Data Retrieval and Handling Tools for the PBIL Gene Family Databases

Guy Perrière [(1)], Jean-François Dufayard [(2)], Simon Penel [(1)], Julien Grassot [(3)],
Laurent Duret [(1)] and Manolo Gouy [(1)]

[(1)] Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard – Lyon 1,
43 bd. du 11 Novembre 1918, 69622 VILLEURBANNE Cedex – France
{perriere,penel,duret,mgouy}@biomserv.univ-lyon1.fr
[(2)] Unité de Recherche INRIA Rhône-Alpes, 655 av. de l'Europe, MONTBONNOT
38334 SAINT ISMIER Cedex – France
Jean-Francois.Dufayard@inrialpes.fr
[(3)] Centre de Génétique Moléculaire et Cellulaire, UMR CNRS 5534, Université Claude Bernard – Lyon 1,
43 bd. du 11 Novembre 1918, 69622 VILLEURBANNE Cedex – France
grassot@biomserv.univ-lyon1.fr

Keywords: Genomics, Databases, Phylogeny

Introduction

Different homologous gene family databases have been developed and are maintained at the PBIL (Pôle Bioinformatique Lyonnais). Historically, the first one was HOVERGEN, a collection of homologous genes from vertebrates [1]. Since, we have developed HOBACGEN [2], devoted to prokaryotic organisms, NuReBase [3], for nuclear receptors in metazoans, and RTKdb [4], for tyrosine kinase receptors. What makes these databases especially useful is the fact that the gene sequences they contain are clustered into homologous families (or sub-families in the case of NuReBase and RTKdb), and that we provide the corresponding multiple alignments and phylogenetic trees. Due to the peculiar nature of the information available in these systems, we have developed specific retrieval and handling tools allowing to browse the data and to select subsets of families.

Some details

1. PBIL server

The easiest way to access these databases is through the PBIL World-Wide Web server (http://pbil.univ-lyon1.fr). On this server, we have recently implemented a set of programs allowing to perform queries centered on families. For instance it is possible to retrieve all families that are shared by a given set of taxa and that are not present in a second set. Any taxonomic level can be used and mixed to compose the query (*e.g.*, *Homo sapiens*, Primates, *Mammalia*). The first set of taxa can be used for an "inclusive" or an "exclusive" selection of families. With the inclusive search, any family presenting at least one sequence of each taxa of the list will be selected and with the exclusive search, the families presenting exclusively taxa of the list will be selected. Moreover families can be pre-selected according to the number of sequences and/or the number of species. For example, it is possible to retrieve with this system all families of bacterial genes that are specific of a pathogenic strain of *Escherichia coli*.
As these databases also include alignments and trees for each family, different possibilities are provided to represent and handle these kinds of data. First, alignments can be simply displayed with a coloring scheme in an HTML document or can be viewed with the JalView applet (http://www.ebi.ac.uk/~michele/jalview/contents.html). Trees can be displayed with the ATV (A Tree Viewer) applet [5]. Second, it is possible to use helper applications as specific MIME-types were defined for these documents [6]. Recommended helper for the

alignments is SeaView [7], while NJplot [6] is a good option to visualize phylogenetic trees. Lastly, it is possible to store the corresponding files locally and to visualize them with *ad hoc* programs. The format used is CLUSTAL for the alignments and NEWICK for the trees.

2. FamFetch interface

For more specialized use, we have also developed the FamFetch client [2], a Java application allowing to query our gene family databases on remote servers (http://pbil.univ-lyon1.fr/software/famfetch.html). This application provides the same facilities as the Web server, but with a greater interactivity: sequences/alignments/trees retrieval and handling, plus families selection facilities. Among the recent improvements brought by the new version of FamFetch is the integration of a tree pattern search facility allowing to retrieve families for which the tree topology matches a given pattern (Fig. 1). For that purpose, we have implemented an algorithm allowing to solve the problem of the unordered tree pattern matching [8, 9] in the special case of phylogenetic trees.

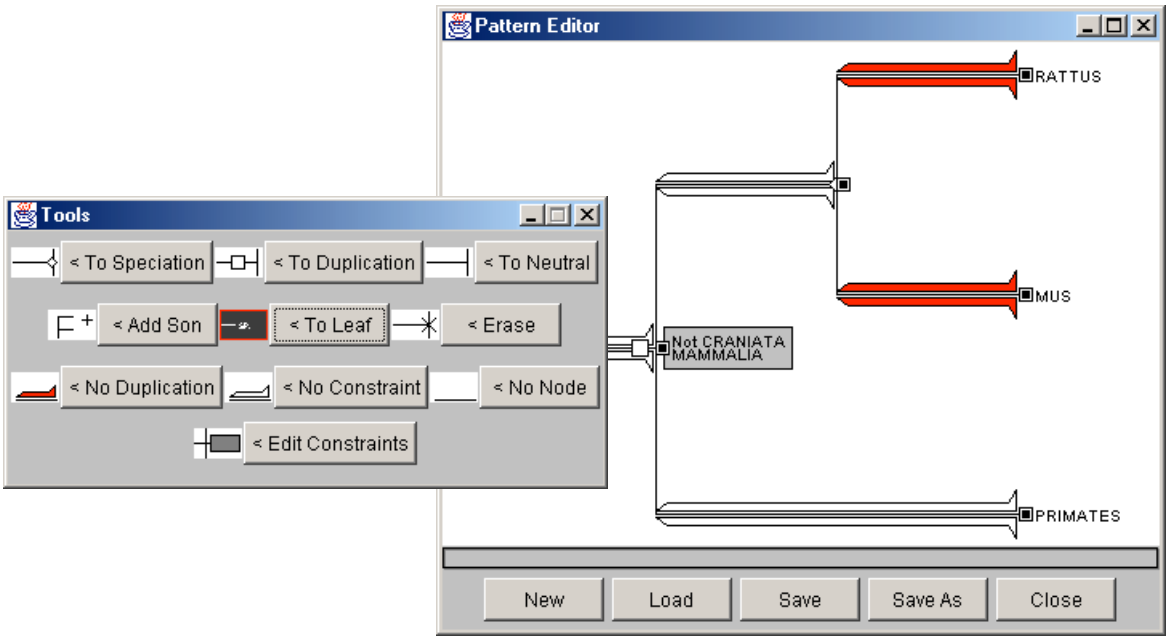

Fig. 1 Screen shots of the tree pattern editor implemented in FamFetch.

This new functionality is especially important as it helps identify orthologous pairs in multigenic families. This distinction between orthologs and paralogs is important to predict the function of a new gene by homology with already characterized genes, because orthologous sequences are more reliable predictors than paralogous sequences. The distinction between orthologs and paralogs is also absolutely necessary for phylogenetic analyses where only orthologs should be used in order to infer a species phylogeny from a gene phylogeny.

Our tree pattern matching implementation presents advantages and inconvenients, as compared with manual search. Indeed, visual inspection allows to detect anomalies on phylogenetic trees and brings a better flexibility in the search. Also, with trees that are affected by reconstruction artefacts (such as poorly chosen phylogenetic roots for deep patterns), the algorithm cannot perform efficiently. On the other hand, the program implemented is very fast, so the search for a given pattern on an entire database is compatible with an interactive application. For instance, pattern search into the HOVERGEN or the HOBACGEN databases (which contain tenth of thousands of trees) takes less than 30 seconds on the PBIL server.

References

[1] L. Duret, D. Mouchiroud and M. Gouy M., HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.*, 22:2360-2365, 1994.

[2] G. Perrière, L. Duret and M. Gouy, HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, 10:379-385, 2000.

[3] J. Duarte, G. Perrière, V. Laudet and M. Robinson-Rechavi, NuReBase: database of nuclear hormone receptors. *Nucleic Acids Res.*, 30:364-368, 2002.

[4] J. Grassot, G. Mouchiroud and G. Perrière, RTKdb: database of receptor tyrosine kinase. *Nucleic Acids Res.*, 31:353-358, 2003.

[5] C.M. Zmasek and S.R. Eddy, ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17:383-384, 2001.

[6] G. Perrière and M. Gouy, WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie*, 78:364-369, 1996.

[7] N. Galtier, M. Gouy and C. Gautier, SeaView and Phylo_win: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Applic. Biosci.*, 12:543-548, 1996.

[8] A.V. Aho, M. Ganapathi and S.W.K. Tjiang, Code generation using tree matching and dynamic programming. *ACM Trans. Program. Lang. Syst.*, 11:491-516, 1989.

[9] P. Kilpeläinen and H. Mannila, Retrieval from hierarchical texts by partial patterns, in R. Korfhage, E.M. Rasmussen and P. Willett (eds.), *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 214-222, ACM Press, New York, 1993.