Nucleic Acids Research, 2004, Vol. 32, Database issue ***D307–D310***
DOI: 10.1093/nar/gkh114

MolliGen, a database dedicated to the comparative genomics of Mollicutes

Aurélien Barré[1], Antoine de Daruvar[1] and Alain Blanchard*

INRA—Université de Bordeaux 2, IBVM, Bordeaux, France and [1]Université de Bordeaux 2, Centre de Bioinformatique de Bordeaux, Bordeaux, France

Received August 14, 2003; Revised and Accepted October 15, 2003

ABSTRACT

Bacteria belonging to the class Mollicutes were among the first ones to be selected for complete genome sequencing because of the minimal size of their genomes and their pathogenicity for humans and a broad range of animals and plants. At this time six genome sequences have been publicly released (*Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Ureaplasma urealyticum-parvum*, *Mycoplasma pulmonis*, *Mycoplasma penetrans* and *Mycoplasma gallisepticum*) and as the number of available mollicute genomes increases, comparative genomics analysis within this model group of organisms becomes more and more instructive. However, such an analysis is difficult to carry out without a suitable platform gathering not only the original annotations but also relevant information available in public databases or obtained by applying common bioinformatics methods. With the aim of solving these difficulties, we have developed a web-accessible database named MolliGen (http://cbi.labri.fr/outils/molligen/). After selecting a set of genomes the user can launch various types of search based on annotation, position on the chromosomes or sequence similarity. In addition, relationships of putative orthology have been precomputed to allow differential genome queries. The results are presented in table format with multiple links to public databases and to bioinformatic analyses such as multiple alignments or BLAST search. Specific tools were also developed for the graphical visualization of the results, including a multigenome browser for displaying dynamic pictures with clickable objects and for viewing relationships of precomputed similarity. MolliGen is designed to integrate all the complete genomes of mollicutes as they become available.

INTRODUCTION

Bacteria belonging to the class Mollicutes are characterized by the small size of their genomes, between 580 and 2200 kb. Although they lack peptidoglycan, they are phylogenetically related to Gram-positive bacteria with low-G+C% genomes. A large number of mollicutes infect animals, with initial colonization of mucosal surfaces such as the respiratory and urogenital tracts. In farm animals, associated diseases have a high economic impact especially for cattle, small ruminant, pig and poultry industries (1). In humans, four mollicutes (*Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Mycoplasma hominis* and *Ureaplasma urealyticum*) are associated with diseases characterized by high morbidity and low mortality; infections can be systemic in neonates and in immunocompromised patients (2). Two *Spiroplasma* species and the phytoplasmas are phloem restricted, transmitted by insect vectors and are pathogenic for plants (3).

Unlike other bacteria with small genomes, such as rickettsias and chlamydias, the mollicutes (except phytoplasmas) can be cultured in acellular medium and are thus considered as the best representatives of the concept of the minimal cell (4). Genome sequencing has been completed for six mollicutes and is in the finishing phase for several others. Because mollicutes represent a phylogenetically coherent group comprised of pathogens, colonizing a broad range of different hosts and body sites, they offer, as such, an ideal set of model organisms for comparative genomic studies. However, *in silico* comparative genomics have been hampered by several factors: (i) the lack of a suitable platform to formulate queries in comparative genomics, (ii) inconsistencies in the annotation of the genomes, (iii) the lack of integration between annotation data and other sources of information. We present here MolliGen, a web-accessible database that is intended to solve these difficulties.

MOLLIGEN DATA

Information extraction from primary data sources

The same primary data source is used for all the genomes that are loaded in Molligen in order to get information as homogeneous and consistent as possible. This source is the complete genome entry from the NCBI Reference Sequence collection (5), available on the NCBI ftp server (ftp://ftp.

*To whom correspondence should be addressed at INRA, Institut de Biologie Végétale Moléculaire, UMR Génomique Développement Pouvoir Pathogène, 71 avenue Edouard Bourlaux, BP 81, F-33883 Villenave D'Ornon Cedex, France. Tel: +33 5 57 12 23 93; Fax: +33 5 57 12 23 69; Email: ablancha@bordeaux.inra.fr ncbi.nih.gov/genomes/Bacteria/) as individual files in GenBank format. Dedicated Perl scripts were written to retrieve various attributes of genome features, in particular the mnemonics, gene names, GI, description and sequence for all putative proteins.

This information is stored as structured data in MolliGen relational database together with the complete sequence of each genome. This process of information extraction involves careful manual checking and data curation. Indeed, there is still some heterogeneity in the way the information is stored within the source files. For example, depending on the genome, the genes coding for tRNA correspond either to features of type 'gene' (i.e. *M.genitalium* entry NCBI Accession NC_000908.1) or to features of the type 'tRNA' (i.e. *M.pneumoniae* entry NCBI Accession NC_000912.1). Detection and correction of such data inconsistencies are absolutely crucial for comparative genomic studies. Furthermore, the accuracy and completeness of the data collected in the database are also essential to establish links towards external public data sources.

Enrichment with complementary information from public databases

Once the batch of initial data is loaded in the MolliGen database, it is enriched with additional information which is retrieved from public repositories and relates to the functional annotation of the proteins coded by each genome.

Two main sources are used:

(i) COG: the membership of each protein to a Cluster of Orthologous Groups (6) is retrieved from NCBI. These data are used in MolliGen for comparative purposes given that two proteins from different genomes that belong to the same COG are putative orthologues.

(ii) KEGG: in a similar way, EC numbers and relevant pathway information are retrieved from KEGG databases (7).

Local computation of additional data

In order to enrich and to homogenize information available for each CoDing Sequence (CDS), various bioinformatics tools are systematically applied:

(i) TMAP (8) software from the EMBOSS package is used for the prediction of transmembrane segments;

(ii) protein motifs are identified with ScanProsite, a program for the detection of motifs recorded in the PROSITE database (9);

(iii) BLAST (10) is used to search for homologues within all the mollicute proteomes and within the complete Swiss-Prot database;

(iv) in addition, Bi Directional Best Hits (BDBH) are computed using a proprietary method (BDBH criteria and tables are available as additional material at http://cbi.labri.fr/outils/molligen/NAR/sup.php) in order to build tables of putative orthologous pairs of proteins.

QUERY, BROWSE AND COMPARE GENOMES

Query

MolliGen provides access to integrated data via multiple methods and formats. Primary access is provided through a web form in which a query is dynamically built by the user with keywords or relative genome positions on selected genomes (Fig. 1A). The user can then choose by a click to display the obtained results either for only one species or globally which is appropriate for comparative purposes. This interface includes various links to other types of information such as a MolliGen summary for each CDS, NCBI and SwissProt entry, and to bioinformatic methods (Fig. 1B–D). A second access offers the BLAST program to find homologues for a nucleic or proteic sequence. The targeted genomes and other parameters are selectable. A third access allows users to find genes having a defined pattern series (up to three), allowing mismatches.

Browse the genomes

The user can navigate through all genomes integrated in MolliGen by using two complementary programs: (i) the Generic Genome Browser (11), which allows a single genome with all associated features (genes, RNAs, repeat region …) to be viewed, (ii) the multi-genome browser, which was developed for MolliGen allows more than one genome to be browsed through and relationships between CDSs of these genomes to be displayed (Fig. 1D). The picture is dynamic, integrates both global and individual zoom funtionalities and the objects are clickable. Finally, to visualize relationships between two genomes over their full length, genome alignment can be performed and viewed as a clickable dot-plot representation. Regions of interest can then be directly selected and viewed using the multi-genome browser.

Comparative genomics

To compare mollicute genomes, various tools were developed and integrated in the web interface.

(i) Based on KEGG predictions for enzymic functions, a metabolic pathway viewer has been developed. This tool allows us to find, within a selected pathway, the enzymes that are common or different among two genome sets; each set consists of one or more genomes.

(ii) To find genes specific for a genome or a set of genomes, multi proteome differential queries can be performed. It allows genes from a genome group having homologues in a targeted group of other genomes to be found (Fig. 1E–F). A third group of genomes can be selected as an exclusion genome set where no homologues must be found.

System design and implementation

The MolliGen database consists of two components:

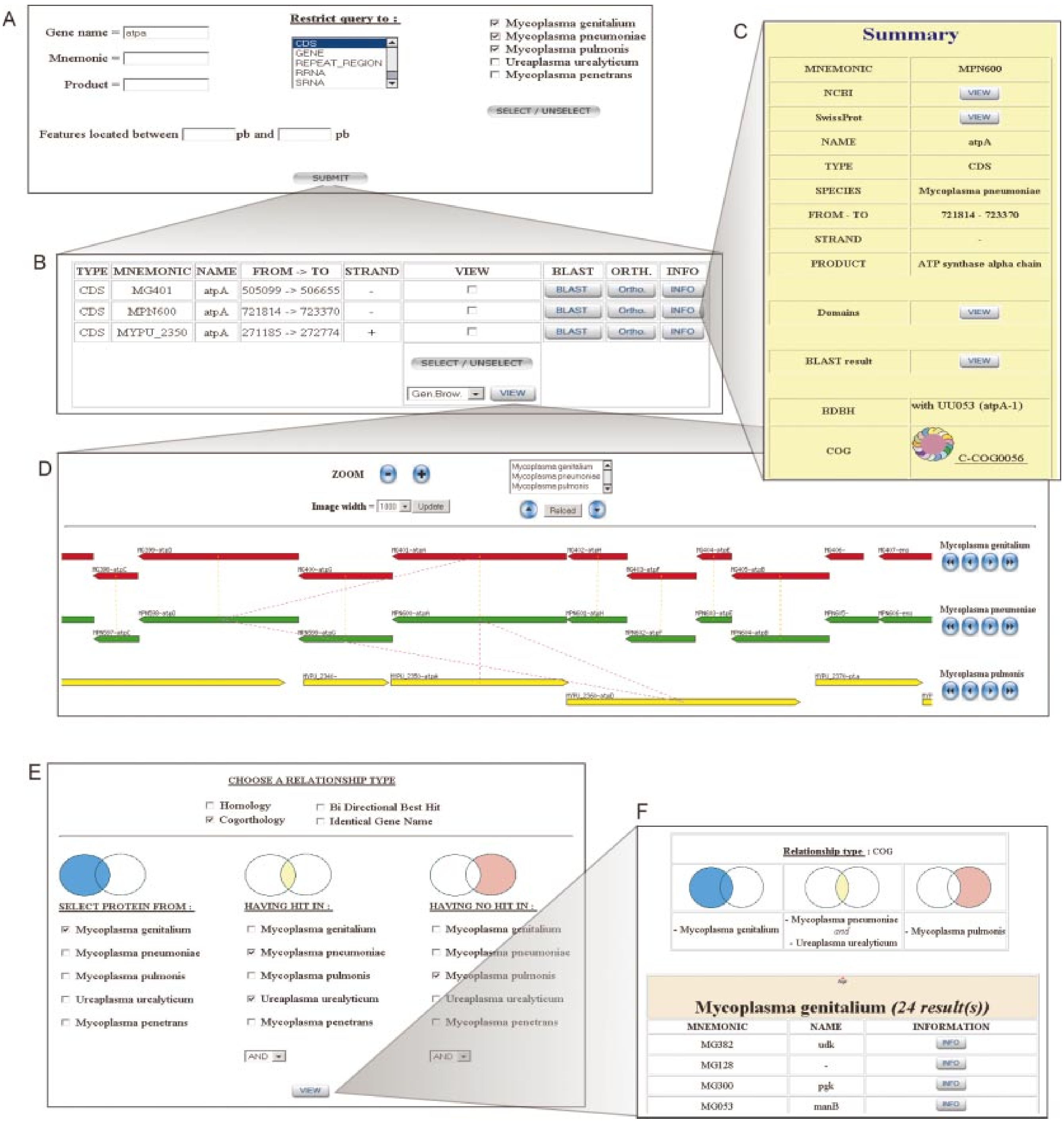
(i) extracted or computed information stored in a relational database using the PostgreSQL management system (a database schema is available upon request);

(ii) native output of bioinformatics methods (BLAST, TMAP …) are kept as flat files.

Software developments are based on a Perl language for all the data processing, and on PHP for the user interface.

PERSPECTIVES

At the time of writing, five mollicute genomes are available in MolliGen: *M.genitalium, M.pneumoniae*, *Mycoplasma pulmonis*, *Mycoplasma penetrans* and *U.urealyticum*. The recently released genome of *Mycoplasma gallisepticum* will be added to the database and so will all the complete mollicute



Gene name = atpa

Mnemonic =

Product =

Restrict query to :

CDS
GENE
REPEAT_REGION
RRNA
SRNA

☑ Mycoplasma genitalium
☑ Mycoplasma pneumoniae
☑ Mycoplasma pulmonis
☐ Ureaplasma urealyticum
☐ Mycoplasma penetrans

Features located between ____ pb and ____ pb

TYPE	MNEMONIC	NAME	FROM -> TO	STRAND	VIEW	BLAST	ORTH.	INFO
CDS	MG401	atpA	505099 -> 506655	-	☐	BLAST	Ortho.	INFO
CDS	MPN600	atpA	721814 -> 723370	-	☐	BLAST	Ortho.	INFO
CDS	MYPU_2350	atpA	271185 -> 272774	+	☐	BLAST	Ortho.	INFO

Summary

MNEMONIC	MPN600
NCBI	VIEW
SwissProt	VIEW
NAME	atpA
TYPE	CDS
SPECIES	Mycoplasma pneumoniae
FROM - TO	721814 - 723370
STRAND	-
PRODUCT	ATP synthase alpha chain
Domains	VIEW
BLAST result	VIEW
BDBH	with UU053 (atpA-1)
COG	C-COG0056

CHOOSE A RELATIONSHIP TYPE

☐ Homology ☐ Bi Directional Best Hit
☑ Cogorthology ☐ Identical Gene Name

SELECT PROTEIN FROM :	HAVING HIT IN :	HAVING NO HIT IN :
☑ Mycoplasma genitalium	☐ Mycoplasma genitalium	☐ Mycoplasma genitalium
☐ Mycoplasma pneumoniae	☑ Mycoplasma pneumoniae	☐ Mycoplasma pneumoniae
☐ Mycoplasma pulmonis	☐ Mycoplasma pulmonis	☑ Mycoplasma pulmonis
☐ Ureaplasma urealyticum	☑ Ureaplasma urealyticum	☐ Ureaplasma urealyticum
☐ Mycoplasma penetrans	☐ Mycoplasma penetrans	☐ Mycoplasma penetrans

Relationship type : COG

- Mycoplasma genitalium	- Mycoplasma pneumoniae *and* - Ureaplasma urealyticum	- Mycoplasma pulmonis

Mycoplasma genitalium *(24 result(s))*

MNEMONIC	NAME	INFORMATION
MG382	udk	INFO
MG128	-	INFO
MG300	pgk	INFO
MG053	manB	INFO

Figure 1. Composite display of screens demonstrating features of the MolliGen interface. (A–D) Following a query based on a simple gene name within selected genomes (**A**), an output table is obtained (**B**). In this table, each line corresponds to a hit with relevant information including a MolliGen summary (**C**). By selecting one or more hit(s) in the table, one can view in particular a dynamic alignment of genome features (**D**) with clickable elements. (E–F) Multi-genome differential queries can be formulated using various precomputed homology relationships (**E**). The results appear as a table of genes sorted by genomes (**F**).

genomes as they become publicly available. Although the integration of a new genome requires careful manual checking of the data, the database and a number of procedures have been designed to facilitate this integration.

As already mentioned, the homogeneity and the consistency of the data are essential for comparative genomics analysis. One of the reasons for inconsistencies in the annotations is that the sequencing of the mollicute genomes is soon going to span a 10 year period. Indeed, the first mollicute sequenced genome [*M.genitalium* (12)] was annotated in 1995 and the last one (*M.gallisepticum*) was released in June 2003. In addition, the genome of *M.pneumoniae*, which was originally annotated in

1996 (13), was re-annotated in 2000 (14). Our aim is to provide users with an even more consistent annotation. The comparison of mollicute genomes provides efficient means to verify the consistency of annotations. In the future, we plan to use MolliGen database to implement functionalities for the re-annotation of these genomes. Since the relationships of similarity/orthology are stored in the database, it would be possible to scan them in order to build automatic procedures to highlight putative annotation inconsistencies. Those can then be examined manually by an expert annotator to provide a revised annotation. This re-annotation will be performed by coordinating with other international annotation initiatives such as HAMAP (15). In addition, since the number of bacterial genomes in the public databases is increasing at a high pace, the number of CDSs annotated as hypothetical on the basis of the lack of significant similarity needs to be re-evaluated. This again, can easily be done using the Molligen database where the list of putative homologues for each CDS is maintained.

ACKNOWLEDGEMENTS

The authors thank their colleagues who kindly shared data and comments: Pascal Sirand-Pugnet, Eduardo Rocha, Cécile Bébéar, David Sherman, Pascal Durrens and Alexis Groppi. This research was supported in part by grants from INRA, Région Aquitaine and University Victor Ségalen Bordeaux 2. A.B. was the recipient of a fellowship from the 'Programme Bioinformatique inter-EPST'.

REFERENCES

1. Frey,J. (2002) Mycoplasmas of animals. In Razin,R. and Herrmann,R. (eds), *Molecular Biology and Pathogenicity of Mycoplasmas*. Kluwer Academic/Plenum Publishers, London, pp. 73–90.
2. Blanchard,A. and Bébéar,C.M. (2002) Mycoplasmas of humans. In Razin,R. and Herrmann,R. (eds), *Molecular Biology and Pathogenicity of Mycoplasmas*. Kluwer Academic/Plenum Publishers, London, pp. 45–71.
3. Bove,J.M., Renaudin,J., Saillard,C., Foissac,X. and Garnier,M. (2003) *Spiroplasma citri*, a plant pathogenic mollicute: Relationships with its two hosts, the plant and the leafhopper vector. *Annu. Rev. Phytopathol.*, **41**, 483–500.
4. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science*, **286**, 2165–2169.
5. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
6. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
7. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
8. Persson,B. and Argos,P. (1994) Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.*, **237**, 182–192.
9. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
10. 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.
11. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
12. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
13. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–4449.
14. Dandekar,T., Huynen,M., Regula,J.T., Ueberle,B., Zimmermann,C.U., Andrade,M.A., Doerks,T., Sanchez-Pulido,L., Snel,B., Suyama,M. *et al.* (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res.*, **28**, 3278–3288.
15. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.