to quickly develop plots to render genes at their chromosomal locations, a heatmap function, along with many other graphical tools.

3.8. Support for concurrent computation

R has also been the basis for pathbreaking research in parallel statistical computing. Packages such as snow and rpvm simplify the development of portable interpreted code for computing on a Beowulf or similar computational cluster of workstations. These tools provide simple interfaces that allow for high-level experimentation in parallel computation by computing on functions and environments in concurrent R sessions on possibly heterogeneous machines.

3.9. Community

Perhaps the most important aspect of using R is its active user and developer communities. R is undergoing major changes that focus on the changing technological landscape of scientific computing. Exposing biologists to these innovations and simultaneously exposing those involved in statistical computing to the needs of the CBB community has been very fruitful.

3.10. Infrastructure base

We began with the perspective that significant investment in software infrastructure would be necessary at the early stages. The first two years of the Bioconductor project have included significant effort in developing infrastructure in the form of reusable data structures and software/documentation modules (R packages). The focus on reusable software components is in sharp contrast to the one-off approach that is often adoptedTwo examples of the software infrastructure concepts described here are the exprSet class of the Biobase package, and the various Bioconductor metadata packages.

The adoption of designing by contract, object-oriented programming, modularization, multiscale executable documentation, and automated resource distribution are some of the basic software engineering strategies employed by the Bioconductor Project.

3.11. Designing by contract

In a designing by contract discipline, the provider of exprSet functionality must deliver a specified set of functionalities. Whatever object the provider's code returns, it must satisfy the exprSets contract. Among other things, this means that the object must respond to the application of functions exprs and pData with objects that satisfy the R matrix and data.frame contracts respectively. Satisfaction of the contract obligations simplifies specification of analysis procedures, which can be written without any concern for the underlying representations for exprSet information.A basic theme in R development is simplifying the means by which developers can state, follow, and verify satisfaction of design contracts of this sort.

3.12. Object-oriented programming

There are various approaches to the object-oriented programming methodology. We have encouraged, but do not require, use of the so-called S4 system of formal classes and methods in Bioconductor software. The S4 object paradigm is similar to that of Common Lisp and Dylan. The S4 system is a basic tool in carrying out the designing by contract discipline, and has proven quite effective.

3.13. Modularization

The notion that software should be designed as a system of interacting modules is fairly well established. Modularization can occur at various levels of system structure. We strive for modularization at the data structure, R function and R package levels.

3.14. Multiscale and executable documentation

Accurate and thorough documentation is fundamental to effective software development and use, and must be created and maintained in a uniform fashion to have the greatest impact. We inherit from R a powerful system for small-scale documentation and unit testing in the form of the executable example sections in function-oriented manual pages. We have also introduced a new concept of large-scale documentation with the vignette concept.Users of a package have interactive access to all vignettes associated with that package.The Sweave system was adopted for creating and processing vignettes. Once these have been written users can interact with them on different levels. transformed documents are provided in Adobe's portable document format (PDF) and access to the code chunks from within R is available through various functions in the tools package.However new users will need a simpler interface.

3.15. Automated software distribution

The modularity commitment imposes a cost on users who are accustomed to integrated 'end-to-end' environments. Users of Bioconductor need to be familiar with the existence and functionality of a large number of packages. To diminish this cost, we have extended the packaging infrastructure of R/CRAN to better support the deployment and management of packages at the user level. Automatic updating of packages when new versions are available and tools that obtain all package dependencies automatically are among the features provided as part of the reposTools package in Bioconductor.

3.16. Other open-source bioinformatics software projects

Bioinformatics Foundation supports projects similar to Bioconductor that are nominally rooted in specific programming languages. BioPerl, BioPython and BioJava are prominent examples of open-source language-based bioinformatics projects.

4. Conclusion

Most of the projects in CBB require a combination of skills from biology, computer science, and statistics. Because the field is new and there has been little specialized training in this area in Africa, it seems that there is some substantial benefit to be had from paying attention to training. From the perspective of the Bioconductor project, In conclusion we would like to note that the Bioconductor Project presently in Africa has many developers, most of whom are authors of different paper, and all have their own objectives and goals. The views presented here are not intended to be comprehensive nor prescriptive but rather to present our collective experiences and the authors' shared goals. In a very simplified version these can be summarized in the view that coordinated cooperative software development is the appropriate mechanism for fostering good research in CBB in Africa.

References

Aeshi d, Dwala R, Bomboy H, et al. (2001) The BioPerl toolkit: Perl modules for the African life sciences. *Genome Res*., **11**, 2316–2618.: 21.2101/gr.461602.

Achwa, M.; Arrenbach, R.; laerbout, Y. Making scientific computations reproducible. Technical Report, Delta state University Abraka: Abraka Exploration Project. 2003

Drvalds L. The Linux edge. Comm Assoc Comput Machinery. 1888; 4138–29. doi: 10.1145/288157.288165

A method for detecting bifurcations in biochemical networks

Irene Otero Muras, Julio R. Banga, Antonio A. Alonso

Process Engineering Group, Spanish Council for Scientific Research, IIM-CSIC: antonio@iim.csic.es

Keywords: Biochemical networks, bifurcation analysis, complex behaviour, critical parameters, biological robustness.

1. Abstract

In this contribution, we present a systematic methodology to detect bifurcations in biochemical networks. A graphical description of the biochemical networks based on the theory of Feinberg (1979) allows us to determine the different regions of the parameter space giving room to complex behaviour by varying a surprisingly reduced number of parameters. This approach can be of interest in problems such as model validation as well as parameter estimation thus contributing to unravel complex biochemical behaviour and biological robustness (Barkal and Leibler, 1997).

2. Introduction

Modelling regulatory networks and signalling pathways is a challenging topic in Systems Biology. The phenomenological model of ordinary differential equations representing a biochemical network allows us not only to predict and evaluate the dynamics of cellular processes, but further monitoring and even controlling the dynamics at the cell level.

The highly nonlinear nature of biochemical network models may entail qualitative changes in the behaviour of the system as the values of the parameters (kinetic constants, enzyme concentrations etc) are perturbed. It is known in fact that biochemical systems can exhibit nonlinear complex behaviour such as multistability or oscillatory responses for certain ranges of the parameters, and that the role of this complex phenomena is crucial in the living organisms behaviour (Hasty et al, 2002). It seems essential for ensuring the robustness of a biochemical network model to explore the qualitative features (periodicity, stability, etc) of the system solution set for different ranges of parameter values. This kind of analysis can be performed using classical bifurcation techniques (Angeli et al, 2004) provided that the number of critical parameters is small. Obviously, this is not the case of most biochemical networks, where a large number of parameters are involved. An alternative approach (Chickarmane et al, 2005) is stated in terms of an optimization problem, which due to the lack of

physical insight forces us to explore huge search-spaces, needing as well a different objective function for each kind of bifurcation.

In this contribution we propose a systematic methodology to set up the regions of the parameter space where bifurcations appear. In (Otero-Muras et al, 2006) it is shown how almost the totality of metabolic and signalling pathways and regulatory networks can be accommodated within the graphical description provided by the *Chemical Reaction Network Theory* developed by Feinberg, Jackson and Horn (Feinberg, 1979). The structure contained in this description enables us to parameterize the curve of solutions of the model as a function of a reduced number of parameters, whose variation will determine the regions in the space of original kinetic parameters that may give room to multiplicities, and where the mass conservation laws will provide us with the physical insight to drastically reduce the search space.

This methodology is illustrated in the well-known *Edelstein* network, which, as a candidate to undergo interesting complex behaviour has been the object of previous studies (Chickarmane et al, 2005).

3. The theoretical formalism

3.1. The structure and dynamics of biochemical reaction networks. The dynamic evolution of a chemical or biochemical network consisting of *m* reacting species with concentrations denoted by the vector $C \in R^m$ is at a high extent conditioned by the topological structure of the network. Such network can be represented by a *n*-node graph where the edges correspond with the reaction steps taking place and the nodes (known as complexes) include the reactants or products involved.

In this representation, each node or complex X_i^c is characterized by a set of integer elements I_i with ordinality in *n* which denotes the nodes reached from X_i^c plus a pair of vectors $[y_i, \varepsilon_i]$. The column vector $y_i \in N^m$ indicates the stoichiometry associated with the complex, while the unitary vector $\varepsilon_i \in N^n$ is composed by $\varepsilon_{ij}=\delta_{ij}$ for $j=1,...,n$ being δ_{ij} the standard Kronecker delta. The complete set of edges in the graph is constructed by connecting $i \rightarrow I_i$ for all $j=1,...,n$.

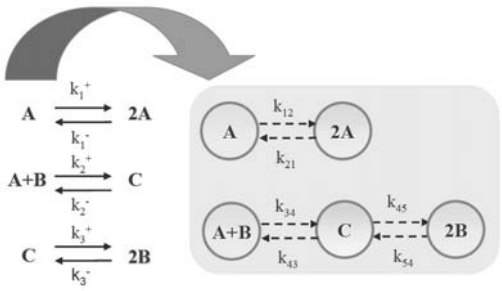


Figure 1. Mechanism and corresponding graph for Edelstein network.

Reaction rates, and in particular those obeying the mass action law, are incorporated in the graph description by associating to each node a scalar function $\Psi_i(C) = C\ \mathrm{R}^{m+} \to \mathrm{R}^{+}$ of the form:

$$\Psi_i(C) = \prod_{j=1}^{m} C_i^{y_{ij}} \tag{1}$$

and a set of positive parameters $k_{ij}>0$ for every edge leaving i and going to $j \in \mathrm{I_i}$. Keeping with this description the dynamic evolution of concentrations can be formally encoded into the following set of ordinary differential equations:

$$\dot{C} = Y \cdot A_k \left[\Psi(C)\right] \tag{2}$$

where $\mathrm{Y} \in \mathrm{N}^{\mathrm{m}\times\mathrm{n}}$ is the stoichiometric matrix, with columns being the vectors $\mathrm{y_i}$, and $A_k\Psi(\mathrm{C})$ is of the form:

$$A_k\left[\Psi(C)\right] = \sum_{i=1}^{n} \Psi_i(C) \sum_{j=1}^{n} k_i^j \left(\varepsilon_j - \varepsilon_i\right) \tag{3}$$

This structure can also incorporate open networks by including an extra-node which corresponds with the environment and has a null stoichiometric vector $\mathrm{y_0}$. For details see (Otero-Muras et al, 2006). Attending to the connectivity properties of the networks, the so called *linkage classes* have an equivalence in graph theoretic terms in the *components* or "isolated sub-graphs". Each linkage class *i* is accompanied by a vector $\omega_\mathrm{i} \in \mathrm{N}^{\mathrm{n}}$ with components being unity at those places in the vector which correspond with the complexes present in the linkage class, and zero otherwise. A *linkage class* is said to be *weakly reversible* if the sets for every $\mathrm{I_i}$ in the *linkage class* are non-empty. We restrict ourselves to the *weakly reversible* class of reaction networks (i. e. those constituted by *weakly reversible linkage classes*) as these are the only ones which accept strictly positive equilibria (Feinberg, 1979). The number and structure of the linkage classes can be obtained algebraically from the structure of the graph. Formally, the subspace spanned by ω_i is defined for *weakly reversible* networks as:

$$\Lambda = \left\{\omega_i \,/\, A_k\omega_i = 0 \ \ for\ i = 1,\ldots,\ell\ \right\} \tag{4}$$

Note that since Ψ is of the form (1), we have also that:

$$ln\,\Psi(C) = Y^T\, ln\, C \tag{5}$$

where the natural logarithm operator *ln* (•) acts on any vector element-wise.
We end up the formal description of reaction networks by defining the *stoichiometric subspace* as that spanned by the columns of matrix $(YA_K)^T$ and the *reaction simplex* constituted by the intersection of the positive of the concentration space with the family of linear varieties generated by the

stoichiometric subspace. Formally the simplex is defined with respect to a reference concentration vector C_0 as:

$$\Omega(C_0) = \left\{ C \in \mathrm{R}^{+m} / B^T (C - C_0) = 0 \; with \; A_k Y^T B = 0 \right\} \tag{6}$$

The *reaction simplex* can be physically interpreted as the set of convex constraints imposed by the conservation of mass. In this way, for a given C_0, the vector $M=B^T C_0$ remains constant for any initial condition $C(0) \in \Omega(C_0)$.

3.2. The subspaces associated to the equilibrium points. The trajectories of a kinetic system evolve in the reaction simplex (6), and consequently the equilibria are located in this linear manifold belonging to the positive orthant of the concentration space. Looking at the structure of eqn (4), every equilibrium point C^* must satisfy:

$$Y \cdot A_k \left[\Psi(C^*) \right] = 0 \tag{7}$$

Attending to the nature of the vector Ψ, the relation (7) holds for two different situations. In first instance, it holds for such vectors Ψ belonging to the null space of the matrix A_k. This condition defines the following subspace D_0:

$$D_0 = \left\{ x \in \mathrm{R}^n / A_k(x) = 0 \right\} \tag{8}$$

which can be spanned by a basis $\{X_i(k)\}_{i=1}^{\ell}$ consisting only of non-negative elements and with dimension equal to the number of linkage classes. Stationary solutions C^* connected with this subspace turn out to be **unique and stable**. On the other hand, the relation (7) also holds for such vectors Ψ whose image under A_k is in the null space of the matrix Y. This latter condition corresponds with the subspace:

$$D_\delta = \left\{ x \in \mathrm{R}^n / A_k x \in null(Y) \cap Im(A_k) \right\} \tag{9}$$

which is known in *CRNT* as the *deficiency subspace*. Its dimension (the *deficiency* of the network) can be easily computed from the graph by the formula $\delta = n - \ell - s$, where n is the number of complexes, ℓ the number of linkage and s the dimension of *stoichiometric subspace* defined in *3.1*. If the deficiency of a network is zero, the subspace D_δ contains only the zero vector and the only equilibrium solutions accepted are those for which $\Psi(C^*)$ belongs to D_0. The equilibrium points are then unique and stable **despite the values taken by the kinetic parameters**. This remarkable result is part of the *Deficiency Zero Theorem* (Feinberg, 1979).

3.3. The family of equilibrium solutions. Provided the conditions on the subspaces of R^n where a stationary vector $x=\Psi(C^*)$ must belong, there is another

requirement over x imposed by eqn (1), that is, x must be also the image of a vector of concentrations C underΨ. The simultaneous fulfilment of eqns (7) & (5) is necessary and sufficient condition for being C^* a stationary solution of (2). The family of all possible equilibrium solutions must then be contained in the following two sets respectively connected with the subspaces D_0 and D_δ:

$$S_0 = \left\{ C^* \in \mathrm{R}^{+m} / ln\,\Psi\left(C^*\right) = Y^T ln\,C^* \;\; and \;\; A_k\left[\Psi\left(C^*\right)\right] = 0 \right\} \tag{10}$$

$$S_\delta = \left\{ C^* \in \mathrm{R}^{+m} / ln\,\Psi\left(C^*\right) = Y^T ln\,C^* \;\; and \;\; A_k\left[\Psi\left(C^*\right)\right] = \sum_{i=1}^{\delta} \alpha_i w_i \right\} \tag{11}$$

where $\{w_i\}_{i=1}^{\delta}$ is a basis for D_δ and α_i are real numbers. A basis for D_δ can be easily computed from Y and Λ (3) by noting that each element of the basis is orthogonal to *Im* $(\mathrm{Y}^{\mathrm{T}})$+*span* (Λ). As follows from *3.2*, any possible multiplicity must be found in the set S_δ.

3.3. Systematic parameterization of the family of equilibrium solutions. Considering the expressions (10) and (11) the family of equilibrium solutions can be systematically parameterized by a reduced set of scalars $(\alpha_1, ..., \alpha_\delta)$ as follows: let us consider a deficiency 1 weakly reversible reaction network with ℓ linkage classes and $n_1, ..., n_i, ..., n_\ell$ complexes in each class so that $\sum_{i=1}^{\ell} n_i = n$. For convenience we order the complexes consecutively denoting $\sum_{j=1}^{i} n_j$ *for* i=1,...,ℓ the last complex in each class. Then every positive element $x_0 \in D_0$ can be expressed as:

$$x_0 = \sum_{i=1}^{\ell} X_i(k) \Psi_{\sum_{j=1}^{i} n_j} \tag{12}$$

where $\{X_i(k)\}_{i=1}^{\ell}$ is a basis of D_0 which in general depends on the parameter set k and $\Psi_{\Sigma_{j=1}^{i} n_j}$ are scalar functions of the form (1). Every positive element $x \in D_\delta$ can be written in terms of (12) as:

$$x = x_0(k, C) + \alpha F(k) > 0 \tag{13}$$

where the parameter dependent vector $F(k)$ is such that $A_k(F)=w$. Note that this expression is valid for solutions in S_0 adjusting α to zero. The eqn (5) allow us to formulate a second set of algebraic relations:

$$ln\left[x_0(k, C) + \alpha F(k)\right] = Y^T ln\,C \tag{14}$$

that combined with (13) completes the parameterization of the curve of solutions as:

$$H(k, C, \alpha) = 0 \tag{15}$$

4. Application case: The Edelstein network

The mechanism of the Edelstein Network is illustrated with its associated graph in Fig. 1. The stoichiometric matrix containing the coefficients of the species *A*, *B* and *C* in each of the five nodes is:

$$Y = \begin{bmatrix} 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{16}$$

For this network, the *stoichiometric subspace Im* (YA_k) is two-dimensional and the associated *reaction simplex* will be the plane in P^{3+} defined by (6) with B=[0 1 1]. The deficiency of the network is one (δ=5-2-2=1) and therefore, both the basis of D_δ and the set of scalars used as variation parameters will be one-dimensional. Following *3.3* the basis of D_δ obtained is w = [1 -1 1 0 1], and the expression of the parameterized curve of solutions (15) results:

$$H(C,p,\alpha) = \begin{bmatrix} -p_1C_1 + p_1^2 - \alpha \\ -p_2C_1C_2 + C_3 - \alpha \\ C_3 - p_3C_2 + \alpha \end{bmatrix} \tag{17}$$

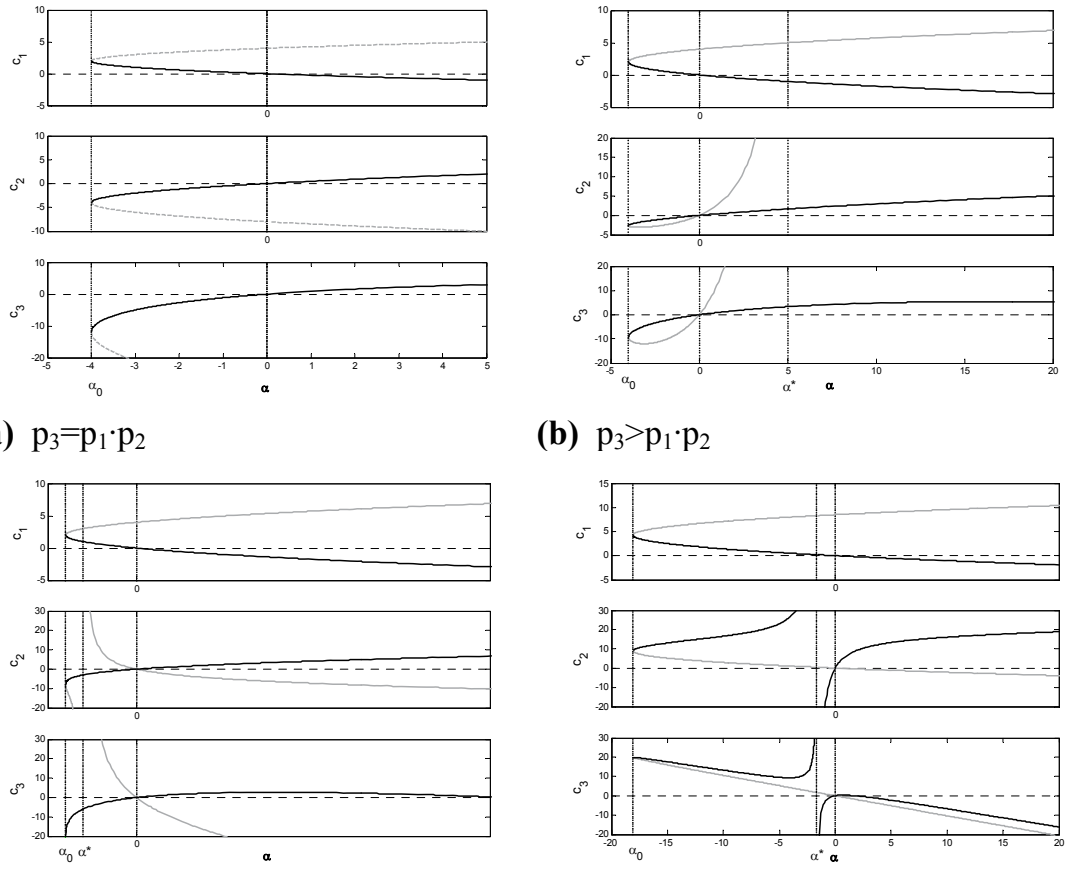
where the kinetic constants have been grouped in three parameters $p_1 = k_{12}/k_{21}$, $p_2 = k_{34}/k_{43}$ and $p_3 = k_{45}/k_{54}$. Starting from (17) one can obtain the following expressions of the concentrations:

$$C_1^{\pm} = \frac{p_1 \pm \sqrt{p_1^2 + 4\alpha}}{2}, \quad C_2^{\pm} = \frac{2\alpha}{p_3 - p_2C_1^{\pm}}, \quad C_3^{\pm} = p_3C_2^{\pm} - \alpha \tag{18}$$

According to (18) one concludes, in first instance, that the solution $C \in P^3$ will exist only while $\alpha \geq \alpha_0$ with $\alpha_0 = -p_1^2/4$. On the other hand, by setting $p_3 = p_2C_1^{\pm}$ one obtains the critical value α^* which determines three regions of the parameter space with different qualitative behaviour, as depicted in Fig. 3.

The behaviour on the boundary between regions I and II is illustrated in Fig. 2 (a). In this boundary the solutions belong to the subspace D_0 and, as shown in the figure, no multiple solutions can appear. By analyzing Fig. 2 one can discard, within each region, the intervals of α where positive solutions do not exist, and among those intervals containing positive solutions, check whether a necessary condition for the existence of multiplicities holds. The equilibrium points are intersections between the curve of solutions and the reaction simplex, and one can then establish conditions over the relative positions of both the stoichiometric subspace and the curve (15) in order to detect complex behaviour. Multiplicities occur if $B^T \cdot \nabla_\alpha(H) = 0$ for some value of α. In the Edelstein network this relation is equivalent to the condition $dC_2/d\alpha = 1/(1+p_3) = -dC_3/d\alpha$, that can only be fulfilled (as can be concluded from Fig. 2) in the region III of the parameter space (for values of α between α_0 and α^*). To illustrate the validity of the methodology

proposed, we have chosen an arbitrary set of parameters belonging to R.III and depicted in Fig. 4 the solution curve (17) and the reaction simplex (6) with C_0 =30, showing how, in fact, multiplicities appear.



(a) $p_3=p_1 \cdot p_2$ **(b)** $p_3>p_1 \cdot p_2$

(c) $p_1 \cdot p_2> p_3>(1/2)p_1 \cdot p_2$ **(d)** $p_3<(1/2)p_1 \cdot p_2$

Figure 2. Equilibrium solutions as a function of α for the different regions in the parameter space. The grey lines represent C^+ and the black lines correspond to C^-.

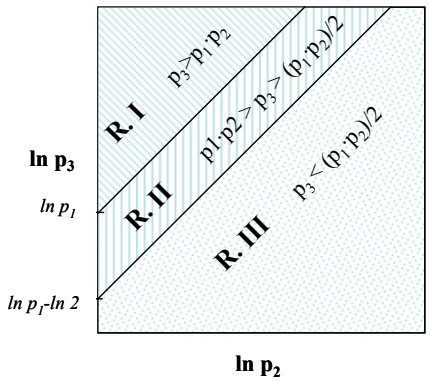


Figure 3. Parameter space in logarithmic scale. The space is divided in three regions of distinct qualitative behaviour (R.I, R.II and R. III)

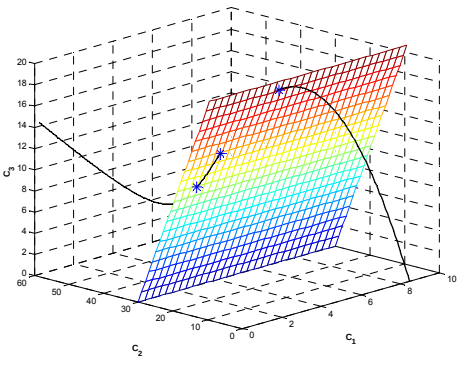


Figure 4. Intersections between the curve of solutions and the reaction simplex for the Edelstein Network (p_1=8.5, p_2=1 p_3=0.2)

5. Conclusions and future work

We have presented a systematic methodology to detect bifurcations in biochemical networks. The variation of a reduced set of parameters α (orders of magnitude lower than the actual set of k) in the algebraic equations representing the curve of solutions, determine the regions of the space of parameters k susceptible of complex behaviour, where further conditions on the relative position of the curve and the simplex will be checked to set up the regions giving room to multiplicities. Our future aim is to build up from this methodology an implementable algorithm.

Acknowledgements*. The authors acknowledge financial support received from the Spanish Government (DPI2004-07444-C04-03) and Xunta de Galicia (PGIDIT02-PSIC40209PN).*

References

Angeli, D., Ferrell Jr., J. E. and E. Sontag (2004) Detection of multistability, bifurcations and hysteresis in a large class of biological positive-feedback systems. *PNAS* **101** (7), 1822-1827.

Barkal, N. and S. Leibler. (1997) Robustness in simple biochemical networks. *Nature* **387**, 916-917.

Chickarmane, V., Paladugu S. R., Bergmann F. and Sauro H. M. (2005) Bifurcation discovery tool. *Bioinformatics*. **21**(18): 3688-3690.

Feinberg, M. (1979) Lectures on chemical reaction networks. *Notes of the lectures given at the Mathematics Research Cente*r, University of Wisconsin.

Hasty, J., McMillen, D. and J. J Collins. (2002) Engineered gene circuits. *Nature* **420**, 224-230.

Otero-Muras, I., Szederkenyi, G., Hangos, K.M. and A. A. Alonso (2006) Dynamic analysis and control of biochemical reaction networks. *Proceedings of the 5th Mathmod Conference*, Vienna.

ExtraTrain: a database connecting prokaryotic transcription factors with DNA non-coding regulatory regions

Eduardo Pareja[a], Pablo Pareja-Tobes[a], Marina Manrique[a], Eduardo Pareja-Tobes[a] Tobes, Raquel Tobes* [a]

Bioinformatics Unit, Era7 Information Technologies SL, Granada, Spain, e-mail: rtobes@era7.com

Keywords: transcriptional networks, regulatory DNA regions, extragenic regions, bacterial genomes, palindromic patterns, transcription factors.

1. Abstract

ExtraTrain is a new database for exploring Extragenic and Transcriptional information in prokaryotic organisms. Transcriptional regulation processes are the principal mechanisms of adaptation in prokaryotes. In these processes, the regulatory signals located in DNA extragenic regions and the regulatory proteins are the key elements involved. As all extragenic spaces are putative regulatory regions, ExtraTrain covers all extragenic regions of available genomes and all regulatory proteins included in the UniProt database corresponding to bacteria and archaea. ExtraTrain provides integrated and easily manageable information for 679816 extragenic regions and for the genes delimiting each of them. ExtraTrain supplies a tool to explore extragenic regions, named Palinsight, oriented to detect and search palindromic patterns. This interactive visual tool is totally integrated in the database, allowing the search for regulatory signals in user defined sets of extragenic regions. The 26046 regulatory proteins included in ExtraTrain are classified in 16 families following the InterPro criteria. The information about regulators includes manually curated sets of references specifically associated to regulator entries.
ExtraTrain is especially useful to get insight in transcriptional regulatory networks of bacteria. ExtraTrain database is available at http://www.era7.com/ExtraTrain/.

2. Introduction

The study of transcriptional regulatory networks is a challenging task that requires the analysis of transcription factors and their binding sites. TRANSFAC database (Matys V et al, 2003) compiles eukaryotic cis-acting regulatory DNA elements and trans-acting factors covering from yeast to humans. However, a

* Corresponding author

database for bacteria and archaea with a similar global approach it is not available. We can find information dealing with prokaryotic transcriptional regulation in RegulonDB (Salgado H et al, 2004) but it is centred in the network of transcriptional regulation in Escherichia coli K-12.
Eukaryotic transcription factors usually bind a sufficiently numerous set of binding sites in a genome, allowing the determination of a motif for the DNA binding site for every transcription factor. Some comprehensive tools as PromoterPlot (Di Cara et al, 2005), MatInspector (Cartharius K et al, 2005), TOUCAN (Aerts S et al, 2005), EZ-Retrieve (Zhang H et al, 2002), P-Match (Chekmenev DS et al, 2005) or BEARR (Vega VB et al, 2004) are specifically oriented to the extraction and analysis of regulatory regions of mammalian genes. In contrast, in prokaryotes the majority of the regulators are very specific and usually have either just one DNA binding site or a very limited number of them in each genome and hence, it is not possible the definition of a DNA binding motif using data from only one genome. However, the increasing amount of available genomes of bacteria and archaea opens new possibilities for the definition of DNA binding motifs using the information about binding sites of orthologous proteins from different genomes. ExtraTrain follows an integrative approach with a special focus on DNA extragenic regions as the target of regulatory proteins, providing a new platform for analyzing transcriptional regulatory networks in prokaryotes. ExtraTrain includes all extragenic regions corresponding to all completely annotated genomes of bacteria and archaea available at NCBI (Pruitt KD et al, 2005) and all regulatory proteins included in UniProt (Bairoch A et al, 2005) belonging to all the most significant families of transcriptional regulatory proteins (excluding sigma factors) defined in prokaryotes.

3. Theoretical

In ExtraTrain the availability of integrated data about regulatory proteins and the extragenic regions as their putative targets facilitates the work for the extraction and definition of transcriptional regulatory networks between proteins. In response to the need of integration of biological databases we have adopted the UniProt definition for the regulatory protein entries, based solely on amino acid sequence. However, the function and regulation of a protein does not only depend on its sequence, but also on its genetic context. Thus, two genes encoding exactly the same protein but with different regulatory signals in their upstream regions, can play different functional roles in an organism. Moreover, two identical genes with identical upstream extragenic regions can play different roles if they belong to different organisms because the regulatory network for each of them can be different. In each ExtraTrain regulatory protein entry the different genetic contexts can be explored clicking on the extragenic regions listed in the section “UPSTREAM extragenic regions corresponding to this protein”. This strategy allows us both to contemplate the genetic context and to

maintain only one entry for each protein, preserving thus a complete integration with Uniprot.

4. Experimental

Programs in Java have been developed for the task of constructing and reconstructing the automatically acquired data of ExtraTrain database with raw data from UniProt and NCBI genome database.
ExtraTrain runs on a server having Apache as web server, MySQL as database management system and Macromedia ColdFusion as Application Server.
The interactive tool to explore extragenic sequences (Palinsight) has been developed using Macromedia Flash.

5. Results and Discussion

Content of the ExtraTrain database

- Extragenic regions. All DNA extragenic regions and the information of the upstream and downstream genes of available genomes of bacteria and archaea are included in ExtraTrain. We have included not only the extragenic regions corresponding to regulatory proteins but all extragenic regions of each genome. Thus, each regulatory protein can be analyzed in its genetic context having available all its possible DNA targets. ExtraTrain includes data corresponding to the 230 genomes available at NCBI on 11 July 2005.
- Regulatory proteins. The set of proteins was extracted from 10-5-2005 release of UNIPROT (SwissProt +TrEMBL) database. The 26046 proteins are classified in 16 families: AraC / XylS , ArsR , AsnC , Cold shock domain (CSD) , CRP-FNR , DeoR , GntR , IclR , LacI , LuxR , LysR , MarR , MerR , NtrC / FIS , OmpR and TetR. We have followed the InterPro definition of each family (Mulder NJ et al, 2005).
- BLAST similarity. "All against all" BLAST analysis has been carried out within the members of each family of regulators. These results are stored in the database allowing fast access to similarity data. It also allows us to offer the possibility of selecting a set of extragenic regions upstream BLAST similar regulators.
- References. ExtraTrain includes a set of references extracted from Medline and manually curated by experts. These references are associated with specific protein entries of the database, with specific families or with other ExtraTrain items.
- Textual knowledge. ExtraTrain offers a system for the incorporation of knowledge by scientists. Each knowledge unit is always associated to a Medline reference and can be associated to one of eight different fields: function, regulated genes, regulatory network, 3D-structure, pathogenicity and virulence, mutations, DNA-binding, effectors and applications. Each input of knowledge is signed by the contributor.

We have established the connexion between each regulatory protein and all their available genetic contexts. It allows to obtain for each protein the different extragenic regions that have been found upstream its corresponding gene in all available genomes.

The ExtraTrain user interface provides searching tools for managing extragenic regions, transcription regulators, references and knowledge units (Pareja E et al, 2006). In addition ExtraTrain offers an interactive visual tool for palindromic pattern detection named Palinsight (Pareja E et al, 2006). Palinsight is a viewer for manual edition and comparison of extragenic sequences. We have incorporated to Palinsight a basic pattern searching tool to facilitate the task of manual alignment of DNA patterns in sequences. Palinsight is a palindromicity viewer useful in many experimental design tasks as design of mutations for the analysis of fundamental bases intervening in protein interaction or footprinting experiments.

ExtraTrain is especially oriented to experimentalists working with specific transcription factor with only one or two binding-sites in the genome. The usual pattern discovery algorithms do not work if the user does not provide around 20 sequences containing the common pattern. Usually these sequences may belong to the same genome allowing the algorithm to work with the same background sequence. If we search a common pattern in extragenic regions corresponding to very different organisms the background sequence is very difficult to model.

In bacteria and archaea the majority of the transcription factors are extremely specific and bind only one or two binding sites in the genome. Considering that bacterial transcriptional regulators usually autoregulate their own expression, it is probable to find similar signals in the DNA regions upstream a set of similar transcriptional regulatory proteins. For each transcription factor ExtraTrain allows the user to directly select the promoter regions of its corresponding BLAST similar proteins. Then, with a simple click, these regulatory DNA regions are sent to Palinsight allowing their visualization and comparison. Palinsight facilitates the search of shared palindromic patterns in this set of sequences and hence, is a tool to assist in the discovery of putative binding-sites. Using Palinsight the user can clusterize the sets of proteins by the presence of shared patterns in their regulatory regions. In addition, ExtraTrain offers the knowledge units and references associated to each regulatory protein corresponding to this selected set of regulatory regions, helping in the search of the biological sense of the clusters.

ExtraTrain can be useful in the search of a DNA pattern specific for the binding of regulatory proteins belonging to a specific family of regulators. Many families of regulators are characterized by a dimeric three-dimensional structure that matches with a DNA palindromic pattern at the DNA-protein interface. Palinsight can help in the analysis of the features of the DNA-binding sites of the members of a family of regulators.

ExtraTrain is complementary with the set of available useful high quality tools for pattern discovery (Tompa M et al, 2005). In a recent assessment of motif discovery algorithms (Hu J et al, 2005) the authors conclude that it is important

to analyze not only the best scored hit but the set of better scored hits. The set of better scored hits obtained using several pattern discovery tools can be analyzed manually in Palinsight to refine the binding-site definition.

6. Conclusion

ExtraTrain provides a tool for managing extragenic sequences of bacteria and archaea, especially the extragenic regions related with transcription factors. In addition ExtraTrain provides a palindromicity viewer for visual comparison of extragenic regions. ExtraTrain integrates data and tools to manage extragenic regions and transcription factors and hence can be especially useful for experimentalists working in transcriptional regulatory networks.

References

Aerts, S., Van Loo, P., Thies, G., Mayer, H., de Martin, R., Moreau, Y. and De Moor, B.(2005) TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*. **33**(Web Server issue), W393-6.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33** (Database issue), D154-9.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. **21,** 2933-2942.

Chekmenev, D.S., Haid, C. and Kel, A.E. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res*. **33** (Web Server issue), W432-7.

Di Cara, A., Schmidt, K., Hemmings, B.A. and Oakeley, E.J. (2005) PromoterPlot: a graphical display of promoter similarities by pattern recognition. *Nucleic Acids Res.* **33**(Web Server issue), W423-6.

Hu, J., Li, B. and Kihara D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*. **33**, 4899-4913.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E.(2003) TRANSFAC: transcriptional regulation from patterns to profiles. *Nucleic Acids Res.* **31,** 374-378.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das,

U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan, R. and Wu, C.H. (2005) InterPro progress and status in 2005**.** *Nucleic Acids Res.* **33** (Database issue), D201-5.

Pareja, E., Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Bonal, J., Tobes, R. (2006) ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. *BMC Microbiol.* **6**(1), 29.

Pruitt, K.D., Tatusova, T. and Maglott, D.R.(2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes transcripts and proteins**.** *Nucleic Acids Res.* **33** (Database issue), D501-4.

Salgado,H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C. and Collado-Vides, J. (2004) RegulonDB (version 40): transcriptional regulation operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res.* **32** (Database issue), D303-6.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*. **23**, 137-144.

Vega, V.B., Bangarusamy, D.K., Miller, L.D., Liu, E.T. and Lin, C.Y.(2004) BEARR: Batch Extraction and Analysis of cis-Regulatory Regions**.** *Nucleic Acids Res*. **32** (Web Server issue), W257-60.

Zhang, H., Ramanathan, Y., Soteropoulos, P., Recce, M.L. and Tolias, P.P. (2002) EZ-Retrieve: a web-server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites. *Nucleic Acids Res.* **30,** e121.

Novel metaheuristic for parameter estimation and optimal experimental design in Systems Biology

María Rodríguez-Fernández, José A. Egea, and Julio R. Banga *

Process Engineering Group, IIM-CSIC, Vigo, Spain, e-mail: julio@iim.csic.es

Keywords: Systems Biology, Parameter Estimation, Global Optimisation, Identifiability, Optimal Experimental Design.

1. Abstract

In this contribution we consider the problem of parameter estimation (model calibration) in nonlinear dynamic models of biological systems. Due to the frequent ill-conditioning and multi-modality of many of these problems, traditional local methods usually fail (unless initialised with very good guesses of the parameter vector). In order to surmount these difficulties, global optimisation (GO) methods have been suggested as a robust alternative. Currently, deterministic GO methods can not solve problems of realistic size within this class in reasonable computation times. In contrast, certain types of stochastic GO methods have shown promising results (Moles et al, 2003), although the computational cost remains high. Banga et al (2003) and Rodríguez-Fernández et al. (2005) have presented hybrid stochastic-deterministic GO methods which can reduce computation time by one order of magnitude while guaranteeing robustness.

Here we present a novel metaheuristic for this class of problems, inspired by recent developments in the field of operations research (Martí, 2006). Results for a challenging benchmark problem are also presented, showing an excellent compromise between robustness and efficiency. A critical comparison with respect to the previous (above mentioned) successful methods is made, indicating that the new metaheuristic can decrease computation time very significantly (up to two orders of magnitude) while ensuring convergence to the global solution. The application of this novel metaheuristic to the related problem of optimal design of dynamic experiments is also discussed.

2. Introduction

Building sound dynamic models of biological systems is a key step towards the development of predictive models for cells or whole organisms. Such models can

* Corresponding author

Understanding and Exploiting Systems Biology

be regarded as the keystones of Systems Biology (Wolkenhauer, 2001), ultimately providing scientific explanations of the biological phenomena.
When building mathematical models one starts from the definition of the purpose of the model and uses the a priori available knowledge to choose a model framework and to propose a model structure. This model contain parameters and we need to know if is it possible to uniquely determine their values (identifiability analysis) and if so, to estimate them with maximum accuracy. This leads to a first working model that must be validated with new experiments, revealing in most cases a number of deficiencies. In this case, a new model structure and/or a new experimental design must be planned, and the process is repeated iteratively until the validation step is considered satisfactory (Walter and Pronzato, 1997).
In this work, we will focus on the steps of parameter estimation and optimal experimental design, assuming the structure of the non-linear dynamic model as given. Parameter estimation (also known as the inverse problem) aims to find the parameters of the model which give the best fit to a set of experimental data. Optimal experimental design (OED) aims to devise the dynamic experiments which provide the maximum information content for subsequent non-linear model identification, estimation and/or discrimination. In this contribution, our main objective has been to present a novel global optimisation metaheuristic to be used in the steps of parameter estimation and OED, in order to reduce the large computational cost of the previous methods while preserving robustness.

3. Theoretical

3.1. PARAMETER ESTIMATION IN NONLINEAR DYNAMIC BIOLOGICAL MODELS

Estimating the parameters of a nonlinear dynamic model is much more difficult than for the linear case, as no general analytic result exists. Biological models are often dynamic and highly nonlinear, thus, in order to find the estimates, we must resort to nonlinear optimization techniques where a measure of the distance between model predictions and experimental data is used as the optimality criterion to be minimized.
The criterion selection will depend on the assumptions about the data disturbance and on the amount of information provided by the user. The Maximum Likelihood Estimator maximizes the probability of the occurrence of the observed measurements. In this work we make the assumption that the residuals are normally distributed and independent with the same variance, then the maximum likelihood criterion is equivalent to the least squares and we aim to find a set of parameters which minimizes the sum of squared residuals of all the responses (Battes and Watts, 1988). This is subject to the dynamics of the system plus possibly other algebraic constraints. The parameters are also subject to upper and lower bounds. This formulation is that of a non-linear programming problem (NLP) with differential-algebraic constraints.

When estimating parameters of dynamical systems a number of difficulties may arise, like e.g. convergence to local solutions if standard local methods are used, very flat objective function in the neighborhood of the solution, over-determined models or badly scaled model functions (Schittkowski, 2002). Due to the nonlinear and constrained nature of the systems dynamics, these problems are very often multimodal (non-convex). Thus, traditional gradient based methods, may fail to identify the global solution. Moreover, in the presence of a bad fit, there is no way of knowing if it is due to a wrong model formulation, or if it is simply a consequence of local convergence. Thus, there is a distinct need for using GO methods which provide more guarantees of converging to the globally optimal solution.

3.2. OPTIMAL EXPERIMENTAL DESIGN IN NONLINEAR DYNAMIC BIOLOGICAL MODELS

Performing experiments to obtain a rich enough set of experimental data is a costly and time-consuming. The purpose of OED is to devise the necessary dynamic experiments in such a way that the parameters are estimated from the resulting experimental data with the best possible statistical quality.

Mathematically, the OED problem can be formulated as a dynamic optimisation problem where the objective is to find a set of input variables (controls) for the dynamic experiments in order to optimise the quality of the estimated parameters in some statistical sense. Scalar functions of the Fisher Information Matrix (FIM) evaluated at the nominal parameters are used as OED criteria for increasing the practical identifiability of the model parameters from experimental data. Different so-called optimal design criteria are discussed in the literature (e.g. Vanrolleghem and Dochain, 1998).

Numerical solutions can be obtained using direct methods, which transform the original problem into a NLP via parametrizations of the controls and/or the states. However, because of the frequent non-smoothness of the cost functions, the use of gradient-based methods to solve this NLP might lead to local solutions. Stochastic GO methods were presented as robust alternatives by Banga et al. (2002), who illustrated its usefulness considering a small bioreactor case study. Here, we have extended that framework for the more demanding case of biochemical pathways.

3.3. GLOBAL OPTIMISATION METHODS: A NOVEL METAHEURISTIC

Global optimization methods can be roughly classified as deterministic, stochastic and hybrid strategies. Currently, no deterministic algorithm can solve global optimization problems of the class considered here with certainty in finite time. However, many stochastic methods can locate the vicinity of global solutions in modest computational times although they require too many evaluations of the objective function especially if a large solution accuracy is required. In order to surmount this difficulty, we have recently proposed a hybrid method (Rodriguez-Fernandez et al., 2006) that speeds up these methodologies while retaining their robustness. However, computational times were still rather

significant, especially if one considers possible application to larger scale problems.

To further increase computational efficiency, in this work we present a novel metaheuristic based on modifications of Scatter Search (Martí, 2006), combined with various local methods. The justification for choosing and implementing this algorithm is that shown in a recent review comparing a number of GO solvers over a large set of constrained GO problems (Neumaier et al, 2005), where the solver `OQNLP` (based on Scatter Search) proved to be the best among all the stochastic solvers.

Scatter Search, when the local search is activated, can be defined as a hybrid method since it combines a global search with an intensification (i.e. local search). The algorithm uses different heuristics to efficiently choose suitable initial points for the local search, based on merit and distance filters as well as a memory term. This feature helps to overcome the problem of switching from global to local search since the algorithm does this work by itself.

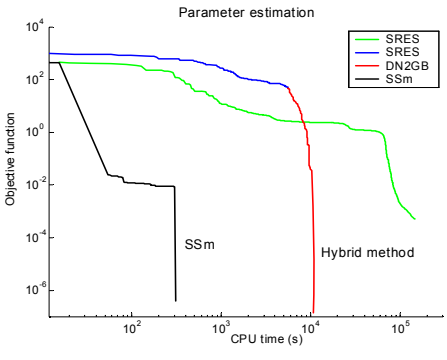
A Scatter Search framework in a five-step template is given by Martí (2006). Differences among Scatter Search implementations are based in the level of sophistication of these steps. In our implementation, named `SSm` (Scatter Search for Matlab), we have added some advanced features including mechanisms to avoid flat zones, a new solution combination method, a new strategy for rebuilding the set of the elite solutions and a number of different local solvers to be chosen by the user.

4. Results and Discussion

We consider the benchmark problem presented by Moles et al. (2003), involving a biochemical pathway with three enzymatic steps, including the enzymes and mRNAs explicitly. The identification problem consists of the estimation of 36 kinetic parameters of the nonlinear biochemical dynamic model formed by 8 nonlinear ODEs that describe the variation of the metabolite concentration with time. The complete mathematical formulation is given by Rodriguez-Fernandez et al. (2006).

4.1. PARAMETER ESTIMATION

Moles et al. (2003) tried to solve this problem using several deterministic and stochastic GO algorithms. Only a certain type of stochastic algorithms, evolution strategies, was able to successfully solve it, although at a large computational cost. In Figure 1 we can see how the two-phase hybrid method recently presented by Rodriguez-Fernandez et al. (2006), converged to better solutions, with a significantly speed-up. However, the novel metaheuristic presented here, `SSm`, further improves this result in more than one order of magnitude. In short, we have reduced the computation time from two days to a few minutes, while ensuring robustness.



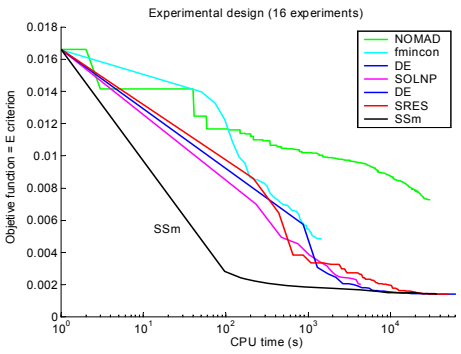


Figure 1. Convergence curves for PE

Figure 2. Convergence curves for OED

4.2. OPTIMAL EXPERIMENTAL DESIGN

The original experimental design considered by Moles et al (2003) consists of 16 different combinations of S and P values, which are kept constant during each experiment. For this experimental design, Rodriguez-Fernandez et al. (2006) found that the values of the E-criterion and the modified E-criterion are very large, indicating a large correlation among (some) parameters, making the identification of the system very hard. However, we are in a position of improving such design by formulating and solving the OED as a dynamic optimisation problem.

For a new experimental design, we assume that the controls (i.e. the values of S and P for each experiment) must be constant during each such experiment and that the time horizon and sampling times are the same as in the original design. Thus, the optimal experimental design problem could be formulated as: given a desired set of N2exp new experiments, find the values of P and S for each one which maximizes or minimizes a certain FIM-based criterion. This leads to a non-linear programming problem with differential constraints, which can be solved as discussed by Banga et al. (2002).

The above problem was solved for the case of the E-criterion and considering N2exp values of 16 and 10 experiments. We have compared various local and global solvers concluding that SSm outperforms all the other methods considered (see Figure 2). The new design of 16 experiments improves the value of the E-criterion (the one used for the optimisation) by one order of magnitude, and simultaneously also improves the others (see Table 1). However, it should be noted that the Modified E-criterion of the new design is still very large, indicating that the identifiability difficulties are still present, although in smaller degree.

Table 1. Comparison of the original vs. two improved experimental designs maximizing the E-criterion.

Criterion	**Original design**	**Improved (16 exp)**	**Improved (10 exp)**
E-criterion = $\lambda_{\max}\left(FIM^{-1}\right)$	**1.658e-02**	**1.404e-03**	**2.586e-03**
Modified E= $\lambda_{\max}(FIM)/\lambda_{\min}(FIM)$	1.682e+06	8.673e+05	1.443e+06
A-criterion = $trace\left(FIM^{-1}\right)$	6.040e-02	6.162e-03	1.181e-02
D-criterion = $\det(FIM)$	2.264e+161	8.799e+185	5.428e+177

5. Conclusion

In this contribution, we have presented a new metaheuristic, based on a modified Scatter Search, which increases very significantly the efficiency of the solution of parameter estimation problems while keeping robustness (i.e. avoiding convergence to spurious local solutions). The performance and capabilities of this new approach were illustrated considering a challenging benchmark problem presented by Moles et al. (2003) and recently studied by Rodriguez-Fernandez et al. (2006). Using the new algorithm (`SSm`) presented here, we were able to obtain better solutions for the estimation problem much faster (up to two orders of magnitude with respect to the best stochastic method).Further, the values of various real-valued functions of the Fisher Information Matrix for the original design revealed correlations among parameters which were causing most of the ill conditioning. By means of new optimal experimental designs, we have shown how this situation can be greatly improved.

Acknowledgements. *This work was supported by the European Community as part of the FP6 COSBICS Project (STREP FP6-512060).*

References

Banga, J.R., Versyck, K.J., Van Impe, J.F. (2002) Computation of optimal identification experiments for nonlinear dynamic process models: an stochastic GO approach. *Ind. Eng. Chem. Res.* **41**, 2425–2430.

Banga, J.R., Moles, C.G. and Alonso, A.A. (2003) Global optimization of bioprocesses using stochastic and hybrid methods. *In* Nonconvex Optimization and Its Applications. **74**, 45-70, Kluwer.

Bates, D. M. and Watts, D. G. (1988) Nonlinear Regression Analysis and its Applications. Wiley.

Martí, R. (2006) Scatter Search: wellsprings and challenges. *EJOR*. **169 (2)**, 351-358.

Moles, C. G., Pedro Mendes and Julio R. Banga (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*. **13 (11)**, 2467-2474.

Neumaier, A., Shcherbina, O., Huyer, W. and Vinko, T. (2005) A comparison of complete global optimization solvers. *Mathematical Programming*. **103(2)**, 335-356.

Rodriguez-Fernandez, M., P. Mendes and J. R. Banga (2006) A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*. **83**, 248-265.

Schittkowski, K. (2002). Numerical data fitting in dynamical systems. Kluwer Academic Publishers.

Vanrolleghem, P.A., Dochain, D. (1998) Bioprocess model identification. *Advanced Instrumentation, Data Interpretation, and Control of Biotechnological Process*. Kluwer Publishers, 251–318.

Walter, E., Pronzato, L. (1997) Identification of Parametric Models from Experimental Data. Springer.

Wolkenhauer, O. (2001) Systems biology: the reincarnation of systems theory applied in biology? *Briefings Bioinf*. **2 (3)**, 258–270.

Robust Stabilization of Inhomogeneous Patterns in a Reaction-Diffusion Biological System

Carlos Vilas, Míriam R. García, Julio R. Banga, and Antonio A. Alonso*

Process Engineering Group, IIM-CSIC, Vigo, Spain, e-mail: antonio@iim.csic.es

Keywords: Biological Waves, Reaction-Diffusion Systems, Robust Nonlinear Control, Reduced Order Models, Proper Orthogonal Decomposition, Spiral Control.

1. Abstract

Many biological phenomena such as neuron firing, cardiac rhythms or visual perception in the retina, as well as catalytic reactions or cellular organization activities, can be described by nonlinear reaction-diffusion (RD) mechanisms. A particular case of RD model is the FitzHugh-Nagumo (FHN). Slight variations of this model have been employed to represent travelling waves that induce the normal heartbeat or the formation of spirals, responsible for the arrhythmia phenomena of the heart. The objective of this paper is to propose a class of nonlinear feed-back controllers which avoid the spiral behaviour by stabilizing the plane front even in the presence of structural uncertainties (robust). In this way, we combine model reduction techniques with classical results on robust control to construct a class of nonlinear feed-back controllers ensuring front stabilization. Robustness and stabilizing properties of the controllers are proved through a numerical simulation experiment.

2. Introduction

Reaction-diffusion (RD) mechanisms are central in modelling a number of physiological systems such as those describing neural communication (Murray, 2002), cardiac rhythms (Fenton et al., 2002), visual perception in the retina (Gorelova and Bures, 1983) or cellular organization (Lebiedz and Maurer, 2004). One such model is the FitzHugh-Nagumo (FHN) system -see (Fitzhugh, 1961; Nagumo et al., 1962)-. This model is able to capture most of the qualitative behaviour of biological phenomena related with the normal operation or with disorders such as arrhythmia associated with the formation of spirals.
Dynamic analysis and control of RD systems has been the subject of intensive research, especially in what refers to bifurcation analysis leading to moving fronts, spiral waves and their stabilization. In (Pumir and Krinsky, 1999; Zykov

* Corresponding author

Understanding and Exploiting Systems Biology

and Engel, 2004) feed-back control schemes were designed to "unpin" spiral waves or to stabilize the spatiotemporal evolution of electrical waves in cardiac tissue.

Regarding theoretical work in the control of non-linear RD systems, it has been mostly focused on the stabilization of given stationary patterns by techniques which combine model reduction with results in non-linear control theory (Alonso and Ydstie, 2001; Christofides, 2001) to develop robust controllers which were able to stabilize arbitrary modes in RD systems. This approach was extended in (Alonso et al., 2004a) to develop robust nonlinear controllers which were able to stabilize arbitrary modes in RD systems and in (Vilas et al., 2006) where the authors developed a control law able to force the FHN system to follow a given non-stationary reference.

The aim of this work is to develop a class of feed-back controllers which, acting on a system exhibiting the spiral behaviour, are able to ensure the stabilization of the travelling plane wave (reference) even in the presence of structural uncertainties (robust). The logic of the controllers is built on the basis of a low dimensional approximation of the reference. Such approximation is constructed using the *Proper Orthogonal Decomposition* (POD) technique. The stabilizing and robustness properties of the control are proved through a simulation experiment. Note however that, although we concentrate on the FHN model, the same methodology could be applied to a wider range of RD systems of relevance in biology.

The paper is structured as follows: In Section 2, we introduce the POD technique. Next, we briefly describe the FHN model, state the control problem we will be dealing with and apply the POD technique to it. Finally, in section 5, we develop the control law and illustrate its performance on a numerical simulation experiment.

3. Low Dimensional Approximation of RD Systems Via the POD technique

Consider the following general parabolic system:

$$\frac{\partial u}{\partial t} = L(u) + \sigma(u)\,, \tag{1}$$

with appropriate boundary and initial conditions. In Eqn (1), u represents the vector field, $L(\cdot)$ a general linear operator and $\sigma(u)$ a given nonlinear function. As it was shown in (Alonso et al., 2004a) the solution (u) of system (1) can be expressed as a convergent infinite series of the form:

$$u(\xi,t) = \sum_{i=1}^{\infty} c_i(t)\phi_i(\xi)\,, \tag{2}$$

where ξ are the spatial coordinates. Time and spatial functions $c_i(t)$ and $\phi_i(\xi)$ are known as *the modes* and *the eigenfunctions*, respectively. In the POD method each $\phi_i(\xi)$ is obtained through measurements of the field (snapshots). Here, we

make use of an alternative to the *direct POD* method (Holmes et al., 1997; Alonso et al.,2004b), known as the *method of snapshots* (Sirovich, 1987), which requires lower computational effort. In this method, each eigenfunction is expressed in terms of the snapshots (u_i) as:

$$\phi_j = \sum_{i=1}^{\ell} z_i^j u_i \ , \tag{3}$$

where the weights z_i^j are calculated by solving the following eigenvalue problem

$$M\, Z_j = \lambda_j Z_j, \ \text{with} \ M_{ij} = \frac{1}{\ell}\langle u_i, u_j \rangle_\Omega \, , \tag{4}$$

where $Z_j = \left[z_j^1, ..., z_j^\ell \right]^T$ and the operator $<\cdot,\cdot>_\Omega$ indicates the spatial integral.
The eigenvalues (λ_j) resulting from Eqn (4) can be ordered so that $\Re(\lambda_1) \geq \Re(\lambda_2)$ (Christofides, 2001), where $\Re(\lambda_j)$ denotes the real part of λ_j. This property allows us to partition the set of modes $C = \{c_i\}_{i=1}^{\infty}$ in two subsets: $c_a = \{c_i\}_{i=1}^{k}$ containing modes with slow dynamics and $c_b = \{c_i\}_{i=k+1}^{\infty}$ containing modes with fast dynamics. The contribution of the fast modes (c_b) to the solution can be neglected so that an approximation (*reduced order model (ROM)*) to system (1) is obtained by projecting Eqn (1) over the set of eigenfunctions $S_a = \{\phi_i\}_{i=1}^{k}$.

$$\frac{dc_a}{dt} = P_{La} c_a + \langle \Phi_a, \sigma(u) \rangle_\Omega \, , \qquad u(\xi,t) \approx \Phi_a c_a \tag{5}$$

Remark that the larger the number of elements (k) in c_a, the better the quality of the approximation.

4. The FitzHugh-Nagumo Model

We are considering a 2D version of the classical FHN model. The spatial domain covers the square $\Omega = \{(x,y) / 0 \leq (x,y) \leq 200\}$ and the model equations are (Fenton et al., 2002):

$$\frac{\partial v}{\partial t} = \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + f(v) - w + p; \qquad f(v) = (\alpha - v)(v-1)v \, , \tag{6}$$

$$\frac{\partial w}{\partial t} = g(w) - \varepsilon\beta v; \qquad g(w) = \varepsilon(\gamma w - \delta) \, , \tag{7}$$

where v (the *activator*) is directly related to the membrane potential while w (the *inhibitor*), collects the contributions of ions such as sodium or potassium to the membrane current (Murray, 2002). The term p corresponds to the control variable which physically might correspond to spatially distributed electrodes supplying currents.

In the context of biological phenomena, such as cardiac or neural activity, the normal behaviour is related with a plane front which moves along the tissue without changing its shape (travelling plane wave). The FHN system (6)-(7) is able to reproduce such behaviour by setting the following parameters: $\alpha = 0.1$, $\varepsilon = 0.01$, $\beta = 0.5$, $\gamma = 1$, $\delta = 0$, and with initial conditions:

$$v(x,y,0) = v_0 = \begin{cases} 1 & if \quad 0 \le x \le 10 \quad \forall y \\ 0 & if \quad 10 < x \le 200 \quad \forall y \end{cases}; \quad w(x,y,0) = w_0 = 0 \ \forall (x,y) \,. \tag{8}$$

A snapshot of this solution is shown in Figure 1(a). Under certain circumstances the plane front can break forcing the front and the back of the wave to meet each other at a given point (Fenton et al., 2002). Thereafter, the broken front rotates around this point resulting in the formation of a spiral. In the context of biological systems, this behaviour can be related to neurological disorders or cardiac dysfunctions such as arrhythmia. The FHN is also able to capture this phenomenon as illustrated in Figure 1(b).

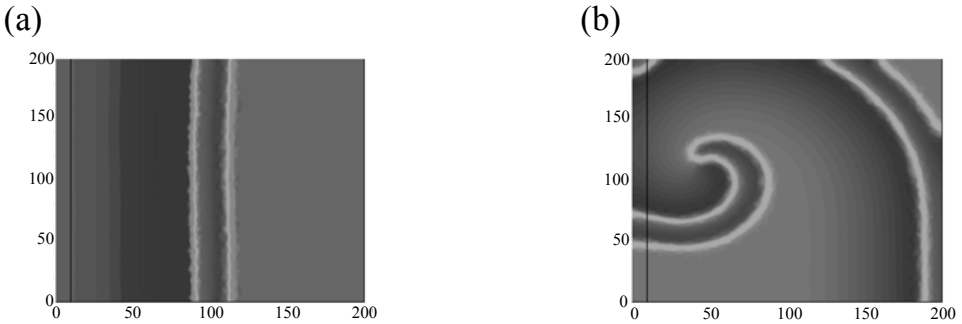


Figure 1. Snapshots for system (6)-(7) corresponding with **(a)** the front solution and **(b)** the spiral induced by the breaking of the front.

The aim of this contribution is *to design a feed-back control scheme* (p) *able to drive the spiral to the plane front reference*. Usually, in biological systems, there is a lack of detailed information on the structure of the nonlinear terms. In this way, our control law has to be able to "do the job" regardless the presence of model uncertainties. The underlying feedback control logic is designed on the basis of a ROM. The derivation of such ROM using the POD technique is described in the remaining of this section.

The first step to construct the ROM is the derivation of a representative set of data. In our case, the snapshots were obtained by solving system (6)-(8) with the finite element method (FEM) with *2342* discretization points (finer grids do not alter significantly the solution). Our representative data set is composed of *400* snapshots collected each *0.8* units of time. The POD basis is calculated using the method described in Section 3. Finally, the ROM is obtained by projecting Eqns (6)-(8) over the PODs.

Table 1 shows the number of equations resulting of using two different ROMs and the FEM. Remark that using ROMs results in reductions of two orders of magnitude.

Table 1. Comparison between the ROMs and the FEM.

Method	**Equations for the *v*-field**	**Equations for the *w*-field**	**Model Equations**
First ROM	20	9	29
Second ROM	32	13	45
FEM	2342	2342	4684

In Figure 2(a) a comparison between the modes obtained with the FEM (lines) and with the ROMs (marks) is depicted. For clarity reasons, only three modes are represented. In this picture, it is shown that the first ROM is able to reproduce only at a qualitative level the system behaviour, while the second ROM results in a much better approximation to the FEM scheme. In fact, by recovering the field from the modes of the second ROM (Figure 2(b)) the essential features of the real model (Figure 1(a)) are preserved.

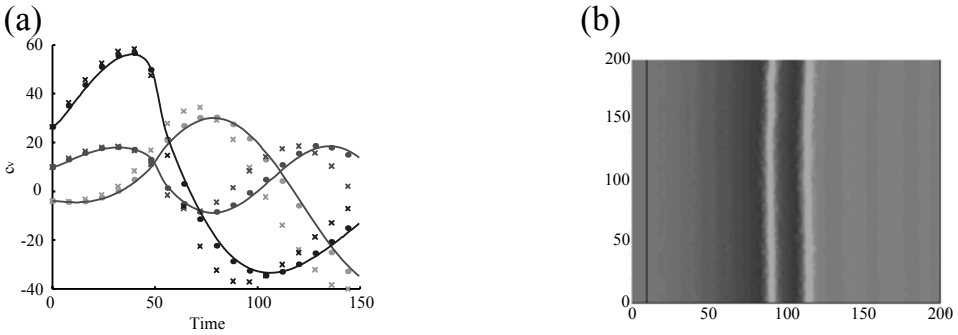


Figure 2. (a) Comparison between the modes obtained with the FEM (lines) and using ROMs with 29 POD (x) and 45 POD (circles). (b) *v*-field spatial distribution obtained with the second ROM.

5. The Control Law

As mentioned in section 4, the spiral behaviour in the FHN system is related with physiological problems such as arrhythmia or neural disorders. On the other hand, the travelling plane wave is associated with, for instance, the normal operation of the heart. The natural control objective should be, therefore, to actuate on a system exhibiting the spiral so as to produce and maintain the plane front.

We have shown in the previous section that the FHN model can be approximated by a low dimensional system of ODEs (ROM). Each ODE of the ROM describes the evolution of a mode. We will call "relevant" modes to those which compose the ROM and "non relevant" to the others. The idea is to use the ROM of the travelling plane wave as the reference trajectory.

The control objective is separated in two: on the one hand to stabilize, in the spiral behaviour, the "non relevant" modes of the reference and, on the other hand, to force the "relevant" modes to follow the reference. Based on classical

results on the robust control of finite dimensional systems, namely the Lyapunov redesign technique (Khalil, 1996), we propose the following control law:

$$\overline{p}_a = \begin{cases} \overline{w}_a - \omega^* \overline{v}_a - \eta^* \dfrac{\overline{v}_a}{\|\overline{v}_a\|_\Omega} & if \quad \eta^* \|\overline{v}_a\|_\Omega \geq \theta_a \\ \overline{w}_a - \omega^* \overline{v}_a - (\eta^*)^2 \dfrac{\overline{v}_a}{\theta_a} & if \quad \eta^* \|\overline{v}_a\|_\Omega < \theta_a \end{cases}, \quad \overline{p}_b = \begin{cases} \overline{w}_b - \omega \overline{v}_b - \eta \dfrac{\overline{v}_b}{\|\overline{v}_b\|_\Omega} & if \quad \eta \|\overline{v}_b\|_\Omega \geq \theta_b \\ \overline{w}_b - \omega \overline{v}_b - \eta^2 \dfrac{\overline{v}_b}{\theta_b} & if \quad \eta \|\overline{v}_b\|_\Omega < \theta_b \end{cases} \tag{9}$$

The sub-indices *a* and *b* indicate that the subfield or the control is associated with the "relevant" and "non relevant" modes, respectively. As mentioned before, the control law is robust, this means that it do not need detailed information on the structure on the nonlinear term. However, some information is required. In this case, the information consists of given bounds on the nonlinear terms. These bounds are included in the control law through η and η^*.

Note that this control law needs to take measurements of the field and to actuate over the whole domain. In this regard, the fields $\overline{v}_a$ and $\overline{v}_b$ are calculated from these measurements through:

$$\overline{v}_a = \sum_{i=1}^{k} \phi_i \langle \phi_i, \overline{v} \rangle_\Omega ; \quad \overline{v}_b = \overline{v} - \overline{v}_a .$$

The price to pay for robustness is that we cannot ensure the convergence to the reference but to a zone around it. It is important to remark that such region can be made arbitrarily small by decreasing the parameters θ_a and θ_b, although at the expense of stronger control effort.

In order to illustrate how the control law (9) stabilizes the front behaviour, we have applied it to the FHN system (6)-(8) in a simulation experiment. Such experiment consisted of introducing, at a given time, a perturbation in order to produce the spiral behaviour and finally, when the spiral is formed, switching on the control law so as to stabilize the front.

The effect of the controllers on the *v*-field is illustrated in Figure 3 where three snapshots are depicted. The system initially evolves as a spiral (Figure 5(a)) and, when the control law is switched on, the reference is reached (Figure 5(c)) after a short transition period (Figure 5(b)).

The differences between the reference and the controlled field can be arbitrarily reduced by decreasing the values of θ_a and θ_b but at the expense of a higher control effort.

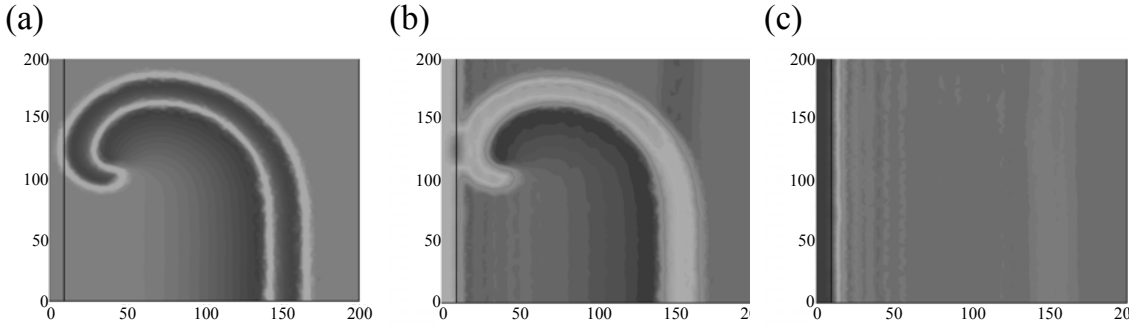


Figure 3. Evolution of the FHN system under the control law (9): (a) in open loop, (b) an instant after entering the control law and (c) under the control law.

6. Conclusions

In this paper, we have designed a class of controllers able to force a system exhibiting the spiral behaviour to follow a reference trajectory given by a travelling plane wave. The control law was designed in such a way that structural uncertainties do not affect to its stability properties. To that purpose, we have combined classical results on robust nonlinear control with the POD technique. On the basis of the ROM, the control law was constructed so as to force the representative modes to follow the reference trajectory, which encodes the plane front, while stabilizing the others. The stability properties were illustrated on a simulation experiment. Although these controllers were applied to the FHN system, they could be easily extended to other class of RD systems.

Acknowledgements. *The authors acknowledge financial support received from the Spanish Government (MCyT Project DPI2004-07444-C04-03) and Xunta de Galicia (PGIDIT02-PXIC40209PN).*

References

Alonso, A.A., Fernández, C.V. and Banga, J.R. (2004a). Dissipative Systems: from physics to robust nonlinear control. *Int. J. Robust Nonlinear Control*. **14** (2)**,** 157-179.

Alonso, A. A., Frouzakis, C.E and Kevrekidis, I.G. (2004b). Optimal sensor placement for state reconstruction of distributed process systems. *AIChE Journal.* **20** (7), 1438-1452.

Alonso, A.A. and Ydstie, B.E. (2001). Stabilization of Distributed Systems Using Irreversible Thermodynamics. *Automatica.* **37** (11), 1739-1755.

Christofides, P. D. (2001). Nonlinear and Robust Control of PDE Systems: Methods and Applications to Transport-Reaction Processes, Birkhäuser, Boston.

Fenton, F. H., Cherry, E.M., Hastings, H.M. and Evans, S.J. (2002). Real-time computer simulations of excitable media: JAVA as a scientific language and as a wrapper for C and FORTRAN programs. *Biosystems*. **64** (1-3): 73-96.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophys. J.* **1** (6), 445-466.

Gorelova, N.A. and Bures, J. (1983). Spiral waves of spreading depression in the isolated chicken retina. *Journal of Neurobiology.* **14** (5), 353-363.

Holmes, P.J., Lumley, J.L., Berkooz, G., Mattingly, J.C. and Wittenberg, R.W. (1997). Low-dimensional models of coherent structures in turbulence. *Physics Reports*. **287** (4), 337-384.

Khalil, H. K. (1996). Nonlinear Systems, Prentice Hall, Upper Saddle River, New Jersey.

Lebiedz, D. and Maurer, H. (2004). External Optimal Control of Self-Organisation Dynamics in a Chemotaxis Reaction Diffusion System. *IEE Systems Biology*. **2**, 222-229.

Murray, J. D. (2002). Mathematical Biology II: Spatial Models and Biomedical Applications. Springer-Verlag, Berlin.

Nagumo, J., Arimoto, S. and Yoshizawa, Y. (1962). Active pulse transmission line simulating nerve axon. *Proc. Inst. Radio. Eng.* **50** (10), 2061-2070.

Pumir, A. and Krinsky, V. (1999). Unpinning of a rotating wave in cardiac muscle by an electric field. *Journal of theoretical biology.* **199** (3), 311-319.

Sirovich, L. (1987). Turbulence and the Dynamics of Coherent Structures. Part I: Coherent Structures. *Quaterly of Appl. Math.* **45** (3), 561-571.

Vilas, C., García, M.R., Banga, J.R. and Alonso, A.A. (2006). Stabilization of inhomogeneous patterns in a diffusion-reaction system under structural and parametric uncertainties. *Journal of Theoretical Biology*. In press. Available on-line on http://www.sciencedirect.com

Zykov, V. and Engel, H. (2004). Feedback-mediated Control of Spiral Waves. *Physica D-Nonlinear phenomena*. **199** (1-2), 243-263.

SYSTEMS BIOLOGY APPLICATIONS: BIOMEDICINE

Computational Design of Optimal Dynamic Experiments in Systems Biology: a Case Study in Cell Signalling

E. Balsa-Canto*, M. Rodríguez-Fernández, A.A. Alonso, and J.R. Banga

Process Engineering Group. IIM- CSIC (Spanish Council for Scientific Research), Vigo, Spain. e-mail: ebalsa@iim.csic.es

Keywords: model calibration, optimal experimental design, cell signalling.

1. Abstract

Mathematical models of complex biological systems, such as cell signalling cascades, usually consist of sets of nonlinear ordinary differential equations which depend on several non measurable parameters that must be estimated by fitting the model to experimental data. This model calibration is performed by minimizing the differences between model predictions and measurements. Optimal experimental design (OED) aims to design an scheme of actuations and measurements which will result in data sets with the maximum amount and/or quality of information, as measured by the Fisher Information Matrix, for the subsequent model calibration.
This work presents new computational procedures for OED in the context of systems biology, with a focus on cell signalling. The OED problem is formulated as a general dynamic optimization problem and its solution is approached using a combination of the control vector parameterization approach and a robust global non-linear programming solver.
The applicability and advantages of using optimal experimental design are illustrated by considering a mitogen-activated protein (MAP) kinase cascade, which is frequently involved in larger cell signalling pathways, and it is known to regulate several cellular processes of major importance.

2. Introduction

More than 10% of the proteins encoded in the human genome are involved in intracellular signalling cascades which regulate the typical cellular responses such as growth, division, differentiation and apoptosis (Pelech, 2004). The malfunction of these signalling pathways, particularly those involving phosphorylation cascades, has a strong relationship with the development of

* Corresponding author

Understanding and Exploiting Systems Biology

diseases including cancer, diabetes, Alzheimer's disease or Parkinson's disease (Olive, 2004).

The aim of modelling cell signalling pathways is to provide a systematic framework to generate hypothesis and make predictions "in silico", in order to get a better insight into the disease process and ultimately to identify potential drug targets (Butcher et al., 2004).

In particular, the modelling and simulation of cellular signalling pathways as networks of biochemical reactions has received major attention during recent years (see the review by Kholodenko, 2006). Most proposed models assume that the system is well-mixed and the mass conservation law results in sets of non-linear ordinary differential equations (ODEs). When the spatial variations of the magnitudes of interest are relevant, compartmental models or partial differential equations have to be considered (see for example, Haugh and Lauffenburger, 1998; Fallon and Lauffenburger, 2000).

These models depend on several parameters (kinetic constants, molecular diffusivity constants, etc.) and probably some initial conditions (initial concentration or number of molecules of some proteins) which are not accessible to experimental determination and must therefore be estimated by fitting the model to experimental data (model calibration).

The model calibration is performed by minimizing a cost function which quantifies the differences between model predictions and measurements. However model calibration may only be performed successfully if the sources of information are of a sufficiently high quality. Unfortunately, experiments in molecular biology are usually time consuming and expensive and rarely produce large and accurate data sets (Kutalik et al., 2004). Concerning this, the following question should be answered: can the parameters be given unique values using a particular experimental procedure? As illustrated later in this contribution the answer to this question may be negative, therefore a careful experimental design is required.

Optimal experimental design consists of the determination of the scheme of measurements that generates the maximum amount of information for the purpose of estimating the parameters with the greatest precision and/or decorrelation (see for example, Asprey and Macchietto, 2002). The amount and quality of information can be measured in terms of a scalar function of the Fisher Information Matrix (FIM) computed for a given (near-optimal) value of parameters. In the context of cell signalling, Faller et al.(2003) made use of simulation based techniques to calculate polynomial optimal input profiles in order to enhance parameter estimation accuracy for a MAP kinase cascade; Kutalik et al. (2004) proposed the calculation of optimal sampling times so as to reduce the variation of the parameter estimates.

Here, the optimal experimental design problem is formulated as a more general dynamic optimisation problem (see for example, Banga et al., 2002, or Asprey and Macchietto, 2002) and its solution is approached using the so called control vector parameterization approach (CVP, Vassiliadis, 1993). The CVP scheme proceeds dividing the duration of the experiment (time horizon) into a number of elements, and approximating the input functions inside these elements using low

order polynomials. As a result, a non-linear programming problem (NLP) is obtained, where the decision variables are the polynomial coefficients plus the sampling times and possibly the experimental initial conditions. The evaluation of the objective function requires the simulation of the system dynamics plus the calculation of the parametric sensitivities to compute the Fisher Information Matrix (FIM). Remark that the non-linear character of the mathematical models of the cell signalling pathways lead to multi-modal NLPs therefore the use of global optimization methods is required.

3. Theoretical

Model development can be regarded as a cycle comprising several phases. Once the model structure has been established based on a priori phenomenological knowledge and hypothesis, experimental data is used to estimate the model unknown parameters. This task is often rather complicated, mainly due to the following reasons (Rodriguez-Fernandez et al, 2006):

- the presence of a large number of parameters (usually dozens, or even hundreds)
- the multimodal character of the optimization problem, i.e. the presence of several sub-optimal solutions
- the presence of identifiability problems, that is, the impossibility of calculating unique values for all parameters.

In order to detect and hopefully reduce such kind of problems, this work proposes an iterative procedure (see Figure 1 for a schematic representation) which involves the use of *parametric sensitivities* to *rank the parameters*; the computation of *collinearity indexes* to evaluate indentifiability problems and finally, the solution of an *optimal experimental design* problem for parameter estimation.

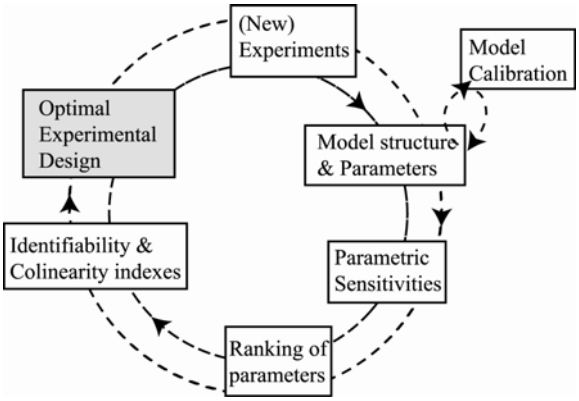


Figure 1: Experimental design phases

3.1. PARAMETRIC SENSITIVITIES AND RANKING OF PARAMETERS

Most of the mathematical models proposed to describe cell signalling behaviour consist on sets of non-linear ordinary differential equations (ODEs) as follows:

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, t) \tag{1}$$

$$\mathbf{x} = \mathbf{h}(\mathbf{y}, \mathbf{u}, \boldsymbol{\theta}, t) \tag{2}$$

where $\mathbf{y} \in \mathbf{R}^n$ represents the vector of states and $\mathbf{x} \in \mathbf{R}^{n_o}$ the vector of observables, usually related to the proteins involved on the cascade, $\mathbf{u} \in \mathbf{R}^{n_u}$ corresponds to the vector of possible stimuli, $\boldsymbol{\theta} \in \mathbf{R}^{n_\theta}$ is the vector of model parameters and t is the time.

Parametric sensitivities measure how the model output is affected by an slight modification of the parameters. Although other possibilities exist, the use of the absolute and relative local parametric sensitivities, as formulated in Eqn.(3), was selected in this work.

$$\text{(a) } S_{ik}(t) = \frac{dx_i}{d\theta_k}; \quad \text{(b) } s_{ik}(t) = \frac{\Delta\theta_k}{\Delta x_i}\frac{dx_i}{d\theta_k} \quad \forall i = 1,\ldots,n_o; \quad \forall k = 1,\ldots,n_\theta \tag{3}$$

Note that a linear (or almost linear) relationship between sensitivities will mean high correlation between the corresponding parameters, i.e. lack of identifiability. Therefore, in order to analyze the parameter identifiability problem in more detail, it would be convenient to plot s_{ij} versus s_{ik} for all i, j and k, so near-linear relationships between parameters could be detected. Unfortunately, this approach is not practical as it would mean an extremely large number of plots even for a reduced number of parameters.

When the number of parameters in the model is relatively large, a typical approach is to partition the vector of parameters $\boldsymbol{\theta}$ into two new vectors $\boldsymbol{\theta}_\kappa$ and $\boldsymbol{\theta}_{\bar{\kappa}}$, where κ is a subset of $\kappa \le n_\theta$ parameters, in such a way that only the components of $\boldsymbol{\theta}_\kappa$ are to be estimated from the experimental data whereas the parameters in set $\boldsymbol{\theta}_{\bar{\kappa}}$ are kept constant. This partitioning is not straightforward, but a ranking of parameters and the evaluation of collinearity indexes may be helpful.

The relative sensitivities can be used to asses the individual local parameter importance, that is to establish a ranking of parameters. Brun and Reichert (2001) suggested several importance factors:

$$\delta_j^{msqr} = \sqrt{\frac{1}{n_s}\sum_{i=1}^{n_s} s_{ij}^2}; \quad \delta_j^{mabs} = \frac{1}{n_s}\sum_{i=1}^{n_s}\left|s_{ij}\right|; \quad \delta_j^{mean} = \frac{1}{n_s}\sum_{i=1}^{n_s} s_{ij}$$
$$\delta_j^{\max} = \max_i s_{ij}; \quad \delta_j^{\min} = \min_i s_{ij} \tag{4}$$

The collinearity indexes (Brun and Reichert, 2001) are used to check the near-linear dependence within the columns of matrix s. Considering the normalized matrix $\bar{s}$, such that $\bar{s}_j = \frac{\mathbf{s}_j}{\|\mathbf{s}\|}$, then the collinearity index results:

$$\gamma_\kappa = \frac{1}{\sqrt{\lambda_\kappa}}, \quad \lambda_\kappa = \min(eig(\bar{\mathbf{s}}^T \bar{\mathbf{s}})) \tag{5}$$

This definition can be interpreted as follows: a change in the output of the model caused by a shift of a parameter on set κ can be approximately compensated by appropriate changes in the other parameters in κ. A parameter subset κ is said to be potentially identifiable if the observed model output is sufficiently sensitive to small changes in the parameters and if the collinearity index does not exceed a critical value (around 20).

3.2. OPTIMAL EXPERIMENTAL DESIGN

The optimal experimental design (OED) problem may be mathematically formulated as a general dynamic optimisation problem:
Find the time-varying stimuli profiles (control variables), sampling times and (possibly) initial conditions, so as to minimize (or maximize) an scalar performance index related to the Fisher Information Matrix:

$$J_{OED} = \phi(FIM) \tag{6}$$

subject to the system dynamics, Eqn.(1), and other algebraic constraints related to proper/safe operation or experimental limitations:

$$\mathbf{g}(\mathbf{y}, \mathbf{u}, \mathbf{v}, t) = 0\,; \tag{7}$$

$$\mathbf{u}^{\mathbf{L}}(t) \le \mathbf{u}(t) \le \mathbf{u}^{\mathbf{U}}(t)\,; \tag{8}$$

$$\mathbf{v}^{\mathbf{L}} \le \mathbf{v} \le \mathbf{v}^{\mathbf{U}} \tag{9}$$

The Fisher information matrix (FIM) may be defined using the first order absolute sensitivities as follows (Ljung,1999):

$$FIM = \sum_{i=1}^{n_o} \sum_{j=1}^{n_s} \nabla x_{ij}^T Q_{ij} \nabla x_{ij} \tag{10}$$

with Q_{ij} a constant matrix representing the variance of each experimental data and $\nabla x_{ij} = \frac{\partial x_i}{\partial \boldsymbol{\theta}}(t_j \mid \boldsymbol{\theta}^*)$, being $\boldsymbol{\theta}^*$ a good estimate for the parameters.

Regarding the scalar function of the FIM, several alternatives have been proposed in the literature (Vanrolleghem and Dochain, 1998), like e.g.:

A- optimality: $\phi_A = trace(FIM^{-1})$

D-optimality: $\phi_D = -\det(FIM)$

E-optimality: $\phi_E = -\lambda_{\min}(FIM)$

Modified E-optimality: $\phi_\varepsilon = abs\left(\frac{\lambda_{\max}(FIM)}{\lambda_{\min}(FIM)}\right)$

being $\boldsymbol{\lambda}$ the vector of eigenvalues of the Fisher information matrix.

The following interpretation can be given to each of these criteria (see for example, Vanrolleghem and Dochain, 1998 or Hidalgo and Ayesa, 2001): A-optimality and D-optimality designs tend to minimize the arithmetic and geometric mean of the identification errors respectively, while the E-optimality aims at minimizing the largest error.
The most widely used are D and E-modified optimality. D-optimality designs result in the smallest volume of the confidence region in the parameter space. It is directly related to the volume of the information hyper-ellipsoid and indicates the quantity of information provided by the experiments. Note that its value increases as the measurements are processed, but it gives no clue about the way the information is distributed among the parameters. The information eccentricity (square root of the E-modified criterion) characterizes this distribution. Geometrically, it represents the relationship between the longest and shortest semi-axes of the information hyper-ellipsoid. The closer its value is to one, the more homogeneous the distribution of the information among the unknown parameters will be. The modified E-criterion has the additional advantage that its theoretical lower bound is known (1.0).

3.2.1. Solution approaches: Control vector parameterisation
The solution algorithms for dynamic optimisation problems can be classified in three main groups: dynamic programming methods, indirect methods, and direct methods. The most popular, direct methods, transform the original infinite dimension dynamic optimisation problem into a finite dimension non-linear programming problem (NLP). Three are the main direct methods: multiple shooting, complete parameterization and control vector parameterisation.
The multiple shooting approach (Bock and Plitt, 1984) divides the duration of the time domain into a number of separate elements, and an initial value problem solver is used to simulate the process within each element. In this formulation, the initial conditions for each element together with the input parameters become the decision variables for the master NLP. In the so called complete parameterisation (CP) technique both the input and state variables are discretised, usually employing a direct collocation approach, so that the coefficients and interval lengths now become the decision variables in a larger NLP (see recent review by Biegler et al., 2002). The control vector parameterisation (CVP) method proceeds dividing the duration of the experiment into a number of elements and approximating the input variables using low order polynomials (Vassiliadis, 1993).
The CVP method is selected in this work due to its ability to handle large dynamic optimisation problems without solving excessively large NLPs (Balsa-Canto et al., 2004). Once the CVP has been applied the general OED problem is transformed into a NLP, being the decision variables the polynomial coefficients, plus the experimental sampling times and initial conditions. Note that the solution of this NLP requires a suitable NLP solver and an IVP solver.

Regarding the NLP solver, both local and global methods may be selected. However several tests performed revealed the multimodality of the OED problem thus the use of global techniques is necessary.
Regarding the IVP solver, a backward differentiation formulae based method is used to compute both states and parametric sensitivities.

4. Case study: Results and Discussion

MAP kinase family members have been found to regulate diverse biological functions by phosphorylation of specific target molecules (such as transcription factors, other kinases, etc.) found in cell membrane, cytoplasm and nucleus, and thereby participate in the regulation of a variety of cellular processes including cell proliferation, differentiation, apoptosis and immuno responses (Seger and Krebs, 1995). We consider here a simple signalling cascade in which stimulation of a receptor leads to a consecutive activation of several downstream protein kinases. The signal output of this pathway is the phosphorylation of the last kinase which can raise a cellular response. Signalling is terminated by phosphatases that dephosphorylate the kinases and by the inactivation of the receptor (Figure 2).

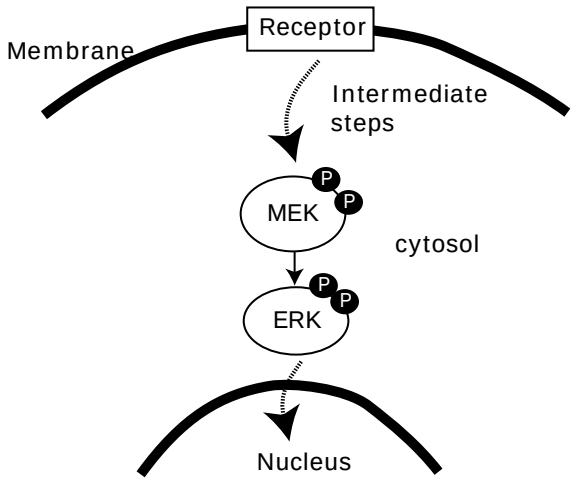


Figure 2: MEK-ERK pathway

The application of the mass action law to each of those reactions result in the following set of non-linear ordinary differential equations (Heinrich, 2002):

$$\frac{d[Erk^{*}]}{dt} = c_1[Erk - Mek^{**}] - a_3[Erk^{*}][Mek^{**}] + b_3[Erk^{*} - Mek^{**}] + c_4[Erk^{**} - Pase] - a_2[Erk^{*}][Pase] + b_2[Erk^{*} - Pase] \quad (11)$$

$$\frac{d[Erk^{**}]}{dt} = c_3[Erk^{*} - Mek^{**}] - a_4[Erk^{**}][Pase] + b_2[Erk^{*} - Pase] \quad (12)$$

$$\frac{d[Erk - Mek^{**}]}{dt} = a_1[Erk][Mek^{**}] - (b_1 + c_1)[Erk - Mek^{**}] \quad (13)$$

$$\frac{d[Erk^{*}-Mek^{**}]}{dt}=a_3[Erk^{*}][Mek^{**}]-(b_3+c_3)[Erk^{*}-Mek^{**}] \quad (14)$$

$$\frac{d[Erk^{*}-Pase]}{dt}=a_2[Erk^{*}][Pase]-(b_2+c_2)[Erk^{*}-Pase] \quad (15)$$

$$\frac{d[Erk^{**}-Pase]}{dt}=a_3[Erk^{**}][Pase]-(b_4+c_4)[Erk^{**}-Pase] \quad (16)$$

since the model assumes total Erk and Pase concentrations are constant, the nonactivated Erk is given in terms of the total Erk concentration and the phosphatase is given in terms of the total Pase, as follows:

$$[Erk](t)=[Erk]_{Total}-[Erk^{*}](t)-[Erk^{**}](t) \quad (17)$$

$$[Pase](t)=[Pase]_{Total}-[Erk^{*}-Pase](t)-[Erk^{**}-Pase](t) \quad (18)$$

The parameters a_i (=0.5) denote the rates at which the substrate binds to the enzyme, b_i (=0.6) denote the corresponding breaking rates, and c_i (=0.9) denote the rate at which the actual activation reaction occurs. The initial concentrations of all phosphorylated Erks and complexes of phosphorilated Erks with Meks or phosphatases are zero. Mek** serves as input, and Erk** as the output of the system, Erk (t=0)=50 and Pase(t=0)=20.

4.1. LOCAL PARAMETRIC SENSITIVITIES AND RANKING OF PARAMETERS

The sensitivities were computed by the numerical solver ODESSA (Leis and Cramer, 1988) which makes use of the direct decoupled approach which basically exploits the fact that the original ordinary differential system and the corresponding parametric sensitivities share the same Jacobian so as to increase efficiency in the simulation process.

Ranking the parameters by one of the values in Eqns. (4), preferably in decreasing order, results in a parameter importance ranking. Note that these importance factors will depend on time, thus two figures are presented bellow, first corresponds to final time, and second to an intermediate instant, particularly when $\delta_j^{msqr}(t)$ achieves its maximum value.

Note that at final time values for the different parameters are quite similar (up to one order of magnitude different for t=1.95) revealing that all parameters have a notable effect on model output. The parameter importance but also the ranking of parameters varies with time, however the output seems to be less sensitive to parameters a_2, b_2, b_3, b_4 and c_2 for any case as illustrated in Figures 3.

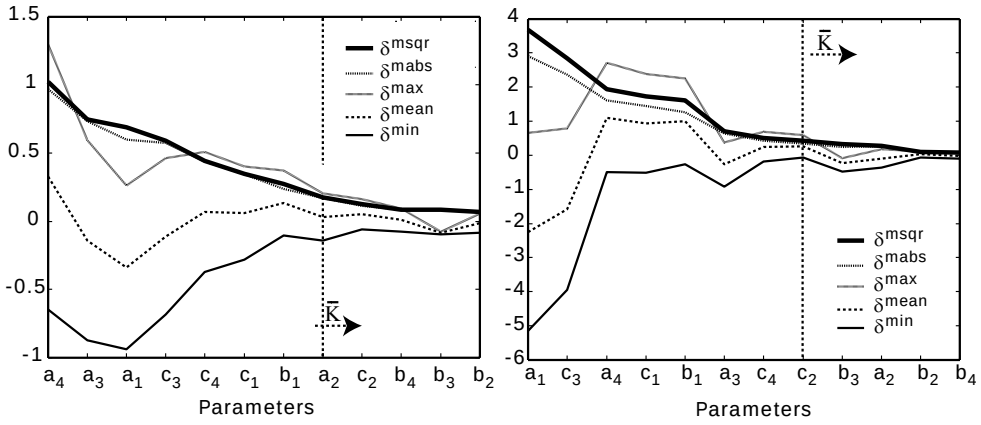


Figure 3: Ranking of parameters: a) At final time, b) At t=1.95, where δ^{msqr} achieves its maximum

As the number of parameters in this case is relatively large (around 4000 possible combinations), only the most relevant (a_1, a_3, a_4, b_1, c_1, c_3, c_4) will be considered. The most correlated parameters resulted to be $a_1$and b_1. Figure 4 shows all possible combinations.

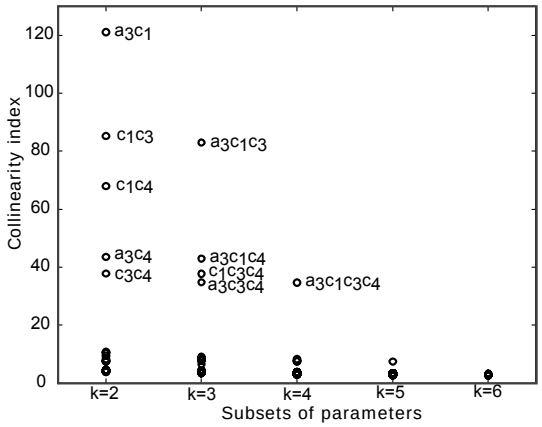


Figure 4: Collinearity indices for all parameter subsets.

From the results it is clear that some couples of parameters will have large identifiability problems, especially a_1b_1 and a_3c_3 (note that these pairs where not represented in the plot due to scale matters) and some combinations of 3 and even 4 parameters.
Considering the previous analysis, and for illustrative purposes, the optimal experimental design for the estimation of the pair a_1b_1 will be considered in following sections.

4.2. OPTIMAL EXPERIMENTAL DESIGN

As it was stated before the design of experiments requires several choices, e.g., how to manipulate the inputs, when to measure, which are the initial conditions, etc. For this example several possibilities will be considered assuming that the length of the experiment and the initial conditions are known:

- Compute the optimal input profile for a given number of equidistant measurements. The control variable will be approximated by ρ = 5 and ρ = 10 different length steps.
- Compute the optimal input profile plus the optimal sampling times for a given number of measurements. The control variable will be approximated by ρ = 5 variable length steps.

As it was mentioned before, the E-modified criterion allows to distribute information between parameters, making the confidence region as round and possible, as the objective ϕ_ε approaches the global solution (=1). The knowledge of the global solution makes this criterion very attractive for OED purposes. Note moreover, that a global optimization method was used here (Differential Evolution, Storn and Price, 1997) so as to avoid possible convergence to local solutions (Banga et al, 2002).

Table 1: Results obtained for the OED problem with different number of equidistant measurements

n_s	ϕ_ε	Confidence intervals	Det(FIM)
200	1.000028	(0.499, 0.501); (0.599, 0.601)	3.61×10^{13}
30	1.000074	(0.495, 0.505); (0.595, 0.605)	3.39×10^{10}
10	1.000265	(0.490, 0.510); (0.590, 0.610)	1.55×10^{9}

Table 2: Results obtained for the OED problem with different number of non equidistant measurements

n_s	ϕ_ε	Confidence intervals	Det(FIM)
10	1.000012	(0.491, 0.509); (0.591, 0.609)	2.08×10^{9}
5	1.000010	(0.484, 0.516); (0.584, 0.616)	2.44×10^{8}
3	1.000001	(0.480, 0.520); (0.580, 0.620)	$9{,}71 \times 10^{7}$

Some remarks:

- The optimization reaches near-global results in all cases. As a consequence, the regions of confidence are all approximately round (that is, the information is equally distributed among parameters). However it is also important to note that, as the flexibility in the stimulus profiles decreases, the E-modified criterion rapidly increases.

- The experiments become less informative as the number of equidistant measurements decreases, which in the end results in larger confidence regions for the parameters.
- From Table 2, it becomes clear that the adequate selection of sampling times increases the information provided by the experiments. Note for example that the use of 10 non equidistant measurements results in a more informative experiment than the one using 30 equidistant measurements. Even the use of 5 non equidistant measurements offers good enough performance with a very small confidence region.

4.3. ADVANTAGES OF USING OPTIMAL EXPERIMENTAL DESIGN

In order to show the advantages of using OED, we consider the problem of estimating model parameters given a sub-optimal experimental design consisting on a constant stimulus Mek** = 4.0 and 10 equidistant sampling times, and the optimal stimulus profile obtained using 5 steps and 10 non equidistant measurements (as illustrated in Figure 5).

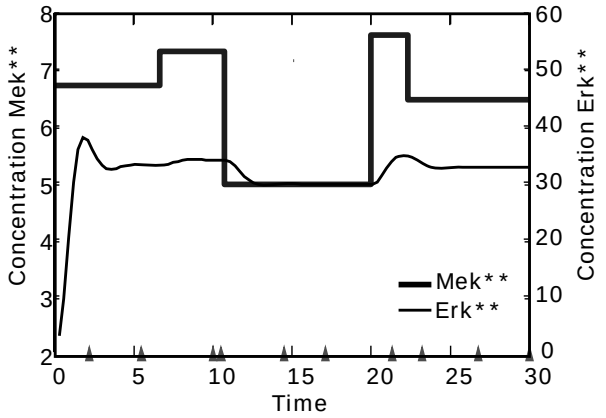


Figure 5: OED using E-modified criterion

The following figures present surface and contour plots of the least squares function to be minimised to estimate the model parameters:

$$J(\boldsymbol{\theta}) = \mathbf{e}(\boldsymbol{\theta})^T \mathbf{Q} \mathbf{e}(\boldsymbol{\theta}); \qquad \mathbf{e}(\boldsymbol{\theta}) = \mathbf{x}(\boldsymbol{\theta}) - \tilde{\mathbf{x}} \tag{17}$$

where $\mathbf{x}$ corresponds to the model predictions and $\tilde{\mathbf{x}}$ to the measured data. Note that log scale has been used for the plots so as to make them more clear.

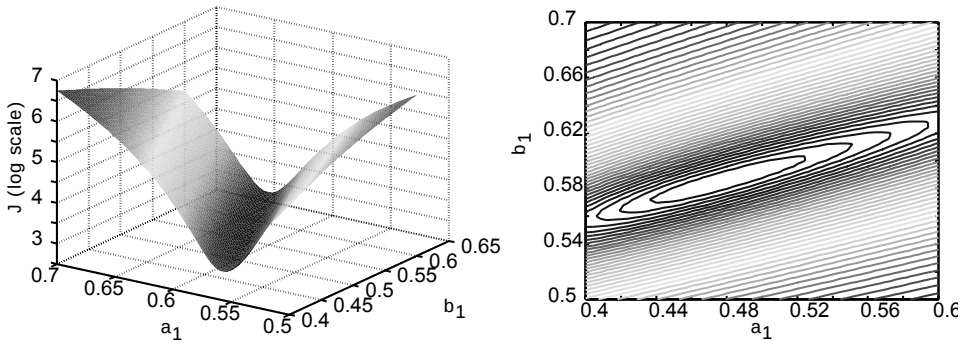


Figure 6: $J(\boldsymbol{\theta}_\kappa)$ for a typical case with Mek** = 4.0 and n_s = 10.

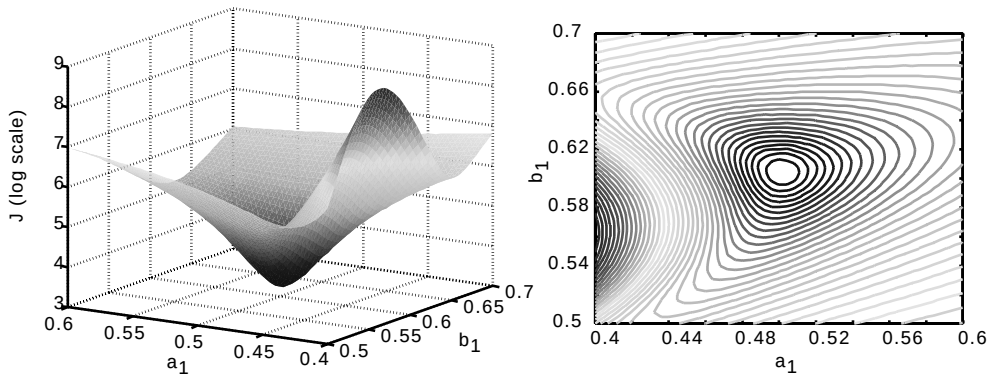


Figure 7: $J(\boldsymbol{\theta}_\kappa)$ for a case under the optimal profile obtained with E modified criterion and $n_s = 10$.

4.4. PARAMETER ESTIMATION USING THE OPTIMAL EXPERIMENTAL CONFIGURATION

In order to check whether the OED obtained will provide successful results in parameter estimation we generate a number of pseudo-experiments and solve the corresponding parameter estimation problems. To generate the pseudo-experimental data, observational noise is introduced to the system in the following manner:

$$\tilde{\mathbf{x}} = \mathbf{x} + \sigma(\mathbf{x} \cdot \mathbf{r}) \tag{20}$$

where **r** represents the normally distributed random variable vector with zero mean and unit standard deviation and σ represents the standard deviations of the observation errors.

As illustrated in Figure 6, for non optimal experimental designs contour plots with long flat valleys are obtained indicating strong dependencies of parameter estimates (poor identifiability). The minimization of $J(\boldsymbol{\theta}_\kappa)$ may result in any combination of the parameters.

However for the case of using E-modified designs, the least squares functions tend to be convex (Figure 7), guarantying an unique optimum to the parameter estimation problem, which can be found even using a local optimization method. Note that the optimal solutions for these particular cases, may or not coincide exactly with the nominal value of the parameters but they are in the so called confidence region.

Following figure illustrates the theoretical and practical confidence region for the E-modified design using 500 pseudo-experiments.

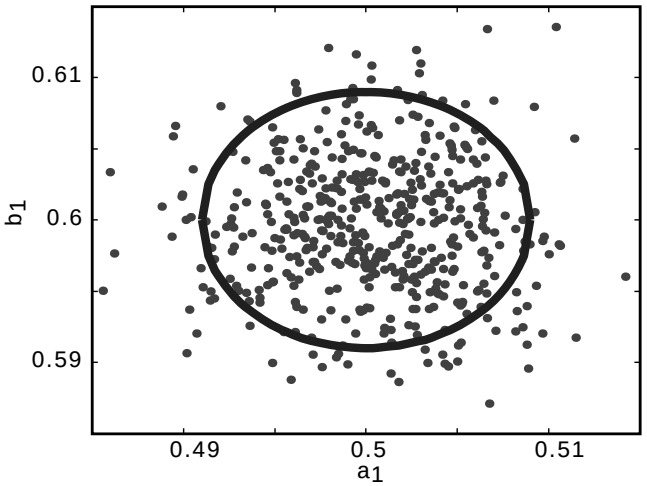


Figure 8: Theoretical (line) and practical (dots) confidence regions for the E modified optimal experimental design

5. Conclusion

In this contribution we illustrate the potential of optimal experimental design of dynamic experiments for parameter estimation (model calibration) in systems biology.
Reliable model calibration in systems biology will largely depend on the quantity and quality of the experimental data. This work proposes the use of an iterative process based on the computation of a ranking of parameters, collinearity indexes and finally optimal experimental designs with the aim of improving model calibration conditions.
The results obtained for a simple signalling pathway clearly indicate that dynamic experiments combined with optimal sampling times yield more information than the classical experiments using constant stimulus and equidistant measurements.
Although the stimulus profile obtained here could be rather difficult to implement in practice, adding constraints to the stimulus and minimum distances between sampling times, as indicated by experimentalists, is straightforward, thus ensuring experimental feasibility.
Future work involves the application of this methodology to real case studies related to cell signalling pathways in cooperation with specialized laboratories.

Acknowledgements. *This work was supported by the European Community as part of the FP6 COSBICS Project (STREP FP6-512060).*

References

Asprey, S.P. and Macchietto, A. (2002) Designing robust optimal dynamic experiments. *J. Process Control*, **12**, 545–556.

Balsa-Canto, E., Banga, J.R., Alonso, A.A. and Vassiliadis, V.S. (2004) Dynamic optimization of distributed parameter processes using second-order directional derivatives. *Ind. Eng. Chem. Res*., **43**, 6756-6765.

Banga, J.R., Versyck, K.J. and Van Impe, J.F. (2002) Computation of optimal identification experiments for nonlinear dynamic process models: an stochastic global optimization approach. *Ind. & Eng. Chem. Res.* **41**, 2425-2430

Biegler, L.T., Cervantes, A.M., and Watcher, A. (2002) Advances in simulaneous strategies for dynamic process optimization. *Chem. Eng. Sci*., **57(4)**, 575–593.

Bock, H.G. and Plitt, K.J. (1984) A multiple shooting algorithm for direct solution of optimal control problems. In Proceedings of the 9th IFAC World Congress, pp: 242–247. Pergamon Press: New York.

Brun, R. and Reichert, P. (2001) Practical identifiability analysis of large environmental simulation models. *Water Resources Res.*, **37**, 1015–1030.

Butcher, E.C., Berg, E. L. and Kunkell, E. J. (2004) Systems Biology in drug discovery. *Nature Biotechnology,* 22(10), 1253-1259.

Faller, D., Klingmüller, U. and Timmer, J. (2003) Simulation methods for optimal experimental design in systems biology. *Simulation*, **79**, 717–725.

Haugh, J.M. and Lauffenburger, D.A. (1998) Analysis of receptor internalisation as a mechanism of modulating signal transduction. *J. Theor. Biol*., **195**, 187–218.

Heinrich, R., Neel, B.G. and Rapaport, T.A. (2002) Mathematical models of protein kinase signal transdcution. *Mollecular Cell*, **9**, 957–970.

Hidalgo, M.E. and Ayesa, E. (2001) Numerical and graphical description of the information matrix in calibration experiments for state-space models. *Wat. Res.,* **35**, 3206–3214.

Kholodenko, B.N. (2006) Cell-signalling dynamics in time and space. *Nature reviews. Molecular Cell Biology*, **7**, 165-176.

Kutalik, Z., Cho, K-H. and Wolkenhauer, O. (2004) Optimal sampling time selection for parameter estimation in dynamic pathway modelling. *BioSystems*, **75**, 43–55.

Fallon, E.M. and Lauffenburger, D.A. (2000). Computational model for effects of ligand/recepter binding properties on interleukin-2 trafficking dynamics and cell proliferation response. *Biotechnol. Prog*., **16**, 905–916.

Leis, J.R. and Kramer, M.A. (1988). Odessa- an ordinary differential-equation solver with explicit simultaneous sensitivity analysis. *ACM Trans. Math. Soft*., **14**, 61–67.

Ljung, L. (1999). System identification: Theory for the user. Prentice Hall, New Jersey.

Moles, C.G., Mendes, P. and Banga, J.R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, **13**, 2467–2474.

Munack, A. Optimization of sampling (1991). In Biotechnology H.J. Rehm, G.Reed, A. Püler and P. Stadler eds. VCH, pp: 251-264.

D.M. Olive. Quantitative methods for the analysis of protein phosphorylation in drug development. Expert Rev. Proteomics 1(3), 327–341, 2004.

S. Pelech. Tracking Cell Signalling Protein Expression and Phosphorylation by Innovative Proteomic Solutions. Current Pharmaceutical Biotechnology, 5, 69-74, 2004.

Rodriguez-Fernandez, M., Mendes, P. and Banga, J.R..(2006) A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, **83** (2-3), 248-265.

Runarsson, T.P. and Yao, X. (2000) Stochastic ranking for constrained evolutionary optimization. *IEEE Transactions on Evolutionary Computation*, **564**, 284–294.

Seger, R. and Krebs, G. (1995) The mapk signalling cascade. *The Faseb Journal,* **9**, 726–735.

Storn, R. and Price, K. (1997) Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J. Global. Optim*., **11**, 341–359.

Vanrolleghem, P.A. and Dochain, D. (1998). Bioprocess model identification, pp. 251–318. Kluwer Academic Publishers.

Vassiliadis, V.S. (1993) Computational Solution of Dynamic Optimization Problems with General Differential-Algebraic Constraints. PhD thesis, Imperial College, University of London, London, U.K.

Establishment of the Dorsal-Ventral Boundary in the *Drosophila* Wing Imaginal Disc

O. Canela-Xandri[#a], H. Herranz[#b], R. Reigada[c], F. Sagués[c], M. Milán[*b], and J. Buceta[*a]

[a]*Parc Científic de Barcelona, Centre de Recerca en Química Teòrica (CeRQT), Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain*
[b]*ICREA and Institut de Recerca Biomèdica (IRB), Parc Científic de Barcelona, Campus Diagonal - Universitat de Barcelona, C/ Josep Samitier 1-5, 08028 Barcelona, Spain*
[c]*Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*

Keywords: Imaginal Discs, Pattern Formation, Regulatory Networks

1. Abstract

Herein, we present a gene-protein regulatory network for the establishment of the dorsal-ventral boundary in the *Drosophila* wing imaginal disc. We perform *in silico* experiments by means of a modelling approach that reduces each transcriptional-translational dynamics into a single effective process where Hill-like functions are assumed as regulatory functions. Thus, we show how short-range (receptor-ligand dynamics) together with long-range (morphogen gradient signalling) interactions shape the border and constitute the gene expression pattern that is observed in *in vivo* experiments. The *in silico* results are complemented with a robustness analysis of the regulatory network.

2. Introduction

As it occurs in most biological phenomena, gene expression underlies morphogenesis. By means of gene expression cells specialize for shaping and organizing tissues. This fact poses the interesting question of how cell fate is determined, i.e., how a given cell and its progeny "know" what genes should and should not express in order to perform a particular task. The latter immediately suggests the concept of information and reveals an additional function that is carried out by gene expression and regulated by cell interactions: gene expression must provide positional information (Wolpert, 1996). Thus, the genetic activity

# Equally contributed
∗ Corresponding authors

Understanding and Exploiting Systems Biology

establishes an expression pattern, a "map", by means of which cell fate is determined depending on the relative positions inside the primordium. This orchestrated plan sets up a dynamical coordinate system that links univocally the expression pattern to the resulting biological structure (Brook et al, 1996).

Within this framework, the model organism in which most of the research has been conducted is *Drosophila* (Lawrence, 1992). Structures such as the eyes, antennae, or wings develop from groups of cells denominated imaginal discs (Held et al, 2002). By using positional information, these potentially-contained structures proliferate in the larvae and, during metamorphosis, the adult insect, imago, is produced and the structures are exposed (Gilbert, 2003). There are nineteen imaginal discs inside the larvae. Two of them give rise to the adult wings. The seminal works of García-Bellido and coworkers showed that the imaginal discs are divided into regions (compartments) that correspond to different patterns of gene expression (García-Bellido et al, 1973; Cohen et al., 1992; Kornberg et al., 1985). The compartments are named after the position that their cells will occupy by the end of the developmental process. Thus, the imaginal disc of the wing is divided into anterior (A), posterior (P), dorsal (D), and ventral (V) compartments as shown in Fig. 1.

The understanding of the mechanisms that underlie the generation of stable and well-defined spatial domains that separate different cell populations (compartmentalization) is crucial for elucidating the relations between pattern formation and positional information. Roughly speaking, one can classify the mechanisms for compartment interactions into short and long ranged. As for the former, cell-cell communication by means of receptor-ligand dynamics is the main source of near-neighbour feedback. On the other hand, morphogens are responsible for long ranged signalling: cells obtain their relative position depending on concentration gradients of diffusive proteins (Teleman et al, 2001). Three proteins qualify as morphogens in the wing imaginal disc development: Hedgehog, Decapentaplegic, and Wingless (Alberts et al, 2002). Note that morphogen signalling is reliable for positional information purposes if proteins diffuse from given spatial positions that function as references in a coordinate system. In the case of the wing disc, those organizing centres are actually located at the compartment boundaries. Thus, from two interacting cell populations, compartments, a third one that presents specific gene trademarks, acts as signalling centre, and controls cell migration between compartments is established: the boundary or axis.

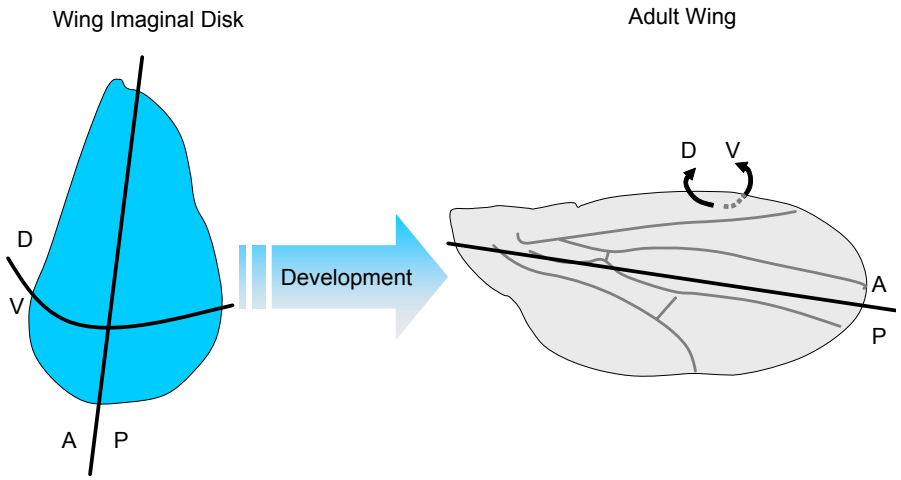


Figure 1. Wing imaginal disc (left) and adult wing (right). The wing disc is divided into regions (compartments). The compartments are named after the region that their cells will occupy by the end of the developmental process. Both short-range (cell-cell communication) and long-range (diffusive morphogens) interactions take place between adjacent compartments. Compartments boundaries are organizing centres that control such interactions (see text).

Herein, we will focus on how short and long ranged compartment interactions shape the formation of the dorsal-ventral boundary. We will show that the concept of *refractoriness*, "blindness", to a particular gene is required for understanding the formation of such organizing axis. Moreover, we will address other open problems in this context. Namely, how the size of the border is regulated (refinement) and how symmetricalness for ligands expression with respect of the border edges is obtained. To this end, we introduce a gene-protein regulatory network and perform *in silico* experiments that we complement by means of robustness analysis.

In order to frame-in appropriately our problem, we briefly review the stages that lead to the establishment of the Dorsal-Ventral (DV) boundary and fix the temporal windows of our interest. We restrict ourselves to larval third instar. Thus, DV compartment subdivisions are primarily established by the activity of the selector gene *apterous* in D cells (reviewed in Blair, 1995). The onset of *apterous* expression in the early wing primordium induces expression of the Notch ligand Serrate in D cells and restricts expression of Delta, another Notch ligand, to V cells (Diaz-Benjumea and Cohen, 1995; Milan and Cohen, 2000). Moreover, due to *apterous* activity, expression of the glycosyltransferase Fringe makes D cells more sensitive to Delta and less sensitive to Serrate (Fleming et al., 1997; Panin et al., 1997). Dorsally expressed *serrate* and ventrally expressed *delta* activate Notch in cells on both sides of the DV compartment boundary (Diaz-Benjumea and Cohen, 1995; de Celis et al., 1996; Doherty et al., 1996). Later, an increase in dLMO levels reduces *apterous* activity in the wing primordium (Milán and Cohen, 2000). Our initial condition refers to this moment when *apterous* activity, that has already caused an asymmetric expression of

delta and *serrate*, ceases and, consequently, Notch equally respond to both ligands (Kim et al, 1995, de Celis et al, 1996).
Notch activation due to short-range interactions between adjacent compartments induces expression of the signaling molecule Wingless in cells along the DV boundary. Wingless induces expression of *serrate* and *delta* in nearby D and V cells that can signal back to Notch. At this stage another set of cell interactions takes over and a stable DV boundary is finally established (de Celis and Bray, 1997; Micchelli et al., 1997). This moment constitutes the endpoint of the developmental process we want to describe. By stable border we mean that a stationary Notch activity is reached for a narrow cell population that separate dorsal and ventral compartments. Moreover, the dynamics of the regulatory network must describe the symmetrization of the expression pattern of Notch ligands at two flanking stripes of the border, the refinement of *notch* expression, and the onset and stabilization of *wingless* expression. The latter, in subsequent developmental stages, organizes pattern and growth of the wing anlage.
The paper is organized as follows. In section 3 we introduce the regulatory network and our modelling approach. In section 4 the results obtained from *in silico* experiments and robustness analysis are shown. Finally, the conclusions of our work are presented in section 5.

3. Regulatory Network for Dorsal-Ventral Boundary Formation

3.1. MODELLING APPROACH

We implement a modelling scheme where the species of interest are the concentration of proteins or protein related products, e.g., a biomolecule resulting from protein cleavage. Our modelling approach reduces each transcriptional-translational dynamics of a gene network, into a single effective process where Hill-like functions, with a given degree of cooperativeness, are assumed as regulatory functions. The resulting differential equations mimic the temporal behaviour for the concentration of proteins in a cell as a consequence of gene interactions. Figure 2 illustrates the interaction between three genes and shows schematic representations of positive and negative regulatory Hill functions.

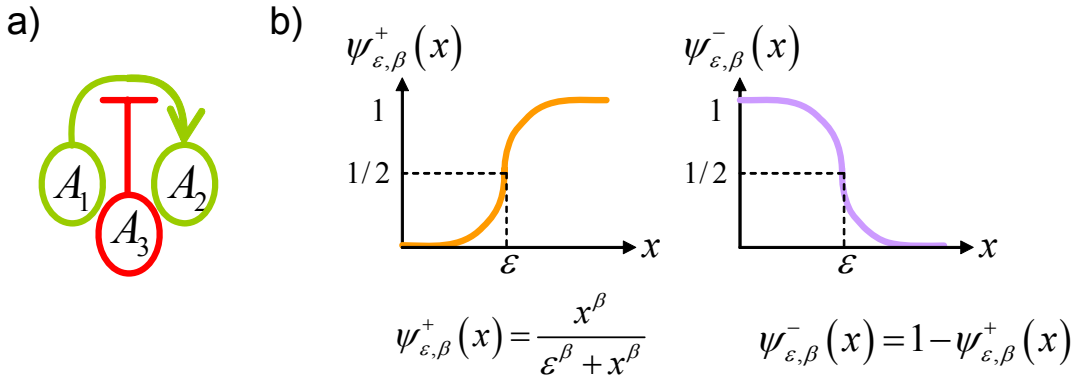


Figure 2. a) Gene-protein A_2 is positively regulated by gene-protein A_1. Alike, such interaction is negatively regulated by gene-protein A_3. Both, positive and negative, regulatory functions depending on the concentration of protein/gene, x,

are schematically represented in b). Hill functions with a given degree of cooperativeness, β, are assumed to effectively model gene-protein regulation. The larger β, the stronger the cooperativeness and the steeper the transition between the states 0 and 1 are. The parameter ε measures the concentration threshold of x for which the regulatory interaction, either positive or negative, reaches *50%* intensity.

Thus, for the network shown in Fig. 2, the regulation of gene-protein A_2 due to interactions with genes-proteins A_1 and A_3 reads,

$$\psi^{+}_{\varepsilon_{A_1A_2}}\left(A_1 \times \psi^{-}_{\varepsilon_{A_3A_2}}\left(A_3\right)\right). \tag{1}$$

We also consider degradation for species by means of exponential decays. Therefore, the modeling differential equation for species A_2 becomes,

$$\frac{dA_2}{dt} = k_{A_2} \times \psi^{+}_{\varepsilon_{A_1A_2}}\left(A_1 \times \psi^{-}_{\varepsilon_{A_3A_2}}\left(A_3\right)\right) - \mu_{A_2} \times A_2, \tag{2}$$

where k_{A_2} and μ_{A_2} stand for the expression and degradation rate constants respectively. As detailed below, we also take into account diffusion and autonomous transcription-translation dynamics for some species. The former is a crucial element of morphogen kinetics whereas the latter is required to reproduce basal levels of expression of genes that are known to occur independently of other genes activities.

Cell-cell interactions due to receptor-ligand dynamics are modelled as follows. In the case we are focusing on, the activation of the receptor in a given cell takes place exclusively when it binds to a ligand that belongs to a (nearest) neighbour cell. Notice that receptor-ligand binding events within the same cell certainly happen. However, in those cases no activation of the receptor is produced and both, receptor and ligand, are "sequestered" and they become useless for further signalling purposes. The activation of the receptor produces proteolytic cleavage of its intracellular part that is transported to the nucleus where induces the expression of downstream genes. Let us denote x_i, x_i^*, and y_j the receptor, its intracellular active part, and the ligand concentrations at cells *i* and *j* respectively. We ignore momentarily upstream and downstream transcription-translation processes and degradation. Accordingly, the receptor-ligand regulatory dynamics reads,

$$\begin{aligned}
\frac{dx_i}{dt} &= -x_i \times \left[k_{binding} \times \sum_{\langle ij \rangle} \frac{y_j}{\varepsilon^2_{binding} + x_j \times y_j + x_i \times y_i} + k_{sequestering} \times \frac{y_i}{\varepsilon^2_{sequestering} + x_i \times \sum_{\langle ij \rangle} y_j + y_i \times \sum_{\langle ij \rangle} x_j} \right], \\
\frac{dy_i}{dt} &= -y_i \times \left[k_{binding} \times \sum_{\langle ij \rangle} \frac{x_j}{\varepsilon^2_{binding} + x_j \times y_j + x_i \times y_i} + k_{sequestering} \times \frac{x_i}{\varepsilon^2_{sequestering} + x_i \times \sum_{\langle ij \rangle} y_j + y_i \times \sum_{\langle ij \rangle} x_j} \right], \\
\frac{dx_i^*}{dt} &= +x_i \times \left[k_{binding} \times \sum_{\langle ij \rangle} \frac{y_j}{\varepsilon^2_{binding} + x_j \times y_j + x_i \times y_i} \right].
\end{aligned} \tag{3}$$

where $\langle ij \rangle$ indicates that the sums in eq.(3) run over all cells *j* that are nearest-neighbours of cell *i*. Note that we keep Hill-like regulatory functions and that the activation of the receptor occurs only due to interaction between different cells. In other case, i.e. same cell coupling, the receptor and the ligand are simply "sequestered" and removed for further signalling. We do not consider unbinding dynamics. Moreover, the activation events are conveniently weighted in the regulatory functions depending on the amount of receptor-ligand couples that can be created within the same cell since they will reduce the probability of successful cell-cell bindings. Likewise, the sequestering events are weighted by taking into account the amount of receptor-ligand couples that are formed between different cells since they decrease the probability of a sequestering event. Altogether, this dynamics may lead to either positive or negative regulation of x_i^* depending on the local concentrations of x_i and y_j, and on the values of the parameters *k* and ε. As shown below, a generalization of the modelling equations when the receptor can be signalled by several ligands is straightforward.

3.2. GENE-PROTEIN NETWORK

Figure 3 shows the gene-protein regulatory network that controls the establishment of the DV boundary. A detailed construction of such network based on *in silico* experiments for wild-type (and mutant) phenotypes, and the comparison with their corresponding *in vivo* counterparts will be published elsewhere (Canela-Xandri et al, 2006). Other networks and modelling approaches have been considered by different authors for this particular problem (Kioda and Kitano, 1999). However, some *in vivo* experimental data are disregarded therein and consequently the establishment of the DV boundary can not be suitably explained. The main characteristics of the proposed network are the following.

Signalled Notch is the "conductor" for the establishment of the DV axis. Notch can be signalled by either Serrate or Delta. At this stage of development, there is no difference on the way the ligands signal to Notch. Therefore, we will not consider any difference in their dynamics apart from the aforementioned asymmetry in the initial condition: *delta* is expressed in ventral cells whereas *serrate* is expressed in dorsal cells. Nonetheless, in the model we distinguish both species to conveniently track how symmetric expression of both ligands is obtained at flanking stripes of the border. As mentioned above, depending on the local concentration of receptor and ligands in cells, their dynamics may lead or not to an effective activation of the receptor. If the receptor is activated, then the transcription-translation of its downstream genes starts. Downstream genes are, or are not, expressed, depending on the level of activity of *notch*. As the level of activation increases, *notch* and the ligands themselves are expressed, afterward *wingless*, and finally *cut*. This ordered sequence of expression as a function of *notch* activity levels, fixes an ordered sequence for the threshold values of the regulatory functions, ε, in our modelling approach. Independently of Notch activation, there is an autonomous off-network Notch transcription-translation dynamics that keeps the protein expression up to a basal level in the disc pouch. Notice also that both Notch ligands are *wingless* downstream genes. The ligands

expression levels due to Wingless are much larger than those due to Notch activation. This may cause downregulation of *notch* pathway due to sequestering effects.

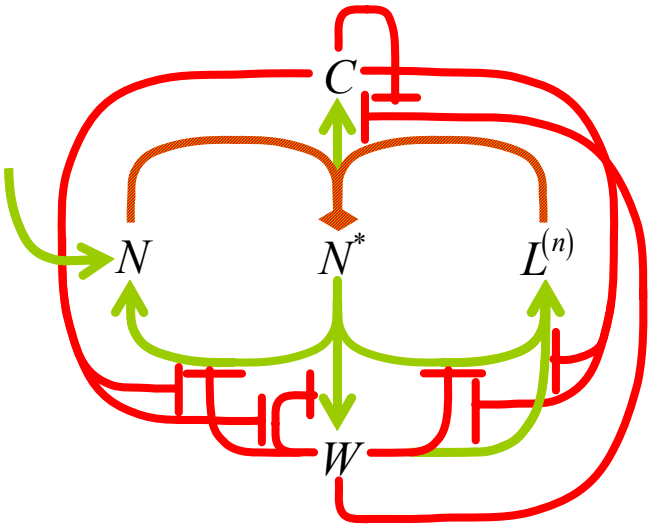


Figure 3. Gene-protein regulatory network for the establishment of the DV boundary at the wing imaginal disc. N stands for Notch receptor and N^* for intracellular active Notch. Two different Notch ligands, Serrate, $L^{(1)}$, and Delta, $L^{(2)}$, are considered. W indicates Wingless diffusive morphogen. Finally C stands for Cut gene-protein. Green and red lines mean positive and negative regulations respectively. The dashed coloured green-red line with a rhombic end, indicates that receptor-ligand dynamics may lead to either positive or negative regulation. Note that Notch has an additional autonomous off-network regulation.

Importantly, there is also a direct mechanism for downregulation of *notch* pathway due to *wingless*. This inhibitory role is mediated by *dishevelled*, a downstream gene of *wingless* (Axelrod et al, 1996). Such negative interaction between *wingless* and *notch* pathways seems to be incompatible with the following reported results for the expression pattern of the border cells. First, the larger the *notch* activity is, the larger the *wingless* expression levels are kept. Second, the expression of ligands is severely reduced at the border cells, where *notch* activity and *wingless* expression are maxima. We have recently shown that a new property is required in order to explain these (apparently) contradictory results: *refractoriness* to Wingless (Canela-Xandri et al, 2006). That is, there must be a mechanism that makes cells expressing *wingless* at high levels, refractory to (all) its effects. In this way, the border cells can simultaneously present a pronounced *wingless-notch* expression pattern and keep the ligands expression levels to a minimum. Such mechanism is mediated by *cut* (a *notch* downstream gene). Hence, *cut*-expressing cells are “blind” to Wingless effects. Notice that since Wingless is a morphogen and diffuses, its effects can be induced at locations where it is not expressed. Therefore ligand expression due to Wingless is produced off the border, at flanking cells, where *cut* is not being expressed. Finally, from flanking cells the ligands signal back to Notch at border

cells. The latter also explains why the size of the border population is kept to two-three cells and the symmetric expression pattern of the ligands at flanking stripes of the organizing axis.

3.3. MODELLING DIFFERENTIAL EQUATIONS, PARAMETER ESTIMATION, AND SIMULATION DETAILS

By taking into account the aforementioned considerations, the differential equations that represent the gene-protein network read,

$$\begin{aligned}
\frac{dN_i}{dt} &= \gamma - \mu N_i - N_i\left[k_1\sum_{\langle ij\rangle}\frac{\sum_n L_j^{(n)}}{\Phi_1\left(L^{(n)}\right)} + k_2\sum_n\frac{L_i^{(n)}}{\Phi_2\left(L^{(n)}\right)}\right] + k_1\chi_{\varepsilon_{N^*N}}\left(N_i^*, W_i, C_i\right)\\
\frac{dN_i^*}{dt} &= -\mu N_i^* + k_1 N_i\sum_{\langle ij\rangle}\frac{\sum_n L_j^{(n)}}{\Phi_1\left(L^{(n)}\right)}\\
\frac{dL_i^{(n)}}{dt} &= -\mu L_i^{(n)} - L_i^{(n)}\left[k_1\sum_{\langle ij\rangle}\frac{N_j}{\Phi_1\left(L^{(n)}\right)} + k_2\frac{N_i}{\Phi_2\left(L^{(n)}\right)}\right] + k_3\chi_{\varepsilon_{N^*L^{(n)}}}\left(N_i^*, W_i, C_i\right) + k_1\Upsilon\left(W_i, C_i\right)\\
\frac{dW_i}{dt} &= -\mu W_i + k_1\chi_{\varepsilon_{N^*W}}\left(N_i^*, W_i, C_i\right) + D\sum_{\langle ji\rangle}\left(W_j - W_i\right)\\
\frac{dC_i}{dt} &= -\mu C_i + k_1\chi_{\varepsilon_{N^*C}}\left(N_i^*, W_i, C_i\right)
\end{aligned} \tag{4}$$

where the following functions have been defined,

$$\begin{aligned}
\Phi_1\left(L^{(n)}\right) &= \varepsilon_{binding}^2 + N_j\sum_n L_j^{(n)} + N_i\sum_n L_i^{(n)}\\
\Phi_2\left(L^{(n)}\right) &= \varepsilon_{sequestering}^2 + N_i\sum_{\langle ji\rangle}\sum_n L_j^{(n)} + L_i^{(n)}\sum_{\langle ji\rangle} N_j\\
\Upsilon\left(W_i, C_i\right) &= \psi^+_{\varepsilon_{WL^{(n)}}}\left(W_i\times\psi^-_{\varepsilon_{CW}}\left(C_i\right)\right)\\
\chi_{\varepsilon_{N^*X}}\left(N_i^*, W_i, C_i\right) &= \psi^+_{\varepsilon_{N^*X}}\left(N_i^*\times\psi^-_{\varepsilon_{WN^*}}\left(W_i\times\psi^-_{\varepsilon_{CW}}\left(C_i\right)\right)\right)
\end{aligned} \tag{5}$$

Note that there are actually 6 differential equations since the superscript *n* takes the values 1 and 2 depending of the ligand. Wingless diffusion has been included by means of a discrete version of a Laplacian operator (see details below). Notice also the first term in the r.h.s. of the equation for Notch dynamics that account for autonomous off-network expression, γ. Finally, we point out that we disregard cell proliferation and motility. The former is known to play a key role in subsequent developmental stages when the border has been already established but can be neglected within the temporal window of our interest. As for the latter, it can be ignored altogether within this context.

For the sake of simplicity, we keep the set of parameters as reduced as possible. Still, we maintain a realistic approach and therefore such set should be large enough to take into account well-known biological facts. For example, as it was mentioned above, ligand expression rates due to *notch* activity are known to be smaller than those due to Wingless. Similarly, sequestering effects play a relevant role in receptor-ligand dynamics when compared to binding. Thus, as

shown in eqs.(4-5), whereas the degradation rate constant, μ, is kept the same for all species, three different regulation rate constants have been used: k_1 (binding and all gene/protein regulatory constants apart from ligand expression due to *notch* activity), k_2 (sequestering), and k_3 (ligand expression due to *notch* activity).

Apart from the Wingless diffusion rate, ~1.4 μm^2/s, and to the best of our knowledge, most of the parameter values that appear in eqs.(4-5) have not being measured. Fortunately, at least the order of magnitude of some of them has been reported for related problems. Thus, the degradation rate of proteins ranges from $10^{-6} s^{-1}$ to $10^{-2} s^{-1}$. We used an intermediate value: $10^{-3} s^{-1}$. As for the effective transcription-translation rates, we estimate them as follows. Suppose a species subjected to regulation and degradation, e.g. eq.(2). The maximum value of its concentration in the steady-state will be given by (note that the regulatory functions are dimensionless),

$$A_s\big|_{\max} = \frac{k_{A_2}}{\mu_{A_2}} \cdot \tag{6}$$

Obviously the minimum value is simply zero. Therefore, if the degradation rates and the steady concentration of protein are known quantities, then the effective transcription-translation rates can be estimated. The number of proteins in a cell commonly ranges from 10^4 to 10^7. By taking into account that the typical diameter of a cell is 10 μm, k∈(~10^{-1}, ~10) *proteins/(*μm^3 *s)*.

Some degree of cooperativeness, β, is mandatory (see robustness analysis results below). However, it can not be too large because in that case the system is too sensitive to the value of the concentrations involved in regulatory tasks: the regulatory functions tend to step functions. We set their value to 2. The thresholds for regulation, ε's, and the finetuning of the parameters where obtained by means of cloning experiments (Canela-Xandri et al, 2006). Such experiments allow us to either knockout or over-express a gene, or a set of genes, for a particular group of cells (and progeny if required). Thus, the observed behaviour within the clones and at neighbouring cells for *in silico* experiments and the comparison with their *in vivo* counterparts, allowed us to check if the gene interactions were appropriately defined and weighted. By testing different clones, we converged to a set of parameter values that reproduces the wild-type behaviour and cloning experiments. Table 1 summarizes the parameter used in our modeling approach for k's and ε's values.

The value of parameter γ associated with *notch* basal transcription-translation is taken as $5 \cdot 10^{-2}\ proteins/(\mu m^3 \cdot s)$ and ensures a minimum amount of Notch protein at each cell, ~50 *proteins/*μm^3. We implement our simulations in a two-dimensional hexagonal lattice by means of an explicit forward-time-centered-space scheme with time step $10^{-4} s$ and size 50×30. Each lattice node represents a cell and therefore the *in silico* disc comprises 1500 cells.

Table 1. *In silico* experiments: k's and ε's values used for numerical simulations of eqs.(4-5).

subscript	N^*N	$N^*L^{(n)}$	N^*W	N^*C	*Bind.*	*Seq.*	WN^*	CW
$\varepsilon\left(\frac{proteins}{\mu m^3}\right)$	*200*	*300*	*400*	*500*	$(50000)^{1/2}$	$(1000)^{1/2}$	*100*	*100*

subscript	1	2	3
$k\left(\frac{proteins}{\mu m^3 \cdot s}\right)$	*1*	*5*	*0.1*

Discretization of the Laplacian operator for such geometry leads to,

$$\tilde{D}\nabla^2\psi \leftrightarrow \frac{2}{3}\frac{\tilde{D}}{l^2}\sum_{\langle ij\rangle}\left(\psi_j - \psi_i\right) \tag{7}$$

where l is the lattice spacing, i.e., the typical cell size, $10\mu m$, and the sum runs over the nearest-neighbours (6 for a two-dimensional hexagonal lattice). Note that we have defined $D = 2\tilde{D}/\left(3l^2\right)$, c.f. eqs. (4) and (7). The value used for D in *in silico* experiments was $D \approx 7\cdot 10^{-3} s^{-1}$, that corresponds to a diffusion coefficient $\tilde{D} = 1\mu m^2/s$.

Figure 4 shows the initial expression pattern, i.e., the initial condition. We initially divide our *in silico* imaginal disc into two domains that correspond to dorsal and ventral compartments. Such division is characterized by the ligand expression pattern: *delta* is expressed in one compartment (V) and *serrate* in the other (D). The value of these initial concentrations in the disc pouch is very small at most cells, $10 proteins/\mu m^3$, but at boundary cells. There, a larger concentration of ligands is expected due to the aforementioned positive feedback induced by the *apterous* onset. Moreover, at boundary cells an initial concentration of Notch and activated Notch is also expected. At all cells Wingless and Cut concentrations are initially set to zero.

4. Results and Discussion

We start by describing the results obtained for *in silico* experiments in regard of the stationary state of the expression pattern. Figure 4 shows both the initial and the final expression patterns of the species involved in the regulatory network indicated in Fig. 3. As mentioned above, the initial expression pattern show an asymmetry in the expression of Notch ligands (dorsal on the right and ventral on the left) and small picks for Notch and *notch* activity. Such pattern evolves up to a stage where the interplay between the border activity and flanking cells signalling is self-sustained and reaches a steady state. Border activity is pointed out by robust active-Notch and *cut* expression patterns, and by the establishment of a Wingless morphogen gradient towards both compartments of the disc. On the other hand, flanking cells signalling is emphasized by a symmetric expression

pattern of both ligands with respect the DV axis. Notch protein also presents a characteristic expression pattern with depressions located at cell positions where ligand expression is pronounced.

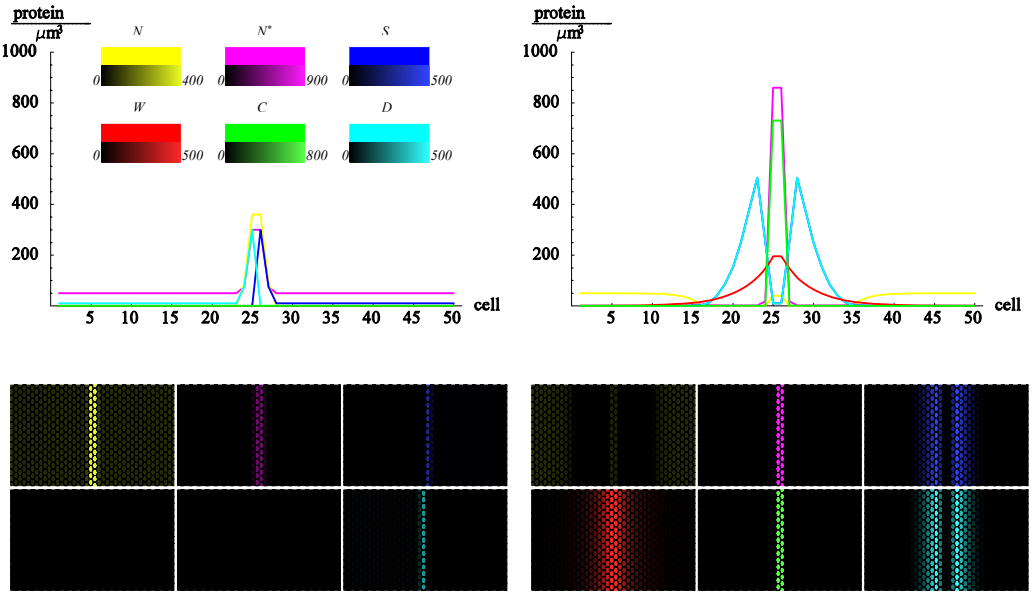


Figure 4. Initial (left) and final (right) expression patterns during the establishment of the DV boundary (*in silico* results). The inset (top-left corner) shows the colour code for both one-dimensional (top) and two-dimensional (bottom) plots. The former represents concentration vs. cell-number along an axis perpendicular to the DV border, i.e., parallel to the AP axis. The latter are concentration density plots for the expression pattern of different species within the imaginal disc pouch. In both cases dorsal is on the right and ventral on the left.

This global patterned state is reached and maintained by means of an orchestrated dynamical process. The incipient *notch* activity in the border induces expression of ligands Serrate and Delta, and Notch protein that generates a positive feedback loop. However, the loop can not be sustained without large ligand expression levels. The latter are provided as a consequence of the *wingless* expression once *notch* activity is large enough. Wingless spreads and begins to establish a morphogen gradient. The cells signalled by Wingless start to express ligands that burst out *notch* activity that subsequently induces *wingless* expression. Such "chain-reaction" broadens the border population and helps to symmetrize ligand expression. The spreading is controlled by downregulation of *notch* pathway due to direct and indirect effects: Wingless inhibitory tasks and ligand-Notch sequestering events respectively. Thus, *notch* downregulation produces a decay of the expression levels of the aforementioned species. However, at boundary cells, *notch* activity has surpassed the *cut* threshold by then, and the latter is being expressed at appreciable levels. Refractoriness to Wingless induced by Cut at border cells builds up a mask at the boundary so that

no downregulation of *notch* pathway is produced and complementary *wingless* downstream genes are not expressed.

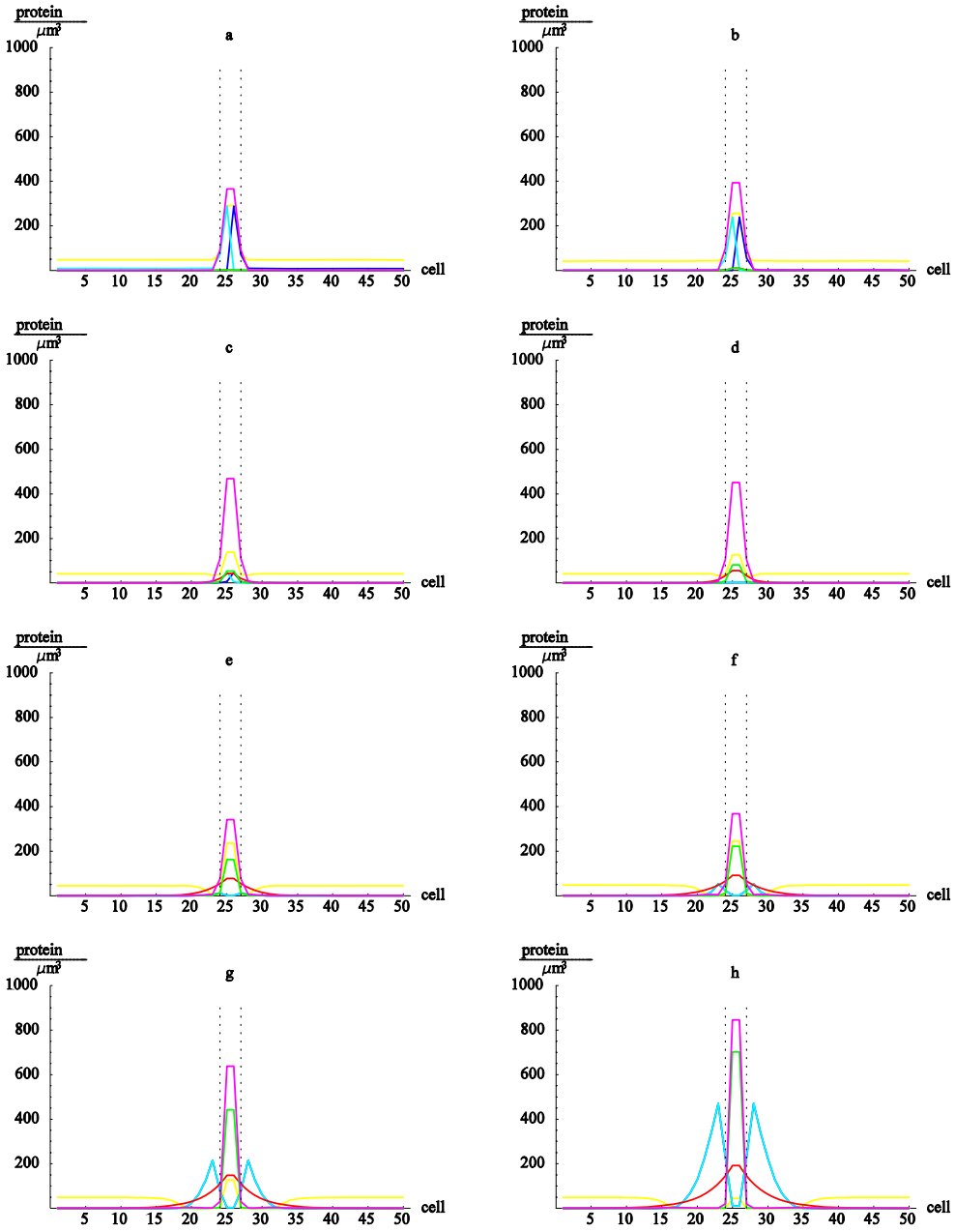


Figure 5. Snapshots of the concentration of species versus cell-number along an axis perpendicular to the DV border (dorsal is on the right and ventral on the right). The sequence shows from the initial condition (a) to the stationary state (h) how the expression pattern is generated. Same colour code for species that in Fig.4 was used. Dotted lines delimit a four-cells-wide region around the DV boundary and highlight the refinement process. Frame-to-frame time lapses are different.

Therefore, at border cells, ligands are not highly expressed but Wingless and *notch* activity levels are kept high at the same time. Outside the border, where

there is no refractoriness (i.e. absence of Cut), *notch* activity and *wingless* expression levels decay and the broad expression pattern begins to shrink. Such spreading-shrinking dynamics is commonly known as refinement and, as shown here, is *cut*-mediated. The border is finally confined to cells where *cut* is being expressed and the morphogen gradient is ultimately shaped.

Note that the levels of expression of the ligands are kept pronounced outside border cells since Wingless effects are noticed there. Thus, the maintenance of the border activity is done by flanking signalling cells: since at the border the ligand expression has been severely reduced, Notch is signalled by flanking cells. Note that receptor-ligand signalling dynamics becomes directional, i.e., one might wonder why flanking cells with large expression levels of ligands signal only *towards* border cells and not also *against*. The answer is the following. Wingless downregulates *notch* pathway outside the border and therefore no activation of the receptor may occur. Complementary, at the boundary there is a lack of ligands due to the refractory effects to Wingless induced by *cut*, and the receptor can be exclusively "fed" by neighbouring flanking cells. The resulting outward/inward polarization dynamics for the receptor/ligand reaches equilibrium as pattern expression and consolidation of the border evolve. Notice also that the width of the cell population is kept to two cells because of the same reason: if the border cell population was larger, then intermediate cells would not have enough ligands around to signal Notch and therefore the border would split.

Figure 5 shows snapshots corresponding to a typical evolution toward the aforementioned equilibrium. From top to bottom the figure show the refinement dynamics and the conformation of the expression pattern. Note that activated Notch refines its expression and becomes restricted to the (two) cells that constitute the DV boundary. Moreover, *notch* activity refinement indeed refines *wingless* expression that becomes restricted to cells where *notch* activity is large. Such process is not clearly shown in Fig.5 since Wingless protein instead of *wingless* expressing cells is depicted there. Observe also the process that leads to the symmetrization of the ligands expression. Summarizing, Notch activity is the "conductor" of the orchestrated plan that establishes the DV boundary where correct maintenance and shaping is mediated by *cut* due to induced refractoriness to Wingless.

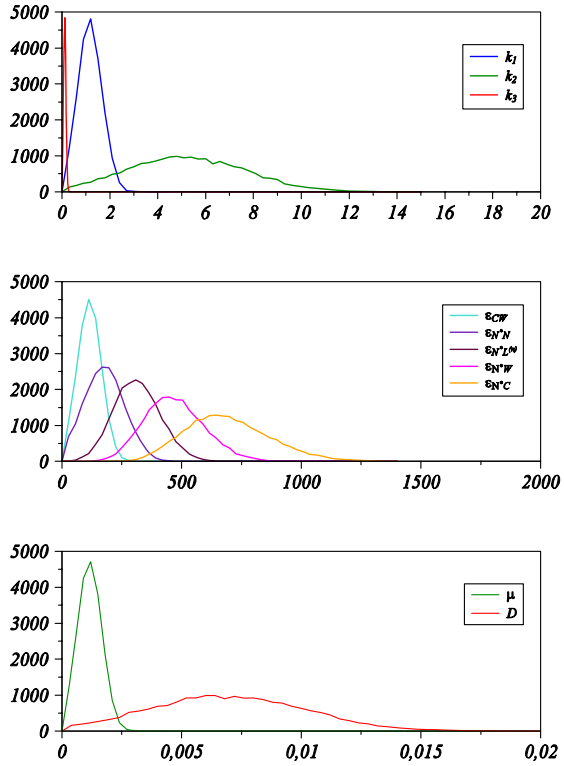


Figure 6. Histograms (count, i.e., no normalization) of the parameter values used in the $\sim 1.5 \cdot 10^4$ *in silico* experiments implemented in the robustness analysis. Units depend on the depicted quantity. The initial condition was also subjected to variation but is not shown in the figure.

Robustness analysis shows that the proposed regulatory network is indeed robust for DV boundary establishment. We implement a Gaussian distributed random variation of each parameter of eqs.(4-5) around the values reported in the previous section. The dispersion typically allows a *50%* variation around the assumed parameter value. We stress that the parameter of each term that appears in eqs.(4-5) is treated separately but consistently, e.g., terms where k_1 appears in eqs.(4-5) are in principle subjected to an independent random variation of k_1 around its mean value $k_1 = 1$ and with statistical properties as described above, however, as indicated in eq.(3), mass conservation in receptor-ligand dynamics enforce the same values of $k_{binding}$, $k_{sequestering}$, $\varepsilon_{binding}$, and, $\varepsilon_{sequestering}$ in all equations.

Since the Gaussian distribution is unbounded, negative values can be certainly obtained: we obviously disregard those in our analysis. Moreover, we check that each parameter set ensures Biological realism in the following sense. The Notch activity thresholds' sequence in order to induce regulation of downstream genes is experimentally well-known and part of the DV boundary formation mechanism. Therefore, a valid parameter set must fulfil the following condition,

$$\varepsilon_{N^*C} > \varepsilon_{N^*W} > \varepsilon_{N^*L^{(n)}} > \varepsilon_{N^*N} \tag{8}$$

This procedure implies that the generated distributions for these quantities are not kept independent that in turn causes their histograms to be shifted to the right as shown in Fig.6. Thus, by keeping unaltered the cooperativeness parameter β, we generate $\sim 1.5 \cdot 10^4$ valid parameter sets. We check for each of them whether the DV boundary and the rest of the expression pattern are correctly obtained or not. Afterwards, we repeat the test for the same sets of parameters but varying β. Figure 6 shows the parameter distributions used in the robustness analysis. The values used for β, apart from $\beta = 2$, either disregard cooperativeness, $\beta = 1$, or overestimate (with respect our original guess) its value: $\beta = 3$. We evaluate the ratio,

$$r_\beta = \frac{\text{successful outcomes}}{\text{number of experiments}} \tag{9}$$

As $r_\beta \to 0/1$ the system becomes less/more robust. We also compute $\left(r_\beta\right)^{1/n}$, where n is the size of our parameter set, i.e., the number of parameters we allow to vary independently: 22 in our case. We note that the initial condition is also taken into account in our robustness analysis. Such quantity, $\left(r_\beta\right)^{1/n}$, provides information on the degree of robustness of each parameter: if all parameters are kept unaltered but one, then it measures the degree of robustness for single parameter variation. We obtain that,

$$\left.\begin{array}{l} r_{\beta=1} \approx 0.0012 \\ r_{\beta=2} \approx 0.1186 \\ r_{\beta=3} \approx 0.1033 \end{array}\right\} \leftrightarrow \left\{\begin{array}{l} \left(r_{\beta=1}\right)^{1/22} \approx 0.74 \\ \left(r_{\beta=2}\right)^{1/22} \approx 0.91 \\ \left(r_{\beta=3}\right)^{1/22} \approx 0.90 \end{array}\right. \tag{10}$$

Interestingly enough, we observe that some degree of cooperativeness is required for a robust DV boundary establishment. Notice also the strong degree of robustness for the regulatory network that, in average, allows a *91%* variation for single parameter variation experiments in the $\beta = 2$ case.

5. Conclusions

Herein, we have presented a gene-protein regulatory network for the establishment of the DV boundary in the *Drosophila* wing imaginal disc. Our modelling approach reduces each transcriptional-translational dynamics into a single process where Hill-like functions, with a given degree of cooperativeness, are assumed as effective regulatory functions. Thus, we have shown by means of *in silico* experiments how short-range (receptor-ligand dynamics) in conjunction with long-range (morphogen gradient signalling) interactions shape the border and establish the observed expression pattern. Moreover, we have shown that a new property, refractoriness to Wingless, is a required element for regulation

within this context. Such property is induced by *cut*. Finally, the robustness analysis reveals that the proposed regulatory network is highly robust to parameter variation.

Acknowledgements*. J.B. and H.H. acknowledge the Ramón y Cajal and the Juan de la Cierva programs that provide their researcher contracts respectively. Partial support was provided by M.E.C. under grants FIS2005-457 (O.C-X. and J.B.), BFU2004-00167/BMC (H.H. and M.M.), and BQU2003-05042-C02-01 (R.R. and F.S), and by DURSI through project 2005SGR00653 (O.C-X., J.B. R.R., and F.S.).*

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002) Molecular Biology of the Cell 4th Ed. (Garland Science, New York).

Axelrod, J.D., Matsuno, K., Artavanis-Tsakonas, S., and Perrimon, N. (1996). Interactions between Wingless and Notch Signalling Pathways Mediated by Dishevelled. Science **271**, 1826:1832.

Blair, S.S. (1995). Compartments and Appendage Development in *Drosophila*. Bioessays **17**, 299:309.

Brook, W.J., Díaz-Benjumea, F.J., and Cohen, S.M. (1996). Organizing Spatial Pattern in Limb Development. Ann. Rev. Cell Dev. Biol. **12**, 161.

Canela-Xandri, O., Herranz, H., Reigada, R., Sagués, F., Milán, M., and Buceta, J. (2006). Cut-induced Refractoriness to Wingless is a Required Element for the Establishment of the Dorsal-Ventral Boundary in the Wing Imaginal Disc of *Drosophila*. *In preparation*.

Cohen, B., McGuffin, M. E., Pfeifle, C., Segal, D., and Cohen, S. M. (1992). Apterous: a Gene Required for Imaginal Disc Development in *Drosophila* Encodes a Member of the LIM Family of Developmental Regulatory Proteins. Genes Dev. **6**, 715:729.

de Celis, J. F., and Bray, S. (1997). Feed-back Mechanisms Affecting Notch Activation at the Dorsoventral Boundary in the *Drosophila* Wing. Development **124**, 3241:3251.

de Celis, J. F., Garcia-Bellido, A., and Bray, S. J. (1996). Activation and Function of Notch at the Dorsal-ventral Boundary of the Wing Imaginal Disc. Development **122**, 359:369.

Diaz-Benjumea, F. J., and Cohen, S. M. (1995). Serrate Signals through Notch to establish a Wingless-dependent Organizer at the Dorsal/ventral Compartment Boundary of the *Drosophila* wing. Development **121**, 4215:4225.

Doherty, D., Fenger, G., Younger-Shepherd, S., Jan, L.-Y., and Jan, Y.-N. (1996). Dorsal and Ventral Cells Respond Differently to the Notch Ligands Delta and Serrate During *Drosophila* Wing Development. Genes Dev. **10**, 421:434.

Fleming, R. J., Gu, Y., and Hukriede, N. A. (1997). Serrate-mediated Activation of Notch is Specifically Blocked by the Product of the Gene Fringe in the Dorsal Compartment of the *Drosophila* Wing Imaginal Disc. Development **124**, 2973:2981.

García-Bellido A., Ripoll, P., and Morata, G. (1973). Developmental Compartimentalization of the Wing Disk of *Drosophila*. Nature New Biol. **245**, 251.

Gilbert, S. (2003) Developmental Biology 7th Ed. (Sinauer, Sunderland MA).

Held Jr., L.I., Bard, J.B.L., Barlow, P.W., and Kirk, D.L.. (2002) Imaginal Disc: The Genetic and Cellular Logic of Pattern Formation (Cambridge University Press, New York).

Kim, J., Irvine, K.D., and Caroll, S.B. (1995) Cell Recognition, Signal Induction, and Symmetrical Gene Activation at the Dorsal-Ventral Boundary of the Developing *Drosophila* Wing, Cell **82**, 795.

Kyoda, K. and Kitano, H. (1999) A Model of Axis Determination for the Drosophila Wing Disc, Lecture Notes in Artificial Inteligence **1674**, 472:476.

Kornberg, T., Siden, I., O'Farrell, P., and Simon, M. (1985). The Engrailed Locus of *Drosophila*: In Situ Localization of Transcripts Reveals Compartment-specific Expression. Cell **40**, 45:53.

Lawrence, P.A. (1992) The Making of a Fly: The Genetics of Animal Design. (Blackwell, Oxford).

Micchelli, C. A., Rulifson, E. J., and Blair, S. S. (1997). The Function and Regulation of Cut Expression on the Wing Margin of *Drosophila*: Notch, Wingless and a Dominant Negative Role for Delta and Serrate. Development **124**, 1485:1495.

Milan, M., and Cohen, S. M. (2000). Temporal Regulation of Apterous Activity During Development of the *Drosophila* Wing. Development **127**, 3069:78.

Panin, V. M., Papayannopoulos, V., Wilson, R., and Irvine, K. D. (1997). Fringe Modulates Notch-ligand Interactions. Nature **387**, 908:913.

Teleman, A.A., Strigini, M., and Cohen, S.M. (2001) Shaping Morphogen Gradients, Cell **105**, 559.

Wolpert, L. (1996) One hundred Years of Positional Information. Trends Genet. **12**, 359.

Caenorhabditis elegans: a gateway to metazoan systems biology

Julián Cerón

Massachusetts General Hospital Cancer Center and Harvard Medical School, Building 149, 13th Street, Charlestown, 02129 MA, USA.

1. Abstract

The recent completion of several metazoan genome sequences presents unprecedented opportunities to researchers studying regulatory networks. The *Caenorhabditis elegans* genome was the first metazoan genome to be sequenced and, as a consequence, researchers using this nematode as a model organism have had a head start in such studies. A few years after the *C. elegans* genome sequence was determined, various genome-wide studies have gathered extensive information on phenotypes, expression profiles and protein-protein interactions. The simple anatomy of the nematode and the stereotyped lineage of its limited number of somatic cells (959) facilitate a holistic view of how molecular networks regulate a multicellular animal. Here I review the feasibility of using *C. elegans* as a model organism for the study of complex biological systems.

2. The model

Sydney Brenner, recently awarded the Nobel Prize, wrote: "*the future lay in tackling more complex biological problems*". In 1965, Brenner selected the nematode *Caenorhabditis elegans* to tackle how genes act to create an organism and a functional nervous system. The first manuscript on *C. elegans* genetics was published in 1974 (Brenner, 1974), and thirty years later thousands of researches around the world benefit from the multiple advantages of this little worm with its simple anatomy and excellent methods for genetic analysis. The 1mm adult animal has 959 somatic cells derived from stereotypical cell lineages that can be traced along development, and hundreds of germ cells (Sulston and Horvitz, 1977). *C. elegans* is diploid and has five pairs of autosomal chromosomes. Gender is determined by sex chromosomes, which are XX in hermaphrodites and XO in males. *C. elegans* is easy to maintain in the lab since it is fed with *E. coli* bacteria and grown on agar Petri plates at temperatures between 15°C and 25°C. The life cycle, from embryo to adult through four larval stages, takes about three days at 20°C. The genetic manipulation is greatly facilitated by its short life cycle and the hermaphrodite's self-fertilization that facilitates generation of homozygous mutant stocks. Its average life span of 2-3 weeks is a feature that, together with the easy genetic manipulation, did not go unnoticed by the many researches working on aging in *C. elegans*. If geneticists have numerous reasons

Understanding and Exploiting Systems Biology

to fall in love with this nematode, cell biologists are flirting with *C. elegans* because its transparent body makes its cells visible under the microscope. In this chapter I encourage scientists working on systems biology to gain appreciation for this worm and seriously consider a date with this model organism.

3. The Genome

The nearly complete *C. elegans* genome sequence was published in 1998, providing for the first time all the genetic information required to make a multicellular organism (Comsortium for sequencing the *C. elegans* genome, 1998). The wormbase consortium (www.wormbase.org) is constantly refining and updating the annotation of the genome (Schwarz et al, 2006), and the latest release (WS156) lists 22,698 genes, including 912 genes encoding RNA transcripts only. These non-coding RNA genes (ncRNAs) are, mentioned in order of abundance, transfer RNA (tRNA) genes, ribosomal RNA (rRNA) genes, trans-spliced leader RNA genes, microRNA (miRNA) genes, spliceosomal RNA genes, and small nucleolar RNAs (snoRNA) genes (Stricklin et al, 2001). The search for new *C. elegans* genes is still ongoing and new gene models are being supported thanks to the improvement of the algorithms for gene prediction, the growth of Expressed Sequence Tag (EST) and ORF sequence tag (OST) databases, and the comparison with other nematodes genomes as *Caenorhabditis briggsae*. The publication of the draft genome sequence of *C. briggsae* (Stein et al, 2003), a soil nematode estimated to have diverged from *C. elegans* approximately 80-100 million years ago, provides a drastic improvement in the annotation of the *C. elegans* genome and will facilitate comparative genomics as well as the study of the evolutionary changes during development (Gupta and Sternberg, 2003). For example, it has been observed that genes located in the center of the chromosome have generally more essential functions and present lower rate of divergence with the *C. briggsae* genome.

Although unusual among animals, the sequence of *C. elegans* and other nematodes revealed the presence of operons, which are polycistronic gene clusters containing two or more genes (Blumenthal et al, 2002; Blumenthal and Gleason 2003). About 15% of all *C. elegans* genes are part of operons and frequently encode proteins related to the basic machinery of gene expression. As genes in operons are co-expressed, they are candidates to be involved in similar biological processes.

Four years after the publication of the *C. elegans* genome, the *Drosophila* and human genome sequence were completed and to the surprise of many people, the number of genes in humans, flies and worms is surprisingly similar, roughly 30,000, 14,000, and 20,000, respectively (Adams et al, 2000; Venter et al, 2001). Thus, the differences in biological complexity existing between humans and invertebrates might be accounted by more alternative splicing, more functional domains and complex control of gene expression rather than the absolute number of genes (Hodgkin, 2001). Interestingly, the worm genome is about 30 times smaller than the human, but has roughly the same number of genes. Moreover,

the human genome is abundant in long intergenic regions that may have regulatory functions.
The scientist community highlighted the day of the publication of the *C. elegans* genome, but we should still remember the words of Sydney Brenner; "The sequence is not the end of the day. It's the beginning of the day".

4. The ORFeome

The ORFeome project aims to clone full-length open reading frames (ORFs) corresponding to all the predicted *C. elegans* protein-coding genes into Gateway donor vectors, which made them easily transferable to any expression vector of interest (Reboul et al, 2003). To date, more than 12,500 ORFs, amplified on basis of predicted gene models, have been cloned in Gateway vectors (Lamesch et al, 2004). The Gateway recombinational cloning system (Hartley et al, 2000) allows efficiency, adaptability and compatibility in the generation of resources for high-throughput approaches. Thus, a given ORF (or any other PCR product) cloned into a universal donor vector could be easily transferred to a variety of destination vectors in parallel, and therefore, generate reagents for different large-scale studies as RNAi or Yeast Two Hybrid (Y2H) libraries (Figure 1).
Moreover, multisite Gateway cloning allows the linking of two or more DNA fragments from different entry clones into the same destination vectors. Such technology could for example be used to link a cell-specific promoter, with (Green Fluorescent Protein) GFP, and with the collection of ORFs to study the subcellular location of thousands of proteins in a g iven cell.
In summary, *C. elegans* researchers can count on the ~12,500 ORFs in a flexible recombinational cloning format for high-throughput operations.

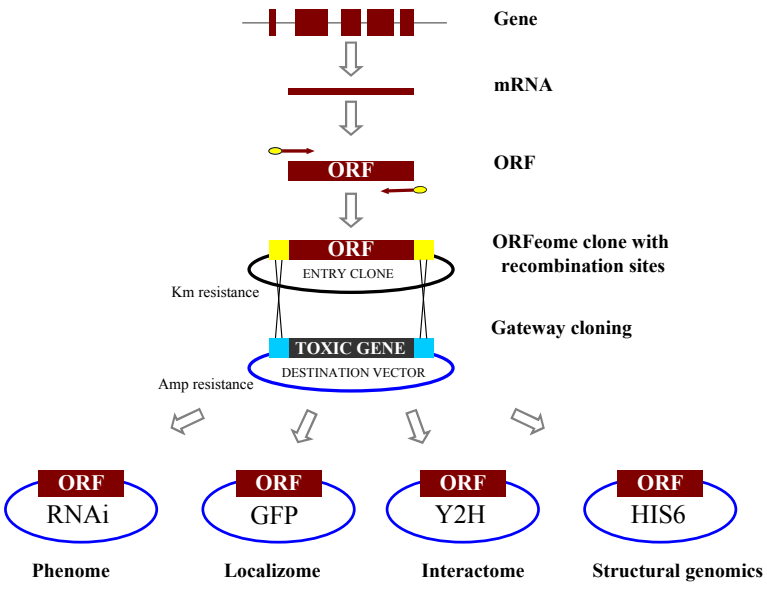


Figure 1.The *C.elegans* ORFeome. Thousands of ORFs are cloned into a Gateway donor vector that allows an efficient transference to destination vectors suitable for diverse functional genomics approaches.

5. Proteome

Ignited by the good quality of the genome sequence, several initiatives are underway to characterize the *C. elegans* proteome. The proteome is being developed in two directions; one is the global annotation and classification of proteins, and the second is the study of protein structures and post-translational modifications in individual proteins. Wormbase version WS156 cataloged 23,086 proteins, including 3,003 alternate splice forms. Importantly, about 80% of all predicted proteins are confirmed or partially confirmed by transcript evidence. In addition, the Wormbase database contains diverse information about proteins, including protein motifs, molecular weight, isoelectric point, and amino acids composition. Another web resource, the Integr8 web portal, provides easy access to integrated information about deciphered genomes, including *C. elegans*, and their corresponding proteomes (Pruess et al, 2005). In this database, for example, we can find a classification of *C. elegans* proteins and learn that there are 535 protein kinases and 216 C2H2-type zinc fingers in the *C. elegans* proteome. An additional *C. elegans* proteome database is WormPD (Costanzo et al, 2001)(see Table 1).

There is a high-throughput proteomic project under way to confirm protein-coding genes by mass spectrometry. This project has already identified 3363 proteins, 121 of which previously had no experimental support (Merrihew, Thomas and MacCoss, unpublished). Mass spectrometry approaches could also inform us about post-translational modifications and protein levels in a particular sample (Venable et al, 2004).

Structural genomic groups are studying individual protein structures in a large-scale format. One of the limiting factors in such studies is the low efficiency in the expression of recombinant proteins. In a large-scale approach, Luan and coworkers (Luan et al, 2004) developed a robotic pipeline for recombinant protein expression, applying the gateway cloning technology to transfer 10,167 ORFs into protein expression vectors in *E. coli*. They observed expression for 4,854 ORFs, and 1,536 were soluble. This group has already determined the crystal structure of 85 proteins or proteins fragments and solved 19 structures to date. Their long-term goal is to solve the three-dimensional structures by x-ray crystallography and NMR of the expressed proteins.

6. Phenome

The function of a gene is commonly inferred from its loss of function phenotype. Genes can be inactivated either by mutation or RNAi. The efforts of individuals and two knockout consortiums have produced mutant alleles many genes, which are available on request (see Table 1). Conveniently, a deletion in a gene of interest can be requested from these knockout consortiums and, frequently, is successfully generated in few months. Although the genome sequence has been of enormous importance in identifying the genes affected by genetic mutations,

1,685 classical genetic loci still remain uncloned, or not mapped to any known molecular loci.

Table 1 Web resources for *C.elegans*

URL	Description
www.wormbase.org	Main database of *C.elegans* biology
http://elegans.swmed.edu/	General resources of interest for *C.elegans* researches
http://biosci.umn.edu/CGC/CGChomepage.htm	Center for collecting, maintaining, and distributing stocks of *C.elegans*
http://celeganskoconsortium.omrf.org/	Consortium that produce deletion alleles at specified gene targets
http://shigen.lab.nig.ac.jp/c.elegans/index.jsp	Project that produce deletion alleles at specified gene targets
http://www.geneservice.co.uk/products/rnai/index.jsp	RNAi library generated in Ahringer lab (Geneservice)
http://www.geneservice.co.uk/products/clones/Celegans_Prom.jsp	Promoterome library (Geneservice)
http://www.geneservice.co.uk/products/cdna/Celegans_ORF.jsp	C.elegans ORFeome library version 1.1
http://www.wormatlas.org/	Database of behavioral and structural anatomy of *C.elegans*
http://sgce.cbse.uab.edu/index.php	Database of crystals and structures of *C.elegans* proteins
http://www.wormbook.org/	Collection of peer-reviewed chapters about *C.elegans* biology
http://www.sanger.ac.uk/cgi-bin/blast/submitblast/c_elegans	*C.elegans* Blast server
http://elegans.swmed.edu/Worm_labs/	Individual *C.elegans* lab servers
http://www.bio.unc.edu/faculty/goldstein/lab/movies.html	*C.elegans* movies
http://vidal.dfci.harvard.edu/interactomedb/i-View/interactomeCurrent.pl	Interactome Database
http://nematode.lab.nig.ac.jp/	Expression Pattern Database. In situ hybridization images.
http://elegans.bcgsc.ca/perl/eprofile/qgene	Expression Pattern Database for *C .elegans* promoter::GFP fusions
http://workhorse.stanford.edu/cgi-bin/genebar/generic_genegraph.pl	Expression levels in experiments related with development, germline, aging, etc..
http://vidal.dfci.harvard.edu/promoteromedb/	Promoterome Database
http://www.textpresso.org/	An information retrieval and extraction system for *C.elegans* literature
http://nematoda.bio.nyu.edu/cgi-bin/rnai/index.cgi	Phenotypic data from RNAi studies in *C. elegans*
http://workhorse.stanford.edu/cgi-bin/gl/gl_mod.cgi	Database for co-expressed genes classified in 48 groups
http://inparanoid.cgb.ki.se/	To search for eukaryotic ortholog groups
http://tenaya.caltech.edu:8000/predict/	Predicitions of C. elegans Genetic Interactions
http://worm-srv1.mpi-cbg.de/dbScreen/index.html	Detailed description of embryonic RNAi phenotypes corresponding to genes on Chr III
https://www.proteome.com/proteome/	Worm Proteome Database (WormPD)

However, RNA interference (RNAi) is the tool that had elevated *C. elegans* to the category of "top model" system. Discovered in *C. elegans*, RNAi induces sequence-specific degradation of homologous mRNA triggered by presence of dsRNA in the cell (Fire et al, 1998). The dsRNA can be administrated by injection (Fire et al, 1998), soaking (Tabara et al, 1998) and, conveniently, by

feeding worms with bacteria expressing dsRNA (Timmons and Fire, 1998). The two existing feeding RNAi libraries, which are reusable, were generated in the laboratories of Julie Ahringer (Cambridge) and Marc Vidal (Boston) and have been validated in genome-wide screens covering ~90% of all the *C. elegans* genes (Kamath et al, 2003; Rual and Ceron et al, 2004).
Interestingly, only about 15% of *C. elegans* genes have been associated with phenotypes, indicating a high presence of functional redundancies. We have addressed this issue at the laboratory of Sander van den Heuvel by screening for synthetic genetic interaction with *lin-35*, which is the single Rb (Retinoblastoma tumor suppressor) related gene in *C elegans*. We used the ORFeome RNAi library to inhibit 10,953 genes by feeding RNAi in wild-type and *lin-35* mutant viable animals and have identified 36 genes that show synthetic or enhanced RNAi phenotype in *lin-35* Rb mutants (Ceron et al, unpublished). Thus, the work of our lab and others indicate that many *C. elegans* functions should be uncovered by inactivation of two or more genes in parallel (Fraser, 2004).
The effect of RNAi is dosage dependent and therefore mutants with higher sensitivity to RNAi could enhance RNAi effects (Simmer et al, 2002). Moreover, RNAi by injection, although more laborious, produces stronger RNAi effect. Hence, RNAi strength can be modulated and the administration regulated by feeding through development. Since cellular exit of dsRNA from normal animal cells has not been directly observed, it is possible to perform RNAi exclusively in a specific tissue by using transgenes expressing hairpin dsRNA under the control of tissue-specific promoters (Tavernarakis et al, 2000; Timmons et al, 2003). This method is commonly used to inactivate genes in neuronal cells, which seem to be resistant to systemic RNAi.

7. Interactome

A physical interaction between two proteins is a strong argument for thinking that those proteins act in related biological process. In order to build a large-scale protein-protein interaction (PPI) map, 1873 ORFs from the *C. elegans* ORFeome library were transferred into Yeast Two Hybrid (Y2H) bait destination vectors that were screened against two different Gal4 activation domain libraries (Li et al, 2004). As result, the initial version of the *C. elegans* interactome contains ~4000 interactions. These interactions can be subdivided into three confidence classes: Core-1, Core-2 and Non-Core of 858, 1299 and 1892 interactions respectively. The overall quality of the dataset was experimentally validated. These ~4000 interactions together with interologs predicted *in silico* and interactions previously known, make for ~5500 interactions available in the currently available version of the Worm interactome (WI5). Interactions present in WI5 generate a network of 2898 nodes and 5460 edges. Still, the number of false positives is a concern but protocols are being optimized to maximize the specificity of Y2H assays (Vidalain et al, 2004). Significant correlation has been observed between interacting protein pairs and *C. elegans* expression profiles (Transcriptome) as well as RNAi phenotypes (Phenome), suggesting that these

interactions are not randomized. The integration of interactome with phenome and transcriptome dataset, has already contributed to the generation of numerous biological hypotheses related to vulva development, DNA damage response, germline formation, and the TGF pathway (Walhout et al, 2000; Boulton et al, 2002; Walhout et al, 2002; Reinke et al, 2004; Tewari et al, 2004)

8. Transcriptome

Microarray technologies allow us to compare the levels of RNA molecules in diverse genetic or environmental conditions. Currently, commercial suppliers offer microarray chips for more than 20,000 *C. elegans* transcripts. Thus, total mRNA of particular worms can be extracted and used in these chips to scan for gene expression differences between two mRNA sets of interest. This approach has been widely used and, as example, facilitated the grouping of genes with predominant expression in the germ line (Reinke et al 2000) or in males (Jiang et al, 2001). Moreover, microarray analyses have also provided insights into cellular pathways identifying downstream genes of *hda-1* Histone Deacetylase –1 (Whetstine et al, 2005), *let-60* RAS (Romagnolo B et al, 2002), or *daf-16* insulin/IGF-1 genes (Murphy et al, 2003).

In a compilation of microarray data, the expression data involving 17,661 genes and 553 microarray experiments were analyzed and genes were clustered in 44 groups based on coexpression in diverse experimental conditions (Kim et al, 2001). Most of these microarray studies used mRNA of the whole organism (embryo, larval or adult), disregarding cell type or tissue specificity. The lack of tissue specific samples was addressed by two methods. First, by collecting specific embryonic cell types labeled with tissue-specific promoters expressing GFP (Green Fluorescent Protein) using FACS (Fluorescence Activated Cell Sorting) (Christensen et al, 2002; Fox et al, 2005). However, only embryos at certain developmental stages can be dissociated. The second approach relies on the expression of a tagged poly(A) binding protein (PABP) in specific cell types, which allows the recovery of cell specific mRNA by inmuno-precipitation (Roy et al, 2002; Pauli et al, 2006).

mRNA expression can also be precisely detected by *in situ* hybridization. The Kohara lab in Japan has performed a large-scale project to localize mRNAs at different stages (Tabara et al, 1996; and table 1). There is a database containing whole-mount *in situ* images corresponding to 11,237 cDNA clones. These images could provide a general view of the expression pattern although there is limited cellular resolution.

Interestingly, and supporting one more time studies in model organisms, analysis of microarray data in different organisms led to the identification of co-regulated gene clusters among yeast, worms, flies and humans (Stuart et al, 2003).

9. Localizome

Where a protein is located at cellular and subcellular level is crucial information to understanding its function. In addition to traditional methods of immunostaining, which are often laborious and dependent on having an immunostaining-friendly antibody, GFP labeling methods are being widely used. Promoter GFP fusions are an alternative method that is more amenable to high-throughput approaches. The Promoterome project (Dupuy et al, 2004) has already released Gateway compatible promoter constructs for ~ 6,500 *C. elegans* genes. Promoter GFP fusions are currently being used for live imaging and ~2000 GFP patterns are freely available on web sites (see table 1). Unfortunately, there are two limitations in this approach: first, the absence of complete and important promoter sequences (as sequences in trans), and second, the expression of the reporter, frequently incorporated in large extrachromosomal arrays, is subjected to germ-line silencing and lost in some of the somatic cells (mosaicism).

10. Integration of "–omes": a tale of nodes and edges

Understanding of the cell machinery might be better achieved by investigating functional modules rather than individual molecules. A functional module is composed of multiple molecules, which all together exhibit properties not found among individual components (Hartwell et al, 1999). Thus, a ribosome and a signal transduction pathway are examples of functional modules. These functional links are commonly represented by diagram of nodes and edges. Nodes represent the components of biological networks (genes, proteins, RNAs, or metabolites) and edges represent interactions between those components. As commented below, vast amount of genomics data have already begun to be successfully integrated to discover functional modules.

Functional modules in early embryogenesis

Gunsalus and coworkers have recently published a predictive model of how molecular nodes are assembled to work in *C. elegans* early embryogenesis (Gunsalus et al, 2005). They selected the first two cell divisions in the embryo as a biological system. The functional network graph for early embryogenesis resulted from the overlapping of specific graphs representing phenotypic correlation (Sonnichsen et al, 2005), physical interaction (Li et al., 2004) and transcriptional profile similarities (Kim et al, 2001)(Figure 2). Transcriptional and phenotypic correlation was considered relevant when the Pearson Correlation Coefficient (PCC) was above a certain statistical threshold. Importantly, they used a high quality phenotypic profiling based on a full-genome RNAi screen (19,075 genes tested) that annotated detailed phenotypic information for 661 genes presenting altered phenotypes in the early embryo (Sonnichsen et al, 20005). The effects of these 661 RNAi experiments in the early embryo were analyzed by time-lapse video recording and 45 defects were annotated. As result

of integrating these genomics dataset, they created a high confidence network containing 305 nodes joined by 1,036 edges, each supported by two or three types of functional evidence. From this multiple support network they predicted several molecular machines. To validate this innovative approach, they studied the expression pattern of several previously uncharacterized genes that were predicted nodes of functional modules and observed expected expression patterns in the early embryo.
In summary, the model resulting from the integrated network suggests that *C. elegans* early embryogenesis is achieved through coordination of a limited set of molecular machines.

Modeling vulval development
There is a challenging ongoing project for computer modeling of several aspects of *C. elegans* development. As starting point, the Stern lab selected the process of cell fate acquisition in vulval precursors cells (VPCs), which is one of the most well known developmental systems in *C. elegans*. Details of this project are described at:
http://www.wisdom.weizmann.ac.il/~kam/CelegansModel/CelegansModel.htm.
This group also plans to extend the project to nerve cells and model behavior. This type of approach, which mainly uses existing genetic information, is of great interest, especially now that an avalanche of functional data becoming available. Thus, the option of performing cyber experiments from a Caribbean beach before taking them to the lab bench does not sound too futurist anymore. (Fisher et al 2005).

Predictions of genetic interactions
In a different bio-informatics approach, interactome data, gene expression data, phenotype data and functional annotation have been computationally integrated to obtain a global view of functional interaction in three different model organisms (*Saccharomyces cerevisiae*, *Drosophila melanogaster* and *C. elegans*). A free access database has been set up for searching genes predicted to interact genetically with your gene of interest (Zhong and Sternberg, 2006) (Table 1).

11. Extending hypotheses to other systems

About 40% of *C. elegans* genes have apparent human homologs and almost all protein domains found in human are present in *C. elegans*. Moreover, the InParanoid algorithm has identified 4558 *C. elegans* orthologs of human genes (O'Brien et al, 2005). Orthologous genes are those whose last common ancestor split into two gene lineages through speciation. Although sequence orthology does not necessarily imply the same function, genes that were shown to be descendants of the same ancestor (orthologous genes) exhibit, in general, retained similar function over the course of evolution. Therefore, as long as we

are able to establish orthology, information obtained for a gene function in one organism is potentially transferable to the other.

12. Concluding Remarks

The nematode *C. elegans* present all the tools required to assemble the numerous fine-tuned mechanisms that let a multicellular organism develop, growth, interact with the environment, and reproduce. Thus, *C. elegans* is leading the research in genomics approaches to understand where, when and how proteins, RNAs and other metabolites act to build functional biological networks in a complex animal. Although improved methods and further studies will be required to cross-validate the data quality for each edge of the network, the actual functional map that can be drawn for *C. elegans* is an excellent tool to generate biological hypothesis about how a biological system works.

Acknowledgements. *I am thankful to Silvia G. Acinas, Mike Boxem and Abha Chandra for critical reading of the manuscript, Sander van den Heuvel, and members of the Heuvel lab for discussions, and Erich Schwarz (Wormbase) for providing helpful information.*

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195.

Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., and Kim, S. K. (2002). A global analysis of *Caenorhabditis elegans* operons. *Nature* **417**, 851-854.

Blumenthal, T., and Gleason, K. S. (2003). *Caenorhabditis elegans* operons: form and function. *Nat Rev Genet* **4**, 112-120.

Boulton, S. J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D. E., and Vidal, M. (2002). Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**, 127-131.

Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71-94.

Christensen, M., Estevez, A., Yin, X., Fox, R., Morrison, R., McDonnell, M., Gleason, C., Miller, D. M., 3rd, and Strange, K. (2002). A primary culture system for functional analysis of *C. elegans* neurons and muscle cells. *Neuron* **33**, 503-514.

Consortium, T. C. e. S. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012-2018.

Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Kranz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., *et al*. (2001). YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75-79.

Dupuy, D., Li, Q. R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I. A., *et al*. (2004). A first version of the *Caenorhabditis elegans* Promoterome. *Genome Res* **14**, 2169-2175.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-811.

Fisher, J., Piterman, N., Hubbard, E. J., Stern, M. J., and Harel, D. (2005). Computational insights into *Caenorhabditis elegans* vulval development. *Proc Natl Acad Sci U S A* **102**, 1951-1956.

Fox, R. M., Von Stetina, S. E., Barlow, S. J., Shaffer, C., Olszewski, K. L., Moore, J. H., Dupuy, D., Vidal, M., and Miller, D. M., 3rd (2005). A gene expression fingerprint of *C. elegans* embryonic motor neurons. *BMC Genomics* **6**, 42.

Fraser, A. (2004). Towards full employment: using RNAi to find roles for the redundant. *Oncogene* **23**, 8346-8352.

Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J. D., Hao, T., Berriz, G. F., Bertin, N., Huang, J., Chuang, L. S., *et al*. (2005). Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861-865.

Gupta, B. P., and Sternberg, P. W. (2003). The draft genome sequence of the nematode Caenorhabditis briggsae, a companion to *C. elegans*. *Genome Biol* **4**, 238.

Hartley, J.L., Temple, G.F., and Brasch, M.A. (2000). DNA cloning using in vitro site-specific recombination. *Genome Res*. **10**, 1788-1795.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.

Hodgkin, J. (2001). What does a worm want with 20,000 genes? *Genome Biol* **2**, COMMENT2008.

Jiang, M., Ryu, J., Kiraly, M., Duke, K., Reinke, V., and Kim, S. K. (2001). Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **98**, 218-223.

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., *et al*. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231-237.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., and Davidson, G. S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087-2092.

Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhaute, J., Hill, D. E., and Vidal, M. (2004). *C.*

elegans ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res* **14**, 2064-2069.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al*. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540-543.

Luan, C. H., Qiu, S., Finley, J. B., Carson, M., Gray, R. J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., *et al*. (2004). High-throughput expression of *C. elegans* proteins. *Genome Res* **14**, 2102-2110.

Murphy, C. T., McCarroll, S. A., Bargmann, C. I., Fraser, A., Kamath, R. S., Ahringer, J., Li, H., and Kenyon, C. (2003). Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424**, 277-283.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33**, D476-480.

Pauli, F., Liu, Y., Kim, Y. A., Chen, P. J., and Kim, S. K. (2006). Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in *C. elegans*. *Development* **133**, 287-295.

Pruess, M., Kersey, P., and Apweiler, R. (2005). The Integr8 project--a resource for genomic and proteomic data. *In Silico Biol* **5**, 179-185.

Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., *et al*. (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet* **34**, 35-41.

Reinke, V., Smith, H. E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S. J., Davis, E. B., Scherer, S., Ward, S., and Kim, S. K. (2000). A global profile of germline gene expression in *C. elegans*. *Mol Cell* **6**, 605-616.

Reinke, V., Gil, I. S., Ward, S., and Kazmer, K. (2004). Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131**, 311-323.

Romagnolo, B., Jiang, M., Kiraly, M., Breton, C., Begley, R., Wang, J., Lund, J., and Kim, S. K. (2002). Downstream targets of let-60 Ras in *Caenorhabditis elegans*. *Dev Biol* **247**, 127-136.

Roy, P. J., Stuart, J. M., Lund, J., and Kim, S. K. (2002). Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418**, 975-979.

Rual, J. F., Ceron, J., Koreth, J., Hao, T., Nicot, A. S., Hirozane-Kishikawa, T., Vandenhaute, J., Orkin, S. H., Hill, D. E., van den Heuvel, S., and Vidal, M. (2004). Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res* **14**, 2162-2168.

Schwarz, E. M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P., et al. (2006). WormBase: better software, richer content. *Nucleic Acids Res* **34**, D475-478.

Simmer, F., Tijsterman, M., Parrish, S., Koushika, S. P., Nonet, M. L., Fire, A., Ahringer, J., and Plasterk, R. H. (2002). Loss of the putative RNA-directed

RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr Biol* **12**, 1317-1319.

Sonnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A. M., Artelt, J., Bettencourt, P., Cassin, E., *et al*. (2005). Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434**, 462-469.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* **1**, E45.

Stricklin, S.L., Griffiths-Jones, S., and Eddy, S.R. *C. elegans* noncoding RNA genes (June 25, 2005), *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.1.1, http://www.wormbook.org.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249-255.

Sulston, J. E., and Horvitz, H. R. (1977). Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev Biol* **56**, 110-156.

Tabara, H., Motohashi, T., and Kohara, Y. (1996). A multi-well version of in situ hybridization on whole mount embryos of *Caenorhabditis elegans*. *Nucleic Acids Res* **24**, 2119-2124.

Tabara, H., Grishik, A. and Mello, C.C. (1998). RNAi in *C. elegans*: soaking in the genome sequence. *Science* **282**, 430-431.

Tavernarakis, N., Wang, S. L., Dorovkov, M., Ryazanov, A., and Driscoll, M. (2000). Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nat Genet* **24**, 180-183.

Tewari, M., Hu, P. J., Ahn, J. S., Ayivi-Guedehoussou, N., Vidalain, P. O., Li, S., Milstein, S., Armstrong, C. M., Boxem, M., Butler, M. D., *et al*. (2004). Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. *Mol Cell* **13**, 469-482.

Timmons, L., Tabara, H., Mello, C. C., and Fire, A. Z. (2003). Inducible systemic RNA silencing in *Caenorhabditis elegans*. *Mol Biol Cell* **14**, 2972-2983.

Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**, 39-45.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001). The sequence of the human genome. *Science* **291**, 1304-1351.

Vidalain, P. O., Boxem, M., Ge, H., Li, S., and Vidal, M. (2004). Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**, 363-370.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in *C.*

elegans using proteins involved in vulval development. *Science* **287**, 116-122.

Walhout, A. J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K. C., Schetter, A. J., *et al*. (2002). Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr Biol* **12**, 1952-1958.

Whetstine, J. R., Ceron, J., Ladd, B., Dufourcq, P., Reinke, V., and Shi, Y. (2005). Regulation of tissue-specific and extracellular matrix-related genes by a class I histone deacetylase. *Mol Cell* **18**, 483-490.

Zhong, W., and Sternberg, P. W. (2006). Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311**, 1481-1484.

Partners of Fate: Robust control of cell commitment in stem cell niches

Juan F. Poyatos

Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain. e-mail:jpoyatos@cnio.es

Keywords: Stem cells, niche, cell competition, bistability, hysteresis, systems biology

1. Abstract

Stem cells are the essential precursors of all cell types in our bodies. A particular tissue location, known as niche, provides the necessary factors for their maintenance in adult organisms, i.e., stem cell self-renewal. How does a stem cell decide to abandon the niche and initiate the program of differentiation? One can readily imagine such commitment to be severely controlled, as sloppy decision-making can seriously risk the survival of the individual. Here, I describe a robust molecular control module regulating stem cell differentiation in the niche of the *Drosophila* ovary. This mode of control is based on a combination of interlinked positive and negative feedback loops which regulate intracellular signal transduction and intercellular competition among stem cells. Consequently, cells within the niche actively determine each other's fate. To completely describe this strategy, I combine the molecular knowledge recently gained in the fields of stem cell and cell competition biology with ideas and tools from the emerging field of systems biology. This type of multidisciplinary approaches, the molecular and the system-level oriented, can fully elucidate how stem cells work and thus start opening unforeseen avenues for the development of novel biomedical strategies.

2. Introduction

Stem cells are the ultimate generalists. As Leonardo da Vinci was capable of making beautiful paintings, sketching sophisticated engineering structures, or dissecting bodies, stem cells are able to transform into very different functioning cells, competent among other things to interpret the external world, carry oxygen to tissues, or protect us against external radiation. This amazing capability is however only found in a type of stem cells, those in charge of making the adult individual, known as embryonic stem cells, and thus embryonic stem cells are considered to be *pluripotent*. This pluripotency is modified along the development process when stem cells are transformed into adult stem cells giving rise to different tissues and organs and being later directly involved in their

posterior maintenance. Initially believed to have very restricted differentiation possibilities, it is now known that some of these adult stem cells exhibit a much wider potential. For these reasons while originally considered as *unipotent* cells they are now generally accepted to be *multipotent*.

Given the importance of adult stem cells for tissue homeostasis and repair, one can anticipate that sophisticated mechanisms must have evolved for their protection and for the regulation of stem-cell self-renewal and differentiation. What type of mechanisms could these be? Two initial hypotheses seem a priori equally valid. Either some cells are specifically programmed to behave as adult stem cells, i.e., they behave as stem cells in an almost cell-autonomously way with the capacity for self-renewal being also somehow genetically programmed, or there exists specific locations regulating the *stemness* of otherwise relatively standard-behaving cells. As in many other cases the truth is somehow in the middle. In particular, locations inside tissues with the role of protect and regulate stemness were initially identified almost four decades ago from studies of transplanted hematopoetic progenitors (Schofield, R., 1978). These locations have been later identified in many different tissues. All these reports contributed to clarify and characterize one of the most important concepts of stem cell biology: *the stem cell niche*.

What is a stem cell niche? Niches are specific microenvironments constituted by a subset of tissue cells and extracellular substrates that can indefinitely support the self-renewing of stem cells. A few points in this definition are worth highlighting: 1) the presence of a given anatomical organization, or niche structure, constituted by one or more specialized cell groups, 2) the localized signalling cells which generally emit a principal signal (*the stem factor*), and, of course, 3) the stem cells. Characterizing stem cell niches *in vivo* remained then the Holy Grail of stem cell biology and a task difficult to achieve. However, during recent years, an increasing number of niches are being localized and classified in different tissues such as testis, skin, or gut crypts (Fuchs, E., *et al*, 2004).

The manuscript is organized as follows. First, I describe the stem cell niche in the *Drosophila* ovary and introduce its basic molecular agents, as this scenario is the focus of my discussions. Second, I introduce basic concepts of cell competition required for the understanding of how stem cell self-renewal is achieved in this niche. Finally, I describe the dynamics of the molecular control module regulating stem cell differentiation by mathematical modelling, and analyse these models in detail. This module induces a two-layer regulation at both the intracellular and intercellular level enhancing the robustness of stem cell fate commitment. With this working example, I hope to illustrate the need to incorporate system-level thinking to standard molecular/genetic approaches to help developing our knowledge of how stem cells work.

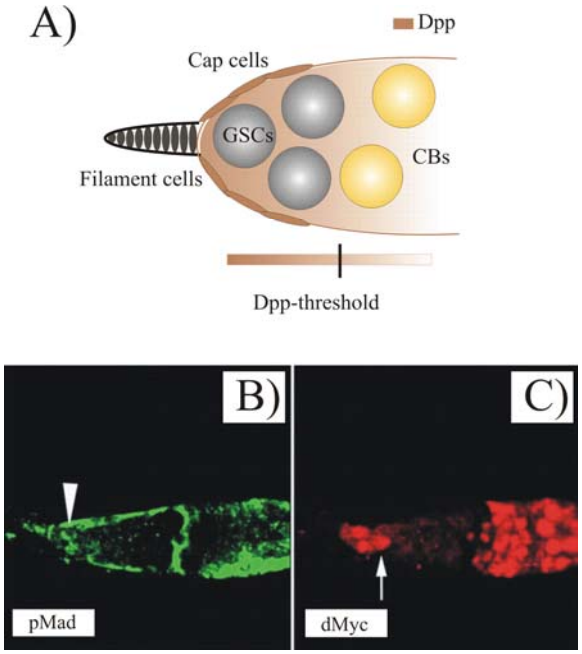


Figure 1. A. Niche of the Drosophila ovary. This structure is mainly constituted by postmitotic somatic cells termed terminal *filament cells* and several epithelial stromal *cap cells*. The principal signal of this niche is the morphogen Decapentaplegic (Dpp). At a given Dpp threshold stem cells (GSCs) differentiate. **B-C.** Staining of Drosophila ovarioles with a d-Myc specific antibody reveals d-Myc expression in the stem cell niche. pMad (green) and d-Myc (red) staining of the same germarium. All pMad positive cells (arrowhead) express high levels of d-Myc (arrow) (B-C. figures courtesy of B. Díaz, I. Fernandez-Ruiz).

3. Stem cell niche in the *Drosophila* ovariole

The Drosophila ovary is constituted of 16-20 functionally equivalent developing egg strings known as *ovarioles*. Within these strings reside two or three cells acting as germ line stem cells (GSCs) whose progeny differentiate into eggs within 8 days as they move along the ovariole. GSCs have been located at the tip of the germarium and this environment was later identified as a stem cell niche. The anatomical organization of the niche structure is mainly constituted by a single stack of postmitotic somatic cells termed terminal *filament cells* and, at the base of these terminal cells, several epithelial stromal *cap cells*. The principal signal of this niche has also been revealed. This stem factor is the morphogen Decapentaplegic (Dpp), the *Drosophila* homologue of the bone morphogenetic protein 2/4 (for details see Spradling, A., *et al*, 2001, and references therein).
GSCs at the niche transduce the highest levels of Dpp and thus continue as stem cells. Excessive Dpp signalling was shown to block germ cell differentiation which suggested that Dpp downregulation is associated to differentiation, i.e., the transformation into a stem cell daughter named cystoblast, CBs (Kai, T. and Spradling, A., 2004). This transformation requires activation of the bag-of-

marbles (*bam*) gene. Dpp signalling acts on Bam as follows: Dpp signal transduction requires phosphorylation of Mad (pMad) and its nuclear translocation. pMad binds to a silencer element which is in turn involved in Bam transcriptional shut-off (Chen D, McKearin D., 2003). This is not the only functional relation between Dpp signalling and Bam. More recent data also supports the down-regulation of Dpp activation by Bam (Casanueva M.O. and Ferguson E.L., 2004). Our experiments (Díaz B. *et al,* 2006) have further extended this core regulatory module by introducing a third molecular player, the protein *Drosophila* Myc (d-Myc, see Results).

4. Cellular competition

The idea that cells of multicellular organisms could compete instead of cooperate among them remained appealing and controversial for many years. Cellular competition was finally experimentally confirmed in *Drosophila*. In an experiment where cells with different metabolic rates were confronted, cells that in isolation shown to be completely viable disappeared due to the additional presence of metabolically more efficient ones. These different metabolic rates were achieved by generating artificially mutations of ribosomal proteins and the corresponding cell mutants are since then termed *Minutes* (Morata, G. and Ripoll, P., 1975). After this seminal work, several studies have revealed new genes able to induce competition (Díaz, B. and Moreno, E., 2005, and references therein). But, isn't it strange that cells compete rather than cooperate for the benefit of the organism? A plausible way out to this paradox could be the idea that competition would act as an efficient strategy to select for cell quality, as the presence of naturally occurring mutants, which are generally less optimal for a given set of attributes, in a normal cell environment would be filtered out by this mechanism.

What are cells competing for? Classical experiments hypothesized that cells compete with each other to fill a limited space that appears to be delineated in advance. Cell selection in this space could be based on the accessibility for a general "growth" factor, which may sometimes also be involved in the shaping of the battlefield itself. Two scenarios can be envisaged. In the first case, the growth factor is available only in small doses, and thus only cells in which the uptake of the growth factor is above a threshold would survive. Alternatively, all cells in the population have enough growth factor to survive, i.e., this factor is not limiting. However, not all of these cells are optimal, and thus the most competitive ones eliminate the others after some "quality comparison" mechanism.

This story took a twist recently when, in a surprising set of discoveries, genes able to induce cell competition above wild-type levels were identified. In particular, genes of the Myc family can transform cells into such *super-competitors* (de la Cova, C., et al., 2004, Moreno, E. and Basler, K., 2004). This phenomenon could in this way be involved in early stages of cancers, in which super-competitor cells were able to invade a particular tissue location by killing

surrounding normal cells. Unbalanced of Myc is common in many cancers and, according to the previous scenario, it could alter tissue balance that in combination to secondary mutations would lead to tumour formation.

5. Results and Discussion

5.1. A CORE REGULATORY MOTIF FOR STEM CELL DIFFERENTIATION.

I have introduced in Sec.2 the basic molecular agents involved in the determination of stem cell fate in the *Drosophila* ovary, i.e., pMad and Bam. By considering this pair we can completely characterize cells as GSCs, (pMad, Bam) = (ON, OFF), or CBs, (pMad, Bam) = (OFF, ON). This behaviour could be established in molecular terms by a simple repression of Bam by pMad. According to this, the transition from pMad-expressing to pMad-non-expressing cells (and viceversa for Bam) would be relatively gradual, following the Dpp gradient. However, this transition is rather observed as all-or-none experimentally (Fig. 1). The presence of a second interaction, this time Bam acting on pMad, can contribute to understand such sharp pattern. Indeed, recent data supports the down-regulation of Dpp activation by Bam (see Sec. 2). This repression has important dynamical consequences since the double-negative motif pMad -| Bam -- Bam -| pMad constitutes a positive loop, a common feedback-based strategy promoting cellular differentiation (Thomas R. and D'Ari R., 1990, Freeman, M., 2000). One could alternatively think of a third circuit: a positive interaction between pMad and Bam. This structure would induce again a graded differentiation but probably with less variation in protein levels with respect to simple pMad -| Bam architecture. This is a consequence of the negative feedback motif pMad -| Bam -- pMad → Bam which would act as a homeostatic regulator of protein levels (Thomas R. and D'Ari R., 1990).

A simple model describing the positive loop pMad -| Bam -- Bam -| pMad reads as

$$\frac{d[pMad]}{dt} = \frac{a_1}{1+\left([Bam]/K_1\right)^n} - b_1[pMad], \quad (1)$$

$$\frac{d[Bam]}{dt} = \frac{a_2}{1+\left([pMad]/K_2\right)^m} - b_2[Bam]. \quad (2)$$

These equations describe the rate of change of the concentrations of each of the molecular components as a function of time. In the equation for pMad, the first term denotes the production of pMad with rate a_1. This production is directly related to Dpp signalling, i.e., more Dpp implies bigger a_1. Bam repression is incorporated by introducing a Michaelis-Menten type repression term (Thomas R. and D'Ari R., 1990) acting directly on pMad production. Note that this is a simplification to the molecular picture yet to be deciphered (but see following

subsection). In this repression term, (K_1,n) are the Michaelis-Menten parameters with K_1 being the concentration of Bam required to half pMad production and n the corresponding steepness of this repression (often associated to cooperativity effects, with $n>1$ denoting cooperativity). Similar terms for production and degradation are considered to describe Bam dynamics. In this case, a_2 (b_2) is the production (degradation) rate, and (K_2,m) the Michaelis-Menten parameters. Cooperative effects are more likely in this case since pMad represses transcription of Bam by forming a complex with the Medea (Med) protein and further recruiting other cofactors (Chen D, McKearin D., 2003).

We can infer some of the dynamical properties of this motif and how it determines cell behaviour. To this aim, I associate cell types to steady states of the dynamical system, i.e., a state where the concentration of each protein does not change with time. The presence of a positive feedback has the potential to yield a molecular switch as the dynamical system can exhibit *bistability*, i.e., the coexistence of two stable steady states. This is in contrast with the other two minimal motif architectures consistent with the cell behaviour in/out the niche, i.e., simple repression and a negative feedback loop, that can only reach a single steady state. Bistability also enables the system to present *hysteresis*: the lack of reversibility as a parameter is changed (for an introduction see Strogatz. S, 2000). Both properties are commonly found in other cellular differentiation scenarios (Xiong, W. and Ferrell, J. E. Jr., 2003, Isaacs, F. J., *et al*, 2003 Bhalla U.S., *et al*, 2002, Acar M, *et al* , 2005).

These two previous features would impose some constraints on the parameters characterizing the interactions associated to the positive loop: they should exhibit a degree of non-linearity, e.g., $m = n = 2$, and proper balanced between both arms, i.e., $a_1 \sim a_2$ (Ferrell J.E., 2002). The transcriptional control of Bam expression by pMad seems to indicate the presence of the required non-linearity. I hypothesise that the strength and non-linearity of the Bam action on pMad is strong enough so that the system acts as a bistable switch. If this core were not bistable but, for instance, *ultrasensitive* the differentiation transition would work in a less robust way, e.g., there might exist some "random" events of de-differentiation. This is shown in Fig.2. where I compare how the steady state of pMad concentration changes for a bistable (continuous black line) vs. a ultrasensitive (dotted blue line) transition as Dpp decreases, i.e., far from the terminal filament cells. This concentration is high in the region with more Dpp. As Dpp decreases, pMad concentration also decreases. Note that in this intermediate Dpp regime, the system is potentially able to reach three distinct states, two stable states (high or low pMad, respectively) and one unstable state (dashed black line). A further decrease in Dpp pushes the system from the upper stem cell state (GSCs) to the lower differentiated state (CBs). While GSCs follow the path denoted with the green arrows to differentiate, CBs follow instead the path denoted with the blue arrows to de-differentiate. De-differentiation occurs at a different Dpp value to that for differentiation, i.e., there exists hysteresis. Unless fluctuations in Dpp are considerably large cell commitment is robust.

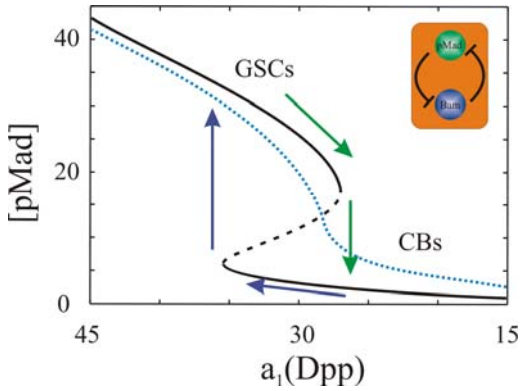


Figure 2. pMad steady states for a core module with (continuous and dashed black lines) and without (dotted blue line) bistability. In the bistable regime the system exhibits three steady states, one of them unstable (dashed line). Bistability enhances the robustness of the differentiation decision as the range of Dpp values to differentiate is different to that to dedifferentiate (hysteresis). Parameters used are: $m = n = 2$, 1 $b_1 = b_2 = 1$, $K_1 = K_2 = 10$, and $a_2 = 30$. Inset. Positive feedback loop architecture between pMad (green circle) and Bam (blue circle).

5.2. D-MYC ENHANCES THE ROBUSTNESS OF STEM CELL DIFFERENTIATION

Since Dpp signalling is a basic element associated to stem cell self-renewal in the *Drosophila* niche and d-Myc was recently proposed to influence Dpp transduction (Díaz, B. and Moreno, E., 2005), we recently studied the possible role of d-Myc in the control of stem cell fate commitment (Díaz B. *et al,* 2006). In short, we have observed a boundary between d-Myc expressing and non-expressing cells precisely located where the decision to differentiate is taking place (Fig.1.C). This suggests that d-Myc is somehow under the influence of the previous motif core and thus plays a role in the control of differentiation. Indeed, our experiments were able to delineate a new functional module constituted by Dpp signaling (pMad), Bam and d-Myc (Inset Fig. 3). The presence of d-Myc enhances the robustness of stem cell differentiation by introducing dual regulatory actions at the intracellular and intercellular level.

Here, I extend the previous mathematical framework to describe these features. Our experiments were able to show several connections to the previously core motif. Firstly, Bam is required for d-Myc repression in CBs cells and thus the presence (absence) of Bam directly implies the absence (presence) of d-Myc. Secondly, high levels of Dpp signalling slightly down-regulate d-Myc independent of Bam and, finally, d-Myc activates Dpp uptake. Thus d-Myc establishes two new feedback loops, one positive (pMad -| Bam -| d-Myc → pMad) and one negative (pMad -| d-Myc → pMad). We can mathematically describe this new control module as follows

$$\frac{d[pMad]}{dt} = \frac{a_1}{1+\left([Bam]/K_1\right)^n} + c_{13}\frac{\left([dMyc]/K_1'\right)^p}{1+\left([dMyc]/K_1'\right)^p} - b_1[pMad], \quad (3)$$

$$\frac{d[Bam]}{dt} = \frac{a_2}{1+\left([pMad]/K_2\right)^m} - b_2[Bam], \quad (4)$$

$$\frac{d[dMyc]}{dt} = \frac{a_3}{1+\left([Bam]/K_3\right)^s} + \frac{c_{31}}{1+\left([pMad]/K_3'\right)^{p'}} - b_3[dMyc], \quad (5)$$

Here, the second term in eq.3 reflects a saturating repression of d-Myc production by pMad with c_{13} being maximal production rate, and (K'_1, p) the corresponding Michaelis-Menten parameters. In eq. 5, a_3 denotes maximal rate of d-Myc production and (K_3, s) the parameters associated to Bam repression. The second term is a saturating repression of d-Myc production by pMad with c_{31} being maximal production rate and (K'_3, $p' = p$) controlling again the shape of this interaction. Finally, b_3 is the rate of degradation of d-Myc. Note that I considered two routes to close the loop between Bam and pMad, d-Myc-independent and d-Myc-dependent, respectively. Further experiments should elucidate whether this is the case or there exists a single d-Myc-dependent loop. This should not modify the main conclusions of the present analysis.

What makes this modular architecture interesting? One can notice the existence of two d-Myc dependent feedback loops, one negative and one positive. In particular, pMad establishes a negative feedback loop with d-Myc which can buffer the system against Dpp signalling fluctuations, as negative feedbacks are generally linked to homeostasis (Thomas R. and D'Ari R., 1990). To show this effect, I compared the behaviour of two systems with d-Myc activated by Dpp with or without a negative feedback with pMad (all other biochemical details being identical). I imagined a situation in which a small perturbation, for instance due to Dpp fluctuations of other sources of biochemical noise, displaces the system from its current steady state, located near the differentiation transition (Fig. 3, colour circles). In Fig. 3, I plot the time it takes to pMad concentration to come back to its previous equilibrium for a module with (Fig. 3 inset dark green curve) and without the pMad/d-Myc feedback (Fig. 3 inset light green curve). This time is shorter when the feedback is present, an indication that in such case the stem cell would be able to avoid unnecessary differentiation due to random signalling fluctuations.

The presence of d-Myc as part of a positive feedback loop induces also interesting effects. As before, it can enable the system to exhibit bistability and hysteresis. In this way, d-Myc modifies the transition from the GSC to the CB state and determines the niche size, i.e., different dosage of d-Myc increases/decreases the number of stem cells in the niche, a phenomenon that we observed experimentally. How does this dosage affect the bistable behaviour?. A higher dosage, i.e., higher a_3 values, shifts the bifurcation diagram to lower Dpp values (lower a_1 values) which implies higher niche sizes but also with the effect of making the switch less robust (the Dpp range where the system exhibits

bistability is smaller). This dynamical instability might avoid the establishment of super-competitor in the niche (see below).

In addition, d-Myc induces cell competition and modifies the classical model of GSC differentiation. In the classic model all cells in the niche differentiate according to the same Dpp threshold value, and Dpp signalling (pMad) decreases following the Dpp gradient. The presence of cell competition modifies pMad levels in GSCs, increasing that of the "winners" and decreasing that of the "losers". As a consequence losers bring forward differentiation, i.e., their Dpp-threshold to differentiate is reached before that of the case without competition. In addition, cell competition decreases the chances of de-differentiation of these very same cells (de-differentiation occurs also at higher levels of Dpp) or even turns it impossible. A cell getting permanently less Dpp compared to its neighbours would express Bam, which at the same time blocks Dpp signalling and d-Myc expression, drastically cutting all possibilities to remain as a GSC. Therefore, d-Myc-induced competition sharpens and promotes the differentiation transition when compared to a hypothetical situation in which cells are just differentiating along a Dpp gradient.

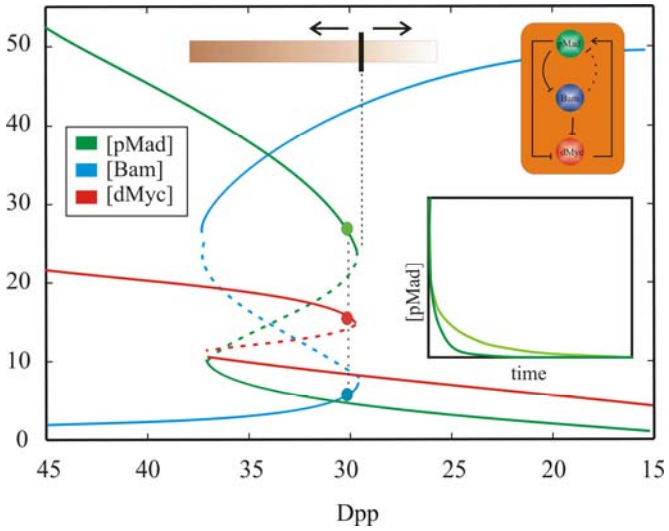


Figure 3. Bifurcation diagram of the cell competition control module. Steady states of the three molecular components, pMad (green), Bam (blue) and d-Myc (red) plotted as a function of Dpp signaling. The left side of the figure describes the situation at the tip of the germarium (Dpp levels denoted by the gradient bar). In this case, cells remain as stem cells with high levels of pMad/d-Myc and low levels of Bam. The right side of the figure describes a situation far from the tip of the germarium. Cells are no longer stem cells and are characterized by high levels of Bam. In between, there exists a parameter regime where three steady states of the dynamical system are available, one of them unstable (dashed lines). The differentiation/dedifferentiation transitions occur at different values of Dpp signaling, i.e., there exists hysteresis in the system. Insets. Top: Molecular control module between pMad (green circle), Bam (blue circle) and d-Myc (red circle). Bottom: Concentration of pMad as a function of time for two control modules with (without) the d-Myc/pMad negative feedback. The time necessary to recover the equilibrium value of pMad after a small stochastic concentration

fluctuation is shorter for the system with the feedback loop (dark green curve). See text for details. Parameters as follows: a_1 = Dpp, a_2 = 50, a_3 = 10, $b_1 = b_2 = b_3 = 1$, $n = m = s = p = 2$, $K_1 = K_1' = K_2 = K_3 = K_3' = 10$, $c_{13} = c_{31}$ = Dpp/4.

6. Conclusion

I discussed the behaviour of a molecular module regulating stem cell renewal in the *Drosophila* niche. I highlighted the role played in this system by d-Myc as a key component of a two-layer, intracellular and intercellular, control mechanism. How is the module architecture determining its function? To understand part of the associated complexity, I introduced basic concepts of systems biology, such as feedback-based control, bistability or histeresis. Many of these concepts have been already used in other cellular scenarios (Xiong, W. and Ferrell, J. E. Jr., 2003, Isaacs, F. J., *et al*, 2003 Bhalla U.S., *et al*, 2002, Acar M, *et al* , 2005) and they also seem essential if we want to fix and understand "stem-cell radios" (Lazebnik Y., 2002). Further work is of course necessary to fully unravel control strategies used in other stem cell niches and to determine the role played by cells within the niche to actively influence each other's fate. These studies can have major implications for understanding fundamental cellular processes and for the development of novel tissue replacement therapies.

Acknowledgements*. This work has been partially supported by the Spanish Ramón y Cajal Program. I thank Eduardo Moreno for many stimulating discussions.*

References

Acar M, Becskei A, and van Oudenaarden A. (2005) Enhancement of cellular memory by reducing stochastic transitions. *Nature* **435**, 228-232.

Bhalla U.S., Ram P.T., and Iyengar R. (2002) MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297**, 1018-1023.

Casanueva M.O. and Ferguson E.L. (2004) Germline stem cell number in the Drosophila ovary is regulated by redundant mechanisms that control Dpp signaling. *Development* **131**, 1881-90.

Chen D, McKearin D. (2003) Dpp signaling silences bam transcription directly to establish asymmetric divisions of germline stem cells. *Curr Biol* **13**, 1786-91.

de la Cova, C., Abril, M., Bellosta, P., Gallant, P. and Johnston, L.A. (2004). Drosophila Myc regulates organ size by inducing cell competition. *Cell* **117**, 107-116.

Díaz, B. and Moreno, E. (2005) The competitive nature of cells. *Exp Cell Res* **306**, 317-322.

Díaz, B., Fernández-Ruiz, I., Poyatos J.F. and Moreno E. (2006) Stem cell fate control through programmed cell competition. *Submitted*.

Ferrell JE Jr. (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr Op Chem Biol* **6**, 140-148.

Freeman, M. (2000). Feedback control of intercellular signaling in development. *Nature* **408**, 313-319.

Fuchs, E., Tumbar, T. and Guasch, G. (2004) Socializing with the neighbors: stem cells and their niche. *Cell* **116**, 769-778.

Isaacs, F. J., Hasty, J., Cantor, C. R. and Collins, J. J. (2003) Prediction and measurement of an autoregulatory genetic module. *Proc Natl Acad Sci USA* **100**, 7714−7719.

Kai, T. and Spradling, A. (2004) Differentiating germ cells can revert into functional stem cells in Drosophila melanogaster ovaries. *Nature* **428**, 564-569.

Lazebnik Y. (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. *Cancer Cell*. **2,** 179-182.

Morata, G. and Ripoll, P. (1975) Minutes: mutants of Drosophila autonomously affecting cell division rate. *Dev Biol* **42**, 211–221.

Moreno, E. and Basler, K. (2004) dMyc transforms cells into super-competitors. *Cell* **117**, 117-129.

Schofield, R. (1978) The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cell* **4**, 7-25.

Spradling, A., Drummond-Barbosa, D. and Kai, T. (2001) Stem cells find their niche. *Nature* **414**, 98-104.

Strogatz SH (2000) Nonlinear dynamics and chaos: With applications in physics, biology, chemistry and engineering. Perseus Publishing, Massachusetts, USA.

Thomas, R. and D'Ari, R. (1990) Biological Feedback. CRC Press, Florida USA.

Xiong, W. and Ferrell, J. E. Jr. (2003) A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature* **426**, 460−465.

A power-law model to describe the dynamics of the JAK2-STAT5 signalling pathway

J. Vera[a], J. Bachmann[b], A. C. Pfeifer[b], J.A. Hormiga[c], N.V. Torres Darias[c], U. Klingmüller[b], and O. Wolkenhauer[a].

[a]*Systems Biology and Bioinformatics Group, Department of Computer Science. University of Rostock. Rostock, Germany.*
[b]*Systems Biology of Signal Transduction. German Cancer Research Center (DKFZ) Heidelberg, Germany.*
[c]*Grupo Tecnología Bioquímica y Control Metabólico, Departamento de Bioquímica y Biología Molecular. La Laguna, Spain.*

Keywords: Power-law models, signal transduction, endocytosis.

1. Abstract

Signal transduction is one of the prominent and most promising fields in systems biology. The development of new quantitative experimental techniques allows us to accumulate high-quality quantitative data required for the estimation of numerical parameter values in dynamic pathway models. The present paper presents a power-law approach to modelling signal transduction pathways and applies this concept to the analysis of time course data set for the JAK2-STAT5 pathway. The power-law model offers an intuitive interpretation of biological observations. Our analysis of the experimental data and the model emphasize the role of dephosphorylation and internalisation of the receptor complex in the overall dynamic behaviour of the system.

2. Introduction

Signal transduction is one of the prominent and most promising fields in systems biology (Wolkenhauer et. al 2005). The development of new quantitative experimental techniques allows us to accumulate high-quality quantitative data required for the estimation of numerical parameter values in dynamic pathway models. The present paper extends previous work based on a kinetic model for the JAK5-STAT5 pathway (Swameye et al., 2003). The conserved JAK2-STAT5 pathway is one of the best-studied signalling networks. Its core module is particularly suitable for an investigation by mathematical modelling.
The cascade is activated through various receptors, including tyrosine kinases, G protein-coupled receptors, and hematopoietic cytokine receptors such as the erythropoietin receptor (EpoR). Signal transduction through EpoR is crucial for the formation of mature erythrocytes. Since cytokine receptors utilizing the JAK/STAT pathway lack intrinsic protein kinase activity, cytokine-activated

phosphorylation is mediated by the cytosolic kinase JAK2 which is associated with the cytoplasmic domain of the EpoR. The EpoR exists as a preformed dimer. Upon binding of the hormone erythropoietin (Epo) the receptor-associated JAK2 is activated and phosphorylates various tyrosine residues in the cytoplasmic domain of EpoR. Subsequently, the latent transcription factor STAT5 is recruited via its SH2 domain to the activated receptor, becomes phosphorylated by JAK2, homodimerizes and migrates to the nucleus where it initiates the transcription of various target genes. While the mechanism of EpoR activation is well understood, little is known about downregulation of the activated receptor. Recent studies have suggested endocytosis, proteasomal degradation, the recruitment of phosphatases like SHP-1 and other negative regulators like SOCS proteins as possible mechanisms to control cytokine responses.

The present paper extends previous work on a data-based mathematical model of the core module of the JAK2-STAT5 signalling cascade (Swameye et al., 2003). Since then methods for quantitative data acquisition have been optimized, to be more quantitative and with more time points. The merging of multiple quantitative data sets has been shown to be feasible (Schilling et al., 2005). The previous model consisted of four coupled differential equations for cytoplasmatic STAT5, phosphorylated monomeric STAT5, phosphorylated dimeric STAT5 and nuclear STAT5. The process of translocation of nuclear deactivated monomeric STAT5 to the cytoplasm was modelled with a discrete time-delay. The dynamics of the Epo receptor were not modelled but the data related to its activation were used to describe the input signal of the system during the experiment.

The model has been extended to consider receptor complex endocytosis and degradation. As an alternative to modelling signal transduction systems based on conventional chemical kinetics, we propose here a power-law model allowing for non-integer kinetic orders (Savageau 1969a, b, 1970).

3. Material and Methods

We have used quantitative immunoblotting techniques yielding high-quality data required for mathematical modelling (Schilling et al., 2005). We performed time-course experiments in BaF3 cells stably expressing the EpoR. The cells were transduced with the retroviral expression vector pMOWS-(HA)EpoR. BaF3-EpoR cells were selected in the presence of puromycin. For time-course experiments BaF3-EpoR cells were starved in RPMI-1640 containing 1 mg/ml BSA for 5 h and were stimulated with 5 units/ml Epo. For each time point, 10^7 cells were taken from the pool and lysed by adding the cells to Nonidet P-40 lysis buffer. Following cell lysis, EpoR and STAT5 were immunoprecipitated with anti-EpoR and anti-STAT5 (Santa Cruz Biotechnology). To achieve normalization of signals appropriate calibrator proteins (GST-EpoR and GST-STAT5) were added to lysates prior to immunoprecipitation (IP). IP samples were loaded on the gel (SDS-PAGE) in randomized order to prevent correlated errors evolving from the immunoblotting process. The immunoblots were

performed under standardized conditions, incubated with chemiluminescence substrate (GE healthcare) for 1 min, and exposed for 10 min on a LumiImager (Roche Diagnostics). For quantification, LUMIANALYST software (Roche Diagnostics) was used. Quantitative immunoblot data was processed using the GelInspector software described in Schilling et al. 2005.

Instead of a multiple shooting algorithm used in Swameye et al. 2003, the present paper uses a genetic algorithm for parameter estimation. The algorithm is a simple genetic algorithm that has been adapted and optimised to estimate parameters in power-law models (Hormiga et al 2006). In the estimation process, each element of the population of solutions represents a point in a parameter value space. The initial population is generated through a random exploration of the search space, which is defined using feasible intervals of values for the variables (Table 1). The best individuals of the population are selected in the considered iteration based on the value of the following objective function:

$$d = \sum_{i=1}^{m} \left(\frac{x_i - y_i}{\sigma_i} \right)^2$$

where m is the number of experimental data, x_i the value for the considered data point, y_i the value obtained after numerical integration of the solution, and σ_i the experimental error. An additional fast climbing-stochastic algorithm is applied to the best solutions in each iteration. The stopping criterion depends on a previously established maximum number of iterations or on a minimum level of satisfaction for the objective function. Finally, the solutions obtained are analysed and selected using additional biological criteria.

The JAK2-STAT5 pathway has been modelled as a general mass action (GMA) model, expanded using power-law terms:

$$\frac{dX_i}{dt} = \sum_j c_{ij} \cdot \gamma_j \cdot \prod_{k=1}^{n_d} X_j^{g_{jk}} \qquad i = 1 \ldots n_d$$

Here X_i are the n_d variables of the model, c_{ij} the coefficients of the stoichiometric matrix, γ_j the rate constants and g_{jk} the kinetic orders of the model. The meaning of the rate constants and stoichiometric coefficients is similar to the usual definition for conventional kinetic models. In the case of kinetic orders, the larger the contribution of one variable to the rate of change in signal is, the higher is the value of the associated kinetic order g_{jk}. Kinetic orders can have non-integer values, and it is possible to assign them negative real values representing allosteric inhibition.

Parameters to be estimated in the model will be rate constants, γ_j, (always real positive numbers) and kinetic orders, g_{jk}, (positive or negative real numbers). For the given system, there are no inhobitory feedback mechanisms and hence kinetic orders can have only positive real values. A conventional model, based on chemical kinetics, is a special case of a GMA model, where all kinetic orders are

equal to one, except those describing dimerisation, in which case the order is equal to two.

S-systems are the other possibility to model systems within the setting of Biochemical System Theory (Voit 2000). We have discarded the use of this simpler kind of power-law models in signal transduction pathways due to the adverse effects of flux aggregation, a constitutive property of these models. The effects of this operation can be neglected for other types of biochemical systems operating in a preferential steady-state, e.g. metabolic systems. Signal transduction networks have no such class of usual steady-states and can even operate with some processes switched off. In that case, the aggregation of rates can provoke significant alterations in essential properties of the model reducing the utility of this approach (Atauri et. al 1999) in signaling studies.

Since raw data are expressed in arbitrary units a convenient normalisation of the data was applied to avoid numeric problems during the parameter estimation and simulation but also to facilitate the visualisation of the data (Vera et. al 2003).

The remaining values of the rate constants and kinetic orders were estimated using a genetic algorithm for parameter estimation. Computations were performed on a Sun Fire V880 Server. Four processors UltraSPARC-III (1200 Mhz, 8MB cache each) were used, and the available RAM memory was 32 GB. The algorithm for parameters estimation was implemented in Matlab running under SunOS 5.10. The algorithm was executed with a population of two hundred initial parameter sets, and one hundred and forty generations per solution were computed. This means that twenty-eight thousand points were explored in the parameter space to obtain a solution. The computing time spent for getting a solution was around two minutes. Parameters were computed for mathematically feasible intervals in the parameter space (see Table 1).

Table 1. Feasible intervals of parameter values

Parameter	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Upper Bound	1.25	1.25	1.5	1.5	1.5	1.5	1.5	1.5	1.5
Lower Bound	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Parameter	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9
Upper Bound	0.15	0.15	0.01	0.15	0.01	0.01	0.15	0.15	0.15
Lower Bound	0.001	0.001	0.0001	0.005	0.0001	0.0001	0.005	0.005	0.005

4. Results and discussion

For all biochemical processes included in the model, the Epo receptor and the JAK2 kinase were assumed to form a stable complex. In the current model, two possible states for the EpoR/JAK2 complex were considered:

- Epo not bound to receptor-kinase complex.
- Epo bound to receptor-kinase complex, and therefore the receptor-kinase complex is activated.

When the non-activated complex binds the ligands, it is activated which subsequently leads to the phosphorylation of both parts of the complex. The dephosphorylation of the complex by phosphatases is also included in the model. The structure of the model was completed by including rates that describe the internalisation and subsequent degradation of the EpoR/JAK2 complex, either activated or non-activated. The recruitment of new EpoR/JAK2 complexes to the plasma membrane has also been considered. In the model three possible states have been considered for STAT5:

- Non-activated and monomeric STAT5 in the cytosol.
- Activated and dimerised STAT5 in the cytosol.
- Activated and dimerised STAT5 in the nucleus.

Since the dimerisation of activated STAT5 is considered a very fast process, the variable that quantifies the single-phosphorylated STAT5 is neglected. The model assumes that the protein is dimerised immediately after the activation process driven by pEpoR/pJAK2.

Experimental data describing the processes of translocation of 2(pSTAT5) to the nucleus and its dynamical behaviour inside the nucleus are currently not available, and the differential equations describing such processes have not been formulated in detail. Therefore, we model the fraction of STAT5 inside the nucleus with a single state.

Figure 1 illustrates the structure of the model developed. Only the states of the proteins and the processes that have been justified from a biological point of view were included.

The previous model (Swameye et al., 2003) focussed on the analysis of different states of STAT5. The structure of the current GMA model expands this model by incorporating the description of essential dynamical processes affecting the EpoR/JAK2 complex (recruitment, activation, deactivation and degradation). Since the dimerisation of activated STAT5 is a rapid biochemical process, the model was simplified by neglecting the dynamics of the variable representing monomeric phosphorylated STAT5. Moreover, the dynamics of depletion of the extracellular Epo were also modelled.

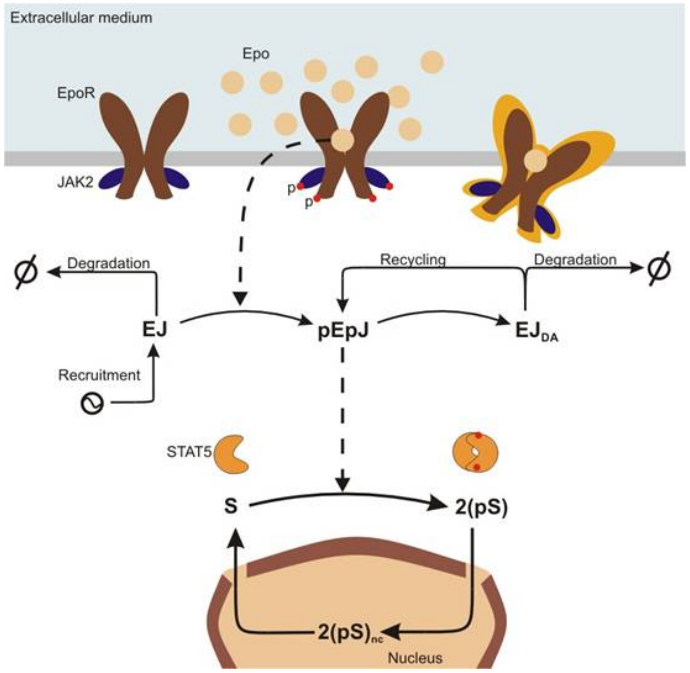


Figure 1. Structure of the JAK2/STAT5 pathway model.

The differential equations for the model were formulated using power-law terms:

$$\frac{dEpo}{dt} = -\gamma_1 \cdot EJ^{g_1} \cdot Epo^{g_2}$$

$$\frac{dEJ}{dt} = \gamma_3 - \gamma_2 \cdot EJ^{g_1} \cdot Epo^{g_2} - \gamma_3 \cdot EJ^{g_3}$$

$$\frac{dpEpJ}{dt} = \gamma_2 \cdot EJ^{g_1} \cdot Epo^{g_2} - \gamma_4 \cdot pEpJ^{g_3} + \gamma_5 \cdot EJ_{DA}^{g_5}$$

$$\frac{dEJ_{DA}}{dt} = \gamma_4 \cdot pEpJ^{g_3} - \gamma_5 \cdot EJ_{DA}^{g_5} - \gamma_6 \cdot EJ_{DA}^{g_6}$$

$$\frac{dS}{dt} = 2 \cdot \gamma_7 \cdot 2(pS)_{nc}^{g_7} - 2 \cdot \gamma_8 \cdot S^{g_8} \cdot pEpJ^{g_9}$$

$$\frac{d2(pS)}{dt} = 2 \cdot \gamma_8 \cdot S^{g_8} \cdot pEpJ^{g_9} - \gamma_9 \cdot 2(pS)^{g_{10}}$$

$$\frac{d2(pS)_{nc}}{dt} = \gamma_9 \cdot 2(pS)^{g_{10}} - \gamma_7 \cdot 2(pS)_{nc}^{g_7}$$

where *Epo* represents the concentration of erytropoietin in the extracellular medium, *EJ* is the fraction of EpoR/JAK2 complex non activated, *pEpJ* the activated fraction of the complex and EJ_{DA} represents the fraction of dephosphorylated complex receptor. Variable *S* represents the fraction of STAT5 non-activated and non-dimerised, *2(pS)* the fraction of STAT5 in the cytosol activated and dimerized, and finally $2(pS)_{nc}$ the fraction of STAT5 inside the nucleus that is activated and dimerized.

In all cases variables were normalised so that the initial total amount of EpoR/JAK2 complex equals one, and the total amount of STAT5 equals one during the experiment. With the exception of *Epo*, values for system variables were taken between zero (no proteins in the considered state) and one (total available amount of protein is this state).
Since there is no complete balance between degradation and recruitment of receptor complex in the equations of the model, the total amount of the EpoR/JAK2 complex can vary during the experiment. However, the total amount of STAT5 is constant, that is, there is an implicit balance between degradation and synthesis. The effect of the dimerisation process was also considered in the formulation of the equations. This leads to factors two in the equations.
The resulting model has seven dependent variables, and nineteen parameters to be estimated (ten kinetic orders and nine rate constants).
In case of the EpoR/JAK2 complex and 2(pSTAT5), data from two replicate experiments were available. Some outliers were detected by visual inspection and discarded. In case of Epo, the initial concentration was known and two more measurements were done at 30 and 180 minutes after stimulation.
Following normalisation, the average of time series is used for parameter estimation. The computed standard deviation was used as a measure of the error in the experiment. The final step was to appropriately scale the data. The current normalised and averaged data are actually not real quantitative data, because they do not relate to the proportion of protein in the considered state. Additional biological assumptions were used to establish the proportion of protein activated in the peaks of stimulation for both variables *pEpJ* and *2(pS)*. Once the proportion of protein in the peaks was known, the rest of data were adequately scaled to obtain the quantitative data shown in Figure 2. In case of Epo, the available normalised data were $Epo(0)=1.0,\ Epo(30)=0.89 \pm 0.024$ and $Epo(180)=0.83 \pm 0.017$.

Additional algebraic equations, reflecting the relation between the measured quantities and the variables, were defined in the model:

$$[\text{Epo}] = Epo$$
$$[\text{pEpoR}] = pEpJ$$
$$[\text{pSTAT5}_{\text{cyt}}] = 2\text{x}2(pS)$$

The variables on the left hand-side represent measured quantities, while the right-hand side represent the variables considered in the model.
The initial state of the system, after starvation and before stimulation, can be approximated by assuming that virtually the entire amount of the EpoR/JAk2 complex and STAT5 was in an inactivated state. This permits the assignment of feasible initial conditions to the variables. Both, experimental data and initial conditions were used to estimate parameter values.
Under the stated biological assumptions, in the peak of activation after stimulation 95% of the EpoR/JAK2 on the plasma membrane was considered activated and 60% of the STAT5 was supposed to be activated and dimerised.

Parameter estimation led to the values summarised in Table 2.

Table 2. Values of the parameters in the selected solution

Parameter	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
Value	0.82	0.75	1.28	1.14	1.02	1.28	0.99	1.02	1.0	1.02
Parameter	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	
Value	0.012	0.1	0.006	0.013	0.004	0.006	0.056	0.092	0.069	

Data fitting of the selected solution is shown in Figure 2. A good fit is obtained in the case of *pEpJ*. However, the model seems not to be able to obtain an activation of 95% of the available EpoR/JAK2 complex at maximum activation. Two possibilities could be considered: i) the model is not able to simulate such a strong activation of the receptor complex or ii) the quantification of the peak is excessive, and in the real system the maximum of activation is lower than the supposed value. In the case of *2(pS)*, the data fitting is not completely satisfactory.

Fig 2.A ***Fig 2.B***

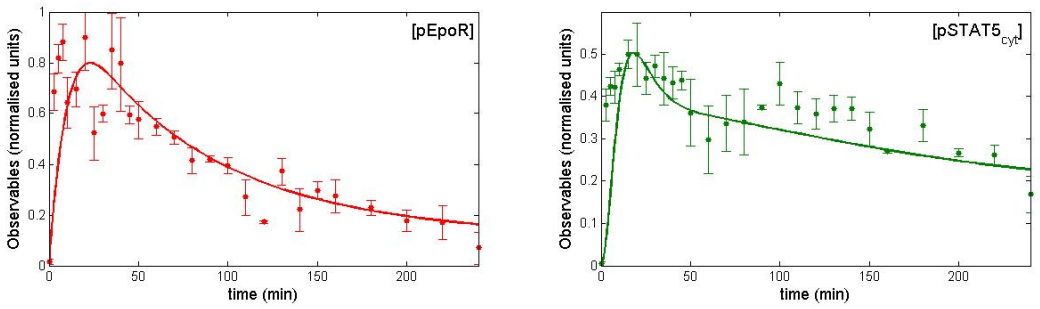


Figure 2. Data fitting of the selected solution. The quantitative data with error bars obtained from the experiments (points) are compared with the data fitting of the solution (lines). Data came from two replicates of the experiment performed using BaF3 cells, which expressed Epo receptor. The samples where stimulated using 5 units/ml of Epo during 240 minutes. Measurements were done using inmunoblotting. 2.A. Normalized quantity of pEpoR/pJAK2. 2.B. Normalized quantity of pSTAT5 in the cytosol.

Figure 3 shows the simulations of the system after the perturbation with Epo. The value of *Epo* decreases very fast until the system reaches the peak of activation. Since *EJ* is recruited very slowly from the endoplasmic reticulum, a strongly reduced quantity of *EJ* is available at the plasma membrane, after the peak of activation. Therefore, the depletion of *Epo* becomes slower because there is no significant amount of *EJ* on the plasma membrane able to bind *Epo*.
The majority of the complex which is initially at the plasma membrane is activated very quickly (approx. 80% after 16 minutes of stimulation). After that,

the concentration of *pEpJ* goes down very fast due to the effects of internalisation and subsequent degradation. The available non-activated EpoR/JAK2 complex, *EJ,* decreases very rapidly after stimulation with *Epo*, and remains low during time frame of the experiment; in that case, the rate of recruitment can not restore the initial concentration of *EJ* due to the effects of fast activation by *Epo*.

The amount of activated STAT5 in the cytosol, *2(pS),* increases very quickly after stimulation. In fact, it reaches the maximum 5 minutes before the maximum activation of *pEpJ*. After that, it maintains a high value during the simulation, even if the value of pEpJ decreases 80% with respect the maximum by the end of the experiment. For the fraction of activated STAT5 in the nucleus, $2(pS)_{NC}$, the increase in the signal after stimulation is delayed. The concentration of non-activated, non-dimerised STAT5 decreases very fast in the beginning of the experiment, but the end of the experiment recovers 40% of its initial value.

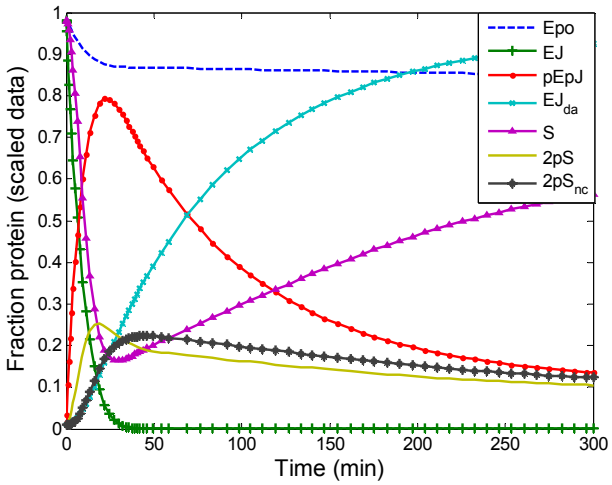


Figure 3. Simulation of the dynamics of system after perturbation including all the variables of the model.

5. Conclusions and Outlook

A mathematical model, based on ordinary differential equations and derived from power-law terms was obtained. The structure of the model was generated using available knowledge and parameters were estimated by fitting the experimental data. The analysis of the model indicates that it can reproduce the main dynamic features of the system adequately. Some interesting dynamic properties were detected after the analysis of the model, which are related to the essential role of low recruitment of receptor and rapid dephosphorylation and internalisation of the EpoR/JAK2 complex to describe the long-term deactivation of the system. These properties were not described with the previous model.

Problems of identifiability were detected for certain parameters in the model. These problems are apparent for some kinetic orders. This leads us to ask whether the inability to estimate these kinetic orders is due to a lack of enough

experimental data or to undesirable structural properties of the power-law models. Only a wider analysis of the problem, the strategy of parameter estimation and the structural properties of the power-law could help us to elucidate this question. We leave this question for further work. Towards this end, work on an extension of the model and further experiments to generate data are currently conduced.

Acknowledgements. *This work was supported by the European Commission 6th Framework program and as part of the COSBICS project under contract LSHG-CT-2004-512060, www.sbi.uni-rostock.de/cosbics.*

References

Alves, R. and Savageau, M.A. (2000a) Comparing systemic properties of ensembles of biological networks by graphical and statistical methods. *Bioinformatics,* **16**, 527-533.

Alves, R. and Savageau, M.A. (2000b) Systemic properties of ensembles of metabolic networks: application of graphical and statistical methods to simple unbranched pathways. *Bioinformatics,* **16**, 534-547.

Atauri, P., Curto, R., Puigjaner, J., Cornish-Bowden, A. and Cascante, M. (1999) Advantages and disadvantages of aggregating fluxes into synthetic and degradative fluxes when modelling metabolic pathways. *Eur. J. Biochem.,* **265**, 671-679.

Hormiga, J.A., Marin-Sanguino, A. and Torres, N.V. (2006) Parameter estimation in power-law models using a modified genetic algorithm (*submitted*).

Savageau, M.A. (1969a) Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reaction. *J. Theor. Biol*., **25**, 365-369.

Savageau, M.A. (1969b) Biochemical systems analysis II. Steady state solutions for an n-poll system using a power-law approximation. *J. Theor. Biol*., **25**, 370-379.

Savageau, M.A. (1970) Biochemical systems analysis: III. Dynamic solutions using a power-law approximation. *J. Theor. Biol*., **26**, 215-226.

Schilling, M., Maiwald, T., Bohl, S., Kollmann, M., Kreutz, C., Timmer, J., Klingmüller, U. (2005) Computational processing and error reduction strategies for standardized quantitative data in biological networks. *FEBS J.,* **272**, 6400-6411.

Swameye, I., Mueller, T.G., Timmer, J., Sandra, O., Klingmueller, U. (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling. *PNAS*, **100**, 1028-1033.

Vera, J., De Atauri, P., Cascante, M. and Torres, N.V. (2003) Multicriteria Optimization of biochemical systems by linear programming. Application

to the ethanol production by *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, **83**, 335-343.

Voit, E.O. (2000)Computational Analysis of Biochemical Systems. *A practical Guide for Biochemists and Molecular Biologists*. Cambridge University Press, Cambridge, UK.

Wolkenhauer, O., Sreenath, S., Wellstead, P., Ullah, M. and, Cho, K.H. (2005) A Systems- and Signal-Oriented Approach to IntraCellular Dynamics. *Biochem.l Soc. Trans*, **33**, 507-515.

Identification of enzyme targets in metabolic diseases by modelling and optimisation. The case of hyperuricemia in humans

Julio Vera-González[a,b], Néstor V. Torres[b,c], Raúl Curto[d] and Marta Cascante[e]

[a]*Systems Biology and Bioinformatics Group, Department of Computer Science, University of Rostock, Albert Einstein Str. 21, 18051 Rostock, Germany. julio.vera@informatik.uni-rostock.de*
[b]*Grupo Tecnología Bioquímica, Departamento de Bioquímica y Biología Molecular, Facultad de Biología, Universidad de La Laguna, 38206 La Laguna, Tenerife, Islas Canarias, Spain. ntorres@ull.es*
[c]*Instituto Canario de Investigación del Cáncer (ICIC), Islas Canarias, Spain*
[d]*University School of Experimental Sciences and Technology (EUCET), Universitat Internacional de Catalunya, Nova Estacio s/n, 43500 Tortosa, Spain.*
[e]*Department of Biochemistry and Molecular Biology, Faculty of Chemistry, and CERQT-Parc Cientific de Barcelona (PCB), University of Barcelona, Marti i Franques 1, 08028 Barcelona, Spain. martacascante@ub.edu*

Keywords: metabolic modeling, optimization, linear programming, S-systems, drug discovery, enzyme targets.

1. Abstract

The discovery of new drugs is a complex scientific task that requires huge amounts of time and money. A very promising approach to this process involves the integration of available biomedical data through mathematical modeling and data mining. We have developed a method known as the optimization drug discovery program (ODDP) that allows new enzyme targets to be identified in metabolic diseases through the integration of mathematical metabolic models and biomedical data in a mathematical optimization program.
The ODDP was used to detect target enzymes in human hyperuricemia. An existing S-system mathematical model and bibliographic information about the disease were used. The method detected six single-target enzyme solutions involving dietary modification: inhibition of adenine phosphoribosyltransferase, AMP deaminase, RNases to AMP and GMP, 5'(3')-nucleotidase, guanine hydrolase and xanthine oxidase. The last detected solution coincides with conventional treatment using allopurinol.

Understanding and Exploiting Systems Biology

2. Introduction

The predictive biosimulation has been suggested to be a promising area of computational biology and is considered to be among the leading technologies for the future of drug discovery. The mathematical models integrate available metabolic, genomic and proteomic information to describe and predict the behaviour of the simulated system. In recent years, several groups have begun to use mathematical modelling of metabolic systems to study biomedical systems (see Curto et al. 1998a,b; Guebel and Torres 2004; Voit 2002).
In classical hyperuricemia, a functional defect in enzyme that controls the synthesis of *de novo* purines increases its activity and leads to an increase in the activity of degradative metabolic fluxes yielding uric acid. The symptoms of the disease include acute episodes of arthritic pain and inflammation and various kinds of nephropathy. Currently, treatment of this disease usually includes a symptomatic treatment for joint pain, a restricted diet that precludes consumption of food with high concentrations of purine precursors, and the prescription of allopurinol. Allopurinol is a specific inhibitor of the enzyme xanthine oxidase that can lead to a drastic reduction in the concentrations of uric acid and urate (Klinenberg 1965).
S-system models were initially developed in the early 1970s by M.A. Savageau to study dynamic and steady-state properties of simple metabolic pathways (Savageau 1969a, 1969b, 1970). They are essentially mathematical models of metabolic networks expanded as ordinary differential equations with a net (aggregated) input flux and a net (aggregated) output flux, both developed using power-law terms. As a consequence of the structural simplicity of S-system models, biological optimization problems in biotechnology and biomedicine can be explored using linear programming (Torres et al. 1998, Torres and Voit 2002).

3. Material and Methods

The aim of the present work was to develop a mathematical approach to the rational identification of active principles for the treatment of metabolic diseases. The approach is based on dynamic mathematical modelling of the metabolic network responsible for the disease and the use of standard optimization routines.
The optimization drug discovery program (ODDP) can be applied to those diseases that meet three general criteria. First, the metabolic disease should be caused by a structural or functional alteration that causes a total or partial loss of catalytic activity in one or more enzymes. Second, the enzyme deficiency should cause physiologically relevant changes in the concentration of one or more metabolic intermediates and/or fluxes. Finally, the changes should be responsible for the symptoms of the disease, either directly or by affecting other related metabolic processes. Thus, the framework for the analysis is the metabolic network involved in the disease rather than a single biochemical process.
The usual steady state of a metabolic network in a typical healthy individual will be referred as the healthy state (HS). This HS is a metabolic configuration in

which metabolite concentrations and fluxes have values that do not lead to metabolic diseases. In contrast, the pathologic state (PS) corresponds to that of an individual suffering from the disease. In this state, the system contains one or more enzymes with significantly altered activity that provokes changes in certain fluxes and concentrations (called *key metabolites and fluxes*) that ultimately lead to the manifestation of the symptoms.
We can alter a metabolic network by using drugs to modulate enzyme activities. In this case, the aim is to alter the values of critical metabolite concentrations and fluxes in the network and shift them towards those encountered in the HS. Eventually we can also change the initial substrate concentrations, e. g. by dietary restriction. Once the metabolic network and variables associated with the PS are known, the ODDP aims to identify modifications of enzyme activities and substrate concentrations that shift the system to a steady state as close as possible to that of the HS, while also verifying all the additionally imposed physiological and biological.

3.1. Mathematical structure of the ODDP method. The ideas proposed above can be translated into a mathematical framework for the identification of new target enzymes for therapeutic treatment of metabolic diseases.
The starting point for any ODDP application is a reliable mathematical model based on ordinary differential equations describing the essentials of the biochemical system. Given that the aim is to shift the PS towards the HS, we look for solutions in which the values of metabolite concentrations and fluxes that play a key role in the manifestation of the disease will be shifted towards those seen in the HS. This aim can be mathematically represented in the following objective function:

$$Min\left[\sum_{j=1}^{l}\lambda_j\left|X_j - X_j^{HS}\right| + \sum_{i=1}^{p}\lambda_i\left|J_i - J_i^{HS}\right|\right]$$

where $\lambda_j, \lambda_i \geq 0$, and X_j and J_i are the key metabolites and fluxes, respectively. The values of λ_j and λ_i are determined by the relative importance of each key metabolite and flux in relation to the symptoms. We model the functional origin of the disease by imposing a characteristic value to the enzyme activities that underlie the PS profile:

$$X_k = X_k^{PS}$$

In this equation, X_k represents the activity value for the deficient enzyme and X_k^{PS} is its characteristic value in the PS. The solution near to the HS will be a stable steady state of the considered metabolic network, which is reflected by the following set of equations:

$$\frac{dX_i}{dt} = 0\,, \quad i = 1,\ldots,n$$

Moreover, we have to assume additional conditions to guarantee physiologically acceptable solutions. These conditions can be modelled with the following set of equations:

$$X_i^{LB} \le X_i \le X_i^{UB} \qquad i = 1...n_D \text{ , } n_D \text{ number of metabolites in the network}$$

where X_i^{LB} and X_i^{UB} establish the physiologically admissible lower and upper bounds for each metabolite and modifiable enzyme activity.

3.2. General structure of the ODDP method. The outlined approach integrates modelling of the system and identification of target enzymes with the analysis of the available information and the choice of the best solutions. The method includes four sequential steps:

Step 1. Information gathering and construction of the disease model. Essential biological information about the system and the disease is collected in the first step. A mathematical model based on differential equations that describe the metabolic system has to be selected. A characterisation of the metabolic basis of the disease is also considered, including essential information about the genetic and functional origin of the disease and the features defining its associated PS. In addition, analysis of the critical interactions between the specific network and general metabolism is important to avoid undesirable interactions.

Step 2. Optimization. This step involves translation of the following information into mathematical terms: i) definition of the pathologic state of the system, assigning an appropriate value for the activity of the deficient enzyme; ii) choice of the set of *key metabolites and fluxes* that deviate from the values seen in healthy individuals, leading to the manifestation of symptoms; *iii)* selection of the set of target enzymes that can be pharmacologically modified; *iv)* establishment of a plausible physiologically admissible interval value for each metabolite and modifiable enzyme; and *v)* definition and implementation of the optimization objective that integrates the available information about the *key metabolites and fluxes*.

Step 3. Computation of solutions. The principle of minimum pharmacological effort. Each time the defined optimization program is carried out a feasible solution is obtained. Each solution consists of a set of predicted values for metabolites, initial substrates and enzyme activities that describes a biologically acceptable steady state of the system that shifts the PS towards the HS.

Step 4. Selection and classification of solutions. Additional selection criteria are necessary in order to select the most feasible solutions from within the generated set. A minimum acceptable value for the objective function can be established to allow selection of the solutions that satisfy this condition. In addition, other biomedical criteria can be mathematically imposed. In this way, a reduced,

biomedically acceptable set of good candidates is obtained for experimental selection using conventional protocols in pharmacological research.
The information obtained allows us to determine which enzymes or set of enzymes in the network are candidates for inhibition and how much we have to alter their activities in order to get the best therapeutic results. The computational core of the ODDP is contained in Step 3. The computational requirements of this step will depend on the characteristics of the selected mathematical model. The computational efficiency of the method is maximized through the use of S-systems models. The sequentially generated optimization programs become linear in the logarithmic space of the variables and can then be solved using the Simplex algorithm (Torres at al 1997).

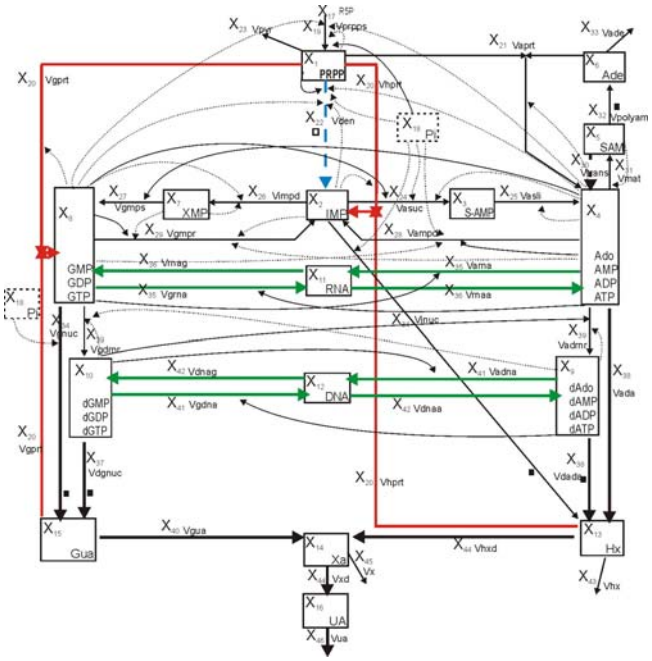


Figure 1. Scheme of the considered model of purines metabolism in humans. Legend: Time dependent metabolites [S-System variable. Name]: X_1. Phosphoribosilpyrophosphate; X_2. Inosine monophosphate; X_3. Adenylsuccinate; X_4. Pool of adenosine derivatives; X_5. S-adenosyl-Lmethionine; X_6. Adenine; X_7. Xanthosine monophosphate; X_8. Pool of guanosine derivatives; X_9. Pool of deoxyadenosine derivatives; X_{10}. Pool of phosphated deoxyguanosine derivatives; X_{11}. Ribonucleic acid; X_{12}. Deoxyribonucleic acid; X_{13}. Pool of inosine derivatives; X_{14}. Xantine; X_{15}. Pool of guanosine derivatives; X_{16}. Uric acid. Non-time dependent metabolites [S-System variable. Name]: X_{17}. Ribose-5-phosphate; X_{18}. Inorganic phosphate. Enzyme Activities [S-System variable (abbreviation)]: X_{19}(Vprpps); X_{20}(Vgprt-hprt); X_{21}(Vaprt); X_{22}(Vden); X_{23}(Vpyr); X_{24}(Vasuc); X_{25}(Vasli); X_{26}Vimpd); X_{27}(Vgmps); X_{28}(Vampd); X_{29}(Vgmpr); X_{30}(Vtrans); X_{31}(Vmat); X_{32}(Vpoliam); X_{33}(Dade); X_{34}(Vinuc-gnuc); X_{35}(Vgrna-arna); X_{36}(Vrnag-rnaa); X_{37}(Vdgnuc); X_{38}(Vada-dada); X_{39}(Vgdrnr-adrnr); X_{40}(Vgua); X_{41}(Dgdna-adna); X_{42}(Vdnag-dnaa); X_{43}(Vhx); X_{44}(Vxd-hxd); X_{45}(Vx); X_{46}(Vua). The complete name and E.C. number of the previously listed enzymes is available in (Curto 1998 and Voit 2000).

4. Results and Discussion

In the following section we apply the ODDP method was applied to the detection of single or combined target enzymes for the treatment of classical hyperuricemia in humans.

Step 1. Modelling purine metabolism in humans. Purine metabolism in humans represents a complex metabolic network that includes the synthesis, recovery and degradation of purine nucleotides. In the present work we have used a modified version of an S-system metabolic model initially developed by the Cascante group (Curto et al. 1997, 1998a,b, Voit 2000, see Figure 1).

Step 2. Optimization program. Using the information available on the disease we can configure the proposed optimization program for drug discovery. The healthy state – HS - of the system can be defined by assigning a value equal to 1 to all of the variables in the model (metabolite concentrations and enzyme activities). In mathematical terms this translates to:

$$X_i^{HS} = 1 \qquad i = 1,...,46$$

Mathematical characterization of the pathologic state. To model the PS, we doubled the value of the variable that represents the activity of PRPPS (X_{19}) with respect to the HS:

$$X_{19}^{PS} = 2 \cdot X_{19}^{HS}$$

In this way, we can model overactivity of the enzyme. The concentration of uric acid (X_{16}) was considered the only *key metabolite* in the ODDP strategy. Accordingly, the *optimization objective* for the ODDP program will be:

$$Min \qquad \left| X_{16} - X_{16}^{HS} \right|$$

Selection of target enzymes. Establishment of plausible, physiologically admissible metabolite and enzyme intervals. In a first approach to the optimization program, we will consider all of the enzymes (except X_{19}) as optimization targets. Thus, the set of *target enzymes* (S^{TE}) is

$$S^{TR} = \{X_{20}, X_{21}, \ldots, X_{46}\}$$

On the basis of previous studies (Torres and Voit 2002, Vera et al. 2003a, Vera et al. 2004), the lower bound is established as half of the HS value for all metabolite concentrations, and the upper bound as up to 10 times the HS value:

$$0.5 \cdot X_i^{HS} \le X_i \le 10 \cdot X_i^{HS} \qquad i = 1,...,16$$

In the model, ribose-5-phosphate (X_{17}) is considered as independent variables representing the network substrates. X_{17} can be partially regulated by controlling the diet. This strategy is called *prescription of diet* and is usually used by physicians to treat gout. Therefore, we model such strategy allowing an interval of feasible values for X_{17} defined at around 50% of the HS value $\left(0.5 \cdot X_{17}^{HS} \le X_{17} \le 1.5 \cdot X_{17}^{HS}\right)$. In terms of the admissible interval of values for the enzyme activities, we only modelled treatments in which drugs act as inhibitors of the target enzymes. We chose a suitable interval for each enzyme activity ranging from 10% of the HS value to the true HS value:

$$0.1 \cdot X_i^{HS} \le X_i \le 1 \cdot X_i^{HS} \qquad i \in TE = \{20,...,46\}$$

Steady-state equations. We guaranteed that the computed solution represents a steady state for the system by introducing the following set of equations:

$$\frac{dX_i}{dt} = 0 \qquad i = 1,...,16$$

Step 3. Computation of solutions. The principle of minimum pharmacological effort. We built and computed one optimization program for each possibility that included the following additional set of equations:

$$0.1 \cdot X_i^{HS} \le X_i \le X_i^{HS}\ i \in \{20,...,46\} \quad X_k = X_k^{HS}\ \forall k \in TE = \{20,...,46\} \wedge k \ne i$$

where we allowed the activity of the target enzyme to vary while the others were fixed at the HS value. The computation of programs was carried out using the Matlab toolbox MetMAP (Vera and Torres 2003). The calculations were computed using a Pentium IV 1.6 GHz processor with 512 Mb RAM. The computing time was in the range of milliseconds per optimization program.

Step 4. Choice of the most biomedically appropriate solutions. Additional biomedical criteria were defined to select the most viable and accurate solutions (Vera et al 2003a, and Vera et al 2003b). Firstly, we imposed an upper limit on the concentration of uric acid (X_{16}) equal to 105% of the HS value:

$$X_{16} \le 1.05 \cdot X_{16}^{HS}$$

The solutions that verify this equation were called *satisfactory solutions*. A second criterion was defined using the Cartesian distance in the space of the metabolite concentrations from the considered solution (X) to the HS (X^{HS}):

$$D\left(X, X^{HS}\right) = \sqrt{\sum_i^{16} \left(X_i - 1\right)^2}$$

We called this the metabolic distance $\left(D\left(X, X^{HS}\right)\right)$, which measures the deviation of the metabolite concentrations from the HS values in the considered solution. Solutions with a high value of $D\left(X, X^{HS}\right)$ would cause metabolic disequilibrium and undesirable physiological side effects. Finally, solutions with the highest

values for the activity of the modified target enzymes ($X_i val$) were chosen. We took into account the fact that higher enzyme activities imply the use of a lower dose of the specific drug, and reduced drug doses minimise adverse side effects.
Six *satisfactory solutions* were obtained: inhibition of adenine phosphoribosyltransferase (V_{den} X_{22}), AMP deaminase (V_{ampd} X_{28}), RNases to AMP and GMP (V_{rgna} V_{rnaa} X_{36}), 5'(3')-Nucleotidase (V_{dgnuc} X_{37}), guanine hydrolase (V_{gua} X_{40}), and xanthine oxidase (V_{xd} V_{hxd} X_{44}). Further details on the selected solutions can be found in Table 1. Examination of the table also shows that all the solutions have significant values of $D(X, X^{HS})$ and require substantial inhibition of the targeted enzyme (low values in $X_i val$).

Table 1. Selected solutions with prescription of diet.

X_i	X_i val	D(X, X^n)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
22	0.22	1.56	2.15	1.58	1.61	1.34	1.06	0.88	1.09	1.48	1.07	1.06	1.07	1.05	0.99	1	1.01	**1**	1.14
28	0.33	1.86	1.03	0.84	1.51	1.99	1.15	2.41	0.65	0.97	1.03	0.97	1.03	1.0	1.14	1	0.83	**1**	0.6
36	0.36	2.07	1.55	1.38	1.15	0.94	0.99	0.69	1.2	1.27	1.03	1.05	2.89	1.03	0.95	1	1.07	**1**	0.67
37	0.18	1.63	1.3	1.20	1.07	0.95	0.99	0.79	1.12	1.16	1.69	2.25	1.02	1.63	0.98	1	1.03	**1**	0.51
40	0.28	4.81	1.11	1.32	1.37	1.22	1.04	1.2	1.03	1.37	1.05	1.05	1.05	1.04	1.76	1	5.7	**1**	0.54
44	0.33	7.73	1.07	1.7	1.49	1.27	1.05	1.3	1.2	1.09	1.02	1.01	1.02	1.01	5.38	7.29	1.24	1	0.5

In Figure 2, this group of solutions is shown in the space of $D(X, X^{HS})$ and $X_i$val. The figure also shows the so-called *utopian solution*, the potential solution which has the lowest computed value of $D(X, X^{HS})$ and the highest value of $X_i val$. The solution inhibiting X_{22} has the lowest metabolic distance ($D(X, X^{HS}) = 1.557$) but the highest enzyme activity value corresponds to X_{36} ($X_i val$=0.356). Examination of Figure 2 shows that the best treatments with prescription of diet are through inhibition of X_{36} or X_{28} (the closest solutions to the *utopian point*).

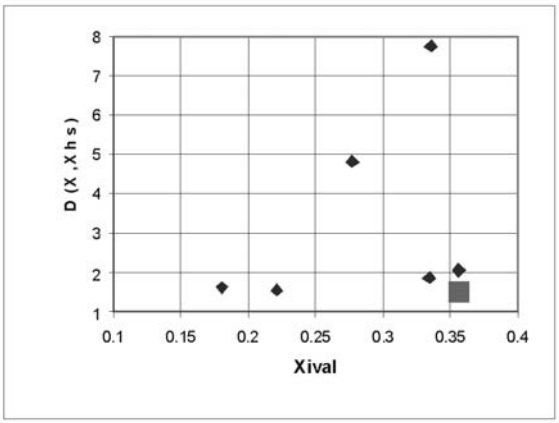


Figure 2. Solutions with a single target enzyme. The large square represents the *utopian point*, while diamonds indicate the selected satisfactory solutions.

Another important observation relates to the solution involving inhibition of X_{44}. This solution coincides with the most common current clinical treatment for hyperuricemia, which is based on the use of a combination of allopurinol and a diet low in purine precursors (see Table 1). Moreover, our method predicts the well-known metabolic side effects of treatment with allopurinol, namely a strong increase in the concentration of xanthine and hypoxanthine (X_{13} and X_{14}). This represents an *a posteriori* pragmatic verification of our model predictions. In fact, the ODDP not only found the current clinical treatment of the disease but also several other possible target enzymes with potentially lower side effects.

5. Conclusions

In this study, we analysed possible treatments involving inhibition of one enzyme in the network without prescription of diet. the ODDP lead us to six solutions based on the inhibition of one enzyme and with dietary restriction. Analysis of these solutions in relation to the values of $D(X, X^{HS})$ and $X_i$val facilitated further classification. The best solutions involved inhibition of X_{22} (without dietary restriction) and X_{28} or X_{36} (with dietary restriction). This finding establishes a rational basis for experimental assays to verify the clinical potential of the proposed solutions. The method predicted a solution that coincided with the current clinical treatment gout (inhibition of X_{44}) and the well-known adverse side effects associated with this therapy.

The ODDP method can be used with any kind of metabolic model based on ordinary differential equations. As we have shown, the computation time associated with the solution of the total amount of optimization programs included in the ODDP when one uses S-system models with an intermediate complexity (<100 variables) and simple treatments (1-2 simultaneous target enzymes) is in the order of 10^1 seconds. In that case, the total computational time is not critic and efforts can be focused on refinement of the optimization program and the biological hypothesis formulated about the disease. Such improvements are not possible in other cases in which each round of optimization would take between 10^4-10^5 seconds.

When we apply the ideas included in the indirect optimization method (IOM; Torres et al 1997; Torres and Voit 2002), the recasting of any kind of model as an S-system allows the described method to be used with its consequent computational advantages.

Acknowledgements. *This work was supported by research grants from the Spanish Ministry of Science and Education (ref. nº BIO2005-08898-C02-02 and SAF2005-01627), by a research grant from the Government of the Canary Islands (ref. nº PI2000-071) and by Fundación la Caixa (ONO3-70-0). Julio Vera was the recipient of a research grant from the Spanish Ministry of Science and Education (ref. PN99-4320298). The authors are grateful to Dr. Julio R. Banga (Spanish Research Council in Vigo, CSIC-Vigo, Spain) for helpful comments.*

References

Curto R., Voit E.O., Sorribas A. and Cascante, M. (1997) Validation and steady-state analysis of a power-law model of purine metabolism in man. *Biochem. J.*, **324**, 761–775.

Curto R., Voit E.O. and Cascante M. (1998a) Analysis of abnormalities in purine metabolism leading to gout and to neurological dysfunction in man. *Biochem. J.,* **320**, 477-487.

Curto R., Voit E.O., Sorribas A. and Cascante M. (1998b) Mathematical models of purine metabolism in man. *Math. Biosci.,* **151**, 1-49.

Guebel D. V. y Torres N. V. (2004) Dynamic of sulphur amino acids in mammalian brain: assesment of the astrocytic-neuronal cysteine interaction by a mathematical model. *Biochim. Biophys. Acta*, **1674**, 12-28.

Klinenberg J.R., Goldfinger S.E. and Seegmiller J.E. (1965). The effectiveness of the xantine oxidase inhibitor allopurinol in the treatment of gout. *Ann Intern Med*., *62*, 639-47.

Savageau, M.A. (1969a) Biochemical systems analysis: I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol*., *25*, 365-369.

Savageau, M.A. (1969b) Biochemical systems analysis II. Steady state solutions for an n- poll system using a power-law approximation. *J. Theor. Biol.*, **25**, 370-379.

Savageau, M.A. (1970) Biochemical systems analysis: III. Dynamic solutions using a power-law approximation. *J. Theor. Biol*., **26**, 215-226.

Torres, N.V., Rodríguez, F., González-Alcón, C. and Voit, E.O. (1997) An Indirect Optimization Method for Biochemical Systems: Description of the Method and Application to Ethanol, Glycerol and Carbohydrates Production in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng*., **55**, 758-772.

Torres N. V. y Voit E. (2002) *Pathways Analysis and Optimization in Metabolic Engineering*. Cambridge University Press, CA.

Vera, J., De Atauri, P., Cascante, M. y Torres, N.V. (2003) Multicriteria Optimization Of Biochemical Systems By Linear Programming. Application To The Ethanol Production By *Saccharomyces Cerevisiae*. *Biotechnol. Bioeng.,* **83**, 335-343.

Vera J., González-Alcón C.M. Torres N.V. (2004) *Recent Research Developments in Applied Microbiology & Biotechnology* Vol. 1. Chapter 5: Multiobjective optimization of biochemical systems. ISBN: 81-271-0043-9. SG Pandalay Editions. Trivandrum, India.

Voit, E. O. (2000) *Computational Analysis of Biochemical Systems.* Cambridge University Press, CA.

Voit, E.O. (2002) Metabolic modelling: a tool of drug discovery in the post-genomic era. *Drug Discovery Today*, **7**, 621-628.

SYSTEMS BIOLOGY APPLICATIONS: BIOPROCESSES

Key enzymes expression and their relationship with energetic coenzyme pools after perturbations in the production of L-carnitine by *Escherichia coli*

M. Cánovas, A. Sevilla, V. Bernal and J.L. Iborra∗.

Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain. E-mail: jliborra@um.es

Keywords: L-carnitine metabolism, optimization, *Escherichia coli*, metabolic perturbation.

1. Abstract

The aim of this work was to understand the steps controlling the biotransformation of trimethylamonium compounds into L(-)-carnitine by *Escherichia coli*. In a high-cell density reactor, steady state levels of carbon source (glycerol) and acetate (fermentation product) were pulsed by increasing five-fold. Following the pulse, the evolution of the enzyme activities involved in the biotransformation process, in the synthesis of acetyl-CoA (ACS: acetyl-CoA synthetase and PTA: ATP: phosphotransferase) and in the distribution of metabolites for the tricarboxylic acids (ICDH: isocitrate dehydrogenase) and glyoxylate (ICL: isocitrate lyase) cycles was monitored. In addition, the levels of carnitine, the cell ATP content and the NADH/NAD^+ ratio were measured in order to assess the importance and participation of these energetic coenzymes in the catabolic system. The results obtained for the NADH/NAD^+ pool indicated that it is correlated with the biotransformation process at the NAD^+ regeneration and ATP production level in anaerobiosis. More importantly, a linear correlation between the NADH/NAD^+ ratio and the levels of the ICDH and ICL (carbon and electron flows) and the PTA and ACS (acetate and ATP production and acetyl-CoA synthesis) activity levels was assessed. The main metabolic pathway operating during cell metabolic perturbation with a pulse of glycerol and acetate in the high-cell density membrane reactor was that related to ICDH and ICL, both of which regulated the carbon metabolism, while PTA and ACS enzymes regulated ATP production. Although varying flux was predominantly caused by changes in the levels of reaction substrate, products and/or allosteric effectors, the pulses showed that repression and derepression/induction of genes or the action of allosteric effectors in driving the system to the steady state. To gain further understanding, we believe that the combination of metabolome analysis with cell-cycle-regulated measurements of enzyme activities, protein levels and gene expression is the correct way.

∗ Corresponding author

Understanding and Exploiting Systems Biology

2. Introduction

In human cells, L-carnitine (R(-)-3-hydroxy-4-trimethylaminobutyrate) transports long-chain fatty acids through the inner mitochondrial membrane, which is why several clinical applications for this compound have been identified. Consequently, the demand for L-carnitine has increased worldwide (Seim et al., 2001) and chemical and biological processes have been developed for its production (Cavazza, 1981; Kulla, 1991; Hoeks et al., 1996; Kleber, 1997). Strains belonging to the genera *Escherichia*, *Proteus* and *Salmonella* racemice D-carnitine, a waste product and an environmental problem resulting from the L-carnitine chemical synthesis, and/or biotransform crotonobetaine (dehydrated D-carnitine) to produce L-carnitine (Kleber, 1997; Castellar et al., 1998; Obon et al., 1999; Cánovas et al., 2002).

In *E. coli*, the genes responsible of L-carnitine metabolism are coded by the *caiTABCDE* and *fixABCX* operons. Both operons are positively modulated by general regulators, such as the cAMP receptor protein (CRP) or the transcriptional regulator responsible for anaerobic induction (FNR), and negatively by the DNA-binding protein H-NS, glucose or nitrate (Unden and Trageser, 1991; Eichler et al., 1994). In addition, it has been proposed that a positively controlled *caiF* gene, found in the 3' adjacent region to the *cai* operon, acts as a specific transcriptional regulator for carnitine metabolism (Eichler et al., 1996). This pathway is detectable not only in cells previously grown anaerobically but also in some species, such as *E. coli* ATCC 25922 and DSM 8828, *P. vulgaris* and *P. mirabilis*, grown under aerobiosis in the presence of inducers such as D,L-carnitine mixture or crotonobetaine (Kleber, 1997; Obon et al., 1999; Elssner et al., 2000; Cánovas et al., 2002). It was first postulated that L-carnitine dehydratase reversibly catalyzed L-carnitine into crotonobetaine and that crotonobetaine reductase non-reversibly transformed crotonobetaine into γ-butyrobetaine as an electron sink (Jung et al., 1989; Roth et al., 1994; Kleber, 1997), even though this latter can be inhibited by fumarate addition as an alternative electron sink (Obon et al., 1999). Now that functions have been assigned to each putative protein of the *cai* operon, it is known that CaiT is an exchanger (antiporter) for carnitine derivates in *E. coli* (Jung et al., 2002) with no energy consume. Further, the irreversible ATP consuming ProU transporter is also present (Verheul et al., 1998; Cánovas et al., 2003a). The enoyl-CoA hydratase (CaiD) requires a CoA-transferase activity (CaiB), since the hydration reaction of crotonobetaine to L-carnitine (CHR) proceeds in two steps at the CoA-level in two steps: the CaiD-catalyzed hydration of crotonobetainyl-CoA to L-carnitinyl-CoA, followed by CoA-transfer from L-carnitinyl-CoA to crotonobetaine, catalyzed by CaiB (Elssner et al., 2001). Thus, the reversible biotransformation of crotonobetaine to L-carnitine requires the presence of a co-substrate, either γ-butyrobetainyl-CoA or crotonobetainyl-CoA (Elssner et al., 2000). CaiD was also postulated to be involved in racemisation of D-carnitine (Eichler et al., 1994). Further, CaiC has been suggested as a CoA-trimethylammonium ligase (Eichler et al., 1994), activating trimethylammonium

compounds upon reaching the cell. The function of protein CaiE is not totally understood and further studies must be undertaken. With all this information, a model to describe the whole activity of *E. coli* able to produce L-carnitine from crotonobetaine under anaerobic conditions has been proposed (Figure 1).

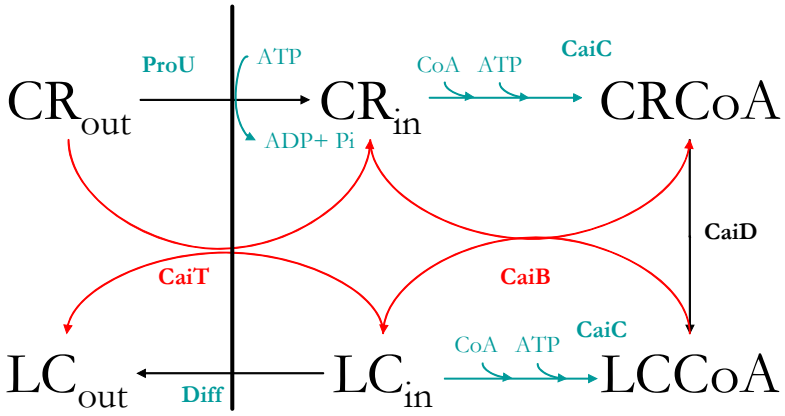


Figure 1. Metabolic pathways involved in the biotransformation of crotonobetaine into L-carnitine in *E. coli*. extracellular crotonobetaine (CR_{out}); extracellular L-carnitine (LC_{out}); intracellular L-carnitine (LC_{in}); intracellular crotonobetaine (CB_{in}); L-carnitinylCoA (LCCoA); crotonobetainylCoA (CRCoA); Acetyl-CoA/HS-CoA transferase (CaiB), Crotonobetaine, L(-)-carnitine or γ-butyrobetaine: acetyl-CoA/HS-CoA ligase (CaiC), Enoyl-CoA hydratase activity (CaiD), L(-)-carnitine protein transporter (CaiT).

Rational optimization of this biotransformation in continuous high-cell density membrane reactors first requires understanding the link between cell carnitine metabolism and the central metabolism in *E. coli.* With this aim, the first approach to link the central carbon or primary metabolism and the metabolism of the secondary trimethylammonium compounds involved in the production of L-carnitine by *E. coli* was performed. The stationary modelling of the whole E. coli central metabolism, including the carnitine metabolism in the growing and resting cell states would allow the design of novel optimization strategies.

3. Theoretical

3.1. MATHEMATICAL MODELLING

Stationary State Approximation.

The carnitine model developed by Cánovas et al. (2003a) was linked to the central large scale stationary model developed by Chassagnole et al. (2002) adapted to represent the central metabolism of *E. coli* under anaerobic conditions and using glycerol as the carbon source (Fig. 2). For the transport of the substrate and the product two systems were considered, the ATP-dependent ProU, which is able to get either crotonobetaine or L-carnitine into the cell in an symport manner

and the non-ATP-dependent CaiT, which operates as an antiport of crotonobetaine/L-carnitine allowing the production of L-carnitine from crotonobetaine without the need of energy (Fig. 1). The first transporter is fully operating under osmotic stress (Cánovas et al., 2003b) together with ProP, another trimethylammonium/H+ antiporter, also energy dependent. The CaiT transporter is fully operating when the carnitine metabolism has been induced in the presence of either L-carnitine or crotonobetaine. The reactions that were taken into account are presented in Appendix 1.

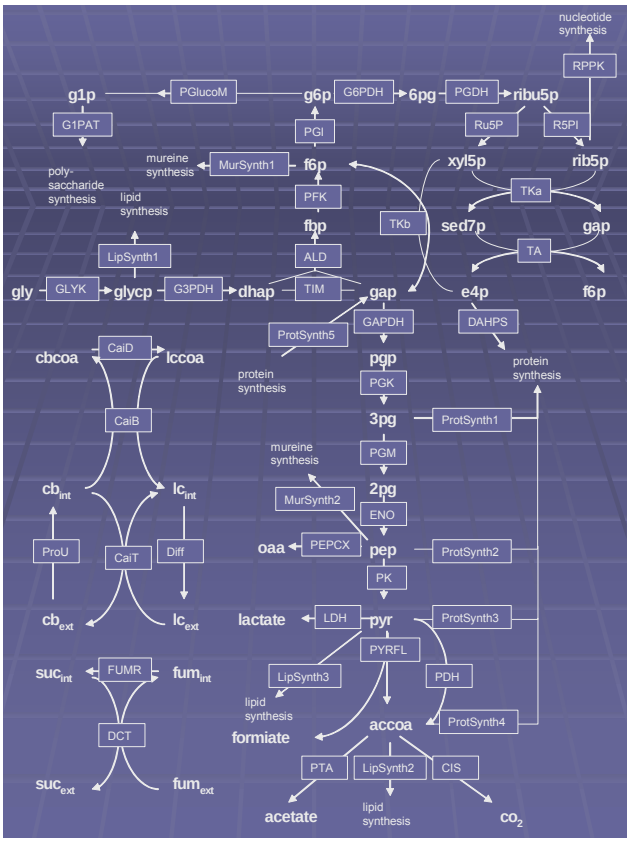


Figure 2. Representation of the anaerobic central metabolism in *E. coli* from model of Chassagnole et al. (2002) linked with the carnitine metabolism model of Cánovas et al. (2003a). The abbreviations are summarized on the Symbols section.

4. Experimental

4.1. Bacterial strain and culture media

The bacterial strain used, *E. coli* O74K74 (DSM 8828), contained the complete *cai* and *fix* operons and was stored in a minimal medium containing glycerol (20%) at –20ºC. The minimal medium (MM) was that described by Obon et al. (1999), while the complex medium (CM) contained (g/L): bacteriological peptone, 20; NaCl, 5; glycerol (carbon source), 12.6; crotonobetaine, 4 and fumarate, 2 (as an electron acceptor to inhibit the crotonobetaine reductase

activity, (Cánovas et al., 2002). The pH of both media was adjusted to 7.5 with 1 *M* KOH prior to autoclaving.

4.2. Growth of the bacteria

Batch and continuous experiments were performed in reactors equipped with temperature, pH and pump controllers (Biostat B, Braun Biotech International GMBH, Melsungen, Germany). A 1 L culture vessel with 0.5-0.8 L working volume was used. *Escherichia coli* O44K74 was grown under different conditions in order to optimise the induction of the carnitine metabolism enzymes. The culture was inoculated with a 3% (v/v) of the liquid culture stored at -20ºC in 20% (v/v) glycerol, while the medium employed was the CM mentioned above. The cells were grown under anaerobiosis at 37ºC in batch or continuous reactors. Anaerobic conditions were maintained to induce the enzymes involved in the carnitine metabolism, while D,L-carnitine mixture, D(+)-carnitine or crotonobetaine were supplied as inducers. Nitrogen was used to maintain anaerobiosis during the experiments.

4.3. Membrane reactor operation

The reactor vessel was also coupled to a cross-flow filtration module (Minitan, Millipore, USA) equipped with four 0.1 μm hydrophilic polyvinylidene difluoride Durapore plates of 60 cm^2 area (Millipore, USA) (Cánovas et al., 2002). The cell broth was recycled into the reactor with a peristaltic pump adjusted to a high flow rate (70 mL/min) to minimise membrane fouling. *E. coli* cells for the inoculum were grown as explained previously and transferred to the fermenter. Continuous operation was set at 37 ºC and started by feeding with the CM medium (anaerobically by bubbling nitrogen previously passed through a water trap).

4.4. Pulse experiments and sampling method

The pulse experiments were carried out using an injection pump supplying 20-25 mL (containing the concentrated component in turn being perturbed) in 3 s (7 to 9 $mL.s^{-1}$). Samples of 2 mL for metabolites and 5 mL for enzyme activities were withdraw from the high cell density reactor 1, 2, 3, 4, 5, 20, 50, 70, 90 and 120 min. after the pulse around the steady state. The experiments were carried out using a sampling mode consisting of closing the outlet of the nitrogen applied (to keep anaerobic conditions) within the reactor, while the sampling valve with a minimal dead volume (~300-400 μL) was opened. The sampling time took from 5 to 10 s. The complete procedure was computer controlled. The valve was flushed with water to clean the tubing. The reactor was left to recover and after 20 to 30 reactor residence times to ensure that a new steady state was reached, a new perturbation experiment was carried out. However, steady state conditions were assumed when biomass and glycerol concentrations remained constant during five times the mean residence time. Samples were collected in test tubes kept at -20 ^oC and immediately centrifuged at 16,000x g at 4 ^oC. The rotor was precooled at -20 ^oC. Supernatant was used for the determination of external metabolites, whereas pellets were used for measuring enzyme activity and ATP

cell content and the NADH/NAD^+ ratio. The cells were inactivated in less than one second.

4.5. Enzyme assays

In order to allow a more informative and precise way of comparing results, enzyme activities are represented as normalized values. Enzyme extraction were performed as follows. In each case, reactor bulk liquid samples were withdrawn and centrifuged at 16,000x g at 4 ^oC. The supernatant was removed and cells were re-suspended within the corresponding extraction buffer. Cells were sonicated for 6 cycles (10 s each), at 10 μm amplitude, with a probe of 1 cm diameter and below 20 ^oC. The extract was centrifuged for 15 min at 16,000xg and 4 ^oC to remove cell debris. The protein content was determined using the method of Lowry et al. (1951). The methods of enzyme assays have previously been described (Cánovas et al., 2003a).

4.6. Substrate consumption for growth and biotransformation processes

L(-)-carnitine concentration was determined enzymatically with the carnitine acetyl transferase method (Jung et al., 1989). Glycerol was analysed by HPLC with a Tracer Spherisorb-NH_2 column, 3 μm, 25 x 0.46 cm, supplied by Teknokroma (Barcelona, Spain) as reported (Obon et al., 1999). The isocratic mobile phase was 65% acetonitrile, 35% 50 mM phosphate buffer pH 5.5 at a flow rate of 1 mL/min. Bacterial growth was followed spectrophotometrically at 600 nm, using a Novaspec II from Pharmacia-LKB, (Uppsala, Sweden), and converted to dry weight accordingly.

4.7. Determination of central metabolite concentration

ATP content and NADH/NAD^+ ratio. The energy content per unit of cell was determined as the ATP level and NADH/NAD^+ ratio throughout the experiments. For ATP measurement, the HS II bioluminescence assay kit from Boehringer (Mannhein, Germany) was used. Reducing power as NADH/NAD^+ ratio was calculated as in Snoep et al. (1990). The cell content was determined after biomass optical density transformation as dry weight and assuming either an intracellular volume of 1.63 μL/mg (Emmerling et al., 2000) or 1.72 mLx 10^{-13}/cell (worked out by flow cytometry).

E. coli anaerobic metabolite production. The acetate, fumarate, lactate and formate contents of the bulk liquid reactor were determined by HPLC using a cation exchange Aminex HPX-87H column, supplied by BioRad Labs (Hercules, USA) was used. The isocratic mobile phase was 5 mM H_2SO_4 at a flow rate of 0.5 mL/min. The eluent was monitored using a refractive index detector. Samples were withdrawn from the reactor and centrifuged at 12,000 x g for 10 min at 4 ^oC. The supernatant was filtered and used for analyses.

5. Results

5.1 *IN SILICO* MODELLING OF THE LINK OF THE CARNITINE TO THE CENTRAL METABOLISM OF *E. coli*.

The model was able to describe the integration of the central and carnitine metabolisms. The growth and maintenance of the cells together with biotransformation results were in agreement with the experimental data, as is shown in the distribution of intracellular fluxes (Table 1).

Table 1. Experimental and theoretical fluxes obtained in the biotransformation of crotonobetaine into L-carnitine by *E. coli* in a continuous high cell density reactor

Transformation	Experimental Rate mM/h	Theoretical Rate mM/h
Effluent CR (reactor)	29.49	29.49
Influent CR (reactor)	50.0	50.0
CaiT(cell)	2357.29	2357.24
Pro U (cell)	3387.89	3387.89
CDH	1030.65	1030.65
PDH (cell)	7.21	4.42
Cellular growth (reactor)	0.039	0.039
Influent glycerol (reactor)	100	100
Effluent glycerol (reactor)	43.28	43.27
Effluent acetate (reactor)	27.08	27.07
Effluent lactate (reactor)	26.20	26.19

Based on the stoichiometric calculations and applying the Metabolic Flux Analysis, MFA, a balance on ATP was performed. Glycerol assimilation generated energy equivalents in terms of ATP from ADP as one of the products of the metabolism. The distribution of the relative ATP usage for the main ATP consuming processes in the metabolism is summarized in the Table 2.

Table 2. Distribution of ATP consumption in the central metabolism also including the carnitine metabolism.

Process	% ATP
Futile Cycle	57.24
L-carnitine production	24.83
Biomass production	15.17
Other metabolic processes	2.76

More than half of the energy was possibly wasted in a futile cycle that is associated to the entrance of crotonobetaine by means of ProU with energy consumption and the function of the transporter CaiT that is able to carry L-carnitine inside the cell with simultaneous excretion of crotonobetaine, generating no L-carnitine as the desired product.

5.2 *IN VIVO* DINAMICS OF *E. coli* METABOLISM: PULSE EXPERIMENTS IN HIGH DENSITY CONTINUOUS REACTORS.

In order to gain knowledge on the dynamic behaviour of steady state growing *Escherichia coli* cells, metabolic pulse experiments were performed. Having in mind the high productivities obtained in steady state high cell density membrane reactors, a five-fold increase of the steady state levels of glycerol (carbon source) and acetate (anaerobic metabolism) were performed. Following the pulse, the variation in isocitrate dehydrogenase (ICDH), isocitrate liase (ICL), acetyl-CoA synthetase (ACS), phosphotransacetylase (PTA) and carnitine dehydratase (CDH) activities, extracellular metabolites production and intracellular $NADH/NAD^+$ ratio and ATP were analyzed.

5.2.1. Pulse of glycerol during the steady state

Central metabolism: enzymes expression and coenzyme levels. Glycerol assimilation after the pulse deeply altered the cellular redox state. A sharp and unexpected fall in the $NADH/NAD^+$ ratio was assessed, possibly due to the rapid cellular response increasing fluxes of the cofactor regeneration pathways. In fact, increased synthesis of lactate and ethanol were observed during the first 20 min after the pulse, the reducing power starting to stabilize after 50 min (data not shown). Moreover, anaerobic fumarate respiration also contributed in NADH regeneration. Formate production and ATP content also increased after the pulse as a result of glycerol assimilation and use as energy source.

The opposed effects on ICDH and ICL activities (Fig. 3), suggested that upon increased NADH synthesis the glyoxylate cycle was the preferred pathway for acetyl-CoA consumption because of the lower amount of reducing power generated. Further, the flux towards acetate synthesis also increased as a result of overflow metabolism and the cellular need of ATP synthesis. At the enzyme expression level, ACS and PTA activities reached a sort of regulation/equilibrium which depended on the ATP and acetate level. Remarkably, the ACS level was one hundred times the steady state level while PTA remained almost constant (Fig. 3).

L(-)-carnitine metabolism. Despite glycerol metabolization yielded a higher cellular level of ATP (data not shown), which is necessary for L(-)-carnitine transport and activation, the crotonobetaine hydrating activity (CHR) levels decreased during the first 20 min after the pulse. However, bulk reactor L-carnitine concentration increased after the pulse.

5.2.2. Pulse of acetate during the steady state.

Central metabolism: enzymes expression and coenzyme levels. A sharp decline in the levels of NADH occurred and a corresponding increase of the lactate level was assessed. Ethanol levels decreased during the first minutes recovering afterwards. ATP production first fell, recovering afterwards and approaching the steady state values (data not shown). Net flux for acetate uptake also occasioned the increase in the production of lactate. After 90 min, both lactate and ethanol levels had diminished, compared to initial state while the acetate uptake and assimilation mechanisms were triggered, reaching the initial steady state value and thus achieving the cellular homeostasis.
An increase in both the ICDH and ICL enzyme activities (Fig. 3) was observed, greater in the case of ICL, while high formate levels revealed increased flux through PFL (anaplerotic pathways). Taken together, these results point to the built up of a high pool of intracellular acetyl-CoA during acetate assimilation. Furthermore, the PTA enzyme activity decreased due to the inhibition of the acetate synthesis after the pulse, while ACS increased, being thus this the preferred pathway for acetate uptake (Fig. 3).

L(-)-carnitine metabolism. The decrease in the ATP levels paralleled that in L(-)-carnitine production, despite the increase in CHR activity.

6. Discussion

In this work, the design and the experimental validation of a model which links the carnitine and central metabolisms is presented. This model is based on the central metabolism large scale stationary model of Chassagnole et al. (2002), which was modified by taking into account the anaerobic conditions under which the biotransformation process was carried out and the L-carnitine metabolism. The complete model presented 11 degrees of freedom, 3 internal and 8 external, so that at least 11 system fluxes were required to determine it completely. After the MFA approach was performed, two important results were extracted: (1) The flux from fumarate to succinate compared with that of other reactions was extremely high since it is necessary to assimilate the high reduction power of the C-source, glycerol; (2) The flux through acetate and lactate synthesis was strongly high since acetate is necessary to generate energy without NAD^+ reduction as well as lactate to regenerate the NADH formed during the glycerol assimilation. Besides, an ATP balance was carried out by the determination of the internal fluxes (MFA approach) as well as stoichiometric calculations (results not shown). The result was the same for both approaches: the biotransformation was ATP limited as a consequence of a futile cycle associated with the entrance of crotonobetaine by means of ProU (energy dependent irreversible transporter) and the transporter CaiT (reversible transporter, energy independent) the final result of which was the consumption of ATP. These results are in accordance with the experimental results of the pulses here presented as well as with

previous work in our group (Sevilla et al., 2005) in which it was established that the limiting point of the biotransformation was the transporter step rather than the biochemical reaction.

Additionally, a topological analysis was carried out and 11 elementary flux modes were found. The most representative were obtained for the consumption of glycerol to produce energy and different fermentation products as CO_2, acetate and lactate, cellular growth as well as carnitine synthesis. Besides, it was found that the L-carnitine yield on glycerol could be increased 3-6 times based on that this value was 0.36 mole/mole (far from the calculated of 2 mole/mole) and also the ATP balance (56% of ATP was waste in the previously mentioned futil cycle).

Fast changes in the cellular redox and energetic state were assessed after perturbation experiments. The metabolic state of *E. coli* and the intracellular pools of coenzymes evolved fast in order to deal with the pulsed metabolites. Though a high $NADH/NAD^+$ ratio was expected after the glycerol pulse, metabolic changes affecting NADH regeneration rate were observed to involve fermentation pathways, TCA and glyoxylate shunt (Fig. 3). The ICDH activity diminished to limit the production of reducing power, while ICL increased since redox equivalents are consumed in anabolic reactions (Cánovas et al., 2003a). In addition, a high flux of formate production, probably through pyruvate formate lyase (PFL), was assessed (de Graef et al., 1999).

As already stated, the accumulation of acetate is a consequence of the inability of the cell to deal with a large amount of substrate (Wolfe, 2005). The increased acetate level provoked opposed effects on the expression of acetate metabolism enzymes. Raising the level of the carbon source led to higher cell activity and increased ATP synthesis through the acetate metabolism. Moreover, high $NADH/NAD^+$ ratio also seems to indirectly inhibit PTA and encourage ACS activity, this regulatory mechanism preventing the hindering of normal cell function by consumption of all cellular NAD^+.

In order to deal with glycerol assimilation more efficiently, it seems likely that the expression of secondary metabolism was left apart, since the CHR activity decreased during the first moments of the pulse (Fig. 3). Despite this, the levels of L(-)-carnitine kept on increasing, indicating that CHR activity was not a limiting step for the bioprocess.

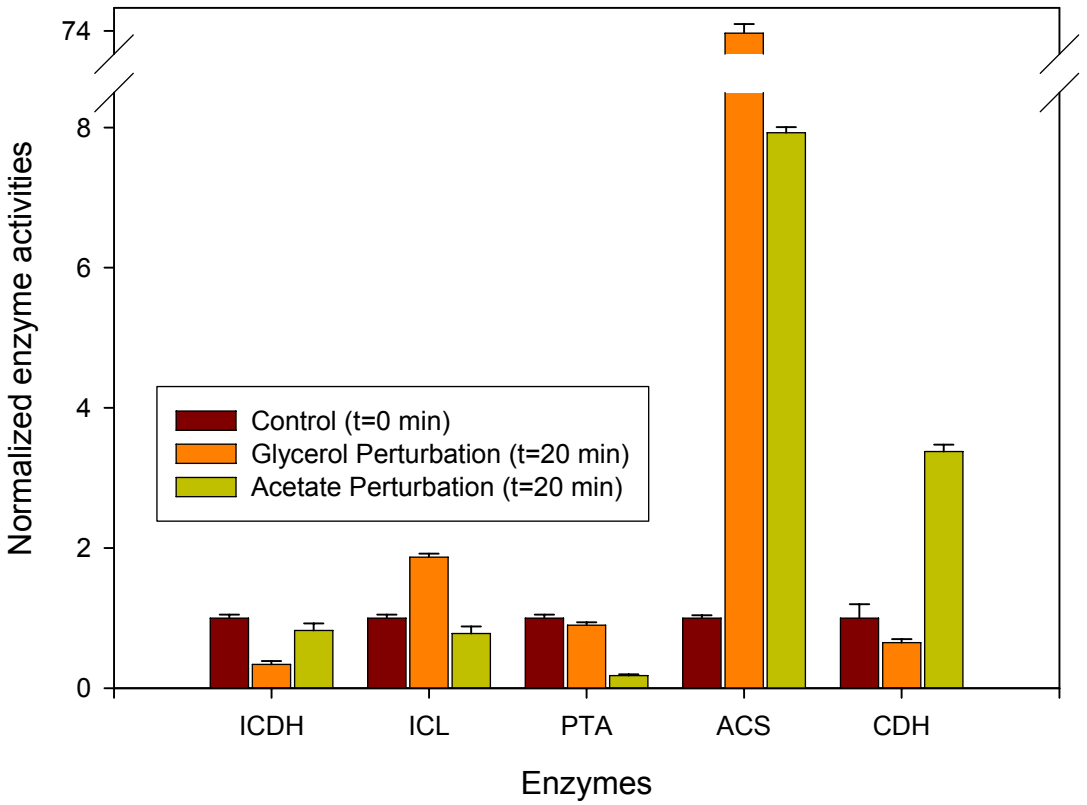


Figure 3. Effect of metabolic pulses of glycerol and acetate in the expression of the anaerobic central metabolism of *E. coli*. The analyzed activities (isocitrate dehydrogenase, ICDH; isocitrate liase, ICL; phosphotransacetylase, PTA; acetyl-CoA synthetase, ACS) have already been shown to be involved in the linking of the carnitine metabolism (carnitine dehydratase, CDH) as proposed in Cánovas et al. (2003a). Enzyme activities are shown normalized to their corresponding steady state value.

After the acetate pulse, the ICDH and ICL enzyme activities increased (Fig. 3) since an increase of the intracellular acetyl-CoA concentration would activate both enzymes. The NADH levels first diminished, leading afterwards to higher NADH and energy levels (TCA versus the glyoxylate shunt) (Fig. 3). The different responses observed in the ethanol and lactate pathways were possibly due to the lower reducing power generated during the acetate assimilation or even to a switch in the metabolic destination of the acetyl-CoA pool. As expected, the pulse also resulted in the increase in ACS activity/expression and inhibition of PTA (Fig. 3) through repression by catabolites (Brown et al., 1977). The L-carnitine level fell despite the increase in the CHR activity (Fig. 3), coinciding its recovery with that in the ATP level and the NADH/NAD^+ ratio. Interestingly, the limiting step in the biotransformation was not really the expression of CHR activity, but more probably the cofactor-dependent processes such as the transport and activation of substrate, which are both ATP-dependent

(Cánovas et al., 2003a). Further, the assimilation of acetate would also alter the CoA pool, probably leading to the accumulation of acetyl-CoA, reducing the amount of free CoA available for substrate activation.

The different carbon number and oxidation state of the pulsed metabolites resulted in different NADH/NAD^+ ratio patterns, since different assimilation pathways were involved. Consequently, distinguishable responses in metabolites and metabolic fluxes were triggered. While the whole glycolytic pathway was perturbed during the glycerol pulse as well as metabolites closely linked to it, the perturbation effected in the acetate pulse was only indirect, via the NADH/NAD^+ ratio and the ATP levels. The levels of ICL and ICDH reflected the key role of the glyoxylate shunt. These two competing pathways are subjected to coordinate functioning, through ICDH-kinase/phosphatase (coexpressesing from the glyoxylate shunt *ace* operon) which inactivates ICDH, improving opportunities for ICL (Cronan and Laporte, 1996). Additionally, in our work, opposite responses were found between these enzymes and the NADH/NAD^+ ratio, reflecting a more complex regulation.

The changes in acetate levels during the pulses suggest that this was produced through the PTA-ACK pathway during *E. coli* perturbations (Kleman and Strohl, 1994) and consumed at the beginning of the transition towards the steady state, by the ACS enzyme, rendering energy and biosynthetic compounds, as previously shown to happen in batch systems (Brown et al., 1977; Kumari et al., 2000). Therefore, the cell metabolism adjusts to nutrient shortages and acetate is consumed. The level of acetate was related to ACS, PTA (Fig. 3) and possibly PFL enzyme activities. The first two enzymes showed a regulatory behaviour, possibly responding to intracellular acetyl-CoA, which effects a feed-back negative regulation on ACS (Kumari et al., 2000). The rest of the metabolites studied provided information on redox cell state, since both lactate and ethanol consume NADH to produce NAD^+, while the glyoxylate shunt generates reducing power.

The existence of general cell regulators allows the coordinate expression of metabolic pathways. Regulation of the carnitine biotransformation pathway has been previously described (Eichler et al., 1994; Kleber, 1997) and depends on general regulators, such as FNR (transcriptional regulator under anaerobic conditions), catabolic repression via cAMP protein receptor (CRP), histones (H-NS) and σ^S (RpoS). Regulation of acetate metabolism has also been shown to depend on these general regulators (Kumari et al., 2000). Further, coordinate functioning of glyoxylate shunt and acetate metabolism relies on IclR, which in its active form represses the expression of ICL and ACS (Shin et al., 1997). In the presence of PEP (Fig. 1), IclR protein is prevented from binding to the promoter region of *aceBAK* (Cortay et al., 1991) allowing the uptake of bulk reactor acetate (Sánchez et al., 2005). Also, other factors such as the two-component ArcAB system (regulator of the anaerobic and aerobic metabolism, inhibiting CAT, the electron transport chain and the PDH activity in anaerobic conditions) activates the PFL enzyme (de Graef et al., 1999), which is also regulated by the one-component protein FNR (fumarate, nitrate reduction). Shalel-Levanon et al. (2005) observed that the internal redox potential, as

reflected by the $NADH/NAD^+$ ratio, was significantly higher in cultures of an *arcA* mutant strain compared with the wild strain. However, the $NADH/NAD^+$ ratio had no influence or specific effect on FNR activity. Furthermore, the $NADH/NAD^+$ ratio regulates the PDH and PFL enzymes (de Graef et al., 1999). Previous results obtained within the research group showed that in batch systems the $NADH/NAD^+$ ratio was higher in anaerobiosis than in aerobiosis while PDH was inhibited (Cánovas et al., 2003a).

The results demonstrate the relationship between the central carbon and the carnitine metabolism under anaerobiosis, and also the importance of the glyoxylate and acetate metabolism during the biotransformation. In a previous work (Cánovas et al., 2003a), the cessation of L(-)-carnitine production was already shown to coincide with the decrease of the ATP pool, which is quite likely to be due to the involvement of ProU during biotransformation (Verheul et al., 1998; Jung et al., 2002) and to substrate activation into crotonobetainyl-CoA and γ-butyrobetainyl-CoA by CaiC activity (Elssner et al., 2001). The importance of the ATP pool was highlighted by the determined futile cycle (Table 2), thus sustaining previous experimental evidences (Cánovas et al., 2003a). Also, the key role of glyoxylate shunt in high cell-density systems after a sudden increase in carbon source (pulse of glycerol and acetate) is highlighted. High ACS activities were observed in both pulses, whereas for the acetate pulse the PTA decreased due to control in the acetate metabolism. More importantly, the highly stringent regulation around the redox cellular state was verified. When the enzyme specific activities of the central metabolism were compared within the different pulses it was seen that: i) the glyoxylate shunt was less active after a glycerol pulse; ii) the flux through the TCA cycle increased under a low $NADH/NAD^+$ ratio and, iii) there was an increase in acetate metabolism enzyme activities, such as PTA and ACS, during the glycerol and acetate pulses, probably induced by the acetate bulk reactor level. Finally, the ATP/cell values were higher after the glycerol pulse.

7. Conclusion

The principal conclusion of the developed model was that more than half of the energy was possibly wasted in a futile cycle This result is in agreement with the previous work (Sevilla et al., 2005), the limiting factor is the transport of substrate, but the explanation at molecular level was found in this work, the simultaneous operation of CaiT and ProU carriers resulted in the waste of ATP in a futile cycle, since both trimethylammonium compound carriers work in the opposite direction leading to a waste of energy. The existence of two kind of transporters is probably due to the double function of L(-)-carnitine in *E. coli*. As an electron acceptor in anaerobic conditions (Kleber, 1997), a fast transporter which interchanges carnitine or crotonobetaine and its reduced form (γ-butirobetaine) is needed (CaiT). As osmoprotector (Verheul et al., 1998) expression of ProU under osmotic stress allows to accumulate trimethylammonium compounds inside the cell (L-carnitine and crotonobetaine)

as a way of protecting itself from the extreme osmotic situation (Verheul et al., 1998). Therefore, two possible metabolic engineering strategies can be applied in order to improve the L(-)-carnitine biosynthesis, increment the ATP levels or abolish the futil cycle generated by the simultaneous use of CaiT and ProU.
The metabolic enzyme activities measured within these work are the result of the whole metabolic function involving not only the mechanisms of regulation of expression, but also those of functioning, such as effectors. The combination of all drive the system back to the steady state. Also, *in vivo* analysis showed that intracellular metabolic pools were efficiently equilibrated by the corresponding enzymes, meaning that enzymes function near equilibrium (Wittmann et al., 2005). A combination of metabolome, enzyme activities, protein and gene expression levels analysis allows the correct understanding of cellular functioning and cell metabolism regulation. In fact, translational and post-translational control on protein expression and turnover as well as enzyme activity is mediated by allosteric interaction and metabolic control.
L(-)-carnitine synthesis depended mainly on the energetic state of the cell, especially on the ATP levels. The production of L(-)-carnitine was observed to be independent on the CHR activity, indicating that this enzyme is not limiting for the process. This study confirms that the ATP level is a critical variable for the biotransformation, not only due to the ProU transporter (Verheul et al., 1998), but, possibly, also because of the postulated trimethylammonium-CoA ligase activity (CaiC, Eichler et al., 1994). The connection of both metabolisms, outlined in Fig. 1, suggests the existence of control points where redirection of metabolic fluxes could be possible. However, further work concerning the linking between primary carbon and the carnitine metabolisms is being performed at our group.

Acknowledgements. *This work has been supported by CICYT projects BIO2002-04157-C02-01, BIO2002-04157-C02-02 and SENECA-CARM PB/10/FS02. A. Sevilla and V. Bernal hold predoctoral research grants from MEC and Fundación CajaMurcia, respectively. We would like to thank Prof. H-P. Kleber (University of Leipzig, Germany) for valuable discussions. Biosint S.p.A.(Italy) is also acknowledged for the kind gift of the substrate*

Symbols

2pg	2-phosphoglycerate
3pg	3-phosphoglycerate
6pg	6-phosphogluconate
accoA	acetyl-coenzyme A
adp	adenosindiphosphate
aicar	5-amino-4-imidazolecarboxamideribotide
akg.	α-ketoglutarate
aki	α-ketoisovalerate
amp	adenosinmonophosphate
arg	arginine
asn	asparagine
asp	aspartate
atp	adenosintriphosphate
carp	carbamoylphosphate
cr	crotonobetaine
crcoa	crotonobetainylCoA
cdp	cytideindiphosphate
cho	chorismate
cmp	cytideinmonophosphate
co2	carbondioxide
coA	coenzyme A
cr	crotonobetaine
crcoa	crotonobetainylCoA
ctp	cytideintriphosphate
cys	cysteine
dala	D-alanine
damp	deoxyadenosinmonophosphate
dcmp	deoxycytideinmonophosphate
dglu	D-glutamate
dgmp	deoxyguanosinmonophosphate
dhap	dihydroxyacetonephosphate
dna	deoxyribonucleic acid
dtmp	deoxythymidinmonophosphate
dipim	diaminopimelate
e4p	Erythrose-4-phosphate
ea	electron aceptor oxidized
eah_2	electron aceptor reduced
etamp	phosphatidyl-ethanolamine
f6p	fructose-6-phosphate
fad	flavin-adenine-dinucleotide, oxidized
fadh2	flavin-adenine-dinucleotide, reduced
$fatty_{n,i}$	fatty acid containing n carbon atoms and i double bonds
fdp	fructse-1,6-bisphosphate
fthf	formyltetrahydrofolate
fum	fumarate
g1p	glucose-1-phosphate
g6p	glucose-6-phosphate
gap	glyceraldehyde-3-phosphate
glc	glucose
gln	Glutamine
gly	glycine
glycp	glycerol-3-phosphate
gmp	guanosinmonophosphate
h	proton
h_2o	water
h	proton
his	histidine
hom	homoserine
ile	isoleucine
imp	inosinmonophosphate
isocit	isocitrate
kival	a-ketoisovalerate
lala	L-alanine
lc	L-carnitine
lccoa	LcarnitinylCoA
leu	leucine
lglu	L-glutamate
lys	lysine
mal	malate
met	methionine
methf	methylentetrahydrofolate
mythf	methyltetrahydrofolate
murunit	subunit of mureine
nad	diphosphopyridindinucleotide, oxidized
nadh	diphosphopyridindinucleotide, reduced
nadp	diphosphopyridindinucleotide-phosphate, oxidized

nadph	diphosphopyridindinucleotide-phosphate, reduced
nh_4	ammonium
oac	oxaloacetate
orn	ornithine
p	inorganic phosphate
pep	phosphoenolpyruvate
phe	phenylalanine
pgp	1,3-diphosphoglycerate
pro	proline
prpp	phosphoribosylpyrophosphate
pyr	pyruvate
Rib5p	ribose-5-phosphate
ribu5p	ribulose-5-phosphate
rna	ribonucleic acid
sed7p	sedoheptulose-7-phosphate
ser	serine
SO_4	sulfate
suc	succinate
succoA	succinyl-coenzyme A
thf	tetrahydrofolate
thr	threonine
trp	tryptophan
tyr	tyrosine
udp	uridindiphosphate
ump	uridinmonophosphate
utp	uridintriphosphate
val	valine
xyl5p	xylulose-5-phosphate

References

Brown, T.D., Jones-Mortimer, M.C., and Konberg, H.L. (1977) The enzymatic interconversion of acetate and acetyl coenzyme A in *Escherichia coli*. *J. Gen. Microbiol.* **102**, 327-336.

Cánovas, M., Bernal, V., Torroglosa, T., Ramírez, J.L. and Iborra, J.L. (2003a) Link between primary and secondary metabolism in the biotransformation of trimethylammonium compounds by *Escherichia coli*. *Biotechnol. Bioeng*. **84**, 689-699.

Cánovas, M., Maiquez, J.R., Obon, J.M., and Iborra, J.L. (2002) Modeling of the biotransformation of crotonobetaine into L-(-)- carnitine by *Escherichia coli* strains. *Biotechnol. Bioeng.* **77**, 764-775.

Cánovas, M., Torroglosa, T., Kleber, H.P., and Iborra, J.L. (2003b) Effect of salt stress on crotonobetaine and D(+)-carnitine biotransformation into L(-)-carnitine by resting cells of *Escherichia coli*. *J. Basic Microbiol.* **43**, 259-268.

Castellar, M.R., Cánovas, M., Kleber, H.P., and Iborra, J.L. (1998) Biotransformation of D(+)-carnitine into L(-)-carnitine by resting cells of *Escherichia coli* O44 K74. *J. Appl. Microbiol.* **85**, 883-890.

Cavazza, C. (1981) D-camphorate of L-carnitinamide and D-camphorate of D-carnitinamide. BE patent 877609 A1.

Chassagnole, C., Noisommit-Rizzi, N., Schmid, J.W., Mauch, K., and Reuss, M. (2002) Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnol. Bioeng.* **79**, 53-73.

Cortay, J.C., Negre, D., Galinier, A., Duclos, B., Perriere, G., and Cozzone, A.J. (1991) Regulation of the Acetate Operon in *Escherichia coli*: Purification and Functional Characterization of the iclR Repressor. *EMBO J.* **10**, 675-679.

Cronan, J.E. and Laporte, D.C. (1996) Tricarboxylic acid cycle and glyoxylate bypass. In: *Escherichia coli and Salmonella: cellular and molecular biology. 2nd ed., vol. 1*. (Neidhardt, F. C., Curtiss, R. III, Ingraham, J. L., Lin, E. C. C., Low, K. B., Magasanik, B., Rezniko, W. S., Riley, M., Schaechter, M., and Umbarger, H. E., Ed.) pp 206-216. Washington D.C: ASM Press.

de Graef, M.R., Alexeeva, S., Snoep, J.L., and Teixeira de Mattos, M.J. (1999) The steady-state internal redox state (NADH/NAD) reflects the external redox state and is correlated with catabolic adaptation in *Escherichia coli*. *J. Bacteriol.* **181**, 2351-2357.

Eichler, K., Bourgis, F., Buchet, A., Kleber, H.P., and Mandrand-Berthelot, M.A. (1994) Molecular characterization of the cai operon necessary for carnitine metabolism in *Escherichia coli*. *Mol. Microbiol.* **13**, 775-786.

Eichler, K., Buchet, A., Lemke, R., Kleber, H.P., and MandrandBerthelot, M.A. (1996) Identification and characterization of the caiF gene encoding a potential transcriptional activator of carnitine metabolism in *Escherichia coli*. *J. Bacteriol.* **178**, 1248-1257.

Elssner, T., Engemann, C., Baumgart, K., and Kleber, H.P. (2001) Involvement of coenzyme A esters and two new enzymes, an enoyl- CoA hydratase and a CoA-transferase, in the hydration of crotonobetaine to L-carnitine by *Escherichia coli*. *Biochem.* **40**, 11140-11148.

Elssner, T., Hennig, L., Frauendorf, H., Haferburg, D., and Kleber, H.P. (2000) Isolation, identification, and synthesis of gamma- butyrobetainyl-CoA and crotonobetainyl-CoA, compounds involved in carnitine metabolism of *E. coli*. *Biochem.* **39**, 10761-10769.

Emmerling, M., Bailey, J.E., and Sauer, U. (2000) Altered regulation of pyruvate kinase or co-overexpression of phosphofructokinase increases glycolytic fluxes in resting *Escherichia coli*. *Biotechnol. Bioeng.* **67**, 623-627.

Hoeks, F.W.J.M., Muhle, J., Bohlen, L., and Psenicka, I. (1996) Process integration aspects for the production of fine chemicals illustrated with the biotransformation of gamma-butyrobetaine into L-carnitine. *Chem. Eng. J. and Biochem. Eng. J.* **61**, 53-61.

Jung, H., Buchholz, M., Clausen, J., Nietschke, M., Revermann, A., Schmid, R., and Jung, K. (2002) CaiT of *Escherichia coli*, a new transporter catalyzing L-carnitine/gamma-butyrobetaine exchange. *J. Biol. Chem.* **277**, 39251-39258.

Jung, H., Jung, K., and Kleber, H.P. (1989) Purification and Properties of Carnitine Dehydratase from *Escherichia coli*: A New Enzyme of Carnitine Metabolization. *Biochim. Biophys. Acta* **1003**, 270-276.

Kleber, H.P. (1997) Bacterial carnitine metabolism. *FEMS Microbiol. Lett.* **147**, 1-9.

Kleman, G.L. and Strohl, W.R. (1994) Acetate Metabolism by *Escherichia coli* in High Cell Density Fermentation. *Appl. Environ. Microbiol.* **60**, 3952-3958.

Kulla, H.G. (1991) Enzymatic Hydroxylations in Industrial Application. *Chimia* **45**, 81-85.

Kumari, S., Beatty, C.M., Browning, D.F., Busby, S.J.W., Simel, E.J., Hovel-Miner, G., and Wolfe, A.J. (2000) Regulation of acetyl coenzyme A synthetase in Escherichia coli. *J. Bacteriol.* **182**, 4173-4179.

Lowry, O.H., Rosebrough, N.J., Farr, A.L., and Randall, R.J. (1951) Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* **193**, 265-275.

Obon, J.M., Maiquez, J.R., Cánovas, M., Kleber, H.P., and Iborra, J.L. (1999) High-density *Escherichia coli* cultures for continuous L(-)- carnitine production. *Appl. Microbiol. Biotechnol.* **51**, 760-764.

Roth, S., Jung, K., Jung, H., Hommel, R.K., and Kleber, H.P. (1994) Crotonobetaine Reductase from *Escherichia coli.* A New Inducible Enzyme of Anaerobic Metabolization of L(-)-Carnitine. *Antonie Van Leeuwenhoek In. J. Gen. Molec. Microbiol.* **65**, 63-69.

Sánchez, A.M., Bennett, G.N., and San, K.Y. (2005) Effect of different levels of NADH availability on metabolic fluxes of *Escherichia coli* chemostat cultures in defined medium. *J. Biotechnol.* **117**, 395-405.

Seim, H., Eichler, K., and Kleber, H.P. (2001) L(-)-carnitine and its precursor g-butirobetaine. In: *Nutraceuticals in health and disease prevention*. (Krämer, K, Hoppe, PL, and Packer, L, Ed.) pp 217-256. Marcel Dekker., New York.

Sevilla, A., Vera, J., Diaz, Z., Cánovas, M., Torres, N.V., and Iborra, J.L. (2005) Design of metabolic engineering strategies for maximizing L-(-)-carnitine production by *Escherichia coli*. Integration of the metabolic and bioreactor levels. *Biotechnol. Prog.* **21**, 329-337.

Shalel-Levanon, S., San, K.Y., and Bennett, G.N. (2005) Effect of ArcA and FNR on the expression of genes related to the oxygen regulation and the glycolysis pathway in *Escherichia coli* under microaerobic growth conditions. *Biotechnol. Bioeng.* **92**, 147-159.

Shin, S., Song, S.G., Lee, D.S., Pan, J.G., and Park, C. (1997) Involvement of iclR and rpoS in the induction of acs, the gene for acetyl coenzyme A synthetase of *Escherichia coli* K-12. *FEMS Microbiol. Lett.* **146**, 103-108.

Snoep, J.L., Demattos, M.J.T., Postma, P.W., and Neijssel, O.M. (1990) Involvement of Pyruvate-Dehydrogenase in Product Formation in Pyruvate-Limited Anaerobic Chemostat Cultures of *Enterococcus Faecalis* Nctc-775. *Arch. Microbiol.* **154**, 50-55.

Unden, G. and Trageser, M. (1991) Oxygen Regulated Gene Expression in *Escherichia coli*. Control of Anaerobic Respiration by the Fnr Protein. *Antonie Van Leeuwenhoek In. J. Gen. Molec. Microbiol.* **59**, 65-76.

Verheul, A., Wouters, J.A., Rombouts, F.M., and Abee, T. (1998) A possible role of ProP, ProU and CaiT in osmoprotection of *Escherichia coli* by carnitine. *J. Appl. Microbiol.* **85**, 1036-1046.

Wittmann, C., Hans, M., van Winden, W.A., Ras, C., and Heijnen, J.J. (2005) Dynamics of intracellular metabolites of glycolysis and TCA cycle during cell cycle related oscillation in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **89**, 839-847.

Wolfe, A.J. (2005) The acetate switch. *Microbiol. Mol. Biol. Rev.* **69**, 12-50.

APPENDIX 1

List of reaction stoichiometries for the steady-state flux model according to ECOCYC (Karp et al., 2004) and Neidhardt et al. (1996). The cometabolites amp, adp and atp are treated as external.
Substances crossing the boundaries of the system were: glycerol, acetate, formiate, lactate, biomass, o_2, co_2, protons, h_2o, L-carnitine, crotonobetaine, fumarate and succinate, *by symport*:

so_4 + h + adp + p ← so_4 + atp + h_2o
2p + h + adp ← p + atp + h_2o
nh_4 + 3h + 3adp + 3p ← nh_4 + 3atp + $3h_2o$
cr + adp + p ← cr + atp + h_2o

Emden-Meyerhof-Parnas pathway:

g6p ↔ g1p
g6p ↔ f6p
f6p + p ←fdp + h_2o
fdp ↔ dhap + gap
dhap ↔ gap
gap + p + nad ↔ pgp + nadh + h
pgp + adp + h ↔ 3pg + atp
3pg ↔ 2pg
2pg ↔ pep + h_2o + h
pep + adp + h → pyr + atp
pyr + coa + nad → accoa + co_2 + nadh

Pentose-phosphate pathway:

g6p + nadp + h_2o → 6pg + nadph + 2h
6pg + nadp → ru5p + nadph + co_2
ribu5p ↔ rib5p
ribu5p ↔ xyl5p
xyl5p + rib5p ↔ gap + sed7p
gap + sed7p ↔ e4p + f6p
xyl5p + e4p ↔ f6p + gap

Tricarboxylic acid cycle:

oac + accoa + h_2o → isocit + coa + h
isocit + nad ↔ akg + nadh + co_2
akg + coa + nad → succoa + nadh + co_2
succoa + adp + p ↔ suc + atp + coa
mal + nad ↔ oac + nadh + h

Phosphoenolpyruvate carboxylase:

pep + co_2 + h_2o → oac + p + h

Amino acid synthesis:

pyr + lglut ↔ lala + akg
lala ↔ dala
$2h_2o + co_2$ + gln + 2atp ↔ carp + lglut + p + 2adp + 3h
2lglut + accoa + atp + nadph + h_2o → orn + akg + coa + adp + p + nadp + acetate + h
p + accoa ↔ acp + coa
acp + adp ↔ acetate + atp
orn + carp + asp + atp + h_2o ↔ arg + fum + amp + 3p + 2h
asp + nh_4 + atp + h_2o ↔ asn + 2h + amp + 2p
oac + lglut ↔ asp + akg
so_4 + 2atp + 4nadph + 2h ↔ sulfide + 4nadp + 2adp + 2p + $2h_2o$
ser + accoa + sulfide + h ↔ cys + coa + acetate
nh_4 + akg + nadph + h ↔ lglut + nadp + h_2o
lglut ↔ dglut
lglut + nh_4 + atp ↔ gln + adp + p + h
ser + thf ↔ gly + methf + h_2o
prpp + gln + atp + 2nad + $5h_2o$ → his + aicar + akg + 2nadh + 7h + 5p
asp + atp + 2nadph + 2h ↔ hom + adp + p + 2nadp
thr + pyr + nadph + lglut + 2h → ileu + nh_4 + nadp + h_2o + co_2 + akg
hom + succoa + cys + mythf + h_2o ↔ met + coa + suc + pyr + nh_4 + h+ thf

asp + pyr + lglut + succoa + atp + 2nadph + h → dipim + akg + suc + coa + adp + 2nadp + p
2pyr + nadph + 2h ↔ aki + co_2 + nadp + h_2o
aki + lglut + accoa + h_2o + nad ↔ leu + akg + coa + nadh + h + co_2
dipim + h ↔ lys + co_2
pep + e_4p + h_2o + h → dahp + p
dahp + pep + nadph + atp → cho + adp + 3p + nadp + h_2o + h
cho + lglut + h → phen + akg + co_2 + h_2o
lglut + atp + 2nadph + 2h → pro + adp + p + h_2o + 2nadp
3pg + lglut + nad + h_2o ↔ ser + akg + p + nadh + 2h
hom + atp + h_2o ↔ thr + adp + p + h
cho + gln + prpp + ser → tryp + 2p + co_2 + gap + lglut + h + pyr + h_2o
cho + lglut + nad → tyr + akg + co_2 + nadh
aki + lglut ↔ val + akg

Nucleotide metabolism:

rib5p + atp → prpp + amp + h
prpp + 2gln + gly + 5atp + asp + fthf + $4h_2o$ + co_2 → aicar + 5adp + 7p+ 2lglut + thf + fum + 9h
aicar + fthf ↔ thf + imp + h_2o
imp + asp + atp ↔ amp + adp + p + fum + 2h
amp + atp ↔ 2adp
h_2o + atp → adp + p + h
imp + nad + atp + gln + $3h_2o$ ↔ gmp + amp + nadh + 2p + 4h + lglut

carp + asp + menaquinone + prpp ↔ ump + menaquinoneh_2 + co_2 + 3 p
ump + atp ↔ udp + adp
udp + atp ↔ utp + adp
utp + gln + atp + h_2o → ctp + adp + p + 2h + lglut
ctp + adp ↔ cdp + atp
cdp + adp ↔ cmp + atp
adp + atp + nadph + h ↔ datp + adp + h_2o + nadp
gmp + 2atp + nadph + h ↔ dgtp + 2adp + nadp + h_2o
cdp + atp + nadph + h ↔ dctp + adp + nadp + h_2o
udp + 3atp + 2nadph + h_2o + methf ↔ dttp + 3adp + 2nadp + 2p + thf

Synthesis of lipid precursors:

dhap + nadh + h ↔ glycp + nad
glycp + ser + ctp + h_2o ↔ ethohaminp + cmp + co_2 + 2p + h
5accoa + 4atp + 8nadph + 3h + h_2o → $fatty_{10,0}$ + 5coa + 4adp + 8nadp+ 4p

6accoa + 5atp + 10nadph + 4h + h_2o → $fatty_{12,0}$ + 6coa + 5adp + 10nadp+ 5p
5accoa + 5atp + 10nadph + 4h + h_2o + succoa + pyr → $fatty_{13,0}$ + 6coa + 5adp + 10nadp + 5p + oac
5accoa + 5atp + 9nadph + 3h + h_2o + succoa + pyr → $fatty_{13,1}$ + 6coa + 5adp + 9nadp + 5p + oac
7accoa + 6atp + 12nadph + 5h + h_2o → $fatty_{14,0}$ + 7coa + 6adp + 12nadp+ 6p
7accoa + 6atp + 1nadph + 4h + h_2o → $fatty_{14,1}$ + 7coa + 6adp + 1nadp + 6p
6accoa + 6atp + 12nadph + 5h + h_2o + succoa + pyr → $fatty_{15,0}$ + 7coa + 6adp + 12nadp + 6p + oac
6accoa + 6atp + 1nadph + 4h + h_2o + succoa + pyr → $fatty_{15,1}$ + 7coa + 6adp + 1nadp + 6p + oac
8accoa + 7atp + 14nadph + 6h + h_2o → $fatty_{16,0}$ + 8coa + 7adp + 14nadp + 7p
8accoa + 7atp + 13nadph + 5h + h_2o → $fatty_{16,1}$ + 8coa + 7adp + 13nadp + 7p
7accoa + 7atp + 14nadph + 6h + h_2o + succoa + pyr → $fatty_{17,0}$ + 8coa + 7adp + 14nadp + 7p + oac
7accoa + 7atp + 13nadph + 5h + h_2o + succoa + pyr → $fatty_{17,1}$ + 8coa + 7adp + 13nadp + 7p + oac
9accoa + 8atp + 16nadph + 7h + h_2o → $fatty_{18,0}$ + 9coa + 8adp + 16nadp + 8p
9accoa + 8atp + 15nadph + 6h + h_2o → $fatty_{18,1}$ + 9coa + 8adp + 15nadp + 8p
8accoa + 8atp + 16nadph + 7h + h_2o + succoa + pyr → $fatty_{190}$ + 9coa + 8adp + 16nadp + 8p + oac

Polymerization reactions:

$\beta_1$lala + $\beta_2$arg + β3asn+ $\beta_4$asp + $\beta_5$cys + $\beta_6$gln + $\beta_7$lgln + $\beta_8$gly + $\beta_9$his + β_{10}ile + β_{11}leu + β_{12}lys + β_{13}met + β_{14}phen + β_{15}pro + β_{16}ser + β_{17}thr + β_{18}trp + β_{19}tyr + β_{20}val + β_{21}atp + $\beta_{22}h_2o$ → protein + β_{21}adp + β_{21}p + β_{21}h

$\chi_1$amp + $\chi_2$gmp + $\chi_3$ump + $\chi_4$cmp + $\chi_5$atp + χ_1h_2o → rna + $\chi_5$adp + χ_5p + χ_5h

$\delta_1$datp + $\delta_2$dgtp + $\delta_3$dttp + $\delta_4$dctp + $\delta_5$atp → dna + $\delta_5$adp + δ_5p + δ_5h

$\varepsilon_1$fatty$_{10,0}$ + $\varepsilon_2$fatty$_{12,0}$ + $\varepsilon_3$fatty$_{13,0}$ + $\varepsilon_4$fatty$_{13,1}$ + $\varepsilon_5$fatty$_{14,0}$ + $\varepsilon_6$fatty$_{14,1}$ + $\varepsilon_7$fatty$_{15,0}$ + $\varepsilon_8$fatty$_{15,1}$ + $\varepsilon_9$fatty$_{16,0}$ + ε_{10}fatty$_{16,1}$ + ε_{11}fatty$_{17,0}$ + ε_{12}fatty$_{17,1}$ + ε_{13}fatty$_{18,0}$ + ε_{14}fatty$_{18,1}$+ ε_{15}fatty$_{19,0}$ + ε_{16}glycp + ε_{17}etamp → lipid + $\varepsilon_{18}h_2o$

(ϕ_1 + 1)g6p + $\phi_1$atp + ϕ_1h_2o → polysaccharides + $\phi_1$adp + ϕ_1h + 2ϕ_1p

2f6p + 2gln + 2accoa + 7atp + pep + nadph + lala + dglut + dipim + dala + 5h_2o → murunit + 2coa + 2lglut + 6adp + amp + nadp + 1p + 9h

$\gamma_1$murunit → murein + γ_2h_2o

values of β_i, χ_i, δ_i, ε_i, ϕ_I and γ_I according to the biomass composition.

Virtual biomass formation reaction:

$\alpha_1$protein + $\alpha_2$rna + $\alpha_3$dna + $\alpha_4$lipids + $\alpha_5$polysaccharides + $\alpha_6$mureine → biomass (values of α_I according to the biomass composition).

Regeneration of C1-transfer cometabolites:

thf + atp + nadh + co_2 ↔ fthf + adp + p + nad

thf + co_2 + 3nadh + 3h ↔ mythf + 3nad + 2h_2o

nadh + nh_4 + co_2 + methf ↔ nad + gly + thf

Anaerobic respiration:

pyr + coa → formate + accoa

pyr + nadh + h ↔ lac + nad

formate + h → co_2 + h_2

Glycerol assimilation

glyc + atp ↔ glycp + adp + h

Fumarate reduction:

fum + eah_2 ↔ suc + ea

fum + $menaquinoneh_2$ ↔ suc + menaquinone

fum + h_2o ↔ mal

nadh + h + ea ↔ nad + eah_2

***Carnitine metabolism*:**

lc + crcoa ↔ lccoa + cr

lccoa ↔ crcoa + h_2o

Evaluation of fluxes of elementary modes through linear programming: Applied to *Corynebacterium glutamicum*

Kalyan Gayen[a] and K. V. Venkatesh[a b*]

[a]*Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai-400076, India.e-mai: gkalyan@iitb.ac.in*
[b]*School of BioSciences & Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai-400076, India. e-mail: venks@che.iitb.ac.in*

Keywords: Metabolic network analysis, Elementary modes, linear optimization, feasible solution space

1. Abstract

Evaluation of the fluxes in a metabolic network is essential for understanding the systematic behavior of the cellular processes. Several steady state methodologies have been developed using metabolic reactions of the network. Elementary mode analysis is a promising approach as it represents minimal subset of reactions in the network connecting the external metabolites. In the present study, we evaluated the fluxes of elementary modes for a well-studied organism, *C. glutamicum*, by imposing linear optimization under the constraint of stoichiometry of elementary modes. Analysis shows that stoichiometric matrix dimension reduced drastically as matrix includes only the coefficients of the external metabolites. Singularity problem will not arise using the basic network of this organism (as observed in the case of metabolic flux analysis). Lesser number of experimental measurements were required when the choice of the objective function was proper. The method was more labile over elemental balance to identify the experimentally measurement errors. Further, the feasible solution space was evaluated using the methodology for various oxygen and nitrogen availability in the environment.

2. Introduction

Recent flood of huge data by virtue of "omics" demands powerful theoretical methods to integrate the systematic analysis of the phenotypic state of the organism, which explain the relationship between the structure, function and regulation of an organism. A rigorous quantitative evaluation of cellular physiology is an essential step in metabolic engineering (Bailey, 1991, 1999). In this regard, intracellular metabolic fluxes are of great importance for metabolic engineers to gain insights into cell functioning and regulation and to derive conclusions on promising strain modifications (Forster et al., 2002). Powerful

modern tools of genetic engineering allow for effective creation of large numbers of mutants. Clearly, adequate quantitative analysis of metabolic properties of these mutants generates a demand for high-throughput experimental tools. Thus, powerful techniques have recently emerged for analysis of intracellular metabolite concentrations (Duetz et al., 1996; Fiehn et al., 2000; Roessner et al., 2001; Soga et al., 2002). For elucidation of metabolic fluxes, different approaches, using ^{13}C-labeled substrates, have been developed and applied to various biological systems (Marx et al., 1996; van Winden et al., 2003; Wittmann and Heinzle, 2002). Unfortunately all these experimental protocols are extensive laborious and cost effective. In metabolic level, the theoretical methods have been developed for simultaneously predicting key aspects of network functionality from network structure. This can be achieved by determining and analyzing the non-decomposable pathways able to operate coherently at steady state. Among the different approaches, elementary flux mode is the most promising which takes the flexibility of the network compared with flux balance (Steffen and Stelling, 2003). Elementary flux modes indicate the all possible routes on which organism can grow, although depending upon the objective of the organism because of environmental conditions specific elementary modes will be active so that organism can maximize the biomass or can maximize one of the products.

In the present work, we have constructed the elementary modes of *Corynebacterium glutamicum* and the fluxes of elementary modes were evaluated using linear optimization technique. Also, a feasible solution space has been constructed by this optimized method varying uptake rates of ammonia and oxygen, which will strengthen the possible strategies required to enhance the lysine productivity.

3. Methodology

For a given network, elementary modes can be generated by convex analysis and there are number of softwares available (Poolman at al., 2004). For the present work python based "ScrumPy" software (http://www.gnu.org/licenses/gpl.html) was used to evaluate the elementary modes connecting the external metabolites. Further, the accumulation rates of external metabolites can be represented using the coefficients of the elementary modes as describe below.

The method for the flux assignment of elementary modes can be demonstrated with the help of a simple illustrative example (Figure 1a). A system boundary (dotted line) is considered around all the internal metabolites and the system is closed for this type of metabolites and the fluxes between the internal metabolites are the internal fluxes. But external metabolites are allowed to enter or exit of that theoretical system boundary and exchange flux is the flux by which one external metabolite can enter into the system or one internal metabolite can exit from the system. The biochemical network consists of three internal metabolites (A, B, C) and three external metabolites (X_O, X_1, X_2). There are three exchange fluxes (one of them is reversible) and three internal fluxes (one of them is

reversible). The system consists of five elementary modes, which are depicted in Figure 1b. The extracellular metabolites are connecting from substrate to the products. Using these elementary modes, balance equations can be written that would be based on the stoichiometry of the reaction network. The fluxes of the external metabolites will be the in terms of fluxes in the elementary modes as given below:

$$-\frac{dX_O}{dt} = v_1 + v_2 + v_3 + v_4$$

$$\frac{dX_1}{dt} = 4v_2 + 4v_4 + 4v_5$$

$$\frac{dX_2}{dt} = v_1 + v_3 - v_4$$

In terms of matrix form, this can be represented as:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 4 & 0 & 4 & 4 \\ 1 & 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} dX_O/dt \\ dX_1/dt \\ dX_2/dt \end{bmatrix}$$

If the objective function is the maximization of dX_2/dt, the problem formulation will be as:

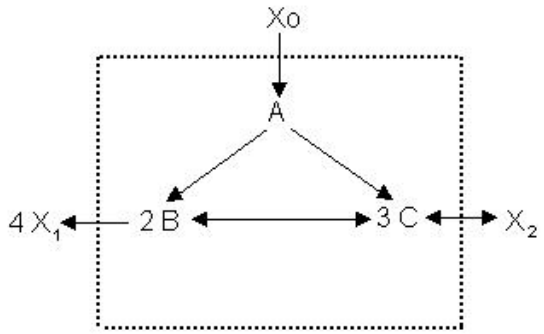
Objective function = maximize $(\frac{dX_2}{dt})$

Subject to

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 4 & 0 & 4 & 4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} dX_O/dt \\ dX_1/dt \end{bmatrix}$$

and $0 \le v_i \le \infty$ for all i^{th} elements

a)



b)

Fluxes of elementary modes	Reaction stoichiometry
v_1	$Xo \longrightarrow X_2$
v_2	$Xo \longrightarrow 4X_1$
v_3	$Xo \longrightarrow X_2$
v_4	$Xo \longrightarrow 4X_1$
v_5	$X_2 \longrightarrow 4X_1$

Figure 1. (**a**) A hypothetical reaction network consisting of three internal metabolites (A, B, C), three external metabolites (X_O, X_1, X_2), three exchange fluxes and three internal fluxes. Double-headed arrows indicate reversible reactions and single headed arrows indicate irreversible reactions. (**b**) The stoichiometric reactions of elementary modes of the hypothetical reaction network.

The right hand side of the matrix equations are the measurable quantities (known parameters), while the fluxes of elementary modes (v_i's) are the unknowns to be evaluated by means of linear programming. The above methodology was used to evaluate the fluxes of elementary modes for the network of *Corynebacterium glutamicum*.

4. Experimental

4.1 ORGANISM AND MATERIAL

Corynebacterium glutamicum (CECT 79) obtained from The Spanish Type Culture Collection (CECT), Valencia, Spain, was used for the experiment. HPLC grade water was purchased form Merck (Mumbai, India). All other chemicals were purchased form Hi-Media (Mumbai, India).

4.2 FERMENTATION PROTOCOL

The strain was cultivated and maintained as reported previously (Vallino, 1991). The seed culture was prepared in medium containing 5 g/L glucose, 5 g/L yeast

extract, 10 g/L tryptone and 5 g/L NaCl. One loop of the organism was inoculated from the slant into 250 ml triple baffle conical flask containing 50 ml of seed media. The seed culture was grown for 10 hrs at 150 rpm maintaining 30 ^{0}C temperature. Then, 15 ml of seed media transferred into 500 ml triple baffle conical flask containing 135 ml of preculturing media as reported by Vallino, (1991) for 8 hrs maintaining the same rpm and temperature. After growth of organism in preculturing medium, the culture was transferred into the fermentation medium (also reported by Vallino, 1991). Fermentation was initiated by inoculating the fermentation medium with 10% v/v precultured seed. Air flow rate was kept at 1 liter per minute per liter reactor volume and stirrer speed of 1000 rpm was maintained though out the experiment. pH was maintained at 7.0 by feeding ammonia.

4.3 ANALYTICAL TECHNIQUES

Samples were drawn in regular intervals during the course of fermentation to analyze the dry cell weight, glucose, trehalose, pyruvate, lysine, ammonium sulphate. Dry cell weight was estimated form the absorbance at 600 nm on spectrometer (V-540, Jasco, Tokyo, Japan). One unit of absorbance was equivalent to 0.28 g/L of dry cell weight. Glucose and trehalose was estimated by RI detector and pyruvate was analyzed by UV detector in HPLC (Hitachi, Merck, KgaA, Darmstndt, Germany) using HP-Aminex-87-H column (Biorad, Inc., Hercules, CA) at 60^0 C. The mobile phase in HPLC was 5 mM sulfuric acid and flow rate was maintained 0.6 mL per minute. The concentration of lysine was analyzed by HPLC method as reported by Pachuski et al. (2002). Ammonium sulfate was measured using ion analyzer (EA940 Ion analyzer, Thermo Orion, Beverly, MA).

5. Results and discussion

5.1 ELEMENTARY MODES OF THE NETWORK

The metabolic network *C. glutamicum* is relatively complex containing core metabolism of glycolytic pathway, tricarboxilic acid (TCA) cycle and Pentose phosphate pathway (PPP). Further, glucose is transported into the organism by active transport and carboxylation reaction from phosphoenol pyruvate (PEP) to oxaloacetate (OAA) plays an important role. Ammonia is consumed through amino acid synthesis and the balancing of NADH/NAD^+ and $FADH_2$/FAD are accounted via oxidative phosphorylation. Glyoxalate shunt is neglected because it is reported that this shunt is not active when this organism grows on glucose. Dinucleotide transhydrogenase reaction is also neglected because this reaction is not active in this organism (Moritz et al., 2002). Chemical reactions of this organism were used to evaluate the elementary modes as listed in appendix 1 and the fluxes of elementary modes were evaluated with the help of experimental accumulation rates of extracellular metabolites.

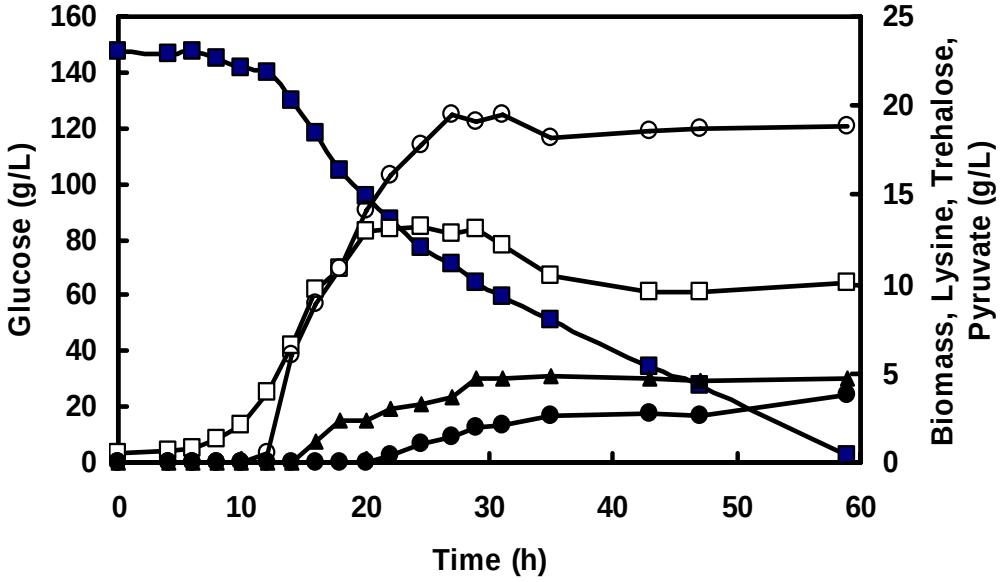


Figure 2. Control lysine fermentation of *Corynebacterium glutamucum* CECT. pH was maintained 7.0 by addition of ammonia. Fermentation media was described by Vallino (1991). Glucose (■), biomass (□), lysine (○), trehalose (▲), pyruvate (●) profiles during the course of fermentation.

Figure 2 shows the concentration profile of the different extracellular metabolites (glucose, biomass, lysine, trehaolse and pyruvate) during the course of fermentation. Biomass concentration reached steady state at 21 h, while lysine production started after 13 h and reached 20 g/L at 25 h. Synthesis of trehalose also started at the same time as lysine with a maximum value of 5 g/L, while pyruvate accumulation started later time (t = 20 h). The accumulation / uptake rates were evaluated by differentiating the concentration with respect to time and these rates were used to analyze the fluxes of the elementary modes.

The fourteen elementary modes of the network have been evaluated using *ScrumPy* software considering the uptakes of the external metabolites as glucose, oxygen, ammonia and accumulation of biomass, lysine, trehalose and carbon dioxide (Appendix 2). It is interesting to note that this organism was not able to operate at anaerobic condition as all the elementary modes are associated with oxygen. Lysine and biomass were simultaneously produced by two elementary modes (2, 11). Similarly, synthesis of lysine + trehalose was associated with three modes (‘1’, ‘9’ and ‘14’) and synthesis of biomass + trehalose was associated with two modes (‘3’, ‘5’). The fluxes of the elementary modes with the accumulation / uptake of the extracellular metabolites (see methodology), were obtained under the criteria of maximization of biomass during the course of fermentation.

Figure 3 shows the flux distribution of the elementary modes towards glucose, biomass and lysine at different time points (t = 11.5 h, 13.5 h, 15.8 h) during the course of fermentation.

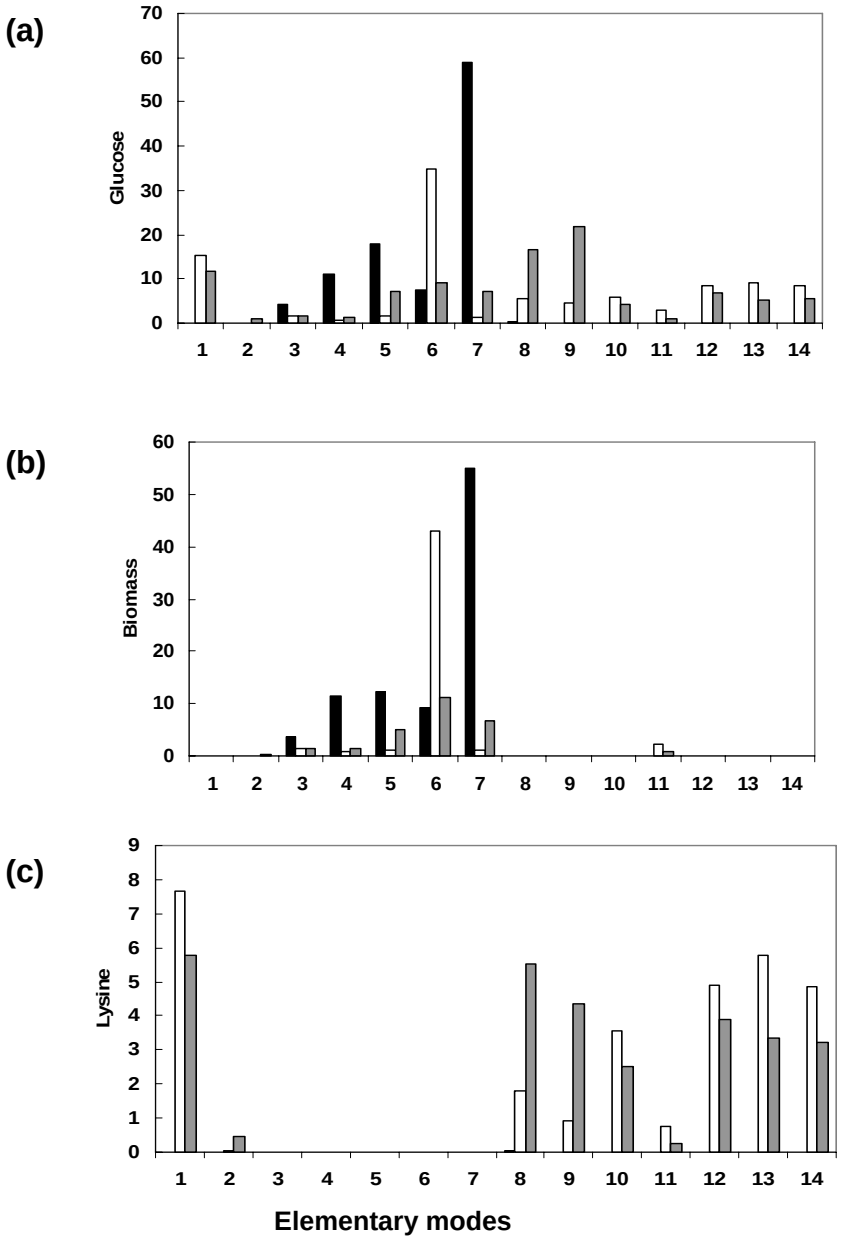


Figure 3. Histogram of the absolute fluxes of elementary modes for the metabolic network of *C. glutamicum*. The fluxes were estimated by linear programming through the objective function of maximization of biomass and stoichiometric coefficient of the elementary modes was as constraint. The fluxes were estimated at different time points (11.5 h, 13.5 h and 15.8 h) during the course of fermentation. Black box indicates the distribution at 11.5 h; white box indicates at 13.5 h and grey box at 15.8. Fluxes are in normalized scale with respect to glucose (100). **(a)** Fluxes of elementary modes associated with glucose; **(b)** Fluxes of elementary modes associated with biomass; **(c)** Fluxes of elementary modes associated with lysine. The number on the x-axis indicates the serial number of the elementary modes as indicated in Appendix 2.

The fluxes were estimated with respect to the normalized value of glucose (glucose = 100). The uptake of glucose was though the elementary modes associated with only the biomass formation at t = 11.5 h and elementary mode '7' (this mode contained glycolysis and TCA cycle) contributed to 65 % of the total glucose uptake rate (Figure 3a). Almost all the modes switched on during the later time of fermentation and the contribution of mode '6' (this mode connected through PPP) was highest (around 36 %) at 13.5 h. The glucose uptake rates depended on the lysine producing modes at 15.8 h indicating that lysine synthesis rate was predominating over biomass formation. Similarly, biomass synthesis rate was depended on the elementary modes associated with biomass formation at t =11.5 h. In this case, mode '7' contributed to 60 % of the biomass formation at t = 11.5 h and mode '6' contributed most of the biomass formation at t = 13.5 h (Figure 3b). The biomass formation also depended on the modes associated with lysine production at later phase of fermentation (t = 15.8 h). The lysine production modes were switched off at early phase of fermentation (t = 11.5), while in later phase of fermentation, lysine production modes contributed dominantly and only biomass formation modes were inactive indicating that there is an on/off switch between lysine producing modes and biomass producing modes.

5.2 SIMULATION OF THE NETWORK

Feasible range of the external metabolites were obtained by varying the normalized oxygen consumption rate or by varying the normalized ammonia consumption rate (assuming glucose consumption rate to be100) under the criteria of an assumed objective function. It was observed that feasible normalized oxygen consumption rate was in the set 146 – 366 and feasible normalized ammonia consumption rate was in the set 40-126.

Figure 4a shows the response of external metabolites with variation of normalized ammonia uptake rate keeping glucose uptake rate constant (100) under the criteria of maximizing biomass. Biomass synthesis increased up to a peak value (121.6) at normalized ammonia consumption rate of 90 and then gradually declined to zero at normalized ammonia uptake rate equal to 126. Lysine synthesis rate started after the normalized ammonia uptake rate was greater than 90 and reached a peak value at normalized ammonia uptake rate equal of 126. This indicates that nitrogen and NADH/ NAD^+ balance can not be compensated by biomass synthesis at high (> 90) ammonia uptake rate resulting in lysine synthesis. Further, trehalose synthesis rate continuously decreased with normalized ammonia uptake rate and reached a zero value at ammonia uptake value of 90. The production rate of carbon dioxide and oxygen were fluctuating depending upon NADH/ NAD^+ load during product synthesis and reached to a minimum when biomass synthesis rate was maximum. It is interesting to note that that carbon dioxide evolution rate was higher compare to the oxygen uptake rate when lysine synthesis started due to the activation of PPP for requirement of NADPH for lysine synthesis. Biomass synthesis was zero when the criterion of maximization of lysine was used and lysine production rate was high at a maximum value of normalized ammonia uptake rate (Figure 4b).

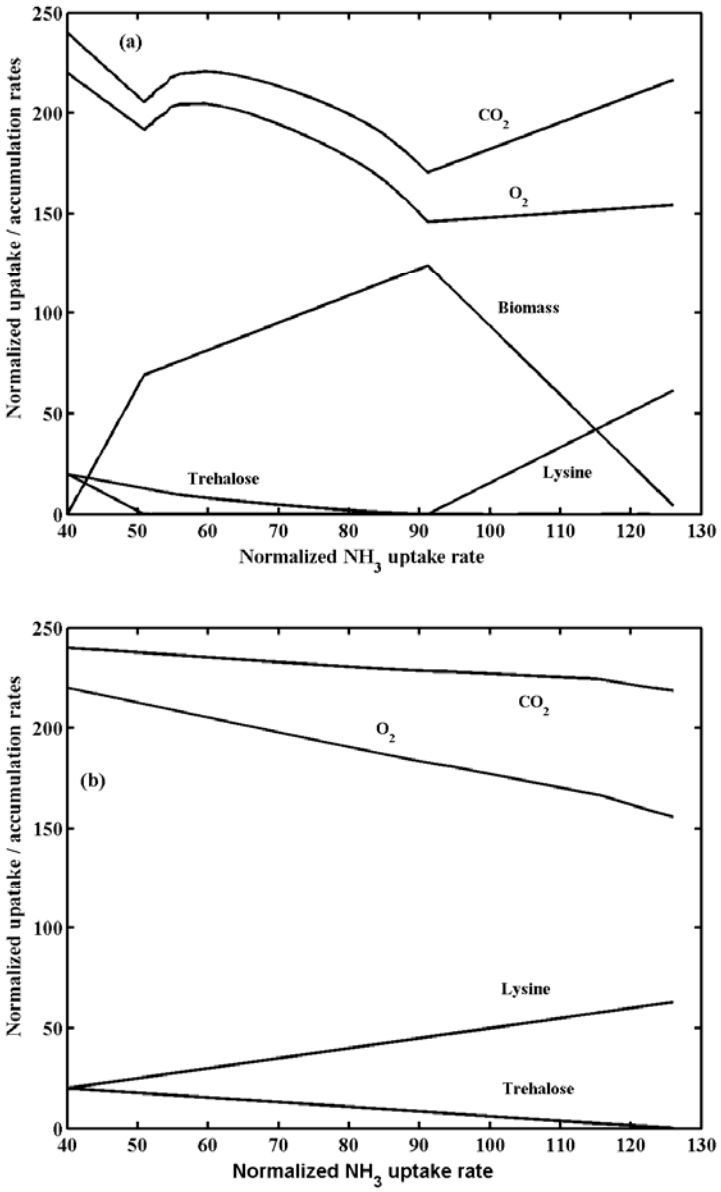


Figure 4. Optimized solution by varying ammonia consumption rate for the different external metabolites of *C. glutamicum.* The inputs of LP optimizer were Glucose and NH_3. **(a)** Objective function: maximization of biomass. **(b)** Objective function: maximization of lysine.

The response of the external metabolites with variation of normalized oxygen uptake rate under the criteria of maximization of biomass synthesis is shown in Figure 5a. In this case, biomass accumulation rate was highest (123) at an uptake rate of normalized oxygen of 146 and deceased to zero at oxygen uptake rate of 336. Lysine production started at normalized oxygen consumption rate equal to 260 and reached 33.3 at maximum feasible value of oxygen uptake. The accumulation rates were also evaluated under the criterion of lysine maximization at various oxygen uptake rates (Figure 5b). Lysine synthesis rate was at the maximum (63.3) and biomass synthesis rate was zero when normalized oxygen consumption rate was equal to 155. The carbon dioxide rate was maximum and reached 400 due to the activation of PPP.

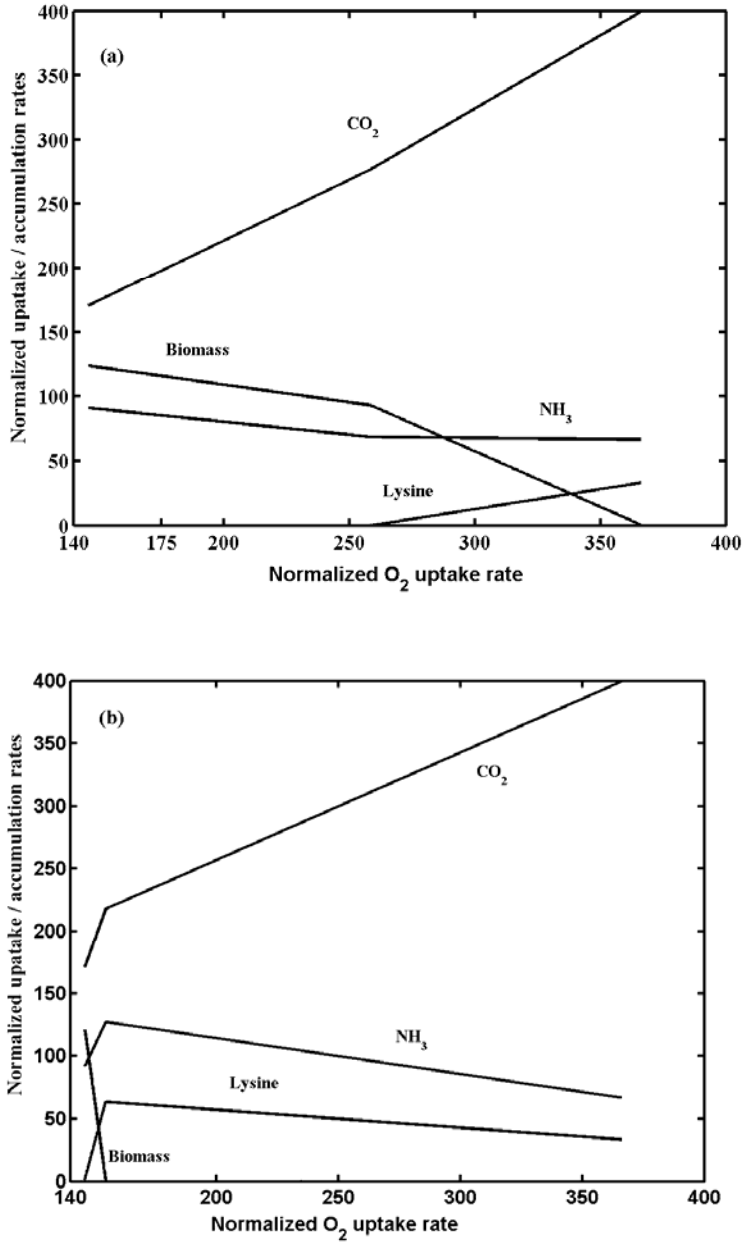


Figure 5. Optimized solution by varying oxygen consumption rate for the different external metabolites of *C. glutamicum.* The inputs of LP optimizer were Glucose and NH_3. **(a)** Objective function: maximization of biomass. **(b)** Objective function: maximization of lysine

6. Conclusion

We have demonstrated the use of elementary modes with the help of linear optimization technique in quantifying metabolic networks and this methodology has been applied for the quantification of the network of *C. glutamicum*. In *C. glutamicum*, the elementary modes associated with biomass formation were operational at the initial experimental growth phase and later phase of fermentation lysine synthesis switch on. The methodology was also used to determine the feasible solution space for a given substrate uptake rate. Such an approach is generic in nature and can be used to determine the optimality of the accumulation rates of a metabolite in any given system.

References

Bailey, J. E. (1991) Toward a science of metabolic engineering. *Science* **252,** 1668–1675.

Bailey, J. E. (1999) Lessons from metabolic engineering for functional genomics and drug discovery. *Nat Biotechnol*. **7,** 616–618.

Duetz, P., Kumps, A. and Mardens, Y. (1996) GC-MS profiling of urinary organic acids evaluated as a quantitative method. *Clin Chem*. **42,** 1609–1615.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N. and Wilmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat Biotechnol*. **18,** 1157–1161.

Fo¨rste,r J., Gombert, A. K. and Nielsen, J. (2002) A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol Bioeng*. **79,** 703–712.

Marx, A., de Graaf, A. A., Wiechert, W., Eggeling, L. and Sahm, H. (1996) Determination of the fluxes in the central metabolism of Corynebacterium glutamicum by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnol Bioeng*. **49,** 111–129.

Moritz, B., Striegel, K., Graaf, A. A. and Sahml, H. (2002) Changes of Pentose Phosphate Pathway Flux in Vivo in Corynebacterium glutamicum during Leucine-Limited Batch Cultivation as Determined from Intracellular Metabolite Concentration Measurements. *Metabolic Engineering*. **4**, 295–305.

Pachuski, J., Fried, B. and Sherma J. (2002) HPTLC analysis of amino acids in *Biomphalaria Glabrata* infected with *Schistosoma mansoni*. *J liquid chromatography & related technology* **25**, 2345-2349.

Poolman, M. G., Venkatesh, K. V., Pidcosk., M. K., and Fell, D. A. (2004) A Method for the Determination of Flux in Elementary Modes, and its Application to *Lactobacillus rhamnosus*. *Biotech. and Bioeng.* **88**, 601-612.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Link, T., Willmitzer, L. and Fernie, A. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*. **13,**11–29.

Soga, T., Ueno, Y., Naraoka, T., Ohashi, Y., Tomita, M. and Nishiokat, T. (2002) Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal Chem*. **74,** 2233–2239.

Steffen, K. and Stelling, J. (2003) Two approaches for metabolic pathway analysis? *Trends in Biotechnology*. **21,** 64-68.

Vallino, J. J., (1991) Identification of branch-point restrictions in microbial metabolism through metabolic flux balance analysis and local network perturbations. *Ph.D. Thesis*. Massachusetts Institute of Technology, USA.

van Winden, W. A., Van Gulik, W. M., Schipper, D., Verheijen, P. J., Krabben, P., Vinke, J. L. and Heijnen, J. J. (2003) Metabolic flux and metabolic

network analysis of Penicillium chrysogenum using 2D [13C, 1H] COSY NMR measurements, and cumulative bondomer simulation. *Biotechnol Bioeng*. **83,** 75–92.

Wittmann, C. and Heinzle, E. (2002) Genealogy profiling through strain improvement using metabolic network analysis—metabolic flux genealogy of several generations of lysine producing corynebacteria. *Appl Environ Microbiol*. **68,** 5843–5859.

Appendix 1

Chemical equations of *C. glutamicum*

The chemical equations representing the metabolic network of *C. glutamicum* as follows, where '=' represents the reversible reaction and '>' represents the irreversible reaction (PhD thesis of Vallino, 1991).

Glucose Phosphotransferase System

1. GLC + PEP > GLC6P + PYR

Storage Compounds; Trehalose

2. 2 GLC6P + ATP = TREHAL + ADP

EMP Pathway

3. GLC6P = FRU6P
4. FRU6P + ATP > 2 GAP + ADP
5. GAP + ADP + NAD = NADH + G3P + ATP
6. G3P = PEP + H2O
7. PEP + ADP > ATP + PYR
8. PYR + NADH = LAC + NAD

Carboxylation reaction

9. PEP + CO2 > OAA

TCA Cycle

10. PYR + COA + NAD > ACCOA + CO2 + NADH
11. ACCOA + OAA + H2O = ISOCIT + COA
12. ISOCIT + NADP = AKG + NADPH + CO2
13. AKG + COA + NAD > SUCCOA + CO2 + NADH
14. SUCCOA + ADP = SUC + COA + ATP
15. SUC + H2O + FAD = MAL + FADH
16. MAL + NAD = OAA + NADH

Acetate Production or Consumption

17. ACCOA + ADP = AC + COA + ATP

Glutamate, Glutamine, Alanine, and Valine Production

18. NH3 + AKG + NADPH = GLUT + H2O + NADP
19. GLUT + NH3 + ATP > GLUM + ADP
20. PYR + GLUT > ALA + AKG
21. 2 PYR + NADPH + GLUT > VAL + CO2 + H2O + NADP + AKG

Pentose Phosphate Pathway

22. GLC6P + H2O + 2 NADP > RIBU5P + CO2 + 2 NADPH
23. RIBU5P = RIB5P
24. RIBU5P = XYL5P
25. XYL5P + RIB5P = FRU6P + E4P
26. XYL5P + E4P = FRU6P + GAP

Oxidation Phosphorylation

27. 2 NADH + O2 + 4 ADP > 2 H2O + 4 ATP + 2 NAD
28. 2 FADH + O2 + 2 ADP > 2 H2O + 2 ATP + 2 FAD

Asparate Amino Acid Family

29. OAA + GLUT = ASP + AKG

30. ASP + PYR + 2 NADPH + ATP > AKP + 2 NADP + ADP + H2O
31. AKP + SUCCOA + H2O + GLUT > MADP + COA + AKG + SUC
32. MDAP > LYSI

ATP Dissipation

33. ATP > ADP

Biomass Synthesis

34. 21 GLC6P + 7 FRU6P + 126 RIB5P + 13 GAP + 150 G3P + 52 PEP + 30 PYR + 332 ACCOA + 80 ASP + 33 LYSI + 446 GLUT + 25 GLUM + 54 ALA + 40 VAL + 100 NADPH > 1000 BIOMASS + 364 AKG + 143 CO2 + 100 NADP

Appendix 2

Sl. No	Reaction Stoichiometry of the elementary modes
1	192 GLC + 336 O2 + 192 NH3 → 12 TREHAL + 96 LYSI + 816 H2O + 432 CO2
2	11892 GLC + 18237 O2 + 14552 NH3 → 6540 LYSI + 49808 H2O + 2000 BIOMAS + 25174 CO2
3	912216 GLC + 1637634 O2 + 562304 NH3 → 86141 TREHAL + 3938420 H2O + 764000 BIOMAS + 1789288 CO2
4	369967 GLC + 818817 O2 + 281152 NH3 → 1969210 H2O + 382000 BIOMAS + 894644 CO2
5	28920 GLC + 55470 O2 + 14720 NH3 → 3725 TREHAL + 125780 H2O + 20000 BIOMAS + 59440 CO2
6	1186920 GLC + 1730295 O2 + 1081920 NH3 → 5020680 H2O + 1470000 BIOMAS + 2022090 CO2
7	10735 GLC + 27735 O2 + 7360 NH3 → 62890 H2O + 10000 BIOMAS + 29720 CO2
8	75 GLC + 275 O2 + 50 NH3 → 25 LYSI + 550 H2O + 300 CO2
9	20 GLC + 44 O2 + 8 X_NH3 → 4 TREHAL + 4 LYSI + 88 H2O + 48 CO2
10	18 GLC + 31 O2 + 22 NH3 → 11 LYSI + 80 H2O + 42 CO2
11	21040 GLC + 31384 O2 + 22112 NH3 → 5168 LYSI + 88608 H2O + 16000 BIOMAS + 39728 CO2
12	56 GLC + 112 O2 + 64 NH3 → 32 LYSI + 272 H2O + 144 CO2
13	44 GLC + 68 O2 + 56 NH3 → 28 LYSI + 184 H2O + 96 CO2
14	38 GLC + 62 O2 + 44 NH3 → TREHAL + 22 LYSI + 160 H2O + 84 CO2

Constraint-based *in silico* modelling of the Fe(III)-reducing bacteria *Geobacter sulfurreducens*: insights into the subsurface microbial activity

R. Mahadevan[a,b,*], A. Esteve-Núñez[a,c,*] , D. R Bond[a] , J. E. Butler[a], M. V. Coppi[a], and D. R. Lovley[a]

[a]*Department of Microbiology, University of Massachusetts, Amherst, MA, US. mahadevan@chem-eng.utoronto.ca*
[b]*Genomatica, San Diego, California, US*
[c]*Centro de Astrobiología, INTA, Madrid, Spain. estevena@inta.es*

Keywords: *in silico* cell model, constraint-based model, genome-scale model, *Geobacter*, iron reducing bacteria, metal bioremediation.

1. Abstract

Here we describe the application of the constraint-based modeling approach, coupled in an iterative fashion with experimental studies, to further elucidate the physiology of *Geobacter sulfurreducens*, a well-studied representative of the Geobacteraceae, which play a critical role in organic matter oxidation coupled to Fe(III) reduction, bioremediation of groundwater contaminated with organics or metals, and electricity production from waste organic matter. The completed reconstructed metabolic network of *G. sulfurreducens* contained 588 genes (or 17% of a total of 3,467 ORFs), 522 biochemical reactions, and 541 unique metabolites. Examination of the reconstructed metabolic network revealed that *G. sulfurreducens* has multiple reactions for acetate utilization, the main electron-donor for these bacteria in the subsurface. Simulations fit well with experimental data obtained from chemostat studies, predicting different flux rates and growth yield under a number of growth rates. Evaluation of the rates of proton production and consumption in the extracellular and cytoplasmic compartments revealed the energy conservation with extracellular elelctron acceptors as Fe(III), was limited compared to intracellular acceptors as fumarate. These results demonstrate that iterative modeling coupled with experimentation can accelerate the understanding of the physiology of poorly studied but environmentally relevant organisms and may help optimize their practical applications.

2. Introduction

The constraint-based approach to modeling microbial metabolism has proven to be an effective strategy for predicting the physiological responses of microorganisms (Price et al, 2003). This approach relies on implementing a

series of physico-chemical constraints including thermodynamic directionality, and enzymatic capacity constraints and reaction stoichiometry constraints arising from the requirement that fluxes consuming and producing both metabolites and protons are balanced. This systems approach to microbial physiology has the ability to predict the metabolic response of organisms to various environmental conditions without the need for information on kinetic parameters for each of the individual reactions (Edwards and Palsson, 2000). Substrates that can be metabolized and the nutrients that are required from the environment to support growth can be successfully predicted, as can growth rates under various conditions.

Although all those models have been limited to *Escherichia coli* and pathogens (Edward and Palsson, 1999; Edward and Palsson, 2000; Schilling et al., 2002) , this methodology should be able to predict the behaviour of microorganisms in more remote environments where they are of geomicrobiological relevance. Here we describe the application of this constraint-based modeling approach, coupled in an iterative fashion with experimental studies, to further elucidate the physiology of *Geobacter* species (Mahadevan *et al*., 2006*)*, the first organisms found to have the ability to conserve energy for growth by completely oxidizing organic compounds to carbon dioxide with Fe(III) serving as the electron acceptor (Lovley et al., 1987; Caccavo et al., 1994). In addition to transferring electrons to *Fe(III), *Geobacter* species can also reduce a variety of toxic and radioactive metals (Lovley et al., 1991; Lloyd et al., 2000; Ortiz-Bernad et al., 2004). Moreover, stimulating the activity of *Geobacter* species in the subsurface is an effective strategy for removing such contaminants from groundwater (Lovley et al. 1994). Another practical application of *Geobacter* species is their ability to oxidize organic compounds with an electrode serving as the electron acceptor (Bond et al., 2002; Bond and Lovley, 2003), which makes it possible to harvest electricity from waste organic matter.

To develop this kind of model, a complete sequenced genome of the microorganism is required, thus *Geobacter sulfurreducens* is the best candidate to be modelled because it is closely related to the environmental strains isolated from the subsurface and its genome had been recently sequenced (Methe et al., 2003). In addition, a chemostat system has been developed (Esteve-Núñez et al., 2005), to further evaluate *in silico* predictions with well established growth conditions. Modelling growth and metabolism under relevant environmental conditions could provide an insight into the factors that might be limiting the rate and extent of bioremediation processes at contaminated sites.

3. Theoretical

This section provides a brief introduction to the constraint-based modelling approach that has been extensively reviewed elsewhere (Price et al., 2004). In this work, we have used the flux balance analysis approach which assumes that

* These two authors have equally contributed to this manuscript.

the cellular objective is growth maximization, to calculate the flux distribution in the metabolic network given the input and output fluxes of substrates exchanged across the membrane.

3.1. Flux Balance Analysis (FBA):

FBA is an analysis tool to quantitatively investigate the systemic properties of a metabolic network. It is based on material balances for each of the internal metabolites and the assumption of optimal growth as the objective of the cell. The details of FBA and the significance of the objective function have been reviewed earlier . The FBA formulation includes a series of linear equations (material balances) and a linear objective function with flux through the reactions as the independent variables as shown below:

$$\text{Max} \quad c^T v$$

$$S.v = 0$$

$$\alpha \le v \le \beta$$

These equations (see symbols section) are solved along with constraints on the fluxes and an objective defined in terms of the biomass growth rate (based on the biomass composition) using Linear Programming (LP) techniques in the SimPheny platform.

4. Experimental

4.1. Strain and Culturing Conditions: Wild type *Geobacter sulfurreducens* (ATCC 51573) was obtained from our laboratory collection. *G. sulfurreducens* was grown in ia chemostat under continuous culture and strict anaerobic conditions at 30 °C using previously described method (Esteve-Núñez et al., 2005). Sodium acetate (5.5mM) was used as sole electron donor, and either sodium fumarate (30mM) or Fe(III)-citrate (60mM) were used as electron acceptor in a bicarbonate-buffered freshwater medium. Organic acids content in culture supernatant were monitored by HPLC as previously described (Esteve-Núñez et al., 2005), and Fe(II) was determined as described by Lovley and Phillips (1982).

4.2. The genome-scale metabolic model for *G. sulfurreducens* was developed using the constraint-based modeling approach (Bonarius *et al*., 1997) and the SimPheny^{TM} (Genomatica, San Diego, CA) platform (Mahadevan *et al*. 2006). BLAST searches of publicly available databases (Overbeek et al., 2000) resulted in the identification of 588 genes (or 17% of a total of 3467 ORFs). The completed reconstructed metabolic network contained 522 biochemical reactions, and 541 unique metabolites. These reactions were further refined using published biochemical and physiological information. To allow full stoichometric balancing, all reactions were entered into the model database as balanced reactions, including the net charge of each metabolite or cofactor and the localization (cytoplasmic or extracellular) of reactants and products. For all simulations presented in this report, all genes included in the network were

assumed to be expressed and their associated reactions functional. Maximization of biomass production (growth) was the objective for all the simulations. The complete list of genes, reactions, applied constraints, and confidence scores is available at the following website (http://www.geobacter.org/).

5. Results and Discussion

5.1. Evaluation of the proton translocation stoichiometry give insights into the growth yield during Fe(III) and fumarate reduction.

Geobacter sulfurreducens is able to use either the metal Fe(III) or the dicarboxylate acid fumarate to conserve energy from acetate oxidation. However, these two respiratory mechanism are quite different, fumarate reduction is an intracellular and well characterized process catalyzed by the enzyme FrdCAB (Butler et al., 2006), while the biochemical mechanism responsible of Fe(III) reduction is much more complex with a high number of cytochomes c involved in electron transport(Leang et al, 2001; Lloyd et al., 2000; Butler et al., 2004; Kim et al., 2005, 2006).

Fe(III) reduction was first modelled as a reaction that occurred outside the cell, consistent with the fact that insoluble Fe(III) oxides are the predominant form of Fe(III) in most soils and sediments (Lovley 1991). Under Fe(III)-reducing conditions, the TCA cycle operated as a closed loop (Galushko and Schink, 2000) and produced 8 electrons per mole of acetate oxidized. However, model simulations using this electron transport scheme indicated that cells would not be capable of growth (*in silico)* under Fe(III)-reducing conditions.

The inability of a single $2H^+/2e^-$ NADH dehydrogenase coupling site to support simulated Fe(III)-dependent growth was traced to the fact that the site of Fe(III)-reduction was extracellular. The cytoplasmic protons that were produced from each mole of acetate oxidized in the cytoplasm were consumed in the cytoplasm when fumarate was the electron acceptor (Figure 1). In contrast, during Fe(III) reduction, electrons were transported outside the cell, while leaving protons in the cytoplasm, effectively dissipating the membrane potential and acidifying the cytoplasm (Figure 1). In order to generate sufficient energy to compensate for the production of protons in the cytoplasm, an additional coupling step was required.

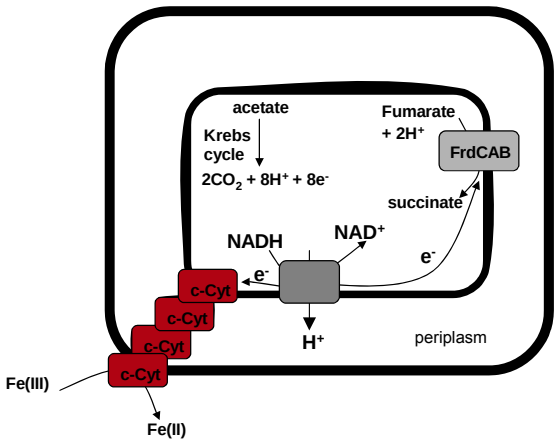


Figure 1. A model for proton and electron consumption in fumarate and Fe(III) reduction by *Geobacter sulfurreducens*

The most likely mechanism for additional membrane potential generation during Fe(III)-reduction was during transfer of electrons into the periplasmic cytochrome pool. Based on the fact that cytochromes implicated in Fe(III) reduction have midpoint potentials in the range of –190 mV (*omcB*) (Magnuson et al., 2000) and –136 to –155 mV (*ppcA*) (Lloyd et al., 2002), the energy available for coupling at this site could support translocation of $1H^+/2e^-$. This reaction was modeled as the release of menaquinol protons back to the cytoplasm by a protein capable of translocating 1 H^+ per pair of electrons transferred to the cytochrome pool. Inclusion of this reaction and accounting for all the protons produced and consumed during metabolism, resulted in a theoretical maximum yield with Fe(III) as the electron acceptor of 0.5 mol ATP/mol acetate as compared to the 1.5 mol ATP/mol acetate during fumarate reduction. This output of the model provides an explanation for the experimental finding that growth yields of *G. sulfurreducens* are ca. three-fold higher when fumarate (E_h=0.03V) serves as the terminal electron acceptor versus growth with Fe(III)-citrate (E_h=0.37V) (Esteve-Núñez et al., 2004), in spite of the higher redox potential of the metal. This result is unexpected because it is generally accepted that those electron acceptors with higher redox potential show a more negative Gibbs free energy and subsequently support higher yield (Unde and Bongaerts, 1997).
These results suggest that reducing extracellular electron acceptors such as Fe(III) oxides, Fe(III)-citrate, elemental sulfur (S°), or electrodes will result in the generation of less biomass per electron transferred than growth with intracellularly reduced electron acceptors. This may be an important consideration for applications such as bioremediation and electricity harvesting from waste organic matter, in which electron transfer to metals or electrodes, rather than production of biomass, is the primary goal.

5.2. Growth yield predictions fit with the experimental data.
The metabolic reaction network, combined with demand reactions for biomass synthesis, correctly predicted growth yields and acetate consumption rates for growth in standard acetate-limited chemostats with Fe(III)-citrate or fumarate as the electron acceptor (Table 1).

Table 1: *in silico* prediction and experimental values for growth parameters of *G. sulfurreducens* growing under Fe(III)/fumarate-respiring conditions.

Growth parameter with Fe(III) as TEA	***in silico*** **($0.05h^{-1}$)**	**experimental ($0.05h^{-1}$)**
$Y_{acetate}$(gdw /mol acetate)*10^3	4.5	3.5
$q_{electron}$ (mol/g dw h) *10^3	83.2	107.61

Growth parameter with Fumarate as TEA	***in silico*** **($0.05h^{-1}$)**	**experimental ($0.05h^{-1}$)**
$Y_{acetate}$(gdw /mol acetate) *10^3	11. 5	11.5
$q_{fumarate}$ (mol/g dw h) *10^3	16.425	19.21

Perturbations in variables used to construct the model, such as the biomass composition, which was derived from batch cultures of fumarate grown cells, had minimal effect on predicted acetate consumption for *G. sulfurreducens* under a number of growth rates. For instance, when a range of biomass composition equations (e.g., reflecting a range from 0.40 g protein/ g dw to 0.55 g protein/ g dw), were incorporated into the model, predicted yields were not significantly affected (1.5-2.5 % differences) (Fig. 2). This revealed that the model was robust to changes in biomass composition and nutrient availability, and was consistent with other work showing that variations in biomass composition produce only subtle effects on predicted growth yields or fluxes through central metabolic pathways (Pramanik and Keasling, 1998; Daae et al., 1999). Hence, it is possible to assume that even significant changes (10-20 %) in biomass composition would not affect the nature of metabolic predictions.

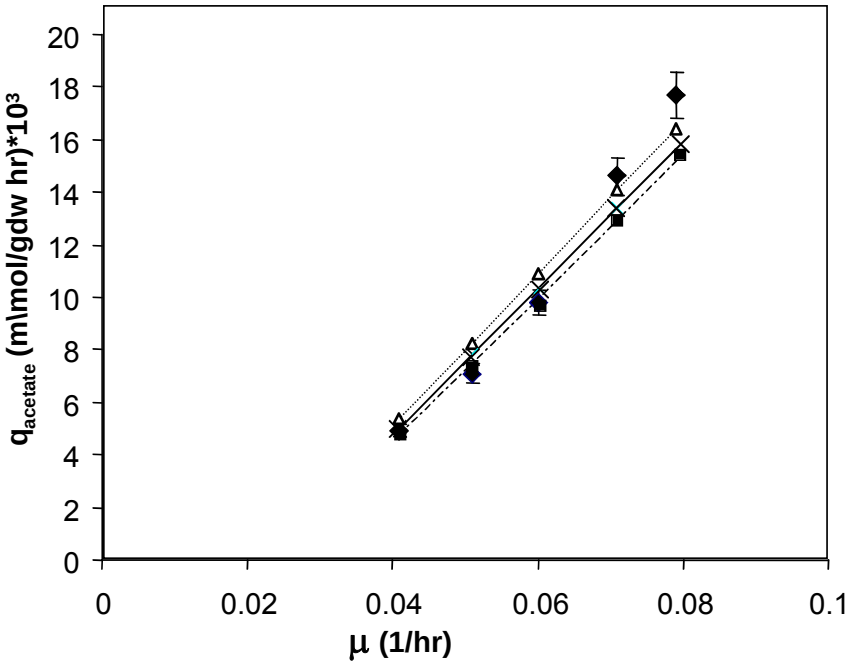


Figure 2. Acetate uptake predictions under different growth rate with Fe(III) as TEA. Experimental data: (◆), model prediction 40% protein (✕), model prediction 46% protein (■), model prediction 55% protein (△),

5.3. Model-based characterization of acetate metabolism in *Geobacter sulfurreducens*

The ability to oxidize acetate is important because acetate is the central intermediate in the anaerobic degradation of organic matter in sedimentary environments (Lovley and Chapelle, 1995). *Geobacter* species metabolize acetate via the tricarboxylic acid cycle (TCA) cycle (Champine and Goodwin, 1991; Galushko and Schink, 2000). In addition, it has been found that injection of acetate into groundwater to stimulate the uranium bioremediation activity of

Geobacter results in this microbial genus becoming the most abundant one in those environments (Anderson et al., 2003; Vrionis et al, 2005). Thus, an extensive analysis of acetate metabolism in *Geobacter* is desired.

Examination of the reconstructed metabolic network revealed that *G. sulfurreducens* has multiple pathways for acetate utilization (acetyl CoA transferase, acetate kinase, and phosphotransacetylase), interconversion of pyruvate to acetyl-CoA (pyruvate formate lyase, pyruvate ferredoxin oxidoreductase, and pyruvate dehydrogenase), and anapleurotic reactions (phosphoenolpyruvate carboxykinase and pyruvate carboxylase), as shown in Figure 3.

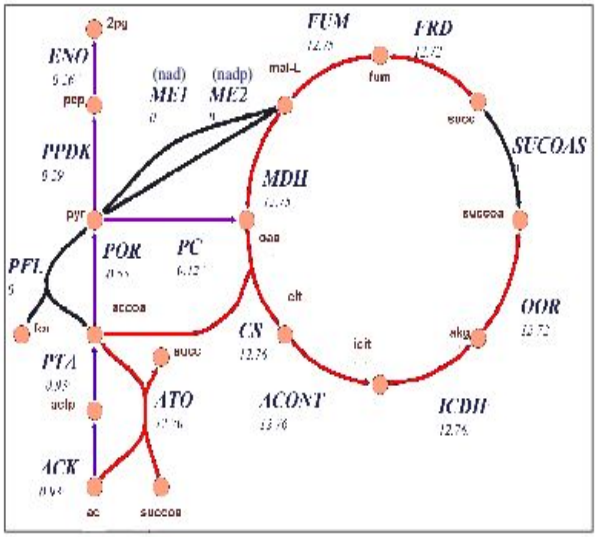


Figure 3. Predicted flux distribution (mmol/gdw h) through central metabolism in *G. sulfurreducens* during *in silico* growth with limiting acetate and excess Fe(III)-citrate. Ac, acetate; fum, fumarate; succ, succinate; succoA, succinyl-CoA;. actp, acetylphosphate; for, formate; accoA, acetyl-CoA;pyr. Pyruvate; pep, phosphoenolpyruvate; cit, citrate; icit, isocitrate; akg, alpha-ketoglutarate; mal, malate; oaa, oxalacetate; ACK, acetate kinase; ATO, acetyl-CoA transferase; PTA, phosphate transacetylase; PFL, pyruvate formate lyase; POR, pyruvate oxidoreductase; PC, pyruvate kinase; PPDK, pyruvate phospate dikinase; ENO, enolase; CS, citrate synthase; ACONT, aconitase; ICDH, isocitrate dehydrogenase; OOR, oxoglutarate oxidoreductase; SUCOAS, succinyl-CoA synthetase; FRD, fumarate reductase; FUM, fumarase; ME, malic enzyme.

Flux from acetyl-CoA to pyruvate via pyruvate-ferredoxin oxidoreductase was predicted to be the sole source of carbon fixation in *G. sulfurreducens*, and *in silico*, 4% of consumed acetate (0.55 mmol/g dw/h) was utilized in this fixation reaction when Fe(III)-citrate was the electron acceptor. Simulations predicted that during acetate-limited growth with Fe(III)-citrate (acetate uptake rate of 13.63 mmol/gdwh for a growth rate of 0.06 hr^{-1}), 93.6 % of all acetate transported into the cell was utilized for oxidation and ATP generation via the TCA cycle which fit well with experimental data from chemostat cultures (Esteve-Núñez et al., 2005).

5.4. Functional analysis of *G. sulfurreducens* mutant phenotypes

The availability of a genome scale model also enabled the characterization of systems level properties of the metabolic network. One such property is the set of genes and reactions that are essential to support growth in a defined medium. This information is important for genetic investigations as it can provide insight into which mutations may or may not have an observable phenotype.

In silico deletion analysis (Edward and Palsson, 2000) for growth with acetate as the electron donor and Fe(III)-citrate or fumarate as the electron acceptor indicated that most mutations were predicted to have either lethal (139 for fumarate, 143 for Fe(III)) or silent phenotypes (440 for fumarate, 437 for Fe(III)) (Table 2). Lethal mutations (*e.g.*, deletion of acetyl-CoA transferase and pyruvate carboxylase) reflected the inability of the perturbed network to synthesize essential components from acetate, a relatively simple two-carbon compound, or the fact that a non-fermentable substrate such as acetate presents few alternative energy-yielding oxidative mechanisms.

Some silent phenotypes predicted by this analysis corresponded to reactions associated with seemingly redundant enzymes. The presence of functionally similar (but non-orthologous) enzymes could be due to selection for genetic robustness, in order to protect against mutations in essential reactions. Alternatively, this redundancy could reflect a need for metabolic robustness, where different enzymes are needed to favor flux in opposite directions, or are optimized for oxidation of different substrates. For instance, model simulations indicated that a mutation in any component of pyruvate-ferredoxin oxidoreductase would be compensated by activity of pyruvate dehydrogenase or pyruvate-formate-lyase. However, as pyruvate-formate-lyase strongly favors function in the oxidative direction, it is unlikely that this enzyme can substitute for pyruvate-ferredoxin oxidoreductase *in vivo*, and the redundancy at this node likely reflects the presence of enzymes specialized for different tasks. Mutational and biochemical investigations are underway to test these hypotheses.

Table 2. Impact of *in silico* deletion of entire reactions, on predicted growth rate of *G. sulfurreducens*.

Growth conditions	% lethal deletions	% intermediate deletions	% silent deletions
Acetate&fumarate	40	4	58
Acetate&Fe(III)	41	3	58

6. Conclusion

These results suggest that genome-based *in silico* modelling can provide important insights into the physiology of environmentally relevant organisms, such as *Geobacter* species. Not only may such *in silico* models aid in understanding the likely physiological responses of *Geobacter* species in environments in which they are important, but the models can serve as a guide for evaluating the likely outcome of various possible strategies for genetically engineering *Geobacter* species in order to improve practical applications such as bioremediation and electricity production. Furthermore, the coupling of genome-based *in silico* models with hydrological/geochemical models may make it possible to predictively model subsurface bioremediation strategies prior to implementation (Lovley 2003) and coupling such models with electrochemical models is likely to enhance the development of microbial fuel cells (Lovley, 2006).

Acknowledgements*. This research was funded by the Genomics: GTL Program, US Department of Energy (Grant DE-FC02-02ER63446). A.E.N. was the recipient of a Postdoctoral Fellowship from the Secretaría de Estado de Educación y Universidades (Spain), co-funded by the European Social Fund.*

Symbols

TEA	Terminal Electron Acceptor	
$q_{electron}$	Respiration rate	mol/g dw h
$q_{acetate}$	Acetate consumption rate	mol/gdw h
$Y_{acetate}$	Growth Yield	g dw/ mol acetate
S	Stoichometric matrix	
v	Vector of the reaction fluxes	

GREEK LETTERS

μ	Growth rate	h^{-1}
α	Lower bound	
β	Upper bound	

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.

Anderson, R.T., Vrionis, H.A., Ortiz-Bernad, I., Resch, C.T., Long, P.E., Dayvault, R., Karp, K., Marutzky, S., Metzler, D.R., Peacock, A., White, D.C., Lowe, M., and Lovley, D.R. (2003) Stimulating the *in situ* activity of *Geobacter* species to remove uranium from the groundwater of a uranium-contaminated aquifer. *Appl.Environ.Microbiol.* **69,** 5884-5891.

Bonarius,. P. J., Schmid, G., and Tramper, J. (1997) Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. *Trends Biotechnol*. **15**, 308-314.

Bond, D.R., Holmes, D.E., Tender, L.M., and Lovley D.R.(2002) Electrode-reducing microorganisms that harvest energy from marine sediments. *Science* **295,** 483-485.

Bond, D.R., and Lovley, D.R. (2003) Electricity production by *Geobacter sulfurreducens* attached to electrodes. *Appl.Environ.Microbiol.* **69,** 1548-1555.

Butler, J.E., Kaufmann, F., Coppi, M.V,, Núñez, C., and Lovley, D.R. (2004) MacA, a diheme c-type cytochrome involved in Fe(III) reduction by *Geobacter sulfurreducens*. *J Bacteriol*. **186**, 4042-5.

Caccavo, F., Jr., Lonergan, D.J., Lovley, D.R., Davis, M., Stolz, J.F., and McInerney, M.J. (1994) *Geobacter sulfurreducens* sp. nov., a hydrogen- and acetate-oxidizing dissimilatory metal-reducing microorganism. *Appl Environ Microbiol.* **60,** 3752-3759.

Champine, J.E. and Goodwin, S. (1991) Acetate catabolism in the dissimilatory iron-reducing isolate GS-15. *J Bacteriol*. **173**, 2704-2706.

Daae, E.B., and Ison, A.P. (1999) Classification and sensitivity analysis of a proposed primary metabolic reaction network for *Streptomyces lividans*. *Metab Eng* **1**, 153-165.

Edwards, J.S., and Palsson, B.O. (1999)Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem*. **18**,274:17410-17416.

Edwards, J.S. and Palsson, B.O. (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci.* USA **97**, 5528-5533.

Edwards, J. S. and Palsson, B.O. (2000). Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC.Bioinformatics*. 1, 1

Edwards, J.S., and Covert, M., and Palsson, B. (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ.Microbiol.* **4,** 133-140.

Esteve-Núñez, A., Núñez, C., and Lovley, D.R. (2004) Preferential Reduction of Fe(III) over Fumarate by *Geobacter sulfurreducens*. *J Bacteriol.* **186,** 2897-2899.

Esteve-Nunez, A., Rothermich, A., Sharma, M., and Lovley, D.R. (2005) Growth of *Geobacter sulfurreducens* Under Nutrient-Limiting Conditions in Continuous Culture. *Environmental Microbiology* **7,** 641-648.

Galushko, A.S., and Schink, B. (2000) Oxidation of acetate through reactions of the citric acid cycle by *Geobacter sulfurreducens* in pure culture and in syntrophic coculture. *Arch.Microbiol* **174**, 314-321.

Kim, B.C., Leang, C., Ding, Y.H., Glaven, R.H., Coppi, M.V., and Lovley, D.R. (2005) OmcF, a Putative c-Type Monoheme Outer Membrane Cytochrome Required for the Expression of Other Outer Membrane Cytochromes in *Geobacter sulfurreducens*. *J Bacteriol*. **187**, 4505-4513

Kim, B.C., Qian, X., Leang, C., Coppi, M.V., and Lovley, D.R. (2006) Two Putative c-Type Multiheme Cytochromes Required for the Expression of OmcB, an Outer Membrane Protein Essential for Optimal Fe(III) Reduction in *Geobacter sulfurreducens*. *J Bacteriol*. **188**, 3138-3142.

Lloyd, J.R., Sole, V.A., Van Praagh, C.V., and Lovley, D.R. (2000). Direct and Fe(II)-mediated reduction of technetium by Fe(III)-reducing bacteria. *Appl.Environ.Microbiol.* **66,** 3743-3749.

Lloyd, J.R., Leang, C., Hodges Myerson, A.L., Coppi, M.V., Cuifo, S., Methe, B., Sandler, S.J., and Lovley, D.R. (2003) Biochemical and genetic characterization of PpcA, a periplasmic c-type cytochrome in *Geobacter sulfurreducens*. *Biochem.J.*, **369**,153-161

Lovley, D.R. and Phillips, E.J. (1987) Rapid Assay for Microbially Reducible Ferric Iron in Aquatic Sediments. *Appl Environ Microbiol*. **53,**1536-1540.

Lovley, D.R., Stolz, J.F., Nord, G.L. Jr, and Phillips, E.J.P. (1987) Anaerobic Production of Magnetite by a Dissimilatory Iron-Reducing Microorganism *Nature*. **330**, 252-254

Lovley, D.R., Phillips, E.J.P., Gorby, Y.A., and Landa, E.R. (1991) Microbial Reduction of Uranium. *Nature* **350,** 413-416.

Lovley, D.R. (1991). Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiological Reviews* **55**,259-287.

Lovley, D.R., Woodward, J.C., and Chapelle, F.H. (1994). Stimulated anoxic biodegradation of aromatic hydrocarbons using Fe(III) ligands. *Nature*. **370**, 128-131.

Lovley, D.R., and Chapelle, F.H. (1995). Deep Subsurface Microbial Processes. *Rev Geophys*. **33**, 365-381.

Lovley, D.R. (2003). Cleaning Up With Genomics: Applying Molecular Biology to Bioremediation. *Nature Reviews Microbiology* **1**, 36-44.

Lovley D.R., Holmes D.E., and Nevin K.P. (2004) Dissimilatory Fe(III) and Mn(IV) reduction. *Adv.Microb.Physiol*.**49,** 219-286.

Lovley, D. R. (2006) Bug juice: harvesting electricity with microorganisms. *Nature Rev. Microbiol.*(*in press*).

Magnuson, T.S., Hodges-Myerson, A.L., and Lovley, D.R. (2000) Characterization of a membrane-bound NADH-dependent Fe(3+) reductase from the dissimilatory Fe(3+)-reducing bacterium *Geobacter sulfurreducens*. *FEMS Microbiol Lett.* **185**, 205-211.

Mahadevan, R., Bond, D.R., Butler, J.E., Esteve-Núñez, A., Coppi, M.V., Palsson, B.O., Schilling, C.H., Lovley, D.R. (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl Environ Microbiol*. **72**, 1558-1568.

Overbeek, R., Larsen, N. Pusch, G.D., D'Souza, M., Selkov, E. Jr, Kyrpides, N., M. Fonstein, M., Maltsev, N., and Selkov, E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28**, 123-125

Patil, K.R., Akesson, M., and Nielsen, J.(2004)Use of genome-scale microbial models for metabolic engineering. *Curr.Opin.Biotechnol.* **15,** 64-69.

Pramanik, J., and Keasling, J.D. (1998) Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol Bioeng* **60**, 230-238.

Price, N.D., Papin, J.A., Schilling, C.H. and Palsson, B. (2003). Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol*. **21**, 162-169.

Price, N.D., Reed, J.L., and Palsson,B.O. 2004 Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol.* **2**, 886-97.

Reed, J.L., Famili, I., Thiele, I., and Palsson, B.O. (2006) Towards Multidimensional Genome Annotation. *Nature Reviews Genetics.* **7**, 130-141.

Schilling, C.H., Covert, M.W., Famili, I., Church, G.M., Edwards, J.S., and Palsson, B.O. (2002). Genome-scale metabolic model of *Helicobacter pylori* 26695. *J Bacteriol*. **184**, 4582-4593.

Unden, G., and Bongaerts, J. (1997). Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors . *Biochim Biophys Acta*. **1320**, 217-34.

Vrionis, H.A., Anderson, R.T., Ortiz-Bernad, I., O'Neill, K.R., Resch, C.T., Peacock, A.D, Dayvault, R., White, D.C., Long, P.E., Lovley, D.R. (2005) Microbiological and geochemical heterogeneity in an in situ uranium bioremediation field site. *Appl Environ Microbiol*. **71**, 6308-6318.

Adaptation of *Lactococcus lactis* to stress: integration of transcriptome and stabilome data

Emma REDON, Sandy RAYNAUD, Pascal LOUBIERE, Muriel COCAIGN-BOUSQUET

Laboratoire Biotechnologie-Bioprocédés, UMR 5504 INSA/CNRS & UMR 792 INSA/INRA, Institut National des Sciences Appliquées, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France.

1. Abstract

The micro-organism *Lactococcus lactis,* recently sequenced and recognised as the model of lactic acid bacteria, is encountered in various environments in which it is submitted to multiple growth-limiting stresses. Surprisingly, mechanisms of adaptation against these adverse environments are poorly characterized.
Two stresses of major importance, glucose starvation and auto-acidification, have been investigated. During controlled cultures in fermentor, stresses were progressively imposed to observe and analyse the dynamic adaptation of *L. lactis*. Throughout the culture, whole-genome expression was measured. Approximately 30 % of the genes were shown to be involved in the adaptation, indicating that the transcriptional responses are pleiotropic. The functional analysis of these genes allowed the different types of responses to be identified, providing a better understanding of the mechanisms involved in stress adaptation. In order to evaluate the mRNA turnover impact on the overall regulation, the mRNA stability was investigated at the genomic scale (stabilome) and analyzed together with the transcriptomic data. A formal method allowing the quantification of the relative influences of transcription and degradation on the mRNA pool control was developed. This approach highlighted that stability modulation in response to adverse growth condition can govern gene expression to the same extent as transcription in bacteria.

2. Introduction

Lactic acid bacteria (LAB) are Gram positive microorganisms with a real economic impact due to their use in many food transformation processes. The bacterium *Lactococcus lactis* commonly used as starter bacterium in the manufacture of different dairy products such as cheese, butter and buttermilk, is generally considered as a model for metabolism regulation and genetic studies in LAB. The strain IL1403 was sequenced in 2001 (Bolotin et al., 2001). In industrial conditions, like in natural ecological niches (plants, animals, gastrointestinal tract), different stresses are encountered leading to sub-optimal

Understanding and Exploiting Systems Biology

growth. More particularly, the carbon starvation and the auto-acidification are viewed as two stresses of major importance for *L. Lactis*. Acid stress is specific in that the acidification of the nutritional environment is directly dependent upon the metabolic activity of the bacterium. Energy metabolism of *L. lactis,* a homo-fermentative lactic acid bacterium converting more than 90 % of metabolized sugar into lactic acid, leads to the accumulation of high concentrations of lactic acid coincident with a progressive acidification of the growth medium and subsequent growth inhibition (Even et al., 2002). Similarly carbon starvation is one of the most drastic stress encountered by *L. Lactis* since catabolism and thus energy supply and anabolism should completely be arrested due to carbon exhaustion (Kunji et al., 1993). However in natural ecosystems, short periods allowing growth alternated with long non-growth period caused by carbon starvation. Consequently, cells should have naturally evolved towards an improved adaptation to this particular nutritional stress.

The response of a bacterium against a particular stress involves various physiological adaptations (growth, catabolism, particular metabolism, cellular morphology…). However fundamental knowledge is still very fragmented in *L. lactis* since studies have generally focused on bacterial survival or particular metabolic pathways. Recently, the development of DNA-array technology has enabled the mRNAs of the entire genome to be quantified simultaneously, easily providing an exhaustive picture of the cellular adaptation. This technology is thus particularly adapted for pleiotropic responses such as stress responses. Nevertheless, certain aspects such as mRNA degradation should be taken into account. Because cellular mRNA concentration provided by transcriptome depends on the relative rates of synthesis and decay, changes in the mRNA pool can occur either by transcriptional control or by modification of the mRNA degradation. However, mRNA stability has been rarely investigated at the genomic scale. Messenger half-lives have been measured in order to identify stability determinants in only two different bacteria, *Escherichia coli* and *Bacillus subtilis* (Bernstein et al., 2002; Hambraeus et al., 2003; Selinger et al., 2003). Furthermore, no study concerning the evolution of the mRNA stability at the genomic scale during response to environmental changes can be found in the literature. Thus, despite mRNA stability can potentially participate to the expression of the genome metabolic functions, the extent to which mRNA half-life regulation contributes to the modulation of gene expression has not yet been quantitatively determined.

In this study, carbon starvation was investigated in a chemically defined medium specifically designed in order to provoke the natural exhaustion of the sugar in an excess of any other nutriment. The acid stress was studied in milk and associated with a cold shock at low pH in order to mimic cheese-making processes. In both cases, the stress was progressively imposed in controlled cultures in fermentor in order to observe and analyze the dynamic adaptation of *L. lactis.* In order to provide an exhaustive picture of the adaptation, whole-transcriptome analysis have been performed. Transcriptome was quantified in the various phases of the cultures allowing the gene with transient expressions to be detected and thus the cellular response to be fully characterized. The identification of complete

stimulons is a prerequisite of regulation network analysis. Furthermore in the case of carbon starvation, mRNA half-lives were measured all along the culture and analyzed with transcriptomic data. A formal method derived from metabolic control analysis was developed and allowed the relative influence of mRNA degradation and transcription to be evaluated during the response of *L. lactis* to carbon starvation.

3. Experimental

3.1. Organism and growth conditions

Two strains of *Lactococcus lactis* ssp. *lactis* were used throughout this study: *L. lactis* ssp. *lactis* IL1403 whose genome was entirely sequenced (Bolotin et al., 2001) and *L. lactis* ssp. *lactis* biovar *diacetylactis* dairy strain, LD61, provided by Soredab-Bongrain. For the study of carbon starvation, the strain IL 1403 was grown on the chemically defined CDM medium (Otto et al., 1983; Poolman and Koenings, 1988) complemented by glucose (55 mM) as the sole carbon source. Cultures were grown under anaerobic conditions in a 2-l fermentor (Setric Génie Industriel, Toulouse, France) at a constant temperature of 30 °C and agitation speed of 250 rpm. The pH was maintained at 6.6 by automatic addition of KOH (10 N). The strain LD61, which contains plasmids allowing optimal growth in milk (lactose, protease and citrate utilization), was grown in non heat-sterilized skim milk ("Lait G", Standa Industrie) and in anaerobic conditions in a 20-l fermentor (Setric Génie Industriel, Toulouse, France) at agitation speed of 250 rpm. The temperature was maintained at 34 °C until the culture pH reached 5.2 (at 8 h of growth) and then was slowly decreased to 12 °C in approximately 10 hours.

Bacterial growth was estimated by spectrophotometric measurement at 580 nm directly in the case of the CDM medium or after transparisation of the milk culture (as described in Raynaud et al., 2005 and Redon et al., 2005a).

3.2. Transcriptome analysis

Transcriptome analysis were performed in the different phases of the cultures on nylon membranes after hybridization of [^{33}P]-dCTP labeled cDNA with a minimum of 3 independent repetitions. *L. lactis* IL1403 specific PCR products and some plasmidic genes of industrial relevance were provided by Eurogentec and spotted in duplicate on positively charged nylon membranes (4 deposits per spot of PCR at a concentration ranging between 40 and 180 $\mu g.ml^{-1}$) by the Plateforme Génomique (Toulouse). 2053 ORFs upon 2310 identified on the genome (89 %) and 63 plasmidic ORFs were effectively available on these membranes. RNA extraction, cDNA preparation, hybridization and detection were previously described (Redon et al., 2005b; Raynaud et al., 2005).

3.3. Determination of mRNA half-lives

For mRNA half-life quantification, 3 growth conditions (exponential, deceleration, and carbon starvation phases) were studied in independent but

physiologically identical cultures. At the required growth state, transcription was arrested by rifampicin addition to a final concentration of 500 $\mu g.ml^{-1}$. Cell samples were taken over 10 min in exponential phase or 45 min in deceleration and starvation phases. Four different time-points, including the reference sample (before rifampicin addition), were analyzed simultaneously by trancriptome measurement as previously described (Redon et al., 2005b). At least three independent time-courses were analyzed for each condition. mRNA half-lives ($t_{1/2}$) were calculated from the degradation rate constant (k) corresponding to the slope of a semi-logarithmic plot of mRNA amount as a function of time with the relation $t_{1/2} = \ln 2/k$.

4. Results and discussion

4.1. Overview of dynamic analysis

The culture of the IL1403 strain in CDM medium was characterized by an exponential phase associated to nutrient excess (0-5 h), a short deceleration phase due to decreasing glucose concentration and a non-growth stationary phase characterized by glucose exhaustion after 6 h of fermentation. The industrial strain LD61 was grown in conditions as close as possible as that used in some cheese making processes (milk, uncontrolled pH, temperature downshift). Thus, two physico-chemical stresses (acidic stress and cold stress) likely to modulate bacterial growth and metabolism overlapped during the fermentation. The growth phase ended at 11 hours of culture when the temperature was only 27 °C and while the pH had decreased from 6.41 (initial value) to 4.94. The temperature decrease continued after the growth arrest to reach a value of 12 °C after 17.5 hours of culture. During the stationary phase, milk acidification still continued (post-acidification phase), while at a slower rate, leading to a final pH of 4.64 at 180 h of culture.

In order to obtain a chronological view of gene expression, the transcriptome was analyzed comparatively in four samples taken during different stages of the culture. For each of the 2 conditions, cells were collected in the rapid growth phase (reference), then at reduced growth rate due to the progressive imposition of the stress and two times in the stationary phase when no-growth is occurring anymore because of the stress intensity. Transcriptomic data were normalized by the whole membrane intensity and thus corresponded to mRNA abundances in total mRNA population. In each condition, 30 % of the genes was differentially expressed, confirming that the stress responses were highly pleiotropic. Indeed during carbon starvation and acidic conditions, respectively 704 and 702 genes showed at least one significant expression variation (Student's test with P-value below 0.05).

4.2. Functional analysis of transcriptomic data

A functional analysis of the differentially expressed genes was realized taking into account the categories established by Bolotin *et al.* (2001) for the IL1403 strain genome. Among the 704 and 702 selected genes, the proportion of genes

with unknown function was similar to that on the entire *L. lactis* genome (36 %), indicating that unknown and known genes are equally involved in the stress responses. Adaptation of *L. lactis* towards these two stresses is mediated by three different types of transcriptomic responses.

i) global response:

For most of the genes involved in general processes associated to cellular growth, a general decline of expression was observed during the stress responses. This is consistent with growth arrest though not necessarily linked to protein decrease and physiological activities diminution. This general decrease of expression was more markedly observed during the carbon starvation and occurred earlier (at the onset of the decelerating phase) than in the milk culture (in the stationary phase). In the two experiments, the various RNA polymerase subunits encoding genes and most of the genes of translation apparatus were under expressed. Similarly most of genes involved in cell division process were under expressed. Genes involved in purine and pyrimidine biosynthesis were under expressed during carbon starvation culture, though activation of purine metabolism was observed in the milk medium probably due to the poor milk content of these compounds.

ii) specific responses functionally related to the imposed stress:

Alternative carbon sources utilization was favoured during glucose starvation. Induction was observed at 3 different levels: utilization pathways, transport and regulation. For instance, genes specifically involved in galactose, lactose, maltose, ribose and other sugars utilization (*galM*, *lacZ*, *malQ, msmK*, *rbsA*, *rbsC, rbsK*, *uxaC, ygjD*, *yidC*, *yngF*, *ypbD, ypdA, xylX* and *xynT*), polysaccharides degradation (*apu* and *yucG*) or citrate utilization (*citC*, *E* and *F*) were over expressed. Glycerol metabolism seemed to play a crucial role in carbon starvation response since 5 genes involved in this pathway were induced at high levels from deceleration phase: *dhaL* and *M* encoding DHA kinases (respective ratios of 7.1 and 3.9), *glpD* and *K* encoding respectively glycerol-3-P dehydrogenase and glycerol kinase, and *glpF1* encoding glycerol uptake enzyme (ratio of 6 at the onset of carbon starvation). Such induction of various alternative carbon sources metabolism, already observed in *B. subtilis* during carbon starvation (Bernhardt et al., 2003), may be a general response to counteract the carbon starvation.

In both experiments, the expression of most of the genes of the ADI pathway was increased (*arcA*, *B*, *C1* and *argF* during carbon starvation and *arcA*, *B*, *C1*, *C2* and *D1* in acidic conditions). The switch-on of ADI pathway in *L. lactis* enabled the cells to be supplied with maintenance energy since one ATP per arginine is produced and allowed cytoplasm alcalinization by the production of one NH_3 per arginine. Thus this pathway is directly involved in the fighting against both carbon starvation and auto-acidification.

The temperature decrease did not provoke however a strong cold-shock response since none of the 2 *csp* genes was induced, and only 3 genes, *llrC*, *ptsH* and *osmC*, encoding known cold-induced proteins (CIPs) (Wouters et al., 2000) were over-expressed.

Carbon starvation is known to confer an increased resistance towards various stresses such as heat, osmotic, acid, ethanol and oxidative stresses (Hartke et al., 1994). However, unlike in *B. subtilis* (Bernhardt et al., 2003; Hecker and Volker, 2001), the general stress response was not observed in *L. lactis* during carbon starvation (no induction of gene encoding chaperones, Clp proteases or known stress proteins), suggesting that the previously described cross protection should be linked to a different mechanism. At the opposite, during acidic conditions the heat-shock response was partially observed, since some of the genes of the heat-shock regulon were induced (*groES*, *grpE, clpB* and *clpE*). Similarly a massive induction of genes involved in oxygen metabolism and cross protection was observed. Lastly some genes linked to UV stress and DNA reparation or degradation, were also induced probably to fight against mutagen effects of acid stress.

iii) other responses

In the two experiments, various genes linked to competence (mostly in the case of the carbon starvation), phage and prophage related function or ion uptake were over expressed. These responses apparently not related to the stresses were never described previously, but were observed here in the two stressing conditions. Therefore they may be under the control of the same mechanism and thus belong to a general stress response.

4.3. Modulation of mRNA stability in response to carbon starvation

Like transcriptome, stabilome (the whole genome mRNA stability) was examined during the carbon starvation at the same degree of culture advancement. Respectively 817, 452 and 579 messenger half-lives could have been measured in exponential phase, in the decelerating phase and in the stationary phase (corresponding to the first point of transcriptome in the stationary phase). mRNA stability was gene dependent since data of stability were scattered between 1 and more than 30 min for the same culture condition (figure 1A). mRNAs were stabilized in response to carbon starvation as shown by the mRNA distribution moving towards higher half-lives (figure 1A) and the 4-fold increase of median half-lives from 4.2 to 17.3 min (table I). Efficient and sensitive sensors should be involved in this mRNA decay phenomena, since the stabilization occurred before glucose exhaustion (in the decelerating phase). Increase of bulk mRNA half-life has been reported previously in *E. coli* and *Vibrio* S14 (Albertson et al., 1990; Alberston and Nystrom, 1994), suggesting that stabilization mechanism during carbon starvation could be widespread in bacteria.

Table 1: mRNA half-life determination during carbon starvation adaptation of *L. lactis*.

	Exponential phase	Deceleration phase	Starvation phase
Mean half-life (min)	5.8 ± 5.6	17.6 ± 13.0	19.4 ± 10.0
Median half-life (min)	4.2	14.4	17.3

Stabilization factors, expressed as a ratio of half-lives, were calculated for each mRNA between the exponential and the deceleration phases and between the deceleration and stationary phases. As shown in figure 1B the statistical partition of these stabilisation factors differed and median ratio between exponential and deceleration phases was much higher than between deceleration and stationary phases (3.9 and 1.1 respectively). Taking into account the precision in stabilization factor determination, it was estimated that 92 % of transcripts were stabilized during deceleration phase in comparison to exponential phase. The evolution of mRNA half-lives was more contrasted between deceleration and starvation phases, since 37 % of messengers were further stabilized while 28 % were destabilized.

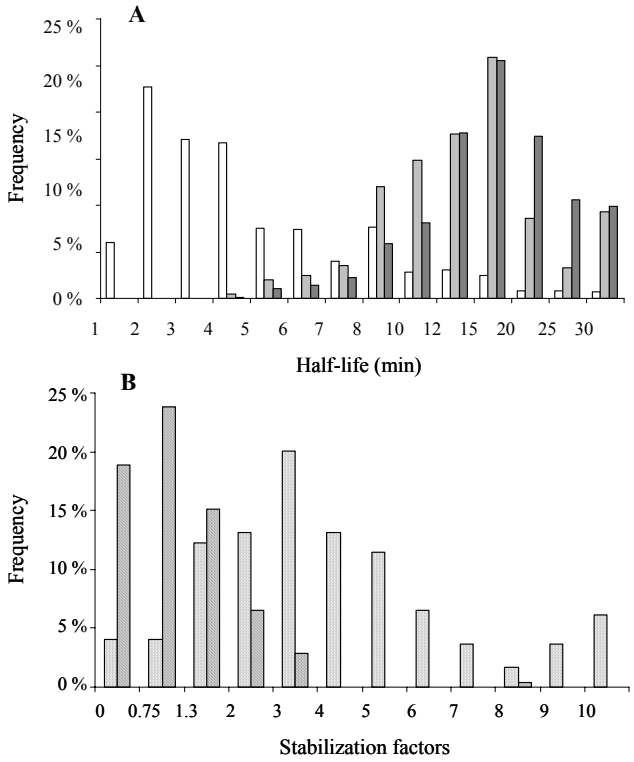


Figure 1: (A) Half-life frequencies distribution in exponential (□), deceleration (▫) and starvation (■) phases during carbon starvation adaptation of *L. lactis*; (B) stabilization factor frequency distribution between exponential and deceleration phases (⚄) and between deceleration and starvation phases (▧) during carbon starvation adaptation of *L. lactis*.

4.4. Stabilome and transcriptome integration: mRNA pool regulation analysis

Stabilome data were compared to trancriptome ones but using mRNA concentrations rather than aboundances. Raw transcriptome data without any

normalisation were corrected by total RNA concentration in each growth conditions (11.7 ± 1.3, 7.6 ± 0.8 and 8.2 ± 1.4 g.(100 g dried cells^{-1}) in exponential, deceleration and starvation phases respectively). mRNA concentrations were showing a general decreasing profile except for a minority of genes (1.6 %) while the profile was more contrasted for abundances.

Changes in mRNA concentrations in the cells can be achieved by changes in transcriptional and/or degradation rates. To quantify the relative importance of the two modes of regulation during carbon starvation adaptation, the approach developed by ter Kuile and Westerhoff (2001) for enzyme regulation was adapted to the level of mRNA. At any time, the transcription rate V_T is equal to the dilution rate due to cellular growth V_μ plus the degradation rate V_D and the time derivative of the mRNA concentration:

$$V_T = V_D + V_\mu + \frac{\partial[mRNA]}{\partial t}$$

The dilution rate V_μ and the degradation rate V_D can be expressed as a function of growth rate (μ) and the degradation constant rate (k) as following:

$$V_\mu = \mu.[\text{mRNA}]$$

$$V_D = k.[\text{mRNA}] \text{ with } k = \ln 2/t_{1/2}$$

Dilution rate of the messengers V_μ could be neglected compared to the degradation rate V_D, since μ was significantly lower than k in the 3 samples explored during the culture. Furthermore, since the time derivative of the mRNA concentration, estimated by the variation of the concentration between the different samples of the culture, was in the mean 26-fold lower than the degradation rate, it could also be neglected. Therefore, [mRNA] could be expressed as a simple function of the rate of transcription and degradation as $V_T = k.[\text{mRNA}]$. And assuming that V_T and k were independent, the derivative of the [mRNA] equation allows the degradation (ρ_D) and transcription (ρ_T) regulation coefficients to be defined as following:

$$\rho_D = -\frac{d\ln k}{d\ln[mRNA]},\ \rho_T = \frac{d\ln V_T}{d\ln[mRNA]},\ \rho_D + \rho_T = 1$$

The degradation regulation coefficient ρ_D was calculated as the opposite slope of the double-logarithmic plot of degradation rate k versus mRNA concentration between exponential and deceleration phases and between deceleration and starvation phases. Between exponential and deceleration phases, 92 % of genes exhibited low ρ_D values ($\rho_D < 0$; Figure 2). This indicated that mRNA concentrations were mostly controlled by transcription but with an antagonist influence of degradation. This transcription control associated with a general mRNA concentration decrease indicated that the transcription rate decreased strongly during the carbon starvation. In this context, mRNA were stabilized to counteract the effects of transcription and limit the drop of mRNA pool in the cells. Among the 7 genes under degradational regulation ($\rho_D > 1$), some of them (*dhaL*, *M* and *msmK*) were previously identified to play a crucial role in the adaptation to carbon starvation (sugar and glycerol metabolism), underlying the importance of decay phenomenon in the response to carbon starvation.

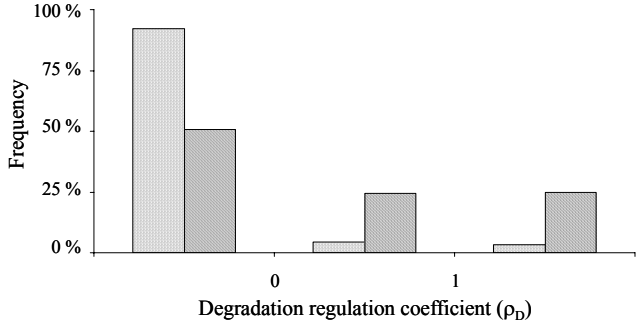


Figure 2: Frequency of genes exhibiting regulation coefficient ρ_D inferior to 0 (transcriptional control), between 0 and 1 (shared control) and superior to 1 (degradation control) between exponential and deceleration phases (⚄) and between deceleration and starvation phases (▨) during carbon starvation adaptation of *L. lactis*.

Between deceleration and starvation phases, the influence of the two modes of regulation was more balanced (figure 2). Indeed, mRNA concentration was still controlled at the transcriptional level for 51 % of genes ($\rho_D < 0$) while 25 % of genes were controlled by degradation ($\rho_D > 1$) and 24 % presented a shared control ($0 < \rho_D < 1$), indicating that mRNA stability plays a significant, if not over-riding effect, on modulating the adaptation of this bacterium to carbon starvation.

5. Conclusion

This control analysis at the genomic scale formally demonstrated that mRNA stability is a significant part of the gene expression regulation in response to adverse conditions, alongside the more classically studied transcriptional phenomenon. Therefore, modulation of gene expression is not necessarily linked only to transcriptional regulations. This important biological result was provided through an integrative approach based on the comparison of transcriptome and stabilome data. This particular domain of System Biology allowing the various levels of observation to be connected will probably in the future offer a more realistic vision of global cellular regulation.

References

Albertson, N., Nystrom, T., and Kjelleberg, S. (1990) Functional mRNA half-lives in the marine Vibrio sp. S14 during starvation and recovery. *J Gen Microbiol.* **136**, 2195-2199.

Albertson, N.H., and Nystrom, T. (1994) Effects of starvation for exogenous carbon on functional mRNA stability and rate of peptide chain elongation in Escherichia coli. *FEMS Microbiol Lett*. **117**, 181-187.

Bernhardt, J., Weibezahn, J., Scharf, C., and Hecker, M. (2003) Bacillus subtilis during feast and famine: visualization of the overall regulation of protein synthesis during glucose starvation by proteome analysis. *Genome Res.* **13**, 224-237.

Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S., and Cohen, S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A.* **99**, 9697-9702.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S.D., and Sorokin, A. (2001) The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403. *Genome Res.* **11**, 731-753.

Even, S., Lindley, N.D., Loubiere, P., and Cocaign-Bousquet, M. (2002) Dynamic response of catabolic pathways to autoacidification in Lactococcus lactis: transcript profiling and stability in relation to metabolic and energetic constraints. *Mol Microbiol.* **45**, 1143-1152.

Hambraeus, G., von Wachenfeldt, C., and Hederstedt, L. (2003) Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs. *Mol Genet Genomics.* **269**, 706-714.

Hartke, A., Bouche, S., Gansel, X., Boutibonnes, P., and Auffray, Y. (1994) Starvtion-induced stress resistance in Lactococcus lactis subsp. lactis IL1403. *Appl Environ Microbiol.* **60**, 3474-3478.

Hecker, M., and Volker, U. (2001) General stress response of Bacillus subtilis and other bacteria. *Adv Microb Physiol.* **44**, 35-91.

Kunji, E.R.S., Ubbink, T., Matin, A., Poolman, B., and Konings, W.N. (1993) Physiological responses of Lactococcus lactis ML3 to alterning conditions of growth and starvation. *Arch Microbiol.* **159**, 372-379.

Otto, R., Ten Brink, B., Veldkamp, H., and Konings, W.N. (1983) The relation between growth rate and electrochemical proton gradient on Streptococcus cremoris. *FEMS Microbiol Lett.* **16**, 69-74.

Poolman, B., and Konings, W.N. (1988) Relation of growth of Streptococcus lactis and Streptococcus cremoris to amino acid transport. *J Bacteriol.* **170**, 700-707.

Raynaud S., Perrin R., Cocaign-Bousquet M., Loubiere P. (2005) Metabolic and transcriptomic adaptation of *Lactococcus lactis* subsp. *lactis* biovar *diacetylactis* strain in response to auto-acidification and temperature downshift in skim milk. *Appl Environ Microbiol.* **71***,* 8016-8023.

Redon E., Loubière P., Cocaign-Bousquet M. (2005a) Transcriptome analysis of the progressive adaptation of *Lactococcus lactis* to carbon starvation. *J Bacteriol.* **187**, 3589-3592

Redon E., Loubière P., Cocaign-Bousquet M. (2005b) Role of mRNA stability during the genome-wide adaptation of *Lactococcus lactis* to carbon starvation. *J Biol Chem.* **280**, 36380-36385.

Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., and Rosenow, C. (2003) Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216-223.

ter Kuile, B.H., and Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500**, 169-171.

Wouters, J.A., Mailhes, M., Rombouts, F.M., de Vos, W.M., Kuipers, O.P., and Abee, T. (2000) Physiological and regulatory effects of controlled overproduction of five cold shock proteins of *Lactococcus lactis* MG1363. *Appl Environ Microbiol.* **66**, 3756-3763.

DYNAMIC MODEL FOR THE OPTIMIZATION OF L(-)-CARNITINE PRODUCTION BY *Escherichia coli*

A. Sevilla, V. Bernal, R. Teruel, C. Bernal, M. Cánovas, J.L. Iborra*.

Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain.

Keywords: L-carnitine metabolism, optimization, *Escherichia coli*, Biochemical System Theory.

1. Abstract

System Biology allows cellular complexity analysis and the optimization of cell metabolic pathways by using cell component enumeration, structured relationship between them, mathematical representation of the metabolic networks, knowledge of the metabolic properties and comparison with experimental outputs of the cell processes involved. In this work metabolic engineering strategies and system biology principles for maximizing L(-)-carnitine production by *E. coli* based on the Biochemical System Theory are presented. The model integrates the metabolic and the bioreactor levels using power-law formalism. Experimental results using a high-cell density reactor were compared with optimized predictions. The model shows control points at macroscopic (reactor operation) and microscopic (molecular) levels where conversion and productivity can be increased. In accordance with the optimized solution, the next logical step to improve the L(-)-carnitine production rate will involve metabolic engineering of the *E. coli* strain by overexpressing the carnitine transferase, CaiB, activity and the protein carrier, CaiT, responsible for substrate and product transport in and out of the cell. By this means, it is predicted that production may be enhanced by up to three times the original value.

2. Introduction

In human cells, L-carnitine (R(-)-3-hydroxy-4-trimethylaminobutyrate) transports long-chain fatty acids through the inner mitochondrial membrane, which is why several clinical applications for L-carnitine have been identified. Consequently, the demand for L-carnitine has increased worldwide (Seim et al., 2001) and chemical and biological processes have been developed for its production (Cavazza, 1981; Kulla, 1991; Hoeks et al., 1996; Kleber, 1997).

* Corresponding author.

Strains belonging to the genera *Escherichia, Proteus* and *Salmonella* racemice D-carnitine, a waste product and an environmental problem resulting from the L-carnitine chemical synthesis, and/or biotransform crotonobetaine (dehydrated D-carnitine) to produce L-carnitine (Kleber, 1997; Castellar et al., 1998; Obon et al., 1999; Canovas et al., 2002).

In *E. coli*, the responsible genes for L-carnitine metabolism are in the operons *caiTABCDE* and *fixABCX*. These operons are modulated positively by general regulators, such as the cAMP receptor protein (CRP) or the transcriptional regulator responsible for anaerobic induction (FNR), and negatively by the DNA-binding protein H-NS, glucose or nitrate (Unden and Trageser, 1991; Eichler et al., 1994). In addition, it has been proposed that a positively controlled *caiF* gene, 3' adjacent region to the cai operon, acts as a specific transcriptional regulator for carnitine metabolism (Eichler et al., 1996). This pathway is detectable not only in cells previously grown anaerobically but also in some species, such as *E. coli* ATCC 25922 and DSM 8828, *P. vulgaris* and *P. mirabilis*, grown under aerobiosis in the presence of inducers such as D-L-carnitine mixture or crotonobetaine (Kleber, 1997; Obon et al., 1999; Elssner et al., 2000; Canovas et al., 2002). It was first postulated that L-carnitine dehydratase reversibly catalyzed L-carnitine into crotonobetaine and that crotonobetaine reductase non-reversibly transformed crotonobetaine into γ-butyrobetaine as an electron sink (Jung et al., 1989; Roth et al., 1994; Kleber, 1997), even though this latter in *E. coli* can be inhibited by fumarate addition as another electron sink (Obon et al., 1999). Now that functions have been assigned to each putative protein of the *cai* operon, it is known that CaiT is an exchanger (antiporter) for carnitine derivates in *E. coli* (Jung et al., 2002) with no energy consume. Another type of transport of these compounds with ATP consume and irreversible is present (Canovas et al., 2003a); this transport is equivalent to the transporter ProU (Verheul et al., 1998). The enoyl-CoA hydratase (CaiD) is composed of two identical subunits, requiring a CoA-transferase activity (CaiB). It has been verified that the hydration reaction of crotonobetaine to L-carnitine (CHR) proceeds at the CoA-level in two steps: the protein CaiD-catalyzed hydration of crotonobetainyl-CoA to L-carnitinyl-CoA, followed by CoA-transfer from L-carnitinyl-CoA to crotonobetaine, catalyzed by CaiB (Elssner et al., 2001). Thus, CaiD and CaiB from *E. coli* have been found to catalyze the reversible biotransformation of crotonobetaine to L-carnitine in the presence of a co-substrate, either γ-butyrobetainyl-CoA or crotonobetainyl-CoA (Elssner et al., 2001). CaiD was also postulated to be involved in racemisation of D-carnitine (Eichler et al., 1996). Further, CaiC has been suggested as a CoA-trimethylammonium ligase (Eichler et al., 1996), activating crotonobetaine/γ-butyrobetaine/L-carnitne when they reach the cell. The function of protein CaiE is not totally understood and further studies must be undertaken. With all this information, we have proposed a model to describe the whole activity of *E. coli* able to produce L-carnitine from crotonobetaine under anaerobic conditions (Figure 1).

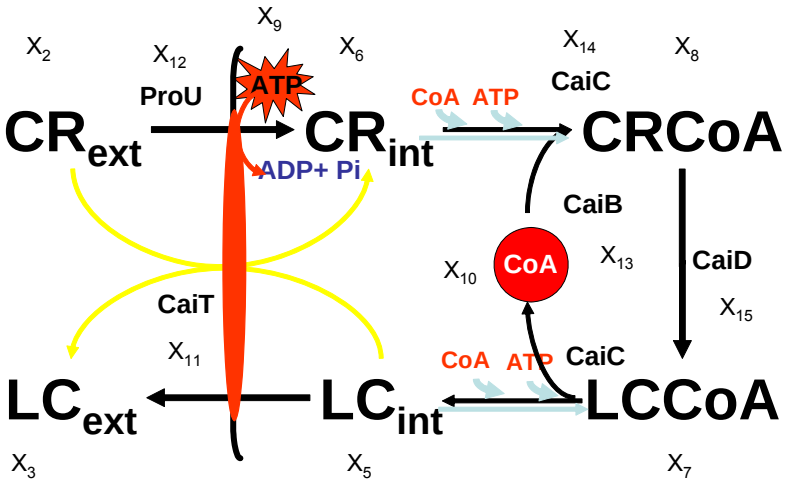


Figure 1. Metabolic pathways involved in the biotransformation of crotonobetaine into L-carnitine in *E. coli*. extracellular crotonobetaine (X_2); extracellular L-carnitine (X_3); intracellular L-carnitine (X_5); intracellular crotonobetaine (X_6); L-carnitinylCoA (X_7); crotonobetainylCoA (X_8); ATP (X_9); CoA (X_{10}); CaiT (X_{11}); ProU (X_{12}); CaiB (X_{13}); CaiC (X_{14}); CaiD (X_{15}).

Rational optimization of this biotransformation in continuous high-cell density membrane reactors first requires understanding the link between cell carnitine metabolism and the connection between the microkinetics of both metabolisms and the macrokinetics of the cell population behaviour in the reactor. From this point a dynamic model was built for the biotransformation of crotonobetaine into L-carnitine including all the genes involved in the carnitine metabolism.

3. Theoretical

3.1. MATHEMATICAL MODELLING

3.1.1. S-System model

Biochemical System Theory offers some choices for the formulation of biochemical systems, among which the most relevant for our purposes is the S-System (Voit, 2000). In an S-System model, each net rate law for synthesis and degradation is represented by a product of power-law functions of the whole set of variables that influence the net rate law in question. For example, the synthesis rate of a given process V_i^+ is written as:

$$V_i^+ = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{i,j}} \qquad (1)$$

where X_j are variables that affect the rate in question. The indices j = 1…n refer

to dependent variables, while j = n+1, …, n+m refer to independent variables. The processes that characterize process forming X_i are aggregated to give a single law for the net synthesis V_i^+. Similarly, those rate laws that characterize reactions removing the same X_i are aggregated to give a single law for net degradation V_i^-. The descriptive equations for a process can then be written in terms of power-law functions as follows:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{i,j}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{i,j}} \qquad (2)$$

$$i = 1,\ldots,n$$

The parameters α_i and $g_{i,j}$ are rate constants and kinetic orders associated with the rate law for net production of X_i . Similarly, β_i and $h_{i,j}$ are associated with the rate laws of net degradation of X_i. The kinetic order parameters $g_{i,j}$ and $h_{i,j}$ are defined as follows:

$$g_{i,j} = \left(\frac{\partial V_i^+}{\partial X_i} \frac{X_i}{V_i^+} \right)_0 \qquad (3)$$

and

$$h_{i,j} = \left(\frac{\partial V_i^-}{\partial X_i} \frac{X_i}{V_i^-} \right)_0 \qquad (4)$$

where subscript 0 indicates that a quantity is evaluated at steady-state stage of the system. Values for the parameters α_i and β_i are determined in such a way that the modelled rate and the power-law approximation are equivalent in steady-state.

4. Experimental

4.1. Chemicals

L(-)-Carnitine and trans-crotonobetaine were gifts from Biosint. S.p.A. (Rome, Italy). Acetyl-CoA, acetyl-phosphate, 5,5´-dithiobis-(2-nitrobenzoic) acid, ATP, D,L-carnitine, carnitine acetyl-transferase, coenzyme A, and thiamine pyrophosphate were from Sigma Chem. Co. (St. Louis, MO. USA). Bacteriological peptone was purchased from Oxoid (Basingstoke, England). All other chemicals employed were of analytical grade.

4.2. Growth of the bacteria

Escherichia coli 044 K74 stored as liquid culture, in glycerol 20% (v/v), at -20 ºC was used to inoculate a culture medium composed of ($g \cdot l^{-1}$): glycerol, 12.60;

KH_2PO_4, 5.44; K_2HPO_4, 10.49; $(NH_4)_2SO_4$, 2.0; $MgSO_4 \cdot 7H_2O$, 0.05; $MnSO_4 \cdot 4H_2O$, 0.05 y $FeSO_4 \cdot 7H_2O$, 0.00013. The pH of the medium was adjusted to 7.5 with 0.1 M KOH prior to autoclaving. Assays were preformed under aerobic conditions in an orbital stirrer (150 rev. min^{-1}) in 250 ml Erlenmeyer flasks containing 100 ml culture medium at 30 ºC that were inoculated with 5% (v/v) culture concentration for 6-8 hours. Cells were grown in the following culture medium ($g \cdot l^{-1}$): bacteriological peptone 20; NaCl, 5; fumarate, 2; glycerol, 12.6 and crotonobetaine, 5 as inducer. The pH of the medium was adjusted to 7.5 with 1 M KOH prior to autoclaving. After inoculation of cultures with 5% (v/v), cultivation was carried out in the high-cell density recycle continuous bioreactor in anaerobic conditions**.**

4.3. Cell recycle bioreactor

The experimental set-up for the membrane cell-recycle system has been presented elsewhere (Obon et al., 1999). The fermentation vessel of 500 ml capacity was coupled to a cross-flow filtration module (Minitan, Millipore, USA) equipped with four 0.1 μm hydrophilic polyvinylidene difluoride Durapore plates of 60-cm^2 surface area (Millipore, USA). The cell broth was recycled into the reactor with a peristaltic pump adjusted to a high flow rate (70 ml/min) to minimize membrane fouling. *E. coli* O44 K74 cells for the inocula were grown as explained elsewhere (Obon et al., 1999) and transferred to the fermenter. Continuous operation was set at 37 ºC and was started up by feeding with the medium. The culture was grown anaerobically by bubbling nitrogen previously passed through a water trap.

4.4. Enzyme assay, metabolites and biomass determination

L(-)-carnitine dehydratase was assayed according to Jung et al. (Jung et al., 1993). ATP was measured by bioluminescence assay (Bioluminescence Assay Kit HS II, Boehringer Mannheim, Germany) using FluoStar (BGP, Germany) without filters. Then intracellular concentration was calculated assuming an intracellular volume of 63 μl·mg-1 (Canovas et al., 2003b). Acetyl-CoA and coenzyme A concentrations were measured by HPLC (Shimadzu Co., Kyoto, Japan) following the method proposed by Debuysere and Olson (1983), using a μ-BondapakTM C18 Millipore (4.5 mm x 25 cm) column with a pre-column (4,5 mm x 4 cm), packed with a C18 phase. Detection was performed at 254 nm. The mobile phase was 0.12 M H_3PO_4, 0,05% ß-mercaptoethanol, (85 v/v) and 15 (v/v), methanol (98%) and chloroform (2%), adjusted to pH 4.0 and the flow rate was 0.8 ml/min. L(-)-carnitine concentration was measured by the carnitine acetyl transferase method (Jung et al., 1989). Glycerol and crotonobetaine concentrations were determined by HPLC using a 25 x 0.46 cm Tracer Spherisorb-NH_2 3 μm column (Tecknokroma, Barcelona, Spain). The mobile phase was acetonitrile/H_3PO_4, 0.05 M, pH 5.5 (65/35), with a flow rate of 1 mL/min. Cell growth was determined spectrophotometrically at 600 nm using a Novaspec II spectrophotometer (Pharmacia-LKB, Uppsala, Sweden) and then translated to dry weigh.

5. Results and Discussion

5.1. S-System Model

Considering cell growth, biotransformation equations and enzyme kinetics we built the following S-System model:

$$
\begin{aligned}
X_1^{'} &= 2.3104 X_{16} X_{17} - 2.3104 X_1^{0.916} X_4^{0.567} X_{16}^{0.432} X_{19}^{0.567} \\
X_2^{'} &= 28.018 X_3^{0.0224} X_4^{0.939} X_6^{0.104} X_{11}^{0.860} X_{16}^{0.0605} X_{18}^{0.0605} \\
&- 28.018 X_2^{0.142} X_4^{0.964} X_{10}^{0.0219} X_{11}^{0.883} X_{12}^{0.0816} X_{16}^{0.0357} \\
X_3^{'} &= 38.848 X_2^{0.111} X_4 X_5^{0.387} X_{11} - 38.848 X_3^{0.393} X_4^{0.974} X_{11}^{0.892} X_{16}^{0.0257} \\
X_4^{'} &= 0.080017 X_1^{0.852} X_4 X_{19} - 0.080017 X_4 \\
X_5^{'} &= 74.572 X_3^{0.165} X_7^{0.563} X_{11}^{0.437} X_{13}^{0.563} - 74.572 X_2^{0.062} X_5 X_{11}^{0.563} X_{13}^{0.437} \\
X_6^{'} &= 50.526 X_2^{0.062} X_8^{0.437} X_{10}^{0.0128} X_{11}^{0.515} X_{12}^{0.0476} X_{13}^{0.437} \\
&- 50.526 X_3^{0.165} X_6^{0.991} X_9^{0.0026} X_{10}^{0.0006} X_{11}^{0.437} X_{13}^{0.554} X_{14}^{0.0091} \\
X_7^{'} &= 167.28\, X_5^{0.437} X_7^{-1.092} X_8^{0.464} X_{13}^{0.437} X_{15}^{0.563} \\
&- 167.28\, X_7^{1.057} X_8^{-0.776} X_{13}^{0.563} X_{15}^{0.437} \\
X_8^{'} &= 162.88\, X_6^{0.554} X_7^{0.494} X_8^{-0.776} X_{13}^{0.554} X_{14}^{0.0091} X_{15}^{0.437} \\
&- 162.88\, X_7^{-1.092} X_8^{0.901} X_{13}^{0.437} X_{15}^{0.563}
\end{aligned}
\qquad (5)
$$

where $X_i^{'} = \frac{dX_i}{dt}$

X_1 is glycerol outside the cell, X_2 crotonobetaine outside the cell , X_3 L-carnitine outside the cell, X_4 cell concentration, X_5 L-carnitine inside the cell, X_6 crotonobetaine inside the cell, X_7 L-carnitinyl-CoA, X_8 crotonobetainyl-CoA, X_9 CoA, X_{10} ATP, X_{11} CaiT, X_{12} Pro U, X_{13} CaiB, X_{14} Cai C, X_{15} CaiD, X_{16} flow rate, X_{17} glycerol inlet, X_{18} crotonobetaine inlet, X_{19} specific grow rate and X_{20} the constant for biomass decay.

The proposed model was useful for representing the whole set of metabolites in the *E. coli* carnitine metabolism, the function of cell transporters (CaiT and ProU), the influence of the pool of energetic compounds (ATP and Acetyl-CoA/HS-CoA ratio). In Table 1, experimental and simulation results for enzyme activities from the secondary and central metabolism of *E. coli* as well as reactor productivity and conversion are presented. Furthermore, although the model is applied to represent the biotransformation of crotonobetaine into L-carnitine, it is also useful for understanding any biotransformation process where a single enzyme or several are involved, even when coenzyme regeneration is required.

Table 1. Comparison between experimental and simulated data from studies carried out with *E. coli* O44 K74 in a minimal medium with glycerol 75 mM and crotonobetaine 50 mM. Productivities and conversion values were calculated at the time of steady-state value.

	Experimental value	Simulated value
Enoyl-CoA hydratase (U)	180	212
Crotonobe-taine reductase(U)	20	27
Acetyl-CoA synthetase (mU/mg prot)	394	405
Isocitrate lyase (mU/mg prot)	5.60	7.0
Productivity (g/L/h)	0.16	0.18
Conversion (%)	25	30

Biotransformation simulation results matched experimental results, meaning that the microkinetics of the cell metabolism matched the macrokinetics of the reactor system. Moreover, simulations allowed the determination of certain keypoints where to improve cell metabolism and the L-carnitine production process by metabolic engineering.

5.2. Dynamic Simulation

The assessment of dynamic responses, such as a transient following a perturbation, can be simulated using PLAS (Ferreira, 1998). For the present purpose, we carried out an extensive series of such analyses, which pointed to a common pattern, that is, the system returned in a short time to the initial steady state and never reached extreme, unfeasible values for any of the intermediate pools. It should be stated that the range of the sudden perturbation was between 20 to 50% of the steady state concentration.

As an illustrative example, Figure 2 shows the dynamics observed after a 50% increase in the CaiT activity.

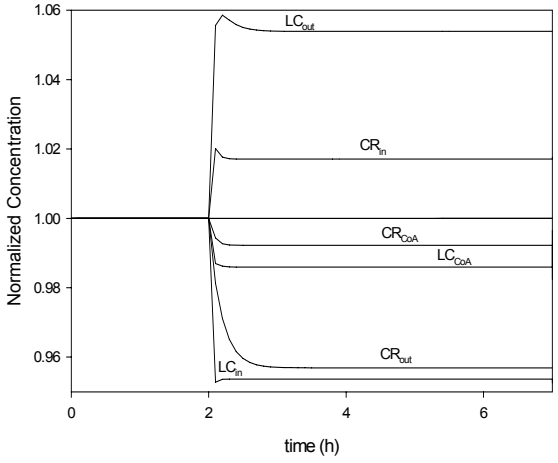


Figure 2. Response of the model to a 50% increase in CaiT activity between 0 and 7 hours. At time 2, a bolus of CaiT activity (X_{12}) was carried out. Time evolutions crotonobetaine outside the cell (CR_{ext}, X_2), L-carnitine outside the cell (LC_{ext}, X_3), L-carnitine inside the cell (LC_{int}, X_5), crotonobetaine inside the cell (CR_{int}, X_6), L-carnitinyl-CoA (LC_{CoA}, X_7), crotonobetainyl-CoA (CR_{CoA}, X_8) are shown. Other intermediates showed no significant variations.

It can be seen a higher efficiency of the system, as a consequence of a higher inlet of crotonobetaine inside the cell as it was reflected in its increment. However, the most important result was the increase in the external concentration of carnitine.

Another example is reflected in the Figure 3. In this pulse the activity of CaiB is increased a 50% from its basal level. The most important consequence is the increment of the L(-)carnitine inside as well as outside the cell as a consequence of a faster interchange of the CoA group carried out by this enzyme.

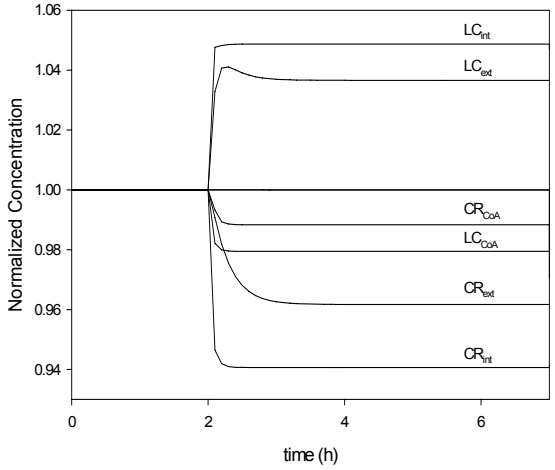


Figure 3. Response of the model to a 20% increase in CaiB activity between 0 and 7 hours. At time 2, a bolus of CaiB activity (X_{13}) was carried out. Time evolutions crotonobetaine outside the cell (CR_{ext}, X_2), L-carnitine outside the cell (LC_{ext}, X_3), L-carnitine inside the cell (LC_{int}, X_5), crotonobetaine inside the cell (CR_{int}, X_6), L-carnitinyl-CoA (LC_{CoA}, X_7), crotonobetainyl-CoA (CR_{CoA}, X_8) are shown. Other intermediates showed no significant variations.

In the Figure 4 is represented the dynamic evolution of the biotransformation of L(-)-carnitine when a bolus of the inlet of the external crotonobetaine was carried out. The concentration of the external L(-)-carnitine was consequently increased. However, the transformation yield dropped dramatically as a probably consequence of the saturation of the L-carnitine biosynthesis at the transport level.

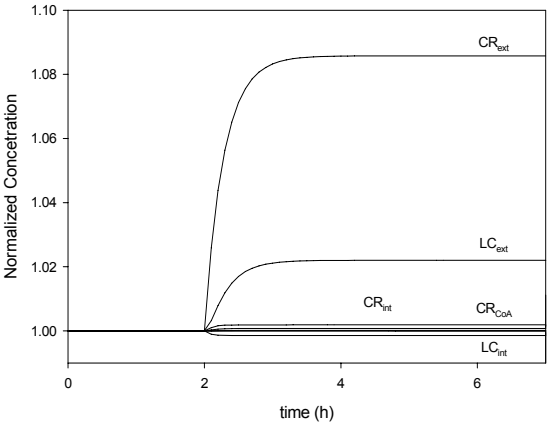


Figure 4. Response of the model to a 20% increase in crotonobetaine inlet between 0 and 7 hours. At time 2, a bolus of the crotonobetaine inlet (X_{18}) was carried out. Time evolutions crotonobetaine outside the cell (CR_{ext}, X_2), L-carnitine outside the cell (LC_{ext}, X_3), L-carnitine inside the cell (LC_{int}, X_5), crotonobetaine inside the cell (CR_{int}, X_6), crotonobetainyl-CoA (CR_{CoA}, X_8) are shown. Other intermediates showed no significant variations.

The obtained results are in agreement with the IOM approach (Marin-Sanguino and Torres, 2003), carried out for the L(-)-canitine metabolism (Sevilla et al., 2005b). This work assessed that it is possible to increase by a factor bigger than 3 the current productivity of L-carnitine. This enhanced productivity can be attained by modifying the current values of five parameters of the system, two bioreactor-operating values and three enzyme activities: the initial concentration of cronotobetaine, the dilution rate, the activities of CaiT and CaiB have to be overexpressed but ProU should be inhibited. The better performance here

encountered is thus due to the role of the transport processes. A possible explanation of this behavior is that any increase in the activity of ProU, a non-reversible trimetilamonium transport systems associated to osmotic stress (Verheul et al., 1998) has a negative influence on the whole carnitine biotransformation due to the negative effect on the combined reaction capacity of CaiT (an antiport crotonobetaine$_{in}$/carnitine$_{out}$) because in this process a futil cycle is generated (Sevilla et al., 2005a). In this process CaiB, an enzyme that catalyzes the transfer of CoA groups between different betaines (Elssner et al., 2001), seems to be the controlling enzyme whereas the control of CaiD (responsible of the biotransformation, Fig. 1) seems to be non significant.
Another conclusion is that the connection of both metabolisms sketched in Figure 1, suggests the existence of control points, not only at the central but also at the carnitine metabolism, where it would be possible to act to redirect the metabolic fluxes (i.e. energy metabolism for transport processes). Therefore, future studies to optimize the biotransformation should also be addressed at redirecting the metabolic fluxes towards an increase in energy levels and the levels of metabolites required for biotransformation by using System Biology. This work means modifying and redirecting pathways by the use of genetic engineering tools and having the image of the complete cell system to maintain cell homeostasis and viability.

6. Conclusion

- The control points at macroscopic (reactor operation) level are the dilution rate and the initial crotonobetaine concentration as well as at microscopic (molecular) level are the carnitine transferase, CaiB, activity and the protein carrier, CaiT, responsible for substrate and product transport. Modifying these points, the production of crotonobetaine biotransformation may be enhanced by up to three times the original value.
- The control of the biotransformation is at the transport level not at the reaction level. This control is exerted by the simultaneous connection of two kind of transporters involved in the carnitine metabolism: CaiT is a reversible non ATP dependent antiporter but ProU summarized the action of a group of irreversible, energy dependent transporters.

Acknowledgements
This work was supported by MEC project BIO2005-08898-C02-01, BioCARM project BIO2005/01-6468 and Séneca project 02928/PI/05. A. Sevilla is recipient of a grant from MEC (Spain). Biosint S.p.A. (Italy) is also acknowledged for the kind gift of the substrate.

References

Canovas, M., Bernal, V., Sevilla, A., and Iborra, J.L. (2003a) Modelling of biotransformation processes in high-density cell recycle membrane reactor: Production of L(-)-carnitine by *E. coli* and *P. mirabilis* strains a case study. 1-3. ECCE-4. 4th European Congress of Chemical Engineering. Abstract Book II. ISBN: 84-88233-33-7.

Canovas, M., Bernal, V., Torroglosa, T., Ramirez, J.L., and Iborra, J.L. (2003b) Link between primary and secondary metabolism in the biotransformation of trimethylammonium compounds by *Escherichia coli*. *Biotechnol. Bioeng.,* **84**, 686-699.

Canovas, M., Maiquez, J.R., Obon, J.M., and Iborra, J.L. (2002) Modeling of the biotransformation of crotonobetaine into L-(-) carnitine by *Escherichia coli* strains. *Biotechnol. Bioeng.,* **77**, 764-775.

Castellar, M.R., Canovas, M., Kleber, H.P., and Iborra, J.L. (1998) Biotransformation of D(+)-carnitine into L(-)-carnitine by resting cells of *Escherichia coli* O44 K74. *J. Appl. Microbiol.,* **85**, 883-890.

Cavazza, C. (1981) D-camphorate of L-carnitinamide and D-camphorate of D-carnitinamide. BE patent 877609 A1.

Debuysere, M.S. and Olson, M.S. (1983) The Analysis of Acyl-Coenzyme A Derivatives by Reverse-Phase High-Performance Liquid-Chromatography. *Anal. Biochem.,* **133**, 373-379.

Eichler, K., Buchet, A., Lemke, R., Kleber, H.P., and MandrandBerthelot, M.A. (1996) Identification and characterization of the caiF gene encoding a potential transcriptional activator of carnitine metabolism in *Escherichia coli*. *J. Bacteriol.,* **178**, 1248-1257.

Eichler, K., Schunck, W.H., Kleber, H.P., and MandrandBerthelot, M.A. (1994) Cloning, Nucleotide-Sequence, and Expression of the *Escherichia coli* Gene Encoding Carnitine Dehydratase. *J. Bacteriol.,* **176**, 2970-2975.

Elssner, T., Engemann, C., Baumgart, K., and Kleber, H.P. (2001) Involvement of coenzyme A esters and two new enzymes, an enoyl-CoA hydratase and a CoA-transferase, in the hydration of crotonobetaine to L-carnitine by *Escherichia coli*. *Biochem.,* **40**, 11140-11148.

Elssner, T., Hennig, L., Frauendorf, H., Haferburg, D., and Kleber, H.P. (2000) Isolation, identification, and synthesis of gamma-butyrobetainyl-CoA and crotonobetainyl-CoA, compounds involved in carnitine metabolism of *E. coli*. *Biochem.,* **39**, 10761-10769.

Ferreira, A. (1998) PLAS: Power law analysis and simulation software, Version 2b.2.

Hoeks, F.W.J.M., Muhle, J., Bohlen, L., and Psenicka, I. (1996) Process integration aspects for the production of fine chemicals illustrated with the biotransformation of gamma-butyrobetaine into L-carnitine. *Chem. Eng. J. and Biochem. Eng. J.,* **61**, 53-61.

Jung, H., Buchholz, M., Clausen, J., Nietschke, M., Revermann, A., Schmid, R., and Jung, K. (2002) CaiT of *Escherichia coli*, a new transporter catalyzing

L-carnitine/gamma-butyrobetaine exchange. *J. Biol. Chem.,* **277**, 39251-39258.

Jung, H., Jung, K., and Kleber, H.P. (1989) Purification and Properties of Carnitine Dehydratase from *Escherichia coli* - A New Enzyme of Carnitine Metabolization. *Biochim. Biophys. Acta,* **1003**, 270-276.

Jung, H., Jung, K., and Kleber, H.P. (1993) Synthesys of L(-)-carnitine by microorganisms and isolated enzymes. *Adv. Biochem. Eng. Biotechnol.,* **50**, 21-44.

Kleber, H.P. (1997) Bacterial carnitine metabolism. *FEMS Microbiol. Lett.,* **147**, 1-9.

Kulla, H.G. (1991) Enzymatic Hydroxylations in Industrial Application. *Chimia,* **45**, 81-85.

Marin-Sanguino, A. and Torres, N.V. (2003) Optimization of biochemical systems by linear programming and general mass action model representations. *Math. Biosci.,* **184**, 187-200.

Obon, J.M., Maiquez, J.R., Canovas, M., Kleber, H.P., and Iborra, J.L. (1999) High-density *Escherichia coli* cultures for continuous L(-)-carnitine production. *Appl. Microbiol. Biotechnol.,* **51**, 760-764.

Roth, S., Jung, K., Jung, H., Hommel, R.K., and Kleber, H.P. (1994) Crotonobetaine Reductase from *Escherichia coli.* A New Inducible Enzyme of Anaerobic Metabolization of L(-)-Carnitine. *Antonie Van Leeuwenhoek In. J. Gen. Molec. Microbiol.,* **65**, 63-69.

Seim, H., Eichler, K., and Kleber, H.P. (2001) L(-)-carnitine and its precursor γ-butirobetaine. In: *Nutraceuticals in health and disease prevention*. (Krämer, K, Hoppe, PL, and Packer, L, Ed.) pp 217-256. Marcel Dekker, New York.

Sevilla, A., Schmid, J.W., Mauch, K., Iborra, J.L., Reuss, M., and Canovas, M. (2005a) Model of central and trimethylammonium metabolism for optimizing L-carnitine production by *E. coli*. *Metab Eng* ,**7**, 401-425.

Sevilla, A., Vera, J., Diaz, Z., Canovas, M., Torres, N.V., and Iborra, J.L. (2005b) Design of metabolic engineering strategies for maximizing L-(-)-carnitine production by *Escherichia coli*. Integration of the metabolic and bioreactor levels. *Biotechnol. Prog.,* **21**, 329-337.

Unden, G. and Trageser, M. (1991) Oxygen Regulated Gene Expression in *Escherichia coli*. Control of Anaerobic Respiration by the Fnr Protein. *Antonie Van Leeuwenhoek In. J. Gen. Molec. Microbiol.,* **59**, 65-76.

Verheul, A., Wouters, J.A., Rombouts, F.M., and Abee, T. (1998) A possible role of ProP, ProU and CaiT in osmoprotection of *Escherichia coli* by carnitine. *J. Appl. Microbiol.,* **85**, 1036-1046.

Voit, E.O. (2000) *Computational analysis of biochemical systems. A practical guide for biochemists and molecular biologists*. Cambridge University Press, Cambridge, U. K.

The Global Biodegradation Network: Who works here?

Trigo A.[a], Cases I. [a], Pazos F. [b], De Lorenzo V. [c], Valencia A. [a]

[a] *Structural Bioinformatics Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain, e-mail: atrigo@cnio.es*
[b] *Protein Design Group, National Center for Biotechnology-CSIC, Cantoblanco, Spain.*
[c] *Microbial Biotechnology Department, National Center for Biotechnology-CSIC, Cantoblanco, Spain.*

1. Abstract

Human activities produce a remarkable amount of compounds, many of them harmful to the natural environment. Microbial communities have developed a natural capacity to degrade xenobiotic and pollutants compound and so to clean-up polluted areas. In an attempt to understand this process better, we have studied the Global Biodegradation Network, which consisting on all biodegradation reactions, regardless of their microbial host, and in which reactions are nodes that are connected if the product of one can be the substrate of the other. More over, we have been able to associate protein sequences to a large number of reactions in the network. This has allowed us, in one hand, to analyze the topological characteristic of network from a reaction-centric view, on the other hand, to localize the regions in which their study has been more intensive. We propose that this improvement in the understanding of the internal structure of the network could help to steer the prospective efforts to complete the global knowledge of the system to a sequence level, and so, to discover functional relations hidden in it.

2. Introduction

Thousands of chemical compounds are been released to the environment every day as a consequence of human activities, ranging from chemical industry to agriculture. Many of these compounds are relatively new in nature (the so called xenobiotic compounds) and microbial communities have had to adapt to them, and develop mechanism for overcome their toxic effect, and even in some cases, have learn to metabolize them in their own benefit. While some xenobiotic compounds are modified until the point in which they can enter into the central metabolism, some others are just partially transformed and during the process, and the whole sequence of reactions can be preformed by several bacterial species. Hundreds of biodegradation reactions have been already experimentally characterized in different conditions, accounting for an interesting collection of

data for analysis, even if they are probably only a small proportion of the larger biological reality. All this area of research is not unrelated to the study of the complete genetic repertory present in a given ecosystems (a field now known as "metagenomics"), whose first spectacular results have been recently published (Venter el al., 2004).

We carried out a first study of the general properties of the known biodegradation network (Pazos et al., 2003), in which we determined the scale-free structure of the network, including the input/output reactions, with characteristics similar to the ones observed for the standard metabolic networks. This similarity fits well with the biological model that describes the collective behavior of the biodegradation networks as similar to the one of single organisms, even if its nature and evolution are necessarily very different. This first analysis of the biodegradation network also allowed us to propose a first model for its evolution. More over, we were able to develop a set of bioinformatic tools (Pazos et al., 2005a) that, beside facilitating storing updating and querying the available information, allowed us for complex data mining. For instance, the system was able to uncover new alternative pathways for the degradation of certain compounds, and compare them to those that appear in bacterial genomes. In this work, we present our new results on the study of global biodegradation network. If the previous analyses were performed over a classical metabolic network formulation, were the compounds are treated as nodes connected by reactions, we have now transformed the network in such a way that now the nodes are reaction that are connected if the product of one can be the substrate of the other. Also, we have been able to associate protein sequences to a large number of biodegradative reactions, a so far pending task. We believe that this new formulation, by focusing in the biological entities (reactions, and the proteins that perform them) instead of chemical entities (the compounds) will be able retrieve new information regarding structure, behavior and evolution of the biodegradation network.

3. Materials and Methods

3.1. Network reconstruction

The biodegradation network introduced here is a directed graph in which the nodes are the reactions. A reaction consists of its substrate(s), the protein that achieve the transformation and the product(s). The edges represent the connection between two reactions in which the product of the first one is the substrate of the second one. When a reaction has more than one substrate or product, all the possible connections are constructed. The initial set of data was obtained from the University of Minnesota Biocatalysis/Biodegradation Database (UMBBD: April 2005 version, Ellis et al., 2003). The chemical compounds that are consider in the UMBBD as cofactors, are not included in the network. All the reactions in UMBBD are included, regardless of their aerobic and anaerobic nature and the organism in which the enzymes are present. A reaction is linked with the Central Metabolism (CM) when its product belongs to it according with

UMBBD. The distance of a given reaction to the CM is defined as the minimum number of steps (edges in the network) to reach it.

3.2. Protein sequence retrieval and association

The sequence of the proteins that participate in each reaction were retrieved by a manual search taking as starting point the bibliographic references included in UMBBD and adding others obtaining from their context in the system. As result, a group of vectors with a different combination of data as the author, organism, operon, pathway, enzymatic activity, etc, were built to query the GenBank Protein Database (Benson et al., 2005) and retrieve the sequences. Result were manually inspected to guarantee consistency with the available literature.

3.3. Data availability

The data used in this work are included in the BioNeMo (Biodegradation Network Modelling) database. This application has been created to store and maintain the enzymatic and regulatory activity of the transformations in biodegradation, in an integrated way. It will be available soon via web server (Trigo et al., 2006).

4. Results and discussion

4.1. Topological properties

Previous studies about biodegradation from the Systems Biology approach have been focused to the understanding of the relations between the compounds that appear in the biodegradation process (the initial and final products of each transformations). In this way, the topological properties of the resulting global network have been discussed (Pazos et al., 2005b). In this work, nevertheless, we have used a different approach. The interest now is concentrated into the biological entities that achieve the transformations (reactions), the proteins. It will allow us to analyze the network in a more biological and no so biochemical way, since the relation between the proteins, their distribution inside the network and their topological properties can be studied. We started by studying the topological properties of this new network. Since the network is a directed graph, two different connectivity properties can be considered, incoming connections, and outgoing connections. As most biological networks, the reaction network reveals a scale-free structure, regardless if you consider incoming, outgoing or the total number of connections. The log-log plot of connectivity against the number of nodes show that the number of compounds (k) and the probability of the number of connections (p(k)) can be expressed as $p(k) \sim k^{-\gamma}$, with an exponent between 2 and 3. This indicates the presence of a few highly connected reactions connecting the mass of poorly connected reactions (Fig. 1).

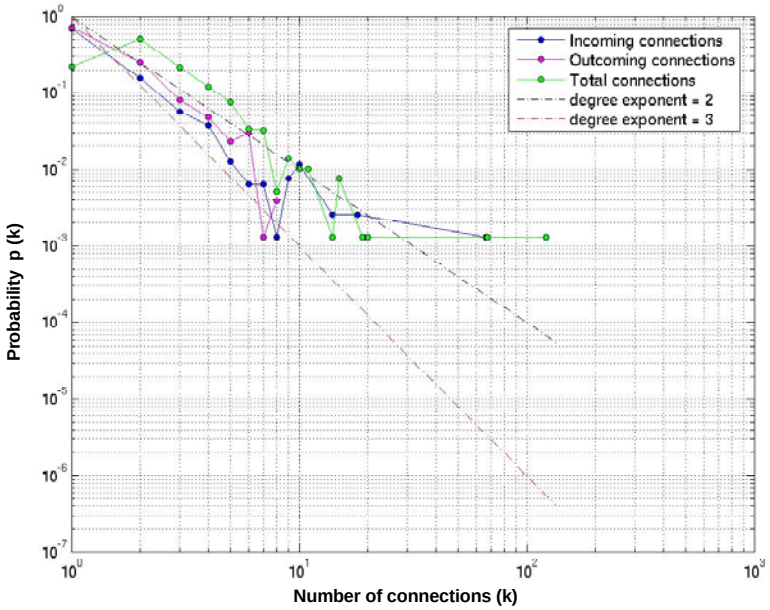


Figure 1. Log-log plots of the number of compounds versus connectivity

One biologically relevant property of the reactions is its connection to the central metabolism (CM). Reactions that are connected to it contribute to the full degradation of compounds, while those that are not connected only contribute to their transformation. Similarly the distance to CM, can also have biological relevance. Thus, we have studied the properties of the nodes according to their distance to the CM. While most of the reactions have only 1 or 2 incoming and 1 or 2 outgoing connections (Fig 2a and 2b), independently of the distance to the CM, there is a clear tendency for the reactions with high inputs or high outputs or high total connectivity, to be closer to the CM and decrease their number according with the distance to the CM

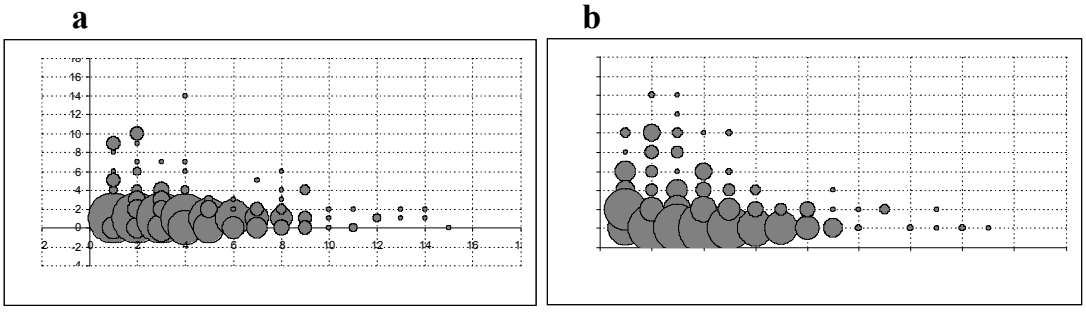


Figure 2. Connections of the reactions. **(a)** Relationship between the number of incoming connections and its distance to the central metabolism (CM). **(b)** Relationship between the number of outgoing connections and its distance to the CM.

Then we studied the relation between the number of incoming (c_i) and outgoing (c_o) connections. This analysis revealed a light "dispersing" structure in the network; there are more nodes with more incoming than outgoing connections (Fig. 3). However, in general, the reactions tend to maintain the flux to the CM, without concentrating or dispersing it, since the more common pair is 1 income, 1 output. It is also noticeable that most of the nodes with no input connection (that is the points were the compound enter the network) has only one output connection. Taken together, all this results display a quite sparse network, with long linear pathways that start to interconnect when close to the central metabolism. This is remarkably different from what was observed when the compound network was studied. In that case, a funnel structure was detected (Pazos et al., 2003).

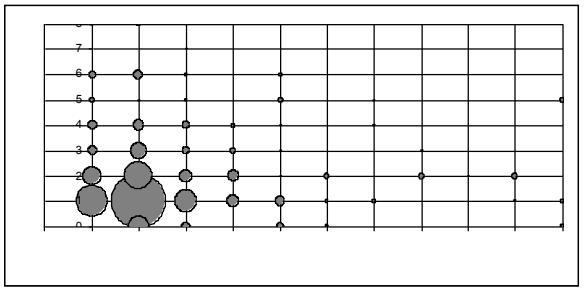


Figure 3. Relationship between the number of incoming connections and the number of outgoing ones.

4.2. Sequence association and its distribution in the network

While it is usual to describe a enzyme by its biochemical activity, i.e. its Enzyme Commission Code (EC Code), this is not enough to uniquely identify the real biological entity that achieves the reaction. More than 60% of the reactions in the biodegradation network share their EC code with other reactions and while in many cases, it does not imply that they can realize the same transformation. This is mainly due to the ambiguity in the definition of the EC code and in its assignation to a reaction. In this manner, to expect that assigning a sequence (or set of homologous sequences) to each reaction would allow a much precise characterization of the reactions of the network. The global biodegradation network consists of 996 reactions, of which, for around 400 have been able to associate a sequence by manual curation of the available bibliography and specific queries to sequence databases. We have focused our analysis in the distribution of the reactions to which we could associate a sequence. This analysis would get insight to the level of knowledge of the system and the regions in which it has been more studied. At the same time, it allows to focus the future work to have a complete description of the network.

A first inspection of the distribution of sequences in the reaction network, revealed that, interestingly, half of the reactions with an associated sequence are completely isolated (42) or in small groups of 2 or 3 elements (140). 215

reactions belong to a interconnected sub network distributed in linear paths often parallel (Fig. 4).

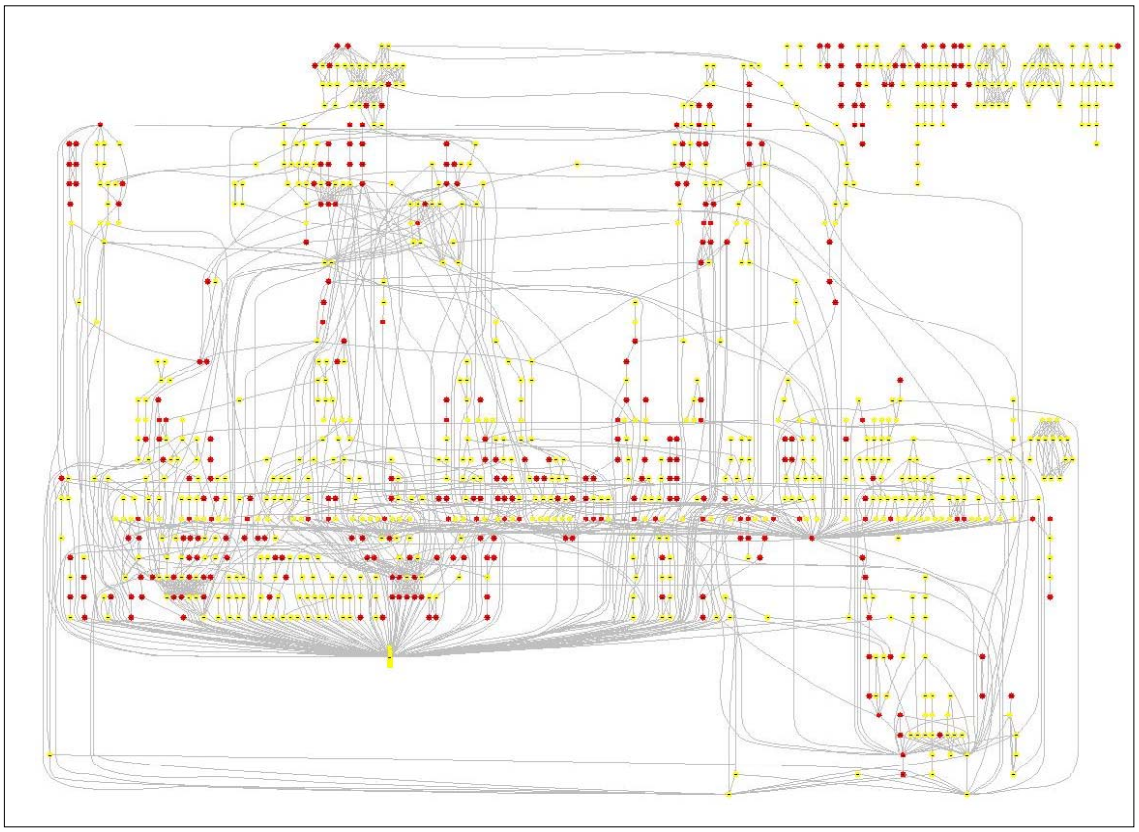


Figure 4. A global view of the biodegradation network from a reaction-centric approach. The dark nodes represent the reaction with a sequence. The rectangle node in the low part of the graph is the entrance to the CM.

In order to get a more detailed analysis of this distribution, we started by taken apart reactions connected to the central metabolism and those not connected (Fig. 5). Nearly two thirds of the reactions reach the CM and for around 45 % of them a sequence has been assigned. This percentage is a 50 % higher than for the nodes without a path to the CM (30%). This suggest a tendency to identify proteins involve in the full degradation of chemicals, as result of the efforts of experimentalist groups to find ways to turn pollutants into innocuous compounds.

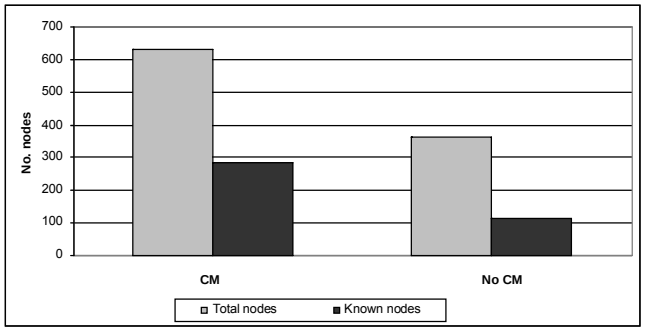


Figure 5. Relationship between reactions with and without a path to the CM and their proportion of total reactions ("total nodes") and reactions with sequence ("known nodes").

Then we turned our attention to the group of reactions connected to the CM, and studied the relation between the distance to the CM and the fact of have an associated protein. The biodegradation network has approximately 80 % of the total number of nodes are less than 6 steps away from the CM. In the same way, 82 % of the reactions with an associated sequence are placed in this region. Making a more detailed study (Fig. 6) we can observe that until a distance 5, 50% of the reactions have an associated sequence. From this point, the percentage decreases and to distances higher than 11 steps, we were unable to associate a sequence to those reactions. The constant proportion of reactions for which a sequence have been obtained at all this distance under 5 can be related to the notion of pathways. It is most likely that sequences are obtained for pathways that appear normally forming operons, or closely associated in DNA fragments, and most biodegradation operons have between 4 and 6 genes (Carbajosa, G. personal communication).

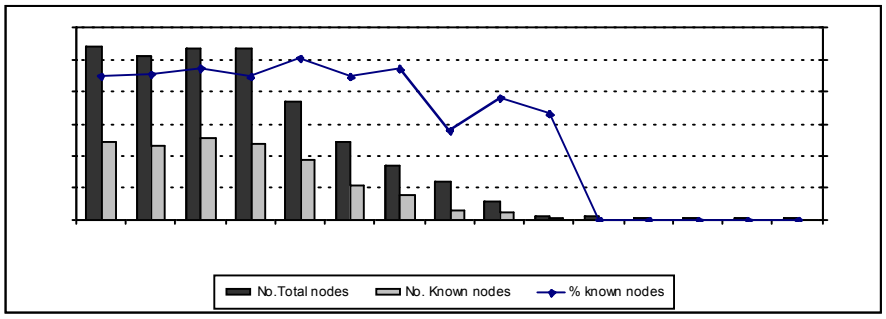


Figure 6. Relationship between the total reactions and those with sequence, and their distance to the CM. The left *y-axis* means the number of nodes (reactions) and the right *y-axis* the percentage of nodes with sequence regarding the total one ("known nodes").

Traditional pathways are a chain of reactions that have been related based in the description from experimental groups, however in many cases their physiological relevance of theses pathaways it is not clear. In fact, our defined network does not incorporate the notion of pathways , we treat the ecosystem as a global no-compartmentalized entity and reactions from different described pathways are allowed to interconnect. However, given the structure of the network described in the previous section and the results regarding the knowledge of sequences relative to distance to CM, we decided to study the distribution of reactions with associated sequences in previously described pathways.

The reactions in our network have been described in 144 pathways, of which around 35% have not any reactions with an associated sequence, while 21% have been completely characterized to sequence level, and for 27% of pathways more than the half of their reactions have an associated sequence. Half of the pathways are characterized more than a 50%, and the other half less than a 50%. Since we have observed a clear turning point in distances around 6 from the CM,

we wondered if there is a correlation between the pathway size and the level of characterization. Out of the 144 pathways included in the network, 75% are shorter than 8 reactions, and 5 is the more common length. Still, there are a small but significant group of 25 pathways longer than 10 reacctions up to 28 (Fig. 7a). When we studied the relation between pathway length and sequence association, we observed a clear tendency of the shorter pathways to be better characterized than the longer ones, and all completely characterized pathways are shorter than 8 reactions (Fig. 7b). A few exceptions can be observed among the long pathways. mainly model pathways, such as the one for degradation of xylene and toluene, which have become the paradigm for molecular biology studies of biodegradation pathways.

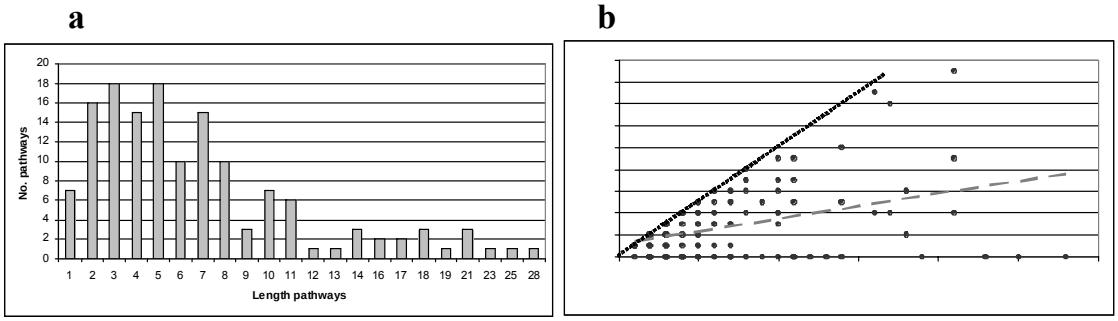


Figure 7. (a) Distribution in the network: number of pathways according with their length. **(b)** Relation between the length of the pathways and their number of reactions with an associated sequence.

5. Conclusions

In this work, we presented a new analysis of the Global Biodegradation with involved a number of improvements over previous ones. First, we have constructed a new formulation of the network in which the reactions occupy the central role, and on top of that we have incorporated sequence information to the description of the reactions. As shown in our topology analysis, this new network have interesting properties, some similar those of the network centred around the chemical compounds, such us the scale-free structure, and other different, such a more linear structure that contrast with the preciously described "funnel" structure. We expect that the analysis of these differences will reveal relevant information about their properties and evolution.

In this first analysis we have also studied the distribution of knowledge in the network, information that we expect will help in focussing subsequent experimental efforts. Our results indicates that so far the experimental work have been centred around reactions connected to the central metabolism, in accordance with the prevalence of the interest for bioremediation (removal of pollutants from the environment by the use of living organisms) in the field of biodegradation. Biotransformation reactions have gained much less attention so far, probably because they do not help to the full mineralization of pollutants. However, given

the actual need of green catalyst and sustainable practices in synthetic chemistry industry, they could be potentially useful, and should therefore be more characterized.

Our results also shown that so far, reactions characterized at the level of sequence are in a large proportion unconnected, restricted to short linear pathways. This makes global analyses more difficult, and constrains our predictive potential on how biodegradation is taken place in the natural environment, where chemical flow almost freely and many different interactions between microorganisms are taking place.

Acknowledgements

This work is funded by the EU COMBIO project (LSHG – CT – 2004 - 503568), and the Fundación Banco Bilbao Vizcaya Argentaria (FBBVA-BIOCON-3). I.C is a member of the Ramón y Cajal Program of the Spanish Ministry of Education and Science.

References

Benson D.A. et al. (2005) GenBank. *Nucleic Acids Res*., **33**, D34-8.

Ellis, L.B., Hou, B.K., Kang, W. & Wackett, L.P., (2003) The University of Minnesota Biocatalysis/Biodegradation Database: Post-Genomic Datamining. *Nucl. Acids Res*., **31**, 262-265.

Pazos, F., Valencia, A. & De Lorenzo, V. (2003) The organization of the Microbial Biodegradation Network from a Systems-Biology perspective. *EMBO Rep*., **4**, 994-999.

Pazos, F., Guijas, D., Valencia, A. & De Lorenzo, V., (2005a) Metarouter: bioinformatics for bioremediation. *Nucl. Acids Res*., **33**, 588-892.

Pazos, F., Guijas, D., Gomez, M.J., Trigo, A., De Lorenzo, V. & Valencia, A., (2005b) The Biodegradation Network a New Scenario for Computational Systems Biology Research. Springer, CMSB 2004. ISBN: 3-540-25375-0, pp. 252–256.

Trigo A., Carbajosa G., Cases I., De Lorenzo V., Valencia A. (2006) BioNeMo database, in preparation.

Venter J.C. et al. (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304**, 66-74.

Wackett, L.P. et al. (2004) Microbial Genomics and the Periodic Table (minireview). *Appl. Environ. Microbiol*., **70**, 647–665.

Metabolic flux improvement through cofactor engineering during L(-)-carnitine production by *E. coli*

V. Bernal*, P. Arense, B. Masdemont, A. Sevilla, M. Cánovas, J. L. Iborra.

Department of Biochemistry and Molecular Biology B and Immunology. Faculty of Chemistry. University of Murcia. Spain. Contact Author: jliborra@um.es. (*)

Keywords: Cofactor engineering, L-carnitinyl-CoA, crotonobetainyl-CoA, CaiB, CaiC, L-carnitine.

1. Abstract

Cofactor level is a control parameter used by cells for flux regulation through metabolic pathways, since it not only affects enzyme activity but also regulates gene expression, finally altering the metabolic state. Perturbation of coenzyme pools, also known as cofactor engineering, is an emerging strategy for metabolic flux redirection with a high potential for metabolic engineering.
Though L(-)-carnitine is a secondary metabolite of *E. coli*, biotransformation occurs at the coenzyme-A level. Coenzyme-A and its thioester derivatives participate in over 100 different reactions in the intermediary metabolism of microorganisms. While coenzyme-A regulates the central and intermediary metabolism, acetyl-CoA has a key role in the link between glycolisis, the Krebs cycle, the glyoxylate shunt and acetate metabolism. The enzymes involved in trimethylammonium compounds activation (CaiC) and CoA transfer between substrates and products (CaiB) are crucial in the biotransformation, suggesting that CoA derivatives act as a feasible bottleneck. In this work, the effect of carnitine:coenzyme-A ligase (CaiC) and carnitine:crotonobetaine:CoA transferase (CaiB) overexpression in *E. coli* LMG194 is analyzed. A three to ten fold increase in the biotransformation yield was assessed. Further, the effect of different carbon sources on enzymes and coenzyme A esters pool of the central metabolic pathways was analyzed. Interrelation between L(-)-carnitine and coenzyme-A metabolism is discussed in relation with energetic primary metabolism.

2. Introduction

L(-)-carnitine (R(-)-3-hydroxy-4-trimethylaminobutyrate) transports long-chain fatty acids through the inner mitochondrial membrane, which is why several clinical applications have been identified for L(-)-carnitine and its demand has increased worldwide, chemical and biological processes having been developed for its production (Kleber, 1997). Strains belonging to the genera *Escherichia*,

Proteus and *Salmonella* racemice D(+)-carnitine or biotransform crotonobetaine, both of which are waste products and represent an environmental problem resulting from the L(-)-carnitine chemical synthesis (Kleber, 1997; Obón et al., 1999; Cánovas et al., 2003). In *E. coli,* the trimethylammonium compounds metabolism has been studied, because of its implication in stress survival and anaerobic respiration, although its role is not totally understood (Eichler et al., 1994; Kleber, 1997; Elssner et al., 2001; Engemann et al., 2005).

Though L(-)-carnitine is a secondary metabolite of *E. coli*, biotransformation occurs at the coenzyme-A (CoA) level (Elssner et al., 2000; Cánovas et al., 2003). In brief, crotonobetaine is transformed into L(-)-carnitine by the involvement of two enzymes, an enoyl-CoA hydratase and a CoA-transferase (Elssner et al., 2001) which are induced anaerobically in the presence of D,L-carnitine mixture and/or crotonobetaine. Using batch and continuous stirred tank reactors with growing and resting *E. coli* cells it was observed that the link between the central and carnitine metabolism was at the level of ATP and the pool of acetyl-CoA/CoA (Cánovas et al., 2003). However, the limitation imposed by the composition of the cellular CoA esters pool in the biotransformation and the need to decipher the limiting steps imposed by the ICDH/ICL and PTA/ACS enzyme ratios still remain unclear. Trimethylammonium compounds activation is performed by a coenzyme-A ligase (CaiC), while CoA transfer between substrates and products is performed by a transferase (CaiB). Both enzymes are crucial in the biotransformation.

Cofactor level is one of the control parameters, which the cell utilizes to regulate fluxes through various metabolic pathways, since not only affects enzyme activity but also regulates gene expression. Coenzyme-A and its thioester derivatives participate in over 100 different reactions in the intermediary metabolism of microorganisms. While CoA regulates the central and intermediary metabolism, acetyl-CoA has a key role in the link between glycolisis, Krebs cycle, glyoxylate shunt and acetate metabolism. Perturbation of coenzyme pools, also known as cofactor engineering, is an emerging strategy for metabolic flux redirection with a high potential for metabolic engineering (San et al., 2002). In fact, the manipulation of the NADH/NAD+ ratio and the CoA esters pool has been used to increase production of industrially useful compounds (Berrios-Rivera et al., 2004; Vadali et al., 2004).

In this work, the effect of carnitine:coenzyme A ligase (CaiC) and carnitine:crotonobetaine:CoA-transferase (CaiB) overexpression in *E. coli* LMG194 is analyzed. Further, the effect of different carbon sources on enzymes and CoA esters pool of the central metabolic pathways was analyzed. Interrelation between L-carnitine and coenzyme A metabolism is discussed.

3. Materials and methods

3.1. Strain and plasmids

E. coli O44 K74 (DSM 8828) and *E. coli* LMG194 (ATCC 47090) were used throughout this study. Both strains contain the complete divergent structural *cai*

and *fix* operons, expressing, carnitine racemase and carnitine dehydratase activities. *E. coli* O44K74 has been isolated as an overexpressing strain for carnitine metabolism (Kleber et al., 1997; Obón et al., 1999). *E. coli* LMG194 [F^- *ΔlacX74 galE galK thi, rpsL ΔphoA (PvuII) Δara714 leu::Tn10*] is defective in L-arabinose metabolism (Guzmán et al., 1995) and was used as expression host. The strains were stored on culture medium containing glycerol (20%) at –20ºC.
Arabinose inducible *pBAD24* was employed as expression vector (Guzmán et al., 1995). *caiB* and *caiC* were PCR-amplified and cloned downstream of the multicloning site of the plasmid. Constructions were verified through sequencing. The construction of pBADcaiB and pBADcaiC was performed employing standard molecular biology techniques (Sambrook et al., 2001), using pBAD24 (Guzmán et al., 1995) as expression vector. For the cloning of *caiB* and *caiC* genes, specific oligonucleotide primers were designed to anneal the 5' and 3' ends of each gene. Futher, specific restriction enzyme cleavage sites for XbaI and PstI were introduced at the ends of the amplified genes and these were employed to perform directed-cloning into the expression vector. Genomic DNA of the L-carnitine overproducing strain *Escherichia coli* O44K74 (DSM 8288) was extracted using Genelute Sigma-Aldrich kit. Plasmid extraction and purification were performed using Qiagen kits.

3.2. Batch cultures

Cells were grown using Miller's LB medium (g/L): 10.0 tryptone, 5.0 yeast extract, 10.0 NaCl. The final pH of the medium was adjusted to 7.5 with KOH. Ampicillin was added at 100 μg/mL. For the biotransformation experiments, 50 mM crotonobetaine was added prior to autoclaving. Anaerobic conditions were maintained to induce the enzymes involved in the carnitine metabolism, while D,L-carnitine mixture, D-carnitine or crotonobetaine were supplied as inducers of *cai* operon, while L-arabinose was used as inducer of the cloned genes at the different concentrations stated in the text.
Batch experiments in anaerobic (under nitrogen atmosphere) assays were performed in reactors equipped with temperature, pH, oxygen and pumps controllers (Biostat B, Braun, Germany). A 1 L culture vessel with 0.5-0.8 L working volume was used.

3.3. Assays

Sample optical density (OD) was followed at 600 nm with a spectrophotometer (Novaspec II, Pharmacia-LKB, Sweden) as a measure of cell concentration. L-carnitine concentration was determined by an enzymatic test (Cánovas et al., 2003), while D,L-carnitine, crotonobetaine and γ-butyrobetaine were determined by HPLC (Obón et al., 1999).

3.4. Enzyme activity determination

The L(-)-carnitine dehydratase and crotonobetaine reductase assays were carried out as previously stated (Cánovas et al., 2003), both using crotonobetaine as substrate. On the other hand, for the determination of D(+)-carnitine racemase activity, D(+)-carnitine was used as the substrate. Enzyme activity was defined

either as the total mmols of substrate consumed per hour (U) or as specific activity, mmol of substrate consumed per hour and mg of protein (mU/mg).

4. Results and Discussion

In order to determine the effect of the pool of coenzyme A derivatives in the biotransformation of L(-)-carnitine, two recombinant strains overexpressing key enzymes in trimethylammonium compounds metabolism were constructed. The genes selected for overexpression experiments were *caiB* and *caiC*, because of their implication in a) coenzyme A transfer between substrate and products and b) synthesis of coenzyme A derivatives of trimethylammonium compounds.
PCR primers were designed upon the database sequences of *caiB* and *caiC* genes (Accession Number: X73904). XbaI and PstI sites were included at the 5' and 3' ends of the genes to be used in direct cloning (Fig. 1). The *ara* promoter based pBAD24 was used as an expression vector (Guzmán et al., 1995). L-arabinose was used as inducer and its presence in the culture medium was assessed by HPLC. All results are thus referred to the *E. coli* LMG194 strain, which carries an almost completely deleted *ara* operon, which was used as cloning and expression host.

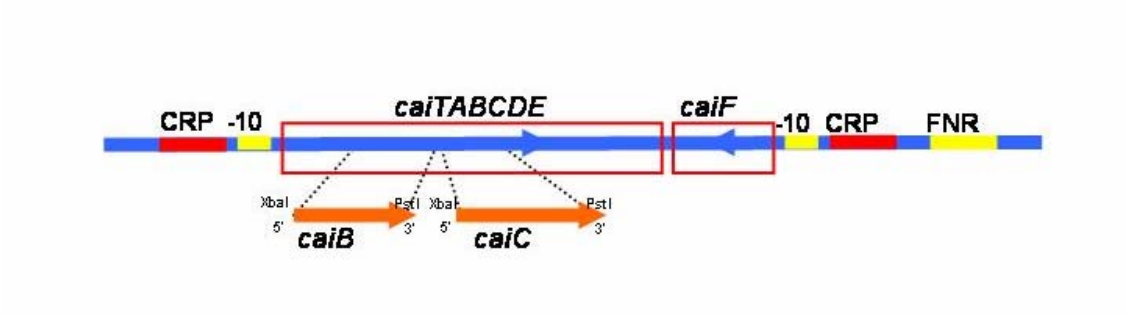


Figure 1. Carnitine operon (*cai*) in *Escherichia coli* (Eichler et al., 1994) and cloned genes. *caiB* and *caiC* ORFs were PCR-amplified and cloned into the arabinose inducible *pBAD24* expression vector (Guzmán et al., 1995).

It has been previously said that at low concentrations of inducer, induction is directly proportional to the concentration. In order to set optimized levels of expression of CaiB and CaiC proteins, the concentration of arabinose was studied. Results showed that optimal concentration of L-arabinose was 0.1-0.2 both in the case of CaiB and CaiC overexpressing strains (Table 1).

Table 1. Optimization of L-arabinose concentration. *E. coli* LMG194 strains were grown in Miller's LB medium. Anaerobiosis was kept in order to ensure efficient expression of carnitine metabolism.

	E. coli LMG194		*E. coli* LMG194 *pBADcaiB*		*E. coli* LMG194 *pBADcaiC*	
[Ara] (%)	DCW (g/L)	[Lcar] (mM)	DCW (g/L)	[Lcar] (mM)	DCW (g/L)	[Lcar] (mM)
0.0001	0.23	0.24	0.25	0.25	0.29	13.03
0.001	0.23	0.40	0.21	0.21	0.28	15.25
0.01	0.26	0.41	0.23	0.23	0.24	16.52
0.1	0.23	0.39	0.28	0.28	0.27	15.12
0.2	0.24	0.50	0.28	0.28	0.20	14.51
1.0	0.20	0.30	0.28	0.28	0.21	11.97

Since crotonobetaine can act as an electron acceptor, the presence of an alternative electron sink in the growth media enhances L-carnitine production by inhibiting the crotonobetaine reductase activity (Kleber, 1997; Cánovas et al., 2003), and thus the γ-butyrobetaine production. Fumarate concentration was also optimized and results showed an enhancement in L-carnitine production not only in the wild type strain, but also in the transformed strains. Maximum production was assessed at 2 g/L for both strains.

Table 2. Optimization of fumarate concentration. *E. coli* LMG194 strains were grown in Miller's LB medium. Arabinose was added at the optimal concentration for each strain. Anaerobiosis was kept in order to ensure efficient expression of carnitine metabolism.

	E. coli LMG194		*E. coli* LMG194 *pBADcaiB*		*E. coli* LMG194 *pBADcaiC*	
[Fum] (g/L)	DCW (g/L)	[Lcar] (mM)	DCW (g/L)	[Lcar] (mM)	DCW (g/L)	[Lcar] (mM)
0.25	0.21	0.97	0.20	3.68	0.26	14.44
0.50	0.21	1.19	0.20	3.21	0.25	15.49
1.00	0.28	2.64	0.20	3.88	0.37	19.03
1.50	0.25	1.78	0.24	3.28	0.34	18.21
2.00	0.27	2.18	0.25	5.33	0.26	20.60
4.00	0.28	1.88	0.23	4.37	0.40	19.54

In order to determine the importance of the existence of a functionally active glyoxylate shunt pathway and the role of the ratio between acetyl-CoA and free CoA, experiments of growth and biotransformation in the presence of pyruvate and acetate were performed. Yield in growth was very different in these two

conditions but in both cases L(-)-carnitine production was inhibited to a great extent (results not shown). The built up of big intracellular pool of acetyl-CoA upon this cultivation conditions, reducing the availability of free CoA for the activation of trimethylammonium compounds would explain this effect assessed.
CaiB protein was the first enzyme in the carnitine metabolism of *E. coli* O44K74 to be characterized. Identification of *caiB* was the milestone which allowed the complete *cai* and *fix* operons sequencing and characterization. On the other hand, although CaiC remains uncharacterized, a possible carnitine/crotonobetaine/γ-butyrobetaine:CoA ligase activity has been proposed on the basis of sequence similarities (Eichler et al., 1994; Engemann et al., 2005). The activation of L(-)-carnitine and derivatives by CaiC is a requirement for the biotransformation, since this metabolism proceeds at the CoA level. Further, the action of CaiB, transferring the CoA moiety between substrate and products allows the biotransformation to proceed through an energetically unexpensive way. Despite this, the fact was that overexpression of CaiC enhanced L-carnitine production with growing *E. coli* cells much more than CaiB.
Addition of fumarate to the growth media allowed to further increase production. It should also be considered the effect of fumarate on central metabolism of *E. coli*. Results recently obtained within our research group pointed to the metabolic modifications suffered by *E. coli* O44K74 upon metabolic pulsing (Cánovas et al., in press).

5. Conclusion

There is a deep relation between coenzyme A availability and carnitine metabolism. The enzymes involved in trimethylammonium compounds activation (CaiC) and CoA transfer between substrates and products (CaiB) are crucial in the biotransformation, suggesting that activation of trimethylammonium compounds into CoA derivatives and CoA transfer between them act as feasible bottlenecks. Further experiments are being accomplished.

Acknowledgements*. This work has been supported by MCYT project BIO2005-08898-C02-01 and CARM project 06 BIO20005/01-6468. V. Bernal is recipient of a predoctoral research grant from Fundación CajaMurcia and A. Sevilla from the Ministerio de Educación y Ciencia. Biosint S.p.A. (Italy) is also acknowledged for the kind gift of the substrate.*

References

Berrios-Rivera, S.J., Sanchez, A.M., Bennett, G.N., San, K.Y. (2004) Effect of different levels of NADH availability on metabolite distribution in *Escherichia coli* fermentation in minimal and complex media. *Appl Microbiol Biotechnol*. **65**, 426-432.

Cánovas, M., Bernal, V., Torroglosa, T., Ramirez, J.L., Iborra, J.L. (2003) Link between primary and secondary metabolism in the biotransformation of trimethylammonium compounds by *Escherichia coli*. *Biotechnol Bioeng*. **84**, 686-699.

Cánovas, M., Sevilla, A., Bernal, V., Leal, R., Iborra, J.L. Role of energetic coenzyme pools in the production of L-carnitine by *Escherichia coli*. *Metab Eng.* (in press)

Eichler, K., Bourgis, F., Buchet, A., Kleber, H-P., Mandrand-Berthelot, M-A. (1994) Molecular characterization of the *cai* operon necessary for carnitine metabolism in *Escherichia coli*. *Mol Microbiol*. **13**, 775-786.

Elssner, T., Hennig, L., Frauendorf, H., Haferburg, D., Kleber, H-P. (2000). Isolation, identification, and synthesis of γ-butyrobetainyl-CoA and crotonobetainyl-CoA, compounds involved in carnitine metabolism of *E. coli*. *Biochemistry* **39**, 10761-10769.

Elssner, T., Engemann, C., Baumgart, K., Kleber, H-P. (2001). Involvement of coenzyme A esters and two new enzymes, an enoyl-CoA hydratase and a CoA-transferase, in the hydration of crotonobetaine to L-carnitine by *Escherichia coli*. *Biochemistry* **40**, 11140-11148.

Engemann, C., Elssner, T., Pfeifer, S., Krumbholz, C., Maier, T., Kleber, H-P. (2005) Identification and functional characterisation of genes and corresponding enzymes involved in carnitine metabolism of *Proteus* sp. *Arch Microbiol*. **183**, 176-189.

Guzman, L.M., Belin, D., Carson, M.J., Beckwith, J. (1995). Tight regulation, modulation, and high level expression by vectors containing the arabinose pBAD promoter. *J Bacteriol*. **177**, 4121-4130.

Kleber, H-P. (1997) Bacterial carnitine metabolism. *FEMS Microbiol Lett.* **147**, 1-9.

Obón, J.M., Máiquez, J.R. Cánovas, M., Kleber, H.P., Iborra, J.L. (1999) High-density *Escherichia coli* cultures for continuous L(-)-carnitine production. *Appl. Microbiol. Biotechnol*. **51**, 760-764.

Sambrook, J., Fritsch, E.F., Maniatis, T. (2001) Molecular cloning: a laboratory manual, 3rd Ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

San, K.Y., Bennett, G.N., Berrios-Rivera, S.J., Vadali, R.V., Yang, Y.T., Horton, E., Rudolph, F.B., Sariyar B, Blackwood K. (2002) Metabolic engineering through cofactor manipulation and its effects on metabolic flux redistribution in *Escherichia coli. Metab Eng.* **4**, 182-192.

Vadali, R.V., Bennett, G.N., San, K-Y. (2004) Cofactor engineering of intracellular CoA/acetyl-CoA and its effect on metabolic flux redistribution in *Escherichia coli*. *Metab Eng.* **6**, 133-139.

THE *TPS2* GENE IS INVOLVED IN THE RESPONSE TO OXIDATIVE STRESS IN *Candida albicans*

E. Martínez-Vicente[1], P. González-Párraga[1], Y. Pedreño[1], M. Martínez-Esparza[2], J.M. Ros[3] and J.C. Argüelles[1*]

[1]Area de Microbiología[1] and Inmunología[2]. Facultad de Biología[1] and Medicina[2]. Universidad de Murcia. E-30071 Murcia. Spain. *e-mail: *arguelle@um.es*
[3]Departamento de Tecnología de los Alimentos, Nutrición y Bromatología. Facultad de Veterinaria. Universidad de Murcia. E-30071 Murcia. Spain.

Keywords: trehalose, trehalose-6P, oxidative stress, HPLC, *Candida albicans*

1. Abstract

The protective role played by disaccharide trehalose against oxidative challenges in *Candida albicans* has been investigated in the homozygous *tps2Δ/tps2Δ* mutant deficient in trehalose-6P-phosphatase activity (encoded by the *TPS2* gene). Whereas growing cultures of the parental strain (SC5314) were able to withstand both moderate (5 mM H_2O_2) and acute oxidative exposures (50 mM H_2O_2), the *tps2Δ* null mutant underwent a marked loss of cell viability. The differential measurement of trehalose and trehalose-6P (T-6P) by a new method based on HPLC analysis, revealed a significant accumulation of T-6P in mutant cells. Remarkably, *tps2Δ* also stored free trehalose, indicating that dephosphorylation of T-6P is rather unspecific. In turn, in parental cells, T-6P was undetectable and the oxidative treatment promoted an additional accumulation of free trehalose. Preliminary analysis revealed a minor resistance of *tps2Δ* cells to lysis mediated by murine macrophages. Collectively, our results strongly support that in *C. albicans, TPS2* gene is involved in the cellular protection against oxidative stress.

2. Introduction

Candida albicans has become the most prevalent opportunistic pathogen fungus in humans, causing from superficial mucosal injuries to life-threatening systemic diseases (Eggimann *et al*., 2003). In the course of an *in vivo* infection, *C. albicans* should counteract the high levels of reactive oxygen species (ROS), which include the radicals O^-_2, H_2O_2 and $^{\cdot}OH$. They are produced from both oxidative metabolism and phagocytic cells. ROS cause oxidative stress and are

Understanding and Exploiting Systems Biology

toxic for essential cellular components (lipids, proteins and nucleic acids), resulting ultimately in cell death (Hohmann and Mager, 2003).
In yeasts, the non-reducing disaccharide trehalose behaves both as a main reserve carbohydrate and as a cellular protector against a variety of nutritional and/or environmental stress challenges (Argüelles, 2000). We have previously demonstrated that trehalose is a specific protector against oxidative damage caused by H_2O_2 (Alvarez-Peral *et al*., 2002). Trehalose is synthesized in two sequential steps: (i) The transfer of a glucosyl unit from UDP-glucose to glucose-6P leads to the formation of trehalose-6P, a reaction catalysed by trehalose synthase (*TPS1*) (Zaragoza *et al.*, 1998); (ii) Dephosphorylation by a specific phosphatase (*TPS2*) gives rise to free trehalose, the physiological stored compound (Van Dijck *et al*., 2002; Zaragoza *et al*., 2002). In turn, the disaccharide mobilisation is brought about by two trehalases: a cytosolic neutral enzyme (Ntc1p) and a cell wall-linked acid trehalase (Atc1p), being the *ATC1* gene cloned by *in silico* screening of a *C. albicans* genome data base (http://genolist.fr/CandidaDB/) (Pedreño *et al*., 2004).
The involvement of trehalose genes as contributory elements in the resistance to oxidative stress has been analyzed. Thus, *tps1Δ* null mutant was very sensitive to *in vitro* oxidative treatments (Álvarez-Peral *et al*., 2002) and to phagocytic lysis carried out by macrophages, whereas *atc1Δ* cells exhibited higher resistance to oxidative exposures concomitant with a lower capacity to undergo dimorphism and a reduced infectivity in a mouse model (Pedreño *et al*., manuscript in preparation). In this study, we have investigated the hypothetical role of *TPS2* gene as a component of the defensive machinery of *C. albicans* against oxidative challenges caused by H_2O_2.

3. Experimental

3.1 Yeast strains and culture conditions

The following strains of *Candida albicans* were used throughout: a wild type SC5314 (SC; TPS2/TPS2/URA$^+$) and its isogenic derivative homozygous mutant *tps2Δ/tps2Δ* (tps2::HISG/tps2::HISG/URA$^+$) deficient in trehalose-6P-phosphatase activity (Tps2p). Yeast cell cultures were grown at 30ºC by shaking in a medium consisting of 2% peptone, 1% yeast extract and 2% glucose (YPD). The strains were maintained by periodic subculturing in solid YPD. Growth was monitored by measuring the changes in optical density of cultures at 600 nm in a Shimadzu U/V spectrophotometer.

3.2 Oxidative stress treatments

Exponentially YPD-growing cultures ($O.D._{600}$ = 0.8-1.2) were divided into several identical aliquots, which were treated with the indicated H_20_2 concentrations (or maintained without H_20_2 as a control) and incubated at 30ºC for one hour. The percentage of cellular viability was determined after appropriate dilution of the samples with sterile water by plating in triplicate on

solid YPD. Between 30 and 300 colonies were counted per plate. Survival was normalized to control samples (100% viability).

3.3 Determination of trehalose and trehalose-6-phosphate (T-6P)

Intracellular trehalose was extracted from 20-50 mg yeast samples and measured following the method described by Alvarez-Peral *et al*. (2002). For T-6P measurements, identical samples were broken by vigorous shaking with Ballotini glass beads (0.45 mm) for 5 min at 4ºC, and the cell-free supernatant boiled during 30 min was resuspended in 2 ml water (milliQ) and analyzed by HPLC using a CHO-682 column with a CHO-682 guard-column, both from Interaction (San Jose, CA, USA). High quality water was used as the eluent at a constant flow of 0.4 ml/min and 80ºC. Detection was carried out by RID. Pure trehalose and trehalose-6P from (Sigma) were used as standards. Supernatants from intracellular extracts were centrifuged (16.390 xg-5 min) to remove insoluble solids, in advance to direct injection for HPLC analysis.

3.4 Quantification of macrophage fungicidal activity

The murine macrophage-like tumour cell line J774 (from a female BALB/c mouse) was obtained from ATCC (Rockville, MD, USA). Adherent cells were cultured at 37ºC in an atmosphere containing 5% CO_2 in DMEM (Biowhittaker, Verviers, Belgium) supplemented with 10% heat-inactivated foetal calf serum (Gibco, NY, USA), 5 mM L-glutamine (Seromed Biochrom), 100 μg/ml streptomycin and 50 μg/ml penicillin (Flow lab., Irvine, UK) here referred as complete medium. For experiments, J774 cells were distributed into 24-well culture plates at $4x10^5$ cells/well. After 18 h, the adherent cells were washed with culture medium and monolayer cells were incubated 2 h with *C. albicans* blastoconidia at 10/1 yeasts/macrophage ratio. Next, the monolayers were washed thoroughly twice with cold PBS, resuspended in 1 ml of sterile distilled water at 37º C for 5 min and vigorously shaken with a micropipette to lyse the cells. Inspection of the initial lysate revealed only single colonies, 98% of which were still in the yeast phase. Finally, appropriate dilutions of lysates were plated and incubated at 28°C for 48 h. The colony-forming units (CFU) were counted and the percentage of surviving yeast was calculated by comparison to the CFU obtained in absence of macrophages.

4. Results and Discussion

4.1 Effect of H_20_2 exposures on cell survival and trehalose storage.

The sensitivity to oxidative stress (H_20_2) is examined in logarithmic yeast cells (blastoconidia) of the wild type (SC) and the trehalose-6P-deficient mutant (*tps2Δ/tps2Δ*). Addition of moderate non-lethal concentration of H_20_2 (5 mM) caused a partial loss of cellular viability compared to control samples, which was more pronounced in mutant cells (Fig. 1). A 10-fold increase (until 50 mM H_20_2) was necessary in order to get a several reduction in cell survival (Fig. 1). Again, the strongest phenotype corresponded to *tps2Δ* mutant. However, the degree of

cell killing recorded was clearly lower respect to that recorded in the trehalose synthase-deficient mutant *tps1Δ* (Alvarez-Peral *et al*., 2002).

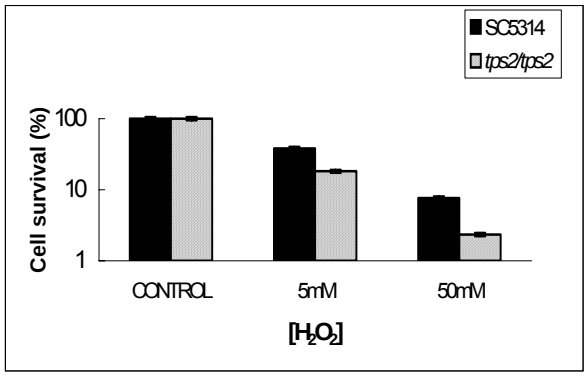


Figure 1. **Levels of cellular survival in parental and *tps2Δ* strains of *C. albicans* after an oxidative stress treatment**. YPD-grown exponential cultures (O.D. 1.0) were divided in three identical samples and subjected for 60 min with 5 mM and 50 mM H_20_2. The percentage of viability is referred to an identical, untreated sample (100% viability). The values are the average of two independent determinations.

The simultaneous determination of endogenous trehalose in SC blastoconidia, revealed a basal level of the disaccharide, that underwent a marked augment upon a subsequent oxidative challenge whether gentle or acute (Table 1).

Table 1. **Changes in the content of intracellular trehalose (T) and T-6P after an oxidative challenge induced by H_20_2**. The same cultures and conditions used for the experiment depicted in Fig.1 were employed for these metabolic determinations.

	TREHALOSE*		TREHALOSE-6P (T-6P)#	
$[H_2O_2]$	**SC5314**	***tps2Δ***	**SC5314**	***tps2Δ***
CONTROL	2.8	1.7	<0.5	0.06
5 mM	7.4	2.6	<0.5	0.31
50 mM	11.2	3.3	<0.5	0.27

*nmoles/mg weight wt.
#mM

Interestingly, in *tps2Δ* cells, significant amounts of trehalose also were accumulated, with a weaker up-shift after oxidative stress (Table 1). Although previously observed (Van Dijck *et al*., 2002), this result is somehow surprising, since this mutant lacks the phosphatase (Tps2p) that renders free trehalose. Apparently, *C. albicans* contains one or several non-specific phosphatases which are able to dephosphorylate T-6P, but with minor efficiency than T-6P phosphatase.

In the course of this study, we have also developed an easy and reliable procedure for the differential measurement of trehalose-6P (T-6P), clearly distinguishable from trehalose (Fig. 2), which is based on HPLC technology (see Methods for details). HPLC analysis of samples from microbial sources revealed

to be a powerful tool (Hellín *et al*., 2001; 2003). Under the chromatographic conditions employed, T-6P was clearly separated from trehalose, with retention times of 10.9 and 17.0 min, respectively (Fig. 2).

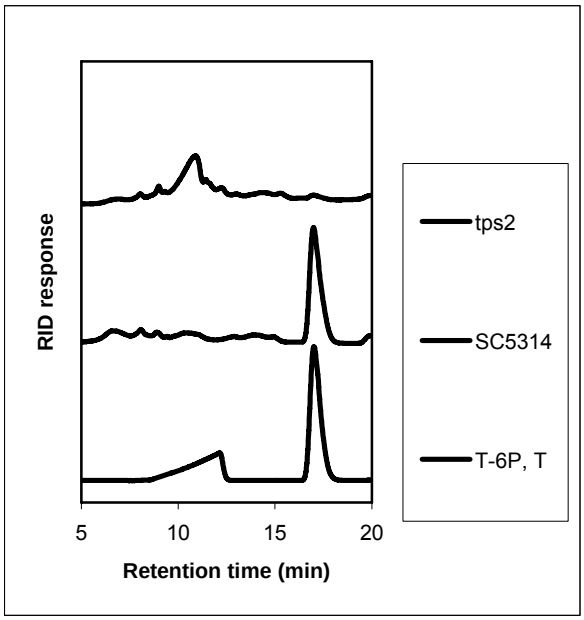


Figure 2. **High performance liquid chromatography elution profile of trehalose and trehalose-6P**. Samples containing: the pure standards trehalose-6P (T-6P) and trehalose (T) (lower plot); the supernatants from cell-free extracts of *C. albicans* parental (SC5314) (middle plot), and of the *tps2Δ* null mutant, were loaded on a CHO-682 column. The latter two samples were subjected to an oxidative stress treatment. Average retention times were 10.9 min for T-6P (trehalose-6P) and 17.0 min for T (trehalose). For other details related to the specific HPLC procedure, see the Experimental section.

The refractive index detector used had a sensibility well enough to measure amounts at the nmol range. A different elution profile between the "natural" T-6P and the "standard" T-6P was observed (Fig. 2). The reason of this behaviour is not yet evident and is under investigation. However, since "standard" T-6P consists in the dipotassium salt form, it is possible that in a solution some kind of aggregation occurred, appearing dimers, trimers, tetramers, etc. of T-6P, which can elute early respect to the "single" monomer of T-6P. This behaviour is consistent with the elution profile of carbohydrates according to the degree of polymerization in the CHO-682 column used. Since "natural" T-6P does not contain the dipotassium salt form like the pure standard, a different HPLC behaviour could be detected.

According to the results obtained (Fig. 2, Table 1), T-6P might also serve as a sensor of oxidative stress, since larger accumulation is achieved after H_20_2 exposures (Table 1). Furthermore, an excess of T-6P induces toxicity (Zaragoza *et al*., 2002); this effect could contribute to the cell susceptibility recorded in *tps2Δ* cells subjected to oxidative treatments (Fig. 1).

4.2 SC5314 is more resistant than tps2Δ/tps2Δ mutant to macrophage killing.

To assess the defensive the role of trehalose in the ability of *C. albicans* to survive the stress conditions triggered during phagocytosis mediated by macrophages, we compared the degree of cell viability in the two *C. albicans* strains engulfed by J774 cells. As shown in Fig. 3, exponential cells from the *tps2Δ* null mutant are less resistant to killing mediated by murine macrophages

than their counterpart SC parental. Although new experiments are currently in progress, this outcome might probably be attributed to the deficient cellular protection against oxidative stress within the phagolysosome, as consequence of *tps2Δ* inability to synthesize substantial levels of endogenous trehalose.

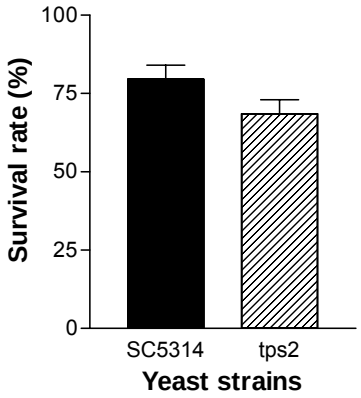


Figure 3. Intracellular yeast survival after phagocytosis. J774 cells were co-cultured at 1:10 cell/yeast ratio, with *C. albicans* SC5314 or *tps1/tps1* strains for 2 h at 37º C. Endocytosed yeasts were recovered from macrophages by osmotic lysis and the number of colony forming units (CFU) was scored after 48 h.

References

Alvarez-Peral, F.J., Zaragoza, O., Pedreño, Y. and Argüelles, J.C. (2002) Protective role of trehalose during severe oxidative stress caused by hydrogen peroxide and the adaptive oxidative stress response in *Candida albicans*. *Microbiology* **148**, 2599-2606.

Argüelles, J.C. (2000) Physiological roles of trehalose in bacteria and yeast: a comparative analysis. *Arch. Microbiol.* **174**, 217-224.

Eggimann, P., Garbino, J. and Pittet. D. (2003) Epidemiology of *Candida* species infections in critically ill non-immunosuppressed patients. *Lancet Infect. Dis*. **3**, 685-702.

Hellín, P., Ros, J.M. and Laencina, J. (2001) Changes in high and low molecular weight carbohydrates during *Rhizopus nigricans* cultivation on lemon peel. *Carbohydr. Polym*. **45**, 169-174.

Hellín, P., Laencina, J. and Ros, J.M. (2003) Isolation and characterisation of resistant-to-fermentation carbohydrate polymers from cultures of *Rhizopus nigricans* grown on agrofood waste materials. *Biotechnol. Lett*. **25**, 1875-1880.

Hohmann, S. and Mager W.H. (Eds.) (2003). Yeast stress responses. Springer-Verlag. Berlin.

Pedreño, Y., Maicas, S., Argüelles, J.C., Sentandreu, R. and Valentín. E. (2004) The *ATC1* gene encodes a cell wall-linked acid trehalase required for growth on trehalose in *Candida albicans*. *J. Biol. Chem*. **279**, 40852-40860.

Van Dijck, P., De Rop, L., Szlufcik, K., Van Ael, E. and Thevelein. J.M. (2002) Disruption of the *Candida albicans TPS2* gene encoding trehalose-6-phosphate phosphatase decreases infectivity without affecting hypha formation. *Infect. Immun*. **70**, 1772-1782.

Zaragoza, O., Blazquez, M.A. and Gancedo. C. (1998) Disruption of the *Candida albicans TPS1* gene encoding trehalose-6P-synthase impairs formation of hyphae and decreases infectivity. *J. Bacteriol*. **180**, 3809-3815.

Zaragoza, O., de Virgilio, C. Ponton, J. and Gancedo, C. (2002) Disruption in *Candida albicans* of the *TPS2* gene encoding trehalose-6-phosphate phosphatase affects cell integrity and decreases infectivity. *Microbiology* **148**, 1281-1290.

Biochemical characterization of mycobacterial phosphoglucose isomerase and its mutants

Divya Mathur, Zaid Ahsan, Madhulika Tiwari and Lalit C. Garg

Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg. New Delhi-110067 divyamathur78@rediffmail, lalitcgarg@yahoo.com.

Keywords: Phosphoglucose isomerase; *Mycobacterium tuberculosis*; Glycolysis; Fructose-6-phosphate**.**

1. Abstract

PGI from *Mycobacterium tuberculosis* H37Rv was cloned and expressed in *E. coli*. Wild type recombinant phosphoglucose isomerase (rPGI) from soluble fraction was purified to near homogeneity by Ni-NTA ion-exchange chromatography. Mycobacterial PGI exhibits catalytic and biochemical properties belonging typically to enzymes of the PGI superfamily. The enzyme is a homodimer and Mass spectrum analysis of the purified rPGI revealed it to be of 61.45 kDa. The K_m of rPGI was determined as 0.27±0.03 mM for fructose-6-phosphate and K_i was 0.75 mM for 6-phosphogluconate. The rPGI had optimal activity at 37°C and pH 9.0 and did not require mono or divalent cations for its activity. The specific activity of recombinant enzyme was 600 U/mg protein. Further, to evaluate the role of crucial amino acid residues, site directed mutagenesis was carried out targeting specific residues to generate mutant proteins.

2. Introduction

Mtb virulence is correlated with a shift from a strict aerobic respiratory mode to anaerobic metabolism. Although ambient expression of glycolytic enzymes is necessary for steady state metabolism, the coordinated increased expression of genes encoding glycolytic enzymes is particularly important for adaptation to hypoxia. *In vivo* growth studies of the organism have indeed indicated that up to 70% of glucose metabolizes through EMP pathway (Ramakrishnan et al, 1962: JayanthiBai et al, 1975). Thus, glycolysis is central to the organism's survival and consequently a potential drug target. Phosphoglucose isomerase (PGI; EC.5.3.1.9) is a key enzyme in glycolysis that functions at the juncture of gluconeogenesis and catabolism. It catalyzes the reversible isomerization of D-glucopyranose-6-phosphate and D-fructofuranose-6-phosphate by promoting the transfer of proton between C1 and C2 and thus exerts considerable control at a pivotal point at the juncture of three metabolic pathways, i.e. the Embden-Meyerhoff-Parnas, the Entner-Doudoroff, and the phosphogluconate pathways.

Further, it also plays a key role in the pathway of cell wall biosynthesis in mycobacteria as glucose-6-phosphate is required for the galactan residue of arabinogalactan (Trejo et al, 1970: Trejo et al, 1971). These key roles that PGI plays in the biology of mycobacterium make it a potential drug target. Hence, a detailed biochemical characterization of *Mtb* PGI is necessary for effective drug designing. In the past, the biochemical characterization of metabolic enzymes has been exploited for the development of potential drug against various pathogens like *Plasmodium falciparum*, Trypanosomes, HIV etc.

3. Theoretical

PGI from *Mycobacterium tuberculosis* (Cole et al, 1998) shows 30% sequence homology with the human PGI. Multiple sequence alignment of the amino acid sequence of the enzyme across different species showed that the two sequences -[DENS]-X-[LIVM]-G-G-R-[FY]-S-[LIVMT]-X-[STA]-[PSAC]-[LIVMA]-G- and [GS]-X- [LIVM]-[LIVMFYW]-XXXX-[FY]-[DN]-Q-X-G-V-E-X-X-K- have remained conserved.
The probable crystal structure of *Mtb* PGI was constructed using PyMOL program (Delano 2002) and on the basis of probable location of these conserved residues four amino acids were chosen for single amino acid mutations on PGI.T212 and G156 lie in the cavity where the substrate sugar is thought to bind. N314 and G360 are surface residues and might be involved in interaction between two subunits. Thus, it is speculated that mutation of these residues can provide us with better insights into factors responsible for enzyme catalysis.

4. Experimental

4.1. Cloning of *pgi* gene in the expression vector, pET-22b(+)

The *pgi* gene was amplified using gene specific primers (Forward- 5'CCCCATATGACCTCCGCGCCAATC-3' and Reverse- 5' CAAACTCGAGTTAGCCCG CGCGGCCACGTT-3') designed on the basis of genome sequence information of *Mtb* H37Rv. *NdeI* and *XhoI* sites (underlined) were introduced in the forward and reverse primers, respectively for convenient cloning in expression vector. BAC clone Rv103 was used as a template. The amplified product of approximately 1.7 Kb corresponding to the complete *pgi* gene was cloned into pGEM-T Easy vector and transformed into *E. coli* DH5α cells. The integrity of the cloned *pgi* gene was verified by automated DNA sequencing.
The 1659 bp *pgi* gene fragment released from the recombinant pGEM-T.*pgi* clone by *XhoI* and *NdeI* digestion, was ligated to plasmid pET-22b(+) digested with the same enzymes. The ligation mixture was transformed into *E. coli* BL21 (DE3) cells (Novagen, USA) and selected on LB-Agar plate containing ampicillin (100 μg/ml). Recombinant colonies were analyzed by restriction digestion with *XhoI* and *NdeI* for the release of the insert.

4.2. Site Directed Mutagenesis

The *in vitro* mutagenesis system Quickchange® (Stratagene, Germany) was used to introduce mutations in the *pgi* gene. Oligonucleotides were designed against the regions surrounding the codons to be mutated. After mutagenesis, plasmid DNA was extracted and purified following the manufacturer's instructions. The clones were screened for the presence of a mutation by direct sequencing.

4.3. Enzyme Preparations

rPGI was purified from the soluble fraction of induced culture by one step Ni-NTA affinity chromatography and eluted in 250 mM imidazole. Proteins were dialyzed against Tris-PO_4 (0.01 M Tris, 0.1 M sodium dihydrogen phosphate) to remove imidazole and checked on 12% SDS-PAGE. BCA protein assay kit (Pierce, USA) was used for protein estimation using bovine serum albumin (BSA) as standard.

4.4. Kinetic measurements

All enzyme kinetics experiments were performed at ambient temperature in 1cm path length quartz cuvettes. Phosphoglucose isomerase activity was determined as described previously (Mathur et al, 2005) by monitoring the increase in absorbance due to the reduction of $NADP^+$ to NADPH at 340 nm. The assay mixture in a total volume of 1ml contained 0.1 mM Tris-chloride buffer (pH 7.6), 2 mM EDTA, 0.5 mM $NADP^+$, 1mM frutose-6-phosphate, 1U glucose-6-phosphate dehydrogenase and ≈0.5 U of the purified recombinant enzyme. The reaction mixture was incubated at 25 °C for 10 minutes and the reaction was initiated by the addition of the enzyme. The reaction was followed for 5 min. The activity was measured by monitoring the change in the absorbance at 340 nm using spectrophotometer Lambda25 (Perkin Elmer, USA).One unit of PGI activity is defined as the amount of enzyme that catalyzes the conversion of 1μM of fructose-6-phosphate to glucose-6-phosphate per minute under the above assay conditions.

5. Results and Discussion

Comparison of amino acid sequence between different species using ClustalW (Thompson et al, 1994) showed that the percentage identities/similarities of the *Mtb* PGI with the PGIs of human, mouse, pig, rabbit, drosophila, leishmania, plasmodium, trypanosoma and bacillus were 46%, 47%, 48%, 49%, 48%, 49%, 37%, 47% and 17%, respectively. The clone containing the *pgi* gene in the pET vector at the *XhoI* and *NdeI* sites was named pET.pgi (Fig.1) and was induced with IPTG. rPGI was purified from the soluble fraction of the induced culture using the one step Ni-NTA affinity chromatography.

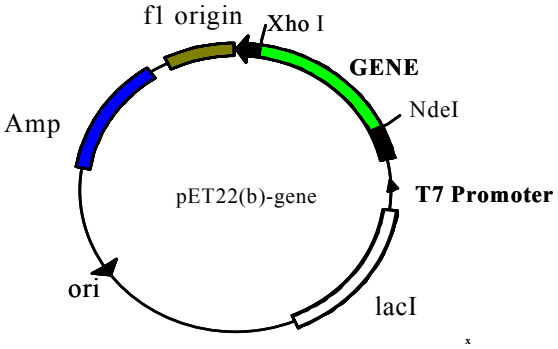


Fig. 1 Schematic representation of cloned *pgi* in the expression vector pET22b(+)

The native molecular mass of the rPGI determined by gel filtration chromatography was ~120 kDa (Fig.2) whereas the molecular mass of the purified rPGI by Mass spectrum analysis was determined to be of 60.266 kDa (Fig.3). These data collectively suggest a homodimeric nature of *Mtb* PGI.

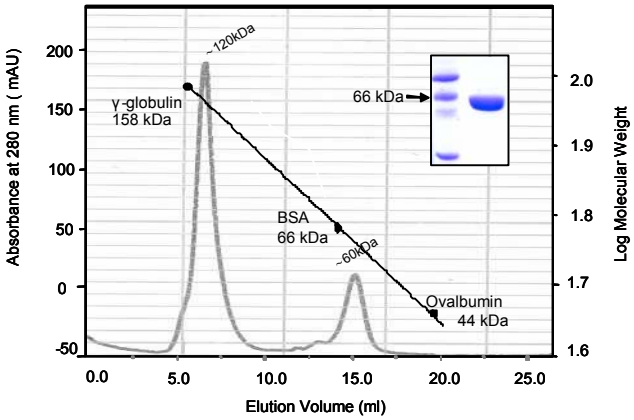


Fig. 2 Gel filtration of recombinant PGI

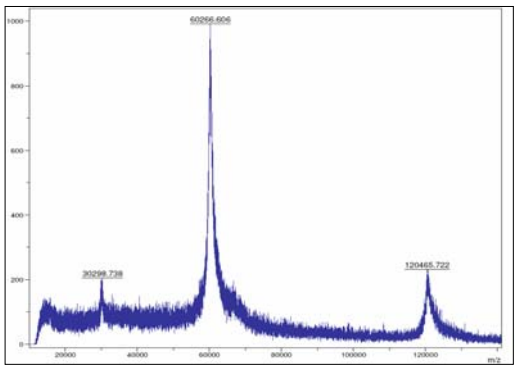


Fig. 3 Mass spectrometry of mycobacterial rPGI

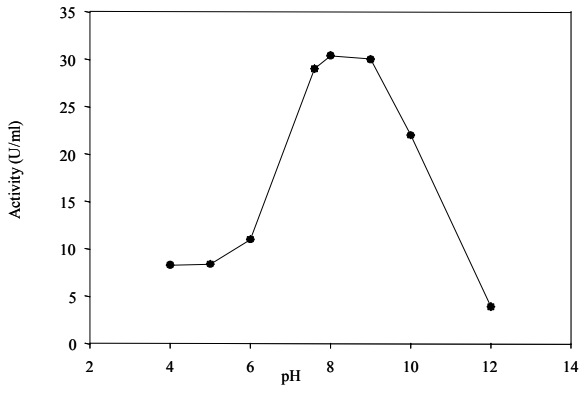


Fig. 4 pH optima of rPGI

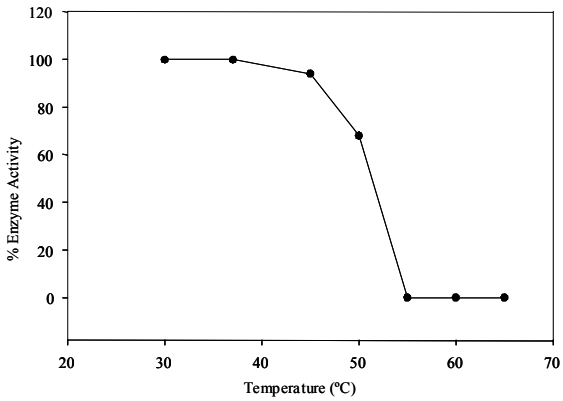


Fig. 5 Thermostability of rPGI

No significant change in the activity of *Mtb* rPGI was observed with change in the concentration of monovalent (Na^+ and K^+) and divalent cations (Mg^{++} and Ca^{++}) (data not shown). The activity of phosphoglucose isomerase from *Mtb* was determined at different pH. The rPGI exhibited bell shaped curve and was active over a broad pH range from 6-11, with maximum activity at pH 9.0.
The enzyme followed Michaelis-Menten kinetics and the K_m and V_{max} were calculated using Lineweaver-Burk plot. At room temperature, the K_m of rPGI (WT) was determined to be 0.27 ± 0.03 mM for fructose-6-phosphate with a V_{max} of 0.032 μmol / min /mg. The inhibition of the rPGI by 6-phosphogluconate was examined and the inhibition constant (K_i) for 6-phosphogluconate was calculated to be 0.75 mM.

Table 1. Oligonucleotides for site directed mutagenesis of rPGI. The residues for mutation are underlined.

Oligonucleotides for mutagenesis of pgi	
Mutation Glycine156-Y	**Pgi1F** 5'-GTCAACATCGGCATCGGTTACTCGGATTTGGGTCCG -3' **Pgi1R** 5'-CGGACCCAAATCCGAGTAACCGATGCCGATGTTGAC -3'
Mutation Threonine212-A	**Pgi2F** 5'-CTTTTCATCGTCGCGTCGAAGGCGTTCTCGACGCTG- 3' **Pgi2R** 5'-CAGCGTCGAGAACGCCTTCGACGCGACGATGAAAAG-3'
Mutation Aspargine314-R	**Pgi3F** 5'-CCGCTGGAATCCCGCGCGCCGGTGCTG -3' **Pgi3R** 5'-CAGCACCGGCGCGCGGGATTCCAGCGG -3'
Mutation Glycine360-E	**Pgi4F** 5'-CAGTTGACCATGGAATCCAACGAGAAGTCCACGCGCG CC -3' **Pgi4R** 5'-GGCGCGCGTGGACTTCTCGTTGGATTCCATGGTCAACT G-3'

Single amino acid replacement mutations were made in the *pgi* gene using PCR based mutagenesis. The oligonucleotides used for the purpose are shown in Table 1. The mutations were confirmed by direct DNA sequencing of the clones. The purification and characterization of the mutants is underway to fathom the affect of these mutations on enzyme activity.

6. Conclusion

The present investigation on *Mtb* rPGI revealed that there are no significant differences in the biochemical/catalytic properties of *Mtb* PGI and the PGIs from other species. Therefore, we can conclude that all PGIs have the same evolutionary origin and employ the same catalytic mechanism.
Crystal structure of PGI from few species has been established. Despite having the similar overall fold, there are significant structural differences mainly in the large domain of the enzyme (Graham Solomons et al, 2004). Subtle differences can also be found in conformation of the small domain of different PGIs (Cordeiro et al, 2004). The structural differences between the host and pathogen PGIs can be exploited for designing drugs, specifically targeting the pathogen. Therefore, elucidation of crystal structure of *Mtb* PGI and its mutants involving structurally and functionally important residues is necessary to identify the characteristics unique to *Mtb* PGI for effective drug designing. We have successfully expressed enzymatically active rPGI from *Mtb* and the crystallization studies are in progress as a step towards the same.

Acknowledgements*. Financial support from the Department of Biotechnology, India is acknowledged.*

References

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E. 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, **393**, 537-544. Erratum in Nature (1998) 396 (6707)190.

Cordeiro, A.T., Michels, P.A.M., Delboni, L.F., Thiemann, O.H. (2004) The crystal structure of glucose-6-phosphate isomerase from *Leishmania mexicana* reveals novel active site features, *Eur. J. Biochem*. **271,** 2765-2772.

DeLano, W.L. (2002) The PyMOL molecular graphics system. DeLano Scientific, San Carlos, CA,USA, http://www.pymol.org.

Graham Solomons, J.T., Zimmerly, E.M., Burns, S., Krishnamurthy N., Swan, M.K., Krings S., Muirhead H., Chirgwin, J., Davies C. (2004) The crystal structure of mouse phosphoglucose isomerase at 1.6A resolution and its complex with glucose 6-phosphate reveals the catalytic mechanism of sugar ring opening*, J Mol. Biol.* **342**, 847-860.

Jayanthi Bai, N., Ramchandran, P.M., Suryanaryana Murthy, P., Venkitasubramanian, T. (1975) Pathways of carbohydrate metabolism in *Mycobacterium tuberculosis* H37Rv. *Can. J. Micro*. **21**, 1688-1691.

Mathur, D., Ahsan, Z., Tiwari, M. and Garg L.C. (2005) Biochemical characterization of recombinant phosphoglucose isomerase of *Mycobacterium tuberculosis. Biochem. Biophy.s Res. Commun.* **337**, 626-632.

Ramakrishnan, T., Indira, M., Maller, R.K. (1962) Evaluation of the routes of glucose utilization in virulent and avirulent strains of *Mycobacterium tuberculosis*. *Biophys. Biochem. Acta*. **59**, 529-532.

Thompson, J.D.,Higgins, D.G.,Gibson, T.J.(1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res*. **22**, 4673-4680.

Trejo, A.J., Chittenden, G.J.F., Buchnan, J.G., Baddiley, J. (1970) Uridine diphosphate alpha-D-galactofuranose, an intermediate in the biosynthesis of galactofuranosyl residues. *Biochem. J.* **117**, 637-639.

Trejo, A.J., Haddock, J.W., Chittenden, G.J.F., Baddiley, J. (1971) The biosynthesis of galactofuranosyl residues in galactocarlose. *Biochem. J.* **122**, 49-57.

The effect of the stringent response regulon and *rpo*35* mutation on mechanism of DNA repair in *E. coli*

Pourahmad Jaktaji R.[1*] and Lloyd R. G.[2]

[1]Biology Department, The University of Shahrekord, Shahrekord 88186/34141, IRAN. Razieh_Jaktaji@yahoo.com,
[2]Institute of Genetics, Queen's Medical Centre, The University of Nottingham, Nottingham NG7 2UH, UK bob.lloyd@nottingham.ac.uk

1. Abstract

Stringent response is one of the global regulatory systems or regulon in bacteria. Nutritional deprivation or other conditions that arrest the growth of cells activated this regulon. This leads to an increase in the level of (p)ppGpp. (p)ppGpp is the modulator of RNA polymerase activity. Certain stringent RNA polymerase mutation, *rpo*35*, that mimics the elevated level of (p)ppGpp enhances survival of UV irradiated *E. coli* cells devoid of RuvABC protein, a key protein in recombination repair pathway. In this work by transposon mutagenesis and transductional analysis I demonstrate that repair pathway in *relA1 spoT207 rpo*35 ruvAC* strain relies on PriA protein and does not require RecBCD. However, it does require UvrABC excision repair pathway. Survival promoted by *rpo** also depends on LexA regulated SOS response and RecF proteins.

2. Introduction

Activation of stringent response following nutritional deprivation or other stressful conditions leads to an increase in the synthesis of (p)ppGpp (an activator of stringent response, derived from GTP) which is mediated by RelA and SpoT proteins. Deletion of RelA and SpoT proteins considerably reduce survival of UV-irradiated *ruv* strain, whereas elevation of (p)ppGpp synthesis promotes survival of *ruv* mutant (McGlynn and Lloyd, 2000). They found stringent RNA polymerase mutants like *rpo*35* in a *relA spot ruv* background which mimics the elevated level of (p)ppGpp suppress UV sensitivity of *ruv* strains. The *rpo*35* mutation and its effect on DNA repair is the subject of this paper.

3. Experimental Procedures

3.1. Media and general methods

LB broth and agar and 56/2 minimal salts media were used for bacterial culture. Media recipes and procedures for strain construction by P1*vir* mediated

transduction, testing sensitivity to mitomycin C and measuring survival of UV-irradiated cells cited in Jaktaji and Lloyd (2003).

3.2. Transposon mutagenesis

Tn*10kan* insertions were generated by infection of strain N4538 with λNK1327 and selected for kanamycin resistant clones at 42º C as described previously (Kleckner *et al*., 1991).

3.3. PCR amplification and DNA sequencing

Chromosomal DNA was extracted as described by Sambrook *et al*. (1989). Mutations in *uvrA*, *uvrC*, *recB* and *priA* genes were identified by sequencing PCR products amplified from chromosomal DNA using *uvrA* (5′-CACACACGGCACGCTTCC-3′), *uvrC* (5′-GATCTTCTGGTCGTTG-3′), *recB* (5′-CCGGCAAACATCTCATCC-3′) and *priA* (5′-CTCCAGCCCACTGGCAGACG-3′) specific primers and IS*10* specific primer 5′-CACCTATGTGTAGAACAGTATA-3′.

4. Results

McGlynn and Lloyd (2000) suggested that RNA polymerase enzymes stalled at lesions in DNA are major obstacles to replication forks progression in UV-irradiated cells. They also proposed that by modulation of RNA polymerase activity, (p)ppGpp and *rpo*35* mutation reduce the incidence of stalled complexes thus reducing the need for RuvABC resolvase to promote survival. To investigate the genetic factors enabling *rpo*35* to promote survival of *ruv* strains, Tn*10kan* was inserted at random into chromosome of the *relA1 spoT ruv rpo*35* strain, N4538. Km^r clones were selected and screened for sensitivity to UV light and mitomycin C (MC). Among fourth clones obtained three were extremely UV sensitive (Table 1) and transcriptional and PCR analysis showed that they had insertions in *uvrA*, *uvrC* and *priA* genes (Fig. 1).

Table 1 Effect of insertion and known mutations on UV survival

Strain	Relevant Genotype[a]	Fraction surviving[b]
MG1655	Wild type	0.9
N4538	*relA1 spoT rpo* ruvAC65*	0.4
RJ1004	*relA1 spoT rpo* ruvAC65 uvrA*	0.00003
RJ1025	*relA1 spoT rpo* ruvAC65 uvrC*	0.000035
RJ1050	*relA1 spoT rpo* ruvAC65 priA*	0.001
RJ1003	*relA1 spoT rpo* ruvAC65 recB*	0.4
RJ1262	*relA1 spoT rpo* ruvAC65 recF143*	0.09
RJ1270	*relA1 spoT rpo* ruvAC65*train *lexA3*	0.07

[a] All clones and mutant strains are MG1655 derivatives.
[b]Strains were irradiated with 45 J/m^2 and survival measured relative to unirradiated controls.

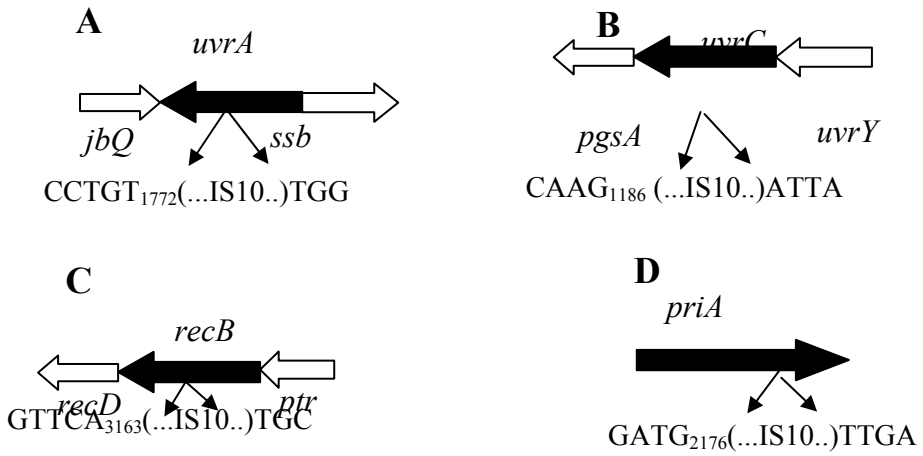


Figure 1. The location of Tn*10kan* insertion in A:RJ1004, B:RJ1025, C:RJ1003 and D:RJ1050.

The last one was hardly more sensitive to UV light than N4538 parent (Table 1), But was sensitive to MC (data not shown). By transcriptional and PCR analysis it was shown that it had insertion in *recB* gene (Fig. 1). These data showed that excision repair complexes and PriA protein are critical for survival of N4538.
Insertion of mutations in genes known to affect DNA repair such as *recF* and *lexA3* in N4538 made this strain UV sensitive (Table 1). Therefore RecF and lexA proteins are also required for survival of N4538.

5. Discussion

We found that survival of N4538 depends critically on the removal of pyrimidine dimmers by UvrABCD excision repair complexes and SOS regulon. Further study on SOS activity in *rpo* ruv* strain supported the idea that *rpo** reduce the stability of RNA polymerase complexes from DNA therefore they are no longer as obstacles to replication fork progression (Jaktaji *et al*., 2005). It was found that PriA is essential for survival implies that replication forks stalls at UV induced lesions and have to be rescued. Further study on interaction of PriA and RecG proteins, together with finding that showed RecBCD is not required, suggested that PriA helicase in conjunction with RecG can promote direct rescue of stalled forks independently of the recombinational pathway promoted by the combined activities of the RuvABC, RecBCD and RecA proteins (Jaktaji and Lloyd, 2003).

References

Jaktaji, R. P. and Lloyd, R. G. (2003) PriA supports two distinct pathways for replication restart in UV-irradiated *Escherichia coli* cells. *Molecular Mic* **47,** 1091-1100.

Kleckner, N., Bender, J. and Gottesman, S. (1991) Uses of transposons with emphasis on Tn*10*. *Methods Enzymol* **20,** 139-180.

McGlynn, P. and Lloyd, R. G. (2000) Modulation of RNA polymerase by (p)ppGpp reveals a RecG-dependent mechanism for replication fork progression. *Cell* **101,** 35-45.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular cloning. A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Trautinger, B. W., Jaktaji, R. P. Rusakova, E., Lloyd, R. G. (2005) RNA polymerase modulators and DNA repair activities resolve conflicts between DNA replication and transcription. *Molecular Cell* **19,** 247-258.

Tunable Promoters for Systems Biology: Applied to Prokaryotic Model Systems

Brian Koebmann, Christian Solem, and Peter Ruhdal Jensen*

Center for Microbial Biotechnology, Biocentrum-DTU, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark, e-mail: prj@biocentrum.dtu.dk

Keywords: Promoter, prokaryote, MCA, glycolysis, *Lactococcus lactis*.

1. Abstract

We present a strategy for construction of synthetic promoter libraries which has been successfully applied to a broad range of model systems. The approach is based on randomization of DNA sequences in the vicinity of fixed consensus boxes in standard prokaryotic promoters, which results in generation of promoters with essentially any strength. Synthetic promoter libraries can be easily placed upstream of a gene or an operon by incorporating a sequence for the randomized promoter region in a primer used for PCR amplification of the gene. Depending on the genetic engineering strategy this allows for either introduction of an extra copy of the gene with modulated expression or for replacement of the native chromosomal promoter with a set of synthetic promoters. Importantly, the synthetic promoter library approach can be used for simultaneous modulation of numerous individual genes in a single cell. The application of synthetic promoter libraries in the studies of biological systems is illustrated by reviewing experimental control analysis studies with the cheese bacterium *Lactococcus lactis*. The importance of phosphoglycerate enolase and the three enzymes encoded by the *las* operon in *L. lactis,* phosphofructokinase, pyruvate kinase and lactate dehydrogenase are quantified in terms of the control exerted by these enzymes on growth rate, glycolytic flux and product distribution.

2. Introduction

Predictive computer models are important goals of systems biology, but since the output from *in silico* models depend entirely on the information provided to them these data must be relevant *in vivo*. The models should therefore be verified by comparing with quantitative experiments and adjusted if necessary. Metabolic Control Analysis (MCA), which was originally developed

* Corresponding author

in the early 1970'es (Kacser and Burns, 1973; Heinrich and Rapoport, 1974), allows for answering questions about system properties such as the control of a flux or a metabolite concentration in terms of flux- and concentration control coefficients which describe the quantitative effect of perturbation of enzyme activities on fluxes and metabolite concentrations. Accurate estimation of these system properties requires *modulation* of enzyme activities in discrete steps over a suitable range of enzyme activities.

In microorganisms, modulation of enzyme activities can conveniently be obtained by tuning the level of transcription of the corresponding gene(s) of the enzyme(s). Traditionally, modulations of transcription levels were performed by the use of inducible promoter systems, and metabolic control analysis have been performed successfully, such as in the study of the H^+-ATPase in *Escherichia coli* (Jensen et al., 1993a, 1993b) and DNA supercoiling in *E. coli* (Jensen et al. 1999; Snoep et al., 2002). However, inducible promoters have severe shortcomings with respect to the use in systems biology and metabolic engineering. Firstly, it may be difficult to obtain subtle tuning of gene expression with inducible promoters due to hypersensitivity to the inducer. Secondly, inducible promoters only allow for the tuning of gene expression within a given range. Thirdly, it is often very difficult to obtain a steady state with fixed gene expression throughout an experiment. Fourthly, inducible promoter systems only allow for tuning of a single gene or a set of genes in parallel, whereas the application of knowledge obtained through systems biology to direct or increase a flux are likely to require simultaneous optimization of expression of several genes due to the homeostatic control mechanisms of the biological system. Finally, from an industrial perspective the use of inducible promoters for industrial fermentation processes may be both costly and difficult to handle.

A major breakthrough for systems biology and metabolic engineering has therefore been the development of synthetic promoter libraries that facilitate a delicate tuning of gene expression in the desired range of expression levels. Such promoters allow for concomitant expression of several genes, which enables metabolic optimization of industrial fermentation processes. This chapter focuses on the construction and use of synthetic promoter libraries in prokaryotes. The application of a synthetic promoter library approach for systems biology is exemplified by recent studies of the phosphoglycerate enolase (PGE) and the *las* operon in *Lactococcus lactis*.

3. Theoretical

Transcription process in prokaryotes. The major player in transcription processes involved in the direction, initiation and elongation of transcription is the RNA polymerase. In prokaryotes RNA polymerase is composed of a so-called core enzyme containing the basic transcription machinery and a sigma factor with the function to direct the core enzyme to sites from which transcription should be initiated. Association of the sigma factor with the core enzyme is referred to as the RNA polymerase holoenzyme. The transcription

process initially depends on the binding of sigma factors to specific DNA sequences in the promoter regions. Traditionally, the consensus boxes -10 (Pribnow box) and -35 have been considered to be highly important for the binding of sigma factors, representing probabilities for which the sigma factors bind to promoters. After binding of the RNA polymerase the sigma factor dissociates and leaves the core enzyme of the RNA polymerase to carry out elongation. The binding of the RNA polymerase holoenzyme to promoters involves local melting of the DNA to form a so-called open promoter complex from which transcription is initiated.

Modulation of gene expression by randomizing the bases in the vicinity of the consensus sequences. Though the consensus boxes -10 and -35 play a dominant role for transcription processes in prokaryotes, the sequences surrounding the consensus sequences may also contribute to the strength of bacterial promoters, for instance in the binding of RNA polymerase or the degree of melting of DNA to form the open complex from which transcription to mRNA can be initiated. The length of the spacer region between the consensus sequences in prokaryotes is usually 17+/-1 bp. An approach to study the relevance of the spacer sequence of prokaryotic promoters and the sequences in the vicinity upstream and downstream to the consensus boxes was published in 1998 (Jensen and Hammer,1998). In this study, the consensus boxes were fixed, while the DNA bases of the surrounding area were randomized (Fig.1A). By randomizing these bases simultaneously it was possible to change the DNA structure, which resulted in promoter libraries with promoters of virtually any activity. Interestingly, though the consensus sequences appear to be almost identical among different prokaryotes, the promoters were found to vary in relative strengths in different microorganisms (Jensen and Hammer, 1998).

Tuning of gene expression by the synthetic promoter libraries. The finding that the bases in the surrounding area of the consensus sequences in prokaryotic promoters affected promoter strengths gave rise to the so-called Jensen-Hammer approach for modulation of gene expression, where the expression of genes were tuned by inserting them after a set of synthetic promoters with varying activities. Modulation of enzyme activity by tuning transcription of the corresponding gene(s) by the Jensen-Hammer approach has been performed on several enzymes. An example is a study of lactate dehydrogenase (LDH) in *L. lactis* from which it was possible to quantify the control of LDH exerted on the product formation (Andersen et al., 2001b). The Jensen-Hammer approach was also used for tuning the level of enzyme complexes of the F_1-ATPase in *E. coli* and *L. lactis* with the purpose to gradually introduce an uncoupled ATPase activity in the cell (Koebmann et al., 2002a, 2002b). From these studies it was possible to quantify the control of the ATP demanding processes on the glycolytic flux.

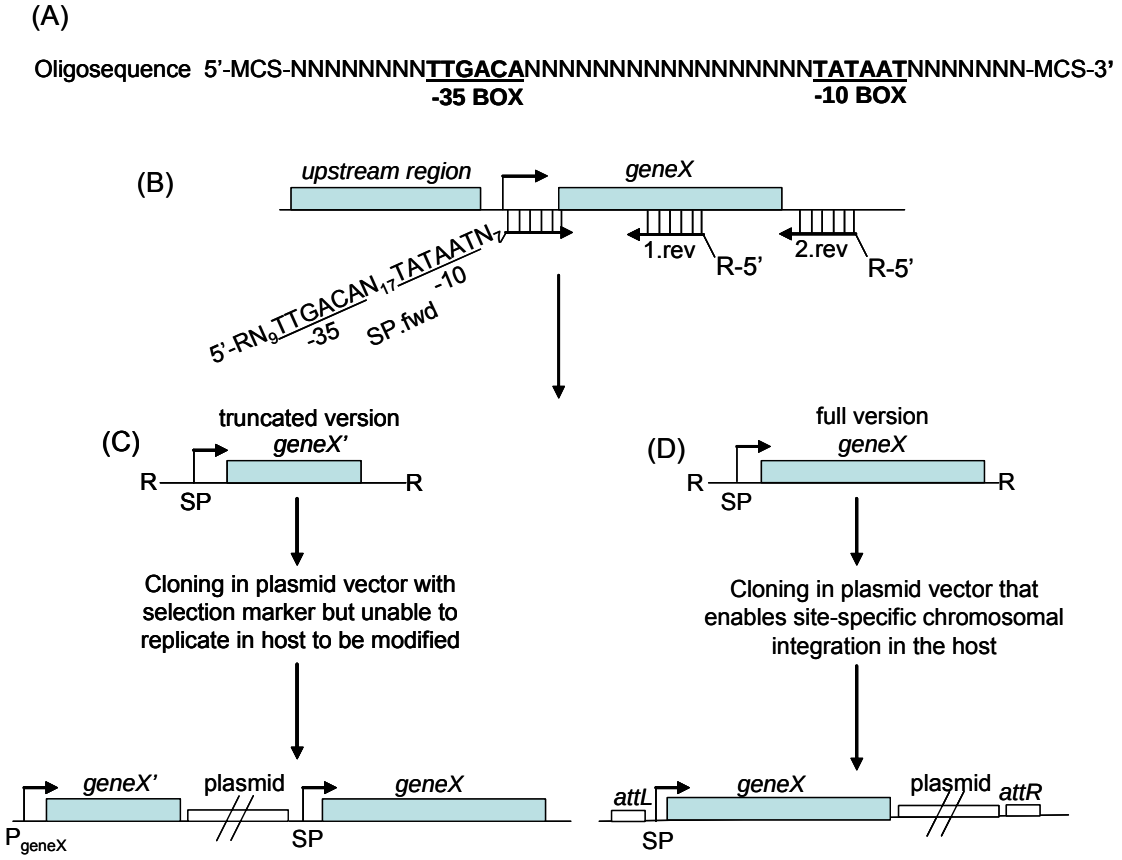


Figure 1. Modulation of gene expression.
(A) Oligosequence for randomized promoters according to the Jensen-Hammer approach. The fixed consensus boxes vary in sequence and length depending on the microorganism (B) Incorporation of randomized promoters through PCR amplification by the Solem-Jensen approach for modulating enzyme activities. (C) Replacement of a native promoter by homogenous recombination with a truncated version of the gene. (D) Introduction of an extra copy of the gene by site-specific recombination. MCS=multiple cloning site, N=random nucleotide, R=restriction site, SP=synthetic promoter.

The Jensen-Hammer method initially required extensive cloning work to modulate the expression of a gene because multiple cloning experiments had to be performed in parallel. In addition, the strengths of the promoters were found to be context dependent which may further increase the number of cloning experiments required. The general concept from the Jensen-Hammer approach to modulate gene expression was therefore used in a new approach to facilitate modulation of gene expression (Solem and Jensen, 2002). The new approach is based on direct incorporation of synthetic promoters with randomized spacers in a PCR amplification of a desired gene, by designing a primer that contains a restriction site, a promoter with randomized spacers, and a sequence with homology to the 5´-end of the mRNA of the target gene (Fig. 1B). The reverse primer also contains a restriction site and a sequence with homology within the

gene or downstream to the gene. The primer set allows for amplification of either a truncated or a full version of the gene, in both cases with incorporated randomized promoters upstream to the gene. The truncated version can be used to replace the native chromosomal promoter with a range of synthetic promoters by homologous recombination, which facilitates a tuning of gene expression below the wildtype level (Fig. 1C; Solem and Jensen, 2002). The full version can then be cloned in a suitable plasmid vector used for introducing an extra copy of the gene in the cell, for instance by site-specific integration on the chromosome (Fig. 1D; Brøndsted and Hammer, 1999; Solem and Jensen, 2002). By using the synthetic promoter library approach to tune transcription of the target gene(s) only clones with appropriate transcription levels will survive providing a library of clones within the window of interest. Another improvement in the new approach, especially when used for operons, is the preservation of the leader sequence of the mRNA, which can be important for the stability of the mRNA. In recent years, the synthetic promoter library approach has been routinely used for tuning of gene expression and is now an important tool for experimental systems biology, metabolic control analysis and metabolic optimization (Solem et al., 2003; Koebmann et al., 2005; Koebmann et al., 2006).

Curve fitting and calculation of control coefficients. In metabolic control analysis the extent to which an enzyme controls a flux or concentration of a metabolite is described by control coefficients. The flux control coefficient C_E^J is defined as $C_E^J = (E/J)*(dJ/dE)$, where J represents the flux and E represents the enzyme activity. A way to estimate flux control coefficients is to fit experimental data points with respect to flux and enzyme activity to appropriate equations by the use of software such as CURVEEXPERT 1.3' (Hyams Development, Hixson, TN, USA). The slope (dJ/dE) can subsequently be estimated by differentiation of the optimal curve fit for the entire range of enzyme activities, which then allows for calculation of flux control coefficients.

4. Experimental

Strains and plasmids. The lactic acid bacteria *L. lactis* subsp. *cremoris* MG1363 (Gasson, 1983) and *L. lactis* subsp. *lactis* IL1403, a plasmid free derivative of strain IL594 (Chopin et al., 1984), were used as model strains for the quantitative analysis of glycolysis. *E. coli* ABLE-C (Stratagene) was used as cloning vector for amplification of plasmid libraries. Plasmid pRC1 (Le Bourgois et al., 1992), which is unable to replicate in *L. lactis*, was used for replacement of native chromosomal promoters with synthetic promoters. Plasmid pCS574 (Solem et al., 2003), a derivative of plasmid pLB85 (Brøndsted and Hammer, 1999), was used for site-specific integration in a phage attachment site as described in (Brøndsted and Hammer, 1999), providing the cell with an extra copy of the gene.

Growth media and growth conditions. The medium for growth experiments with *L. lactis* was based on defined SA medium (Jensen and Hammer, 1993c) modified by inclusion of 2 μg/ml of lipoic acid (SAL) and 20 μg/ml of each of the nucleosides adenosine, guanosine, thymidine, cytidine, uridine, and inosine and exclusion of acetate (SALN) supplemented with glucose (GSALN) as described in (Koebmann et al., 2005, 2006). Glucose consumption and product formation were determined by HPLC taken during the experiments in order to estimate the fluxes.

DNA techniques. Extractions of chromosomal *L. lactis* DNA, PCR amplification, restriction, ligation, transformation and plasmid purification from *E. coli* were performed as described in (Koebmann et al., 2005) and according to the prescription from the manufacturer of the applied enzymes.

Construction of a strain library with modulated phosphoglycerate enolase activity. The *pge* gene coding for phosphoglycerate enolsase (PGE) was amplified from *L. lactis* IL1403 by PCR as described in (Koebmann et al., 2006). The resulting PCR fragment consisting of truncated version of the *pge* gene with randomized promoter regions positioned upstream to the leader sequence was inserted in plasmid vector pRC1 (Le Bourgois et al., 1992) as described in (Solem and Jensen, 2002; Koebmann et al., 2006) and subsequently amplified in *E. coli*. The resulting plasmid library was introduced to *L. lactis* IL1403 and a selection of strains was analyzed with respect to PGE activities as described in (Koebmann et al., 2006).

Construction of strain libraries with tuned expression of the *las* operon. Construction of *L. lactis* MG1363 libraries with modulated activities of the individual *las* enzymes phosphofructokinase (PFK), pyruvate kinase (PYK) and lactate dehydrogenase (LDH) and concomitant modulation of the *las* enzymes were performed as described in (Andersen et al., 2001a, 2001b; Solem et al., 2003; Koebmann et al., 2005).

Measurement of enzyme activities. Measurements of enzyme activities of PGE, PFK, PYK, and LDH were based on assays modified from (Even et al., 2001) and performed as described in (Koebmann et al., 2005; Koebmann et al., 2006). All enzymes used in the enzymatic assays were purchased from Roche A/S (Hvidovre, Denmark).

Curve fitting and control coefficients. Estimation of flux control coefficients were performed by fitting the experimental data points by the least square method using the software program CurveExpert 1.3' (Hyams Development, Hixson, Tn, USA) as described in (Koebmann et al., 2005; Koebmann et al., 2006). The control coefficients were calculated for the entire range of enzyme activities.

Pulse labelling and 2D-gel electrophoresis of proteins. The strains were inoculated at 30°C overnight in SALN medium (Jensen and Hammer, 1993) supplemented with 10 g/L of glucose (GSALN) and 20 μg/ml of methionine. Next day, 0.5 ml culture was diluted in 50ml GSALN medium supplemented with 5 μg/ml of methionine. At an optical density of OD_{450}=0.4 150 μl culture was transferred to 1.5 ml eppendorph tube containing 3 μl ^{35}S-methionin (15 μCi/μl). After 10 min 10 μl methionine (10 mg/ml) was added, and after 12 min 10 μl chloramphenicol (20 mg/ml) was added in order to stop translation. Cell pellets were obtained at 4°C by centrifugation at 10000 *g* for 5 min, subsequently washed in 200 μl ice-cold 0.9% NaCl+30% ethanol, centrifuged and freeze-dried overnight in a speed-vac. Preparation and electrophoresis of protein samples were performed as described in (Guillot et al., 2003). Gel images were developed on X-ray films.

5. Results and Discussion

In the following we present recent studies in which the synthetic promoter libraries were applied for modulation of single genes and for a bacterial operon.

Tuning the expression of the gene encoding phosphoglycerate enolase in *L. lactis*. The most recent example of the use of the synthetic promoter library approach for modulating gene expression is in the study of the glycolytic enzyme phosphoglycerate enolase (PGE) in *L. lactis* IL1403. PGE is positioned late in glycolysis where it converts 2-phosphoglycerate to phosphoenolpyruvate (PEP) (Fig. 2). The product of the enzyme, PEP, has dual functions: PEP is either metabolized to pyruvate by pyruvate kinase (PYK) with the concomitant generation of 1 ATP, or used as a phosphate donor for PTS sugar transport. In order to perform control analysis on PGE, the expression of the corresponding gene (*pge*) was tuned by replacing the native promoter with a library of synthetic promoters: a truncated version of *pge* with incorporated synthetic promoters was inserted in the plasmid vector pRC1 as described previously (Koebmann et al., 2006) and subsequently cloned in *E. coli*. Since the plasmid vector pRC1 is unable to replicate in *L. lactis*, transformation of *L. lactis* with the resulting plasmid library and selection for the plasmid coded resistance marker forced the plasmids to integrate on the chromosome by homologous recombination in the *pge* gene. This resulted in strains in which the native promoter was replaced by synthetic promoters.

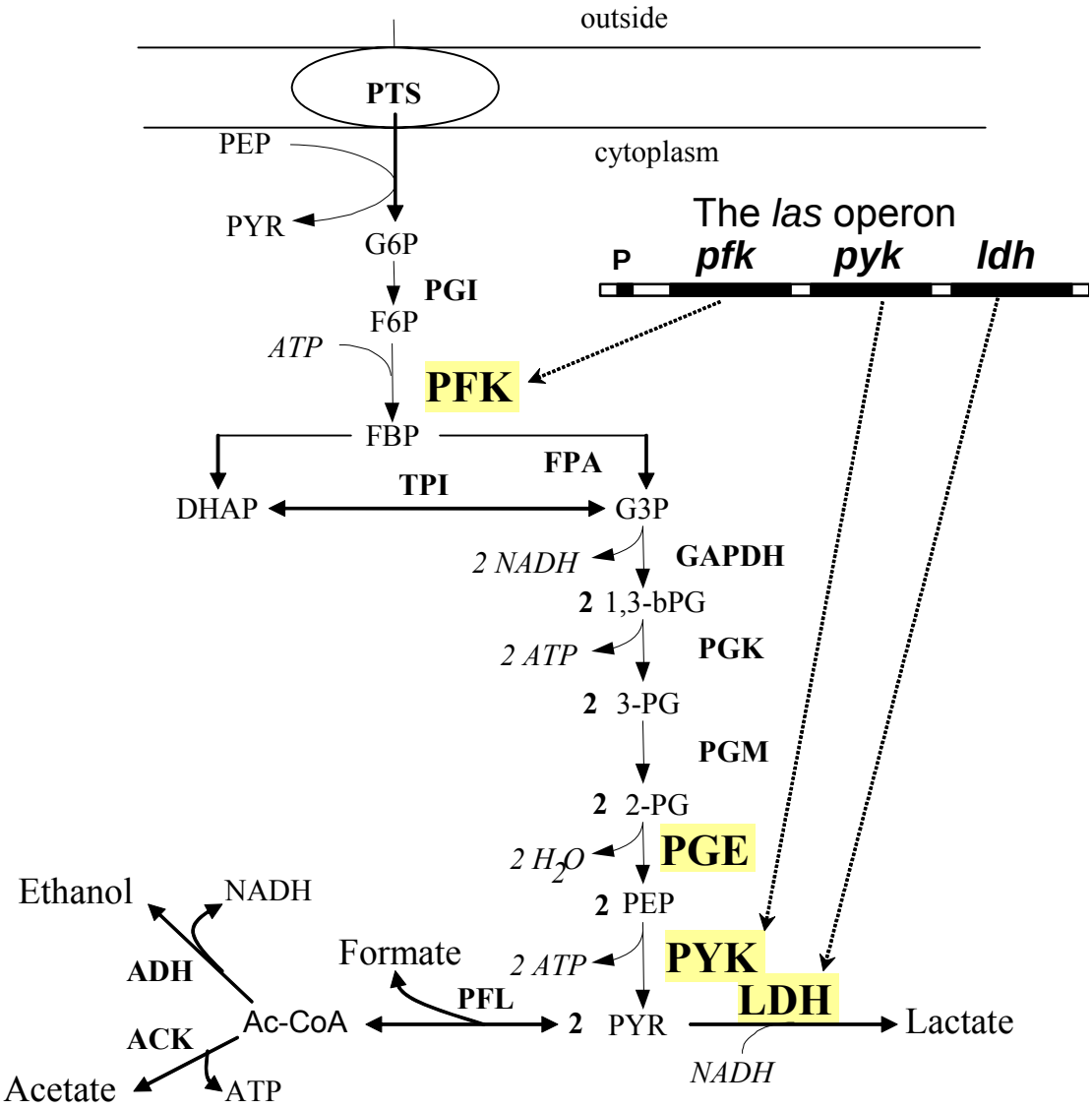


Figure 2. Glycolysis in *Lactococcus lactis*.
Phosphoglycerate enolase (PGE) is positioned late in glycolysis. The *las* operon in *L. lactis* consist of the three genes *pfk*, *pyk* and *ldh*, coding for phosphofructokinase (PFK), pyruvate kinase (PYK) and lactate dehydrogenase (LDH), respectively, positioned early and late in glycolysis (Modified from Koebmann et al., 2005).

From the resulting library of strains it was possible to isolate strains with PGE activities from 36% to 232% of wildtype level and these strains were subsequently studied in growth experiments with measurements of growth rate and metabolic fluxes (Koebmann et al., 2006). The resulting data were used for determining the control by PGE on growth rate and metabolic fluxes (Fig. 3).

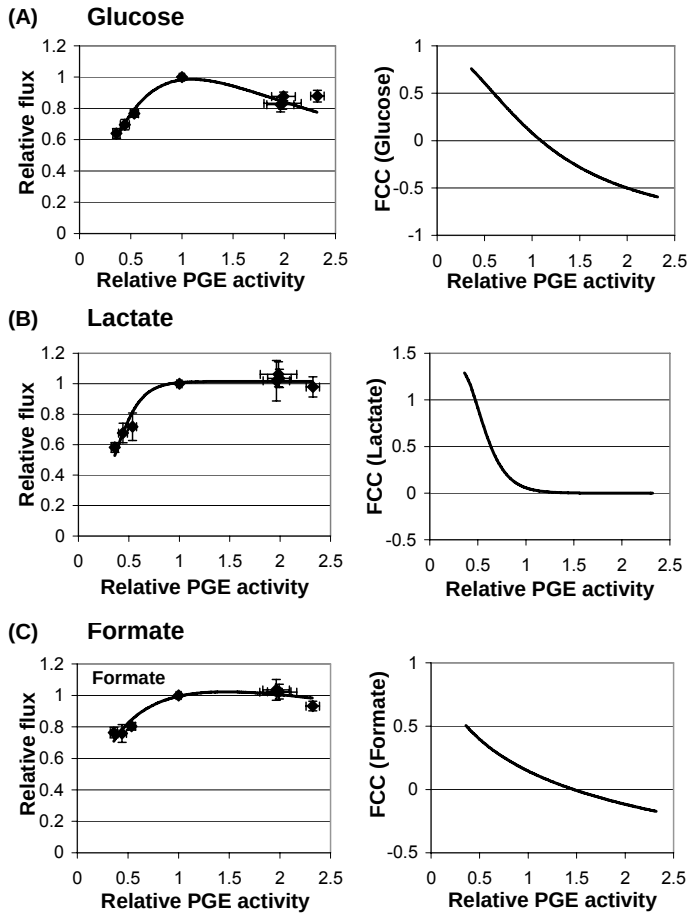


Figure 3. Metabolic control analysis of PGE. Effect of PGE activity on **(A)** Glycolytic flux, **(B)** Lactate production, and **(C)** Formate production. Flux control coefficients (FCC) for PGE were determined from fitted equations (Data obtained from Koebmann et al., 2006).

Curves were fitted to the data points as described in the experimental section and flux control coefficients calculated from these curve fits. At the wildtype enzyme level no control was observed on either growth rate or metabolic fluxes. However, at 36% PGE activity significant high control was observed on growth rate ($C_{PGE}^{\mu} \approx 0.7$), glycolytic flux ($C_{PGE}^{J_g} \approx 0.8$), lactate production ($C_{PGE}^{lactate} \approx 1.3$), while the flux of mixed acid products were lower with formate production ($C_{PGE}^{formate} \approx 0.5$) and acetate production ($C_{PGE}^{acetate} \approx 0.25$). The magnitude of these flux control coefficients showed that *L. lactis* becomes slightly more mixed acid at reduced PGE activities. According to the literature, fructose-di-phosphate (FDP) is an activator of LDH (Crow and Pritchard, 1977), while dihydroxyacetonephosphate (DHAP) and glyceraldehyde-3-phosphate (G3P) are inhibitors of PFL (Garrigues et al., 1997; Takahashi et al., 1982). If we therefore assume that a reduction in PGE activity results in an increased pool of upper metabolites of glycolysis then our data indicates that the regulatory mechanisms involved are more complex.

Experimental control analysis of the enzymes encoded in the *las* operon in *L. lactis*. The synthetic promoter library approach has also been used to study the *las* operon in *L. lactis* MG1363 in which the genes encoding the three enzymes phosphofructokinase (PFK), pyruvate kinase (PYK), and lactate dehydrogenase (LDH) are organized (Fig. 2) (Koebmann et al., 2005). The organization of the

las enzymes indicates a possible role in control and regulation of glycolysis. The three individual enzymes were first studied individually and then together by tuning the entire *las* operon. In a previous study it was found that LDH had no control on the glycolytic flux, but appeared to have a significant negative control on formate production ($C_{LDH}^{formate} < -1$) (Andersen et al., 2001b). The study of PFK and PYK was performed by modulation of individual enzymes by site-specific chromosomal integration of a full version of the corresponding genes. Since PYK is positioned in between PFK and LDH in the *las* operon a modulation of PYK activity below wildtype activity was obtained by site-specific chromosomal integration of the *pyk* gene followed by deletion of the *pyk* gene in the *las* operon by a double cross over. Metabolic control analysis of the individual PFK and PYK showed that neither of the enzymes had control on the glycolytic flux (Fig. 4). For PFK no control was found on the formate production, but for PYK a high positive control on formate production was found ($C_{PYK}^{formate} = 0.9 - 1.1$).

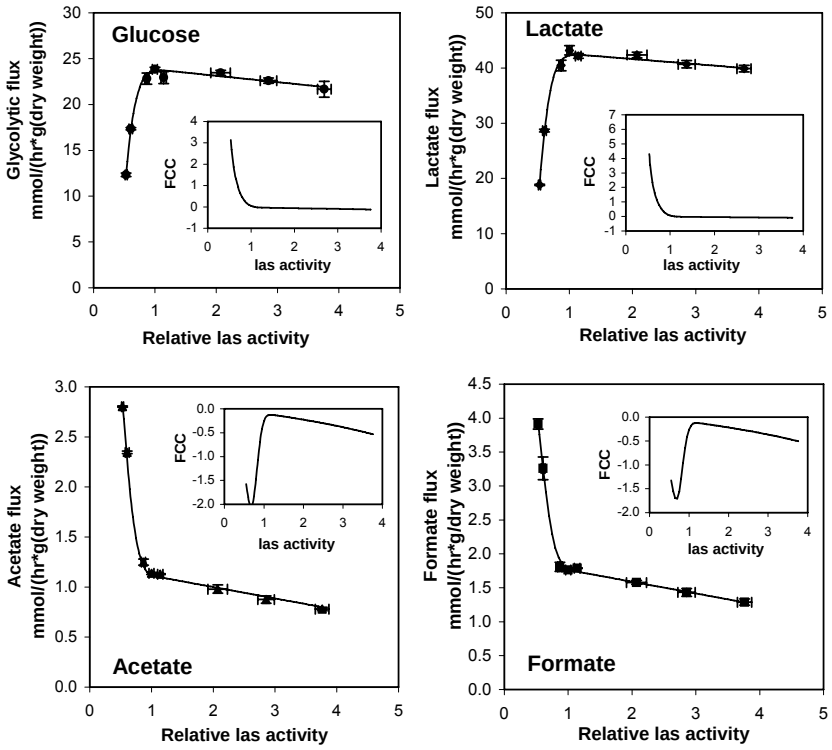


Figure 4. Flux control coefficients for the las enzymes on metabolic fluxes. A selection of strains was analyzed with respect to glycolytic flux and metabolic fluxes. Flux control coefficients (FCC) with respect to glycolysis, lactate, acetate and formate production were determined from fitted equations (From Koebmann et al., 2005).

Tuning of the entire *las* operon was achieved by replacing the native promoter with synthetic promoters, keeping the leader sequence of the resulting mRNA intact, which turned out to be important for keeping the proportional expression of all three las enzymes (Solem and Jensen, 2002). A good correlation among relative enzyme activities was found and ranged between 0.5 – 3.5 times the wildtype level and a selection of strains was chosen for studying the control exerted by the three las enzymes simultaneously. It was found that the control of the las enzymes together at the wildtype level was close to 0. Interestingly, only a

slight reduction in las activity resulted in a significant decrease in growth rate and glycolytic flux, and at 53% las activity the flux control were found to be $C_{las}^{\mu} \approx 3$ and $C_{las}^{J_g} \approx 3$, respectively. With respect to the fermentation pattern, flux control coefficient of $C_{las}^{formate} \approx -0.3$ and $C_{las}^{acetate} \approx -0.3$ were observed for the formate and acetate flux. At reduced *las* activity was observed a strong negative flux control coefficient on the formate and acetate production.

The obtained data from the control studies of the individual *las* enzymes and for the simultaneous tuning of the *las* operon were used for comparison to investigate if $C_{las}^{x} = C_{PFK}^{x} + C_{PYK}^{x} + C_{LDH}^{x}$. With respect to growth rate, glycolytic flux and lactate production all control coefficients were close to 0. More interesting was the comparison of the formate flux, where LDH exerted negative control, PFK no control and PYK positive control. Interestingly, the sum of flux control by the individual las enzymes added up closely to what was found for simultaneous tuning of the entire *las* operon.

Simultaneous tuning of transcription of several individual genes. An important feature of the synthetic promoter library approach is the ability to replace a native promoter with a synthetic promoter with desired strength without leaving residual fragments of the used plasmid vector. This can be accomplished by construction of a plasmid which contains the upstream chromosomal DNA region of the native promoter followed by the gene placed after a synthetic promoter. If the plasmid is unable to replicate in the organism of interest a homologous recombination with double cross-over can be obtained as described by Solem and Jensen (Solem and Jensen, 2002). Since the resulting strain will only be genetically modified in the approximately 60 bp of the promoter region of the gene or operon it is possible to repeat the process for other genes, thereby enabling a simultaneously modulation of several individual genes in a single strain. An attempt to double the activity of an entire pathway in a single cell by the use of the synthetic promoter library approach is currently under way for glycolysis in *L. lactis*. It is the goal to approximately double the activities of all glycolytic enzymes, thereby doubling the capacity of ATP supplying processes. In combination with an additional ATP consuming process and transport processes of substrates and products the cell will have enhanced the entire system to produce and consume ATP which appears necessary for increasing the glycolytic flux in *L. lactis*.

Since perturbation of enzyme activities in microorganisms often result in activation of mechanisms that seeks to counteract the changes (homeostatic control), i.e. by up- or down regulation of the amount or activity of an enzyme, it is desirable to get an overview of the response of the cell to genetic changes. An obvious way to achieve this is by applying global techniques such as transcriptomics, proteomics and metabolomics. A simple way to study the proteome is to make a 2D-protein gel in which the proteins in the cell are separated in two dimensions. The consequences of simultaneous modulation of five glycolytic genes with respect to protein content was investigated based on 2D protein gels from which it appeared that the amounts of the modulated glycolytic enzymes were increased, while only limited changes on other proteins

were observed (Fig. 5). Identification of the glycolytic enzymes was based on a study by Guillot and co-workers (Guillot et al., 2003).

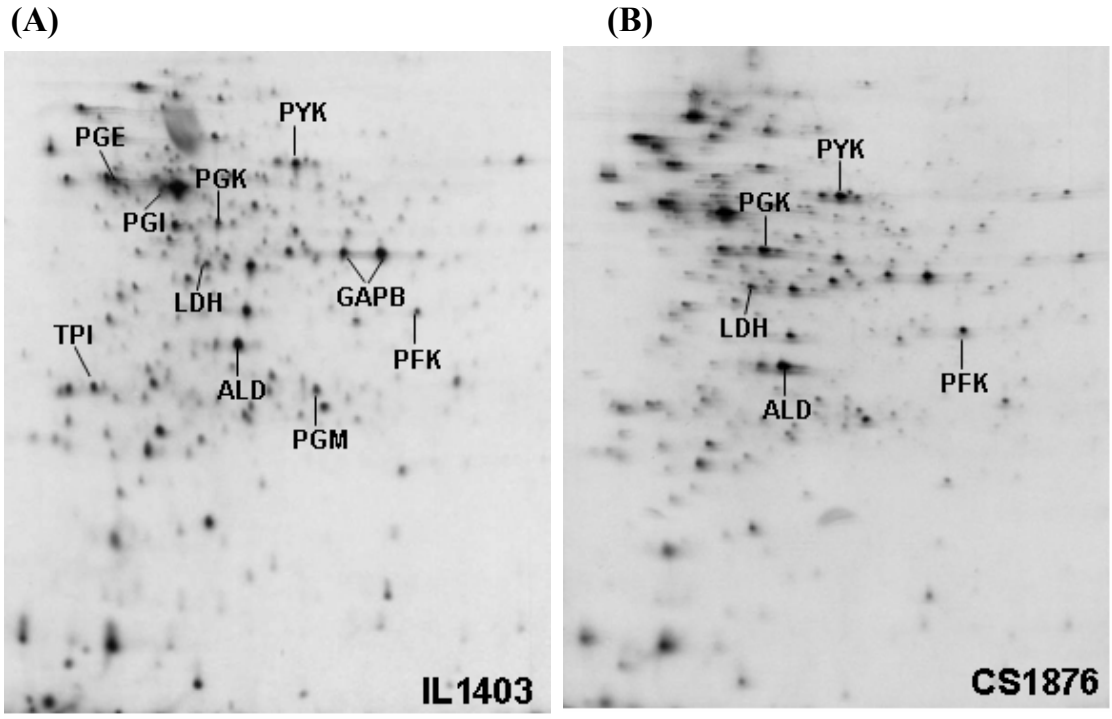


Figure 5. Comparison of 2D protein gels of *L. lactis* IL1403 and CS1876. Wildtype *L. lactis* IL1403, **(B)** *L. lactis* CS1876 in which activities of the five glycolytic enzymes (PFK, PYK, LDH, ALD, PGK) are increased approximately 2-fold relative to IL1403. The protein contents were isolated as described in the experimental section. Spots for the relevant glycolytic enzymes are indicated based on identification described by (Guillot et al., 2003).

Comparison of the synthetic promoter library approach to other approaches for construction of tunable promoters. The primary focus on the tuning of gene expression presented until now has been on modulation of the bases surrounding the consensus boxes -10 and -35. According to literature several approaches to construct tunable promoters have been applied. In a recent study, an *E. coli* P_L-λ promoter sequence was mutated by mutagenic PCR (Alper et al., 2005). This approach also resulted in a pool of promoters with slight sequence variations. A major difference, however, between this approach and the synthetic promoter library approach is in the creation of sequence variability. In the synthetic promoter library approach the consensus boxes are fixed, while in the other approaches many of the resulting promoters will have mutations in the consensus boxes, which may reduce the promoter strength and the frequency of usable promoters.

6. Conclusion

We expect that the ability to tune gene expression will be an important task in the field of systems biology in the future. Due to several disadvantages with the use of inducible promoters both from fundamental and industrial viewpoints there is a need for alternative ways to modulate gene expression. One of the promising approaches is the so-called synthetic promoter library approach, where consensus sequences of promoters are fixed while the bases in the vicinity of the consensus boxes are randomized, a strategy which enables the generation of essentially all values of promoter strengths. By incorporation of the randomized synthetic promoter library sequence in the design of a primer used for PCR amplification of a gene or operon it is possible in a single PCR amplification to generate a library of fragments with promoters of essentially any strengths placed upstream to the gene or operon. Depending on the cloning strategy such fragments can be used for incorporation of an additional copy of the gene in the cell or for substitution of the native promoter with a library of synthetic promoters with varying strengths. Moreover, the approach allows for simultaneous tuning of several genes in a single cell, which is often important for metabolic optimization of a pathway due to distribution of flux control over many enzymes, and to keep metabolite pools at acceptable levels (Kacser and Acerenza, 1993).

Recent examples on the application of the synthetic promoter library approach are in control analysis studies of glycolysis in *L. lactis*. In one study the gene coding for PGE was modulated around wildtype level and the enzyme's control on growth, glycolytic flux and product formation were determined. In another study, the control by the enzymes of the *las* operon in *L. lactis*, PFK, PYK, and LDH, were determined and showed that, whereas none of the enzymes controlled the main glycolytic flux, with respect to formate production LDH exerted negative control, PFK no control and PYK positive control. Interestingly, the sum of flux control by the individual *las* enzymes added up closely to what was found for simultaneous tuning of the entire *las* operon. These examples illustrate that randomizing promoters for tuning gene expression is a valuable tool which is likely to be applied frequently in future quantitative studies of biological systems. It also has the potential to optimize the expression of entire metabolic pathways.

Acknowledgements

The work presented here was financially supported by the Danish National Research Council (SNF), the Danish Center for Advanced Food Studies (LMC) and the Danish Dairy Research Foundation (MFF).

Symbols

C_E^J	Flux control coefficient of enzyme E on flux J
	Glycolytic enzymes
ALD	Fructose-1,6-bisphosphate aldolase
GAPB	Glyceraldehyde-3-phosphate dehydrogenase (*gapB*)
LDH	Lactate dehydrogenase
PFK	Phosphofructokinase
PGE	Phoshoglycerate enolase
PGI	Phosphoglucose isomerase
PGK	Phosphoglycerate kinase
PGM	Phosphoglycerate mutase
PYK	Pyruvate kinase
TPI	Triosephosphate isomerase
	Glycolytic metabolites
FDP	Fructose-di-phosphate
G3P	Glyceraldehyde-3-phosphate
DHAP	Dihydroxyacetonephosphate
PEP	Phosphoenolpyruvate

References

1 Alper, H., Fischer, C., Nevoigt, E. and Stephanopoulos, G. (2005) Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12678-12683.

2 Andersen, H.W., Solem, C., Hammer, K. and Jensen, P.R. (2001a) Twofold reduction of phosphofructokinase activity in *Lactococcus lactis* results in strong decreases in growth rate and in glycolytic flux. *J. Bacteriol.* **183**, 3458-3467.

3 Andersen, H.W., Pedersen, M.B., Hammer, K. and Jensen, P.R. (2001b) Lactate dehydrogenase has no control on lactate production but has a strong negative control on formate production in *Lactococcus lactis*. *Eur. J. Biochem*. **268**, 6379-6389.

4 Brøndsted, L. and Hammer, K. (1999) Use of the integration elements encoded by the temperate lactococcal bacteriophage TP901-1 to obtain chromosomal single-copy transcriptional fusions in *Lactococcus lactis*. *Appl. Environ. Microbiol.* **65**, 752-758.

5 Chopin, A., Chopin, M.C., Moillo-Batt, A. and Langella, P (1984) Two plasmid-determined restriction and modification systems in *Streptococcus lactis*. *Plasmid* **11,** 260-263.

6 Crow, V.L. and Pritchard, G.G. (1977) Fructose 1,6-diphosphate-activated L-lactate dehydrogenase from *Streptococcus lactis*: kinetic properties and factors affecting activation. *J. Bacteriol*. **131**, 82-91.

7 Even, S., Lindley, N.D. and Cocaign-Bousquet, M. (2001) Molecular physiology of sugar metabolism in *Lactococcus lactis* IL1403. *J. Bacteriol.* **183,** 3817-3824.
8 Garrigues, C., Loubiere, P., Lindley, N.D. and Cocaign-Bousquet, M.(1997) Control of the shift from homolactic acid to mixed-acid fermentation *in Lactococcus lactis*: predominant role of the NADH/NAD^+ ratio. *J. Bacteriol*. **179**, 5282-5287.
9 Gasson, M.J. (1983) Plasmid complements of *Streptococcus lactis* NCDO 712 and other lactic streptococci after protoplast-induced curing. *J. Bacteriol*. **154**, 1-9.
10 Guillot, A., Gitton, C., Anglade, P. and Mistou, M.Y. (2003) Proteomic analysis of *Lactococcus lactis*, a lactic acid bacterium. *Proteomics* **3**, 337-354.
11 Heinrich, R. and Rapoport, T.A. (1974) A linear steady-state treatment of enzymatic chains: General properties, control and effector-strength. *Eur. J. Biochem*. **42**, 89-95.
12 Jensen, P.R. and Hammer, K. (1998) The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Appl. Environ. Microbiol.* **64,** 82-87.
13 Jensen, P.R., Westerhoff, H.V. and Michelsen, O. (1993a) The use of *lac*-type promoters in control analysis. *Eur. J. Biochem.* **211**, 181-191.
14 Jensen, P.R., Westerhoff, H.V. and Michelsen, O. (1993b) Excess capacity of H^+-ATPase and inverse respiratory control in *Escherichia coli. EMBO J.* **12**, 1277-1282.
15 Jensen, P.R. and Hammer, K. (1993c) Minimal requirements for exponential growth of *Lactococcus lactis*. *Appl. Environ. Microbiol.* **59,** 4363-4366.
16 Jensen, P.R., Van Der Weijden, C.C., Jensen, L.B., Westerhoff, H.V. and Snoep, J.L. (1999) Extensive regulation compromises the extent to which DNA gyrase controls DNA supercoiling and growth rate of *Escherichia coli. Eur. J. Biochem.* **266**, 865-877.
17 Kacser, H. and Burns, J.A. (1973) The control of flux. *In*: D. D. Davies (ed.), Rate control of biological processes. Cambridge University Press, Cambridge. *Symp. Exp. Biol.* **27**, 65-104.
18 Kaacser, H. and Acerenza, L. (1993) A universal method for achieving increases in metabolite production. *Eur. J. Biochem.* **216**, 361-367.
19 Koebmann, B. J., Westerhoff, H.V., Snoep, J.L., Nilsson, D. and P. R. Jensen. (2002a). The glycolytic flux in *Escherichia coli* is controlled by the demand for ATP. *J. Bacteriol.* **184,** 3909-3916.
20 Koebmann, B.J., Solem, C, Pedersen, M.B., Nilsson, D. and Jensen, P.R. (2002b) Expression of genes encoding F_1-ATPase results in uncoupling of glycolysis from biomass production in *Lactococcus lactis*. *Appl. Environ. Microbiol.* **68,** 4274-4282.
21 Koebmann, B., Solem, C. and Jensen, P.R. (2005) Control analysis as a tool to understand the formation of the *las* operon in *Lactococcus lactis. FEBS Journal* **272,** 2292-2303.

22 Koebmann, B., Solem, C. and Jensen, P.R. (2006) Control analysis of the importance of phosphoglycerate enolase for metabolic fluxes in *Lactococcus lactis* subsp. *lactis* IL1403. *IEE Proceedings Systems Biology*, in press.
23 Le Bourgeois, P., Lautier, M., Mata, M. and Ritzenthaler, P. (1992) New tools for the physical and genetic mapping of *Lactococcus* strains. *Gene* **111**, 109-114.
24 Snoep, J.L., van der Weijden, C.C., Andersen, H.W., Westerhoff, H.V. and Jensen, P.R. (2002) DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. *Eur. J. Biochem.* **269,** 1662-1669.
25 Solem, C. and Jensen, P.R. (2002) Modulation of gene expression made easy. *Appl. Environ. Microbiol*. **68,** 2397-2403.
26 Solem, C., Koebmann, B. J. and Jensen, P.R. (2003) Glyceraldehyde-3-phosphate dehydrogenase has no control over glycolytic flux in *Lactococcus lactis* MG1363. *J. Bacteiol.* **185,** 1564-1571.
27 Takahashi, S., Abbe, K. and Yamada, T. (1982) Purification of pyruvate formate-lyase from *Streptococcus mutans* and its regulatory properties. *J. Bacteriol*. **149**, 1034-1040.

AUTHORS AND KEYWORDS INDEXES

AUTHORS INDEX

KEYWORDS INDEX