GeMCore : a mammalian comparative mapping database

Navratil Vincent [1, 2, *], Bronner Gisèle [1], Spataro Bruno [1], Faraut Thomas [2], Mouchiroud Dominique [1], Gautier Christian [1]

[1] Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Lyon 1
43 Bd du 11 Novembre 1918, 69622 VILLEURBANNE CEDEX

[2] Laboratoire de Génétique Cellulaire, UR444 INRA Toulouse
INRA BP27 31326 CASTANET-TOLOSAN

[*] contact : phone : +33 472 44 80 00 (83659); fax : +33 478 89 27 19; e-mail : navratil@biomserv.univ-lyon1.fr

Keywords : Mammalian comparative mapping, Knowledge representation, Database

Introduction

The aim of **genomic mapping** is to locate genetic information along chromosomes. Mapping area acts as an "anchor" for many biological fields and leads to assign a large amounts of heterogeneous data (gene, tissular expression, polymorphism, ...) to genomic regions, using nucleotide sequences available in many databases (EST, SNP, microsatellite, haplotype, ...).
Many studies have demonstrated the high level of conservation between mammalian genome organization [1], making it clear that **comparative mapping** analysis would become a powerful tool to gather new genomic knowledge. Indeed, comparative mapping provides clues **to understand mammalian genomic evolution** at different level, such as regional variations in G+C content or evolutionary rate [2,3] and clustering of regulated genes [4]. Furthermore, few genomes are currently completely sequenced among mammals, like human [5] and mouse [6] whereas only particular maps (cytogenetical, genetical, RH) are available for animals of agronomical (cattle, pig, sheep) or medical interests (dog, cat). Thus, comparative mapping allows **an exchange of information** between model species (human, mouse, rat) and mammals of agronomical or medical interests in order to map traits of economical value like QTLs [1] or diseases.
However, at the present time comparative mapping tools, taking into account many biological fields and many species for global analysis, are not available. They only deal with specifical biological problems such as genes, markers or regions comparison between a limited number of species.

The GeM project

The goal of the **GeM** (Genomic Mapping) **project** is to provide a general **bioinformatic environment** allowing multi-species retrieval and comparison of heterogeneous biological data (genes, SNP, microsatellites, tissular expression, homology) based on their genomic localization, within a species (paralogy, polymorphism) and between species (orthology, conserved regions).

The GeMCore database

The **GeMCore database** is a part of the GeM environment and will be connected to **graphical user interfaces** devoted to specifical biological problems, like **GeMMe** [7] for "**G**enomic **M**apping and **M**olecular **e**volution" and will be also a starting point for developments of algorithms using markers localization and sequence data.

At the **conceptual level**, the GeMCore database was built using an "Object-Relational" data model. The aim of the GeMCore model [8,9] is to propose a comparative mapping ontology. At present the knowledge representation is structured into object and relational hierarchies and describes several concepts such as:

- genomic mapping through markers (genes, SNP, microsatellites), maps (cytogenetical, genetical, RH, genomic sequence) and localization of markers on maps,
- nucleotide sequences (genomic sequences, CDS),
- homologies (gene family, orthology, conserved regions),
- polymorphism (population, allele frequency).

At the **physical level**, the GeMCore model is implemented using PostgreSQL, an Object-Relational DBMS, supporting almost all SQL constructs, including subselects, transactions, and user-defined types and functions.

At present time, the GeMCore database contains comparative mapping data for three mammalian species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*) and an automatic procedure ensure reliability and maintenance of heterogeneous data on many databases such as :

- Ensembl [10], release 12.3, for complete genome annotations,
- Hovergen [11], release 43, and Orthogen, respectively for homologous families and orthologous groups, using ACNUC [12] for Ensembl and Hovergen nucleotide sequences retrieval.

Currently, the GeMCore **database connection** can be made through a psql client command or a postgresql driver available for many programming languages (java, php, perl, ...).

To conclude, GeMCore is also a part of the bioinformatic project called **"Development of a vertebrate comparative mapping platform"**, in collaboration with many genomic mapping laboratories working on INRA agronomical animals (cattle, pig, chicken, trout, ...) and medical model mammals (dog). The main target is to provide a database service for multiple users, built on a Client-Server model, that will allow mapping data integration and comparative mapping data querying through a graphical user interface.

References

[1] Andersson L. et al. (1996) Comparative Genome Organization of Vertebrates. Mammalian Genome, 7 :717-734.
[2] Matassi G. et al. (1999) Chromosomal location effect on gene sequence evolution in mammals. Curr. Biol. 9 :786:791.
[3] Smith N.G. et al. (2002) Regional similarities in polymorphism in the human genome extend over many megabases. Trend Genet. 18 :281-3.
[4] Lercher M.J. et al. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Gent. 59 :380-386.
[5] Lander E.S. et al. (2001) Initial Sequencing and analysis of the human genome. Nature 409 :860-921.
[6] Waterston R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature. 420 :520-62.
[7] Spataro B. et al. (2000) Analyse comparative des génomes humain et murins avec l'interface GeMMe. JOBIM 2000, receuil des actes (2000) 351-360.
[8] Bronner G. et al. (2001) GeMCore, a knowledge base dedicated to mapping mammalian genomes. Lect. Notes Comp. SCI .2066 :12-13.
[9] Bronner G. et al. (2002) Modeling comparative mappingusing objects and associations. Comput. Chem. 26 :413-420.
[10] Clamp M. et al. (2003) Ensembl 2002 : accommodating comparative genomics. Nucleic Acids Res. 31 :38-42.
[11] Duret L. et al. (1994) Hovergen : comparative analysis of homologous vertebrate genes. In : S. Letovsky (ed.) Bioinformatics : Databases and Systems. Kluwer Academic Publishers, Boston, Dortrecht, London, (1999) 21-36.
[12] Gouy M. et al. (1985) ACNUC - a portable retrieval system for nucleic and sequence databases: logical and physical design and usage. CABIOS 1 :167-172.