Automatic Generation of Cell-Wide Pathway Model from Complete Genome

Kazuharu Arakawa[1,2] gaou@g-language.org
Yohei Yamada[1,3] skipper@g-language.org
Hiromi Komai[1,3] t00547hk@sfc.keio.ac.jp
Kosaku Shinoda[1,4] bonito@g-language.org
Yoichi Nakayama[1,3] ynakayam@sfc.keio.ac.jp
Masaru Tomita[1,3] mt@sfc.keio.ac.jp

[1] Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan
[2] Bioinformatics Program, Graduate School of Media and Governance
[3] Department of Environmental Information
[4] Department of Policy Management

Keywords: genome annotation, pathway reconstruction, simulation, systems biology

1 Introduction

Knowledge in molecular biology is rapidly accumulating in the fields of genome, transcriptome, proteome, and metabolome, demanding for a systems biology approach in order to view the dynamic behavior of a cell as a complex system. However, simulation is a challenging task especially where large-scale modeling is required due to the necessity for vast amount of accurate parameters. E-Cell project [2] estimated the necessary cost for modeling the whole cell of *Escherichia coli* to be at least 1800 man month, from the experience in modeling an *in silico* mitochondria. Therefore a large scale modeling of cell *in silico* demands for a novel high-throughput approach. If successfully integrated, availability of large amount of genome sequence, transcripts and expression data, enzyme reaction data, metabolic pathway maps, and the data of metabolites in cells will create a strong base for a qualitative cell model. The Genome-based E-cell Modeling System (GEM System) developed upon the generic bioinformatics workbench G-language Genome Analysis Environment [1] realizes a fully automatic conversion of genome sequence data into a qualitative in silico cell model, linking information from major public databas such as GenBank, EMBL, SWISS-PROT, KEGG, ARM, Brend, and WIT.

2 Method and Results

The ORFs are matched to the corresponding proteins derivatives and thus to the stoichiometric reactions, through a combined method using the three levels of prediction, annotation, homology, and orthology. Orthology is a key method in this step, because the method is able to identify enzymes with similar functions sharing limited homology in terms of their amino acid sequences. Taking advantage of the biochemical databases available online, the list of enzymes are matched to their metabolic reaction list expressed in their stoicheometry. Thus a static metabolic reaction matrix can be obtained at a cell-wide scale, and the generated reaction network is then checked with KEGG reference pathway maps for false positives and false negatives, and also for connectivity of pathways where applicable.

Additional information about the metabolic reactions is also obtained during the process of pathway reconstruction, including the reaction mechanism, protein localization, reversibility of the reactions, metabolite properties of the involved compounds, and other enzyme properties such as the information on its subunits and effectors.

Using the hybrid dynamic/static simulation method, when all bottleneck reactions are dynamically represented, every other reaction can be statically represented with the same accuracy as completely dynamic simulation. GEM system can semi-automatically generate the static part of this hybrid algorithm, and provides a base for large-scale dynamic simulation. The generated models can be directly simulated using E-Cell, and SBML porting is being developed.

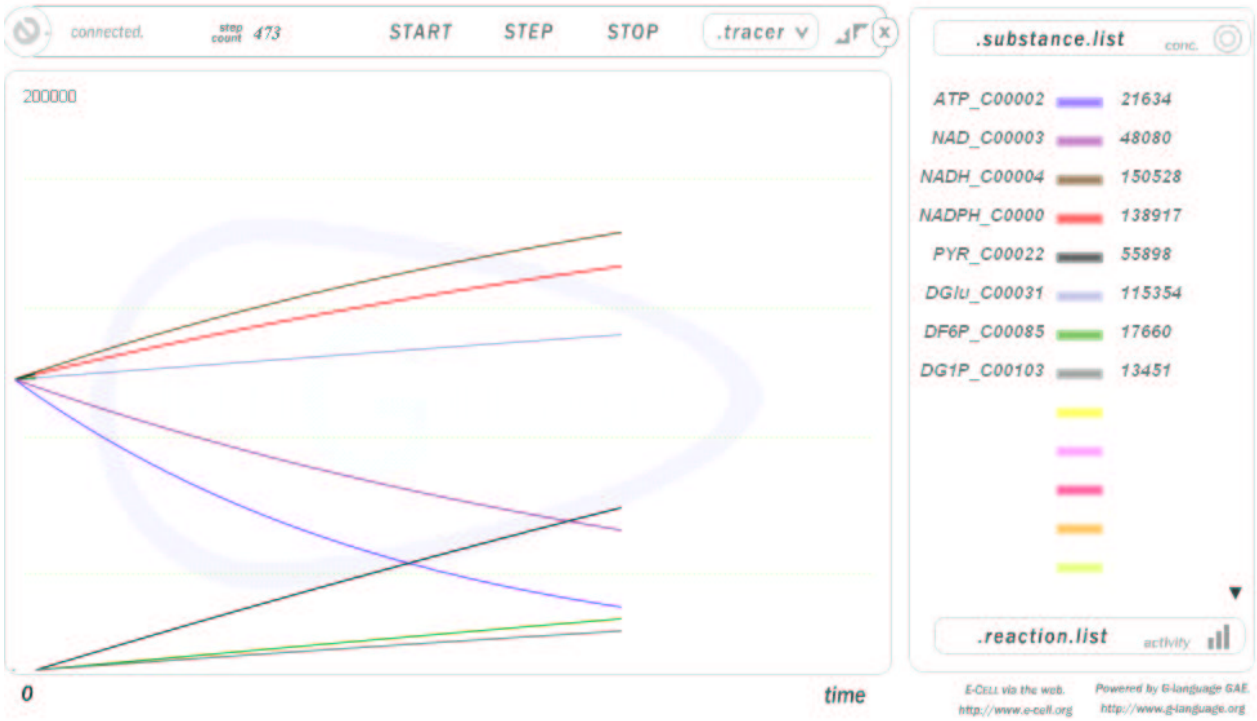

Figure 1: Simulation of Glycolysis pathway of automatically generated *Escherichia coli* model.

3 Discussion

Currently the checking process only employs the KEGG database, but we will utilize other pathway databases including BioCyc. Utilizing other major databases such as TRANSPATH/TRANSFAC, we will include the gene regulatory network and signal transduction pathways generation to the present metabolic pathway reconstruction.

References

[1] Arakawa, K., Mori, K., Ikeda, K., Matsuzaki, T., Kobayashi, Y., and Tomita, M., G-language genome analysis environment: a workbench for nucleotide sequence data mining, *Bioinformatics*, 19:305–6, 2003.

[2] Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T.S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J.C., and Hutchinson, C.A., E-CELL: software environment for whole-cell simulation, *Bioinformatics,* 15:72–84, 1999.