Leproma: A *Mycobacterium leprae* genome browser

L. JONES*, I. MOSZER** & S. T. COLE***
**Service Informatique Scientifique*
***Unite de Genetique des Genomes Bacteriens*
****Unite de Genetique Bacteriens, Institut Pasteur, Paris, France*

Introduction

Leproma is a powerful Web based tool for extracting information about annotations from a *Mycobacterium leprae* genome database. The URL for the Leproma web site is http://genolist.pasteur.fr/Leproma.

With *Leproma*, the user may search the *M. leprae* genome[1] database using several search criteria. One can search by gene name or synonym, by region in the genome, by gene function or classification, by DNA or protein patterns, by a BLAST[2,3] or FASTA[4] search in the DNA sequence or the protein sequences, or by free text.

Search results can be in the form of a list where the columns in the list are set by the user or in the form of a drawing if the search results in a region in the genome. The user can also download or view the DNA sequence or the protein sequence from a single gene or from the list of a search result.

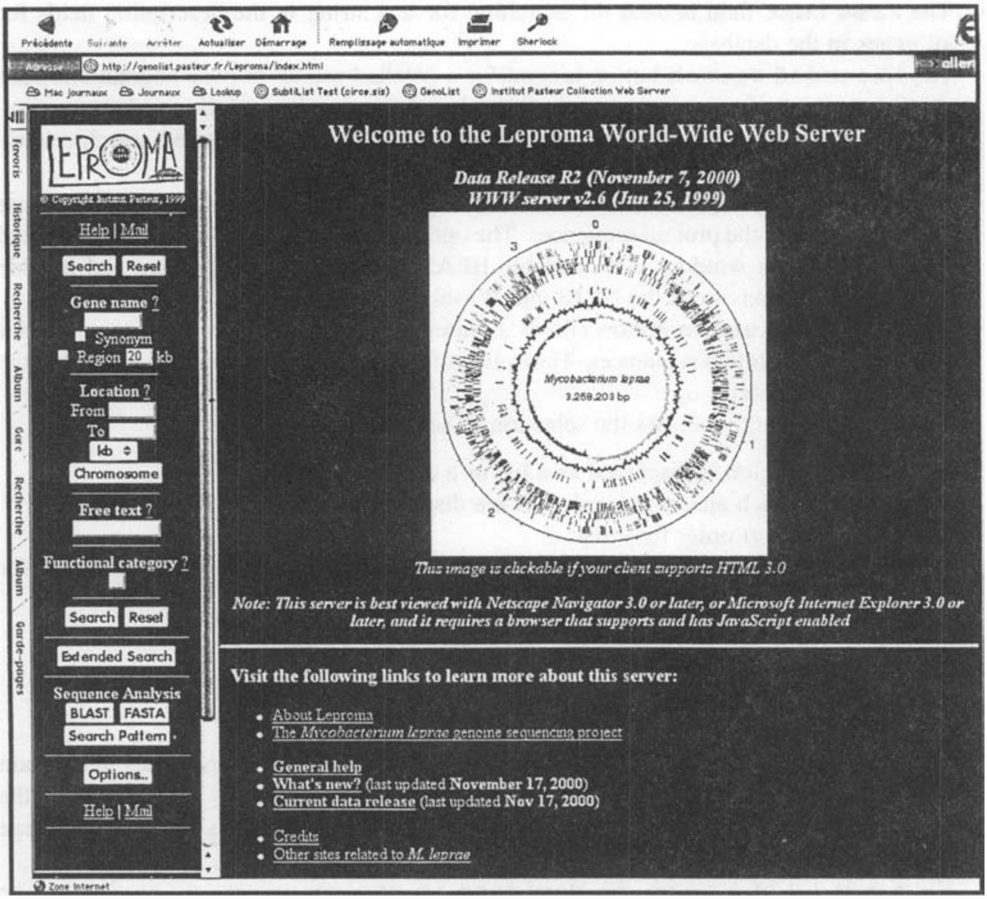
Leproma opening Web page

The *Leproma* web server will show a page (Figure 1) with free frames. The left frame shows the form for entering search criteria. The top right frame is used for showing the search results, either in the form of a list or a drawing (if the search request results in a region). It is also used for entering additional search criteria for a BLAST/FASTA search, a pattern search, or an extended annotation search. The bottom right frame is used for showing detailed information for a gene. Other pages may be shown for the help pages, a view of a DNA or protein sequence, the pre-calculated BLAST results for a gene, or links to other external databases.

The opening Web page (Figure 1; see also accompanying colour poster) shows in the upper right frame, the *Leproma* data version and date of the data release and an image of the *M. leprae* genome showing CDS or ORF positions, gene classification, and GC%. This image is clickable and results in a list (or drawing) of all the genes ± 10 kb in the region

Correspondence to: S. T. Cole (e-mail: stcole@pasteur.fr)

0305-7518/01/072470+07 $1.00 © Lepra



Welcome to the Leproma World-Wide Web Server

Data Release R2 (November 7, 2000)
WWW server v2.6 (Jun 25, 1999)

This image is clickable if your client supports HTML 3.0

Note: This server is best viewed with Netscape Navigator 3.0 or later, or Microsoft Internet Explorer 3.0 or later, and it requires a browser that supports and has JavaScript enabled

Visit the following links to learn more about this server:

- About Leproma
- The *Mycobacterium leprae* genome sequencing project

- General help
- What's new? (last updated November 17, 2000)
- Current data release (last updated Nov 17, 2000)

- Credits
- Other sites related to *M. leprae*

Figure 1. *Leproma* opening Web page (for colour version of this figure see also accompanying colour poster.

around the click point. The bottom right frame gives links to information about the *M. leprae* genome project.

The left side frame contains the form for entering basic search criteria.

- The **`Gene name`** field is used for specifying a gene name. One or more wild-card characters may be used for searching several genes. For example, 'dnaA' will search for the gene '`dnaA`' while '`dna*`' will search for all genes whose name starts with 'dna'. The **Synonym** option allows the use of synonyms or old gene names. The **Region** option gives a list of genes ±20 kb around the specified gene name. This option is ignored if a wild-card character is used in the **`Gene name`** field.
- The **Location** fields **From** and **To** will give a list (or drawing) of all genes within the given genome coordinates. The coordinates are in kilo-bases. The **`Chromosome`** button will display the image of the *M. leprae* genome from the opening page (Figure 1).
- The **`Functional category`** field is used to search for all genes of the given classification code.

- The **Free text** field is used for searching for text string in the description fields for all genes in the database.
- The **Extended Search** button is used for a detailed search of gene annotations such as Accession number, protein length, DNA length, molecular weight, isoelectric point, and a free text search limited to certain annotation fields. The extended search criteria form is displayed in the upper right frame.
- The **BLAST** and **FASTA** buttons will allow a BLAST or FASTA search of the genome DNA sequence or the protein sequences. The options for BLAST or FASTA are presented in the upper right window. All available BLAST programs are presented and include the BLAST versions from the NCBI and Washington University.
- The **Search Pattern** button allows a pattern or motif search of the genome DNA sequence or the protein sequences. The options for the pattern search are presented in the upper right window.
- The **Options** button allows the selection of several default options:
  - Whether a region is displayed as a list or a drawing.
  - For a list, which annotation columns are displayed.
  - The default sort order for a list.
  - How many items to be displayed at once in a list. A **More**... button is displayed for scrolling through the list.

Leproma list display of a region

The list display (Figure 2; see also accompanying colour poster) shows the genes in a certain region or the genes based on the user specified search criteria. A gene name in the **Gene Name** column may be clicked to display the detailed annotations for the given gene in the lower right frame.

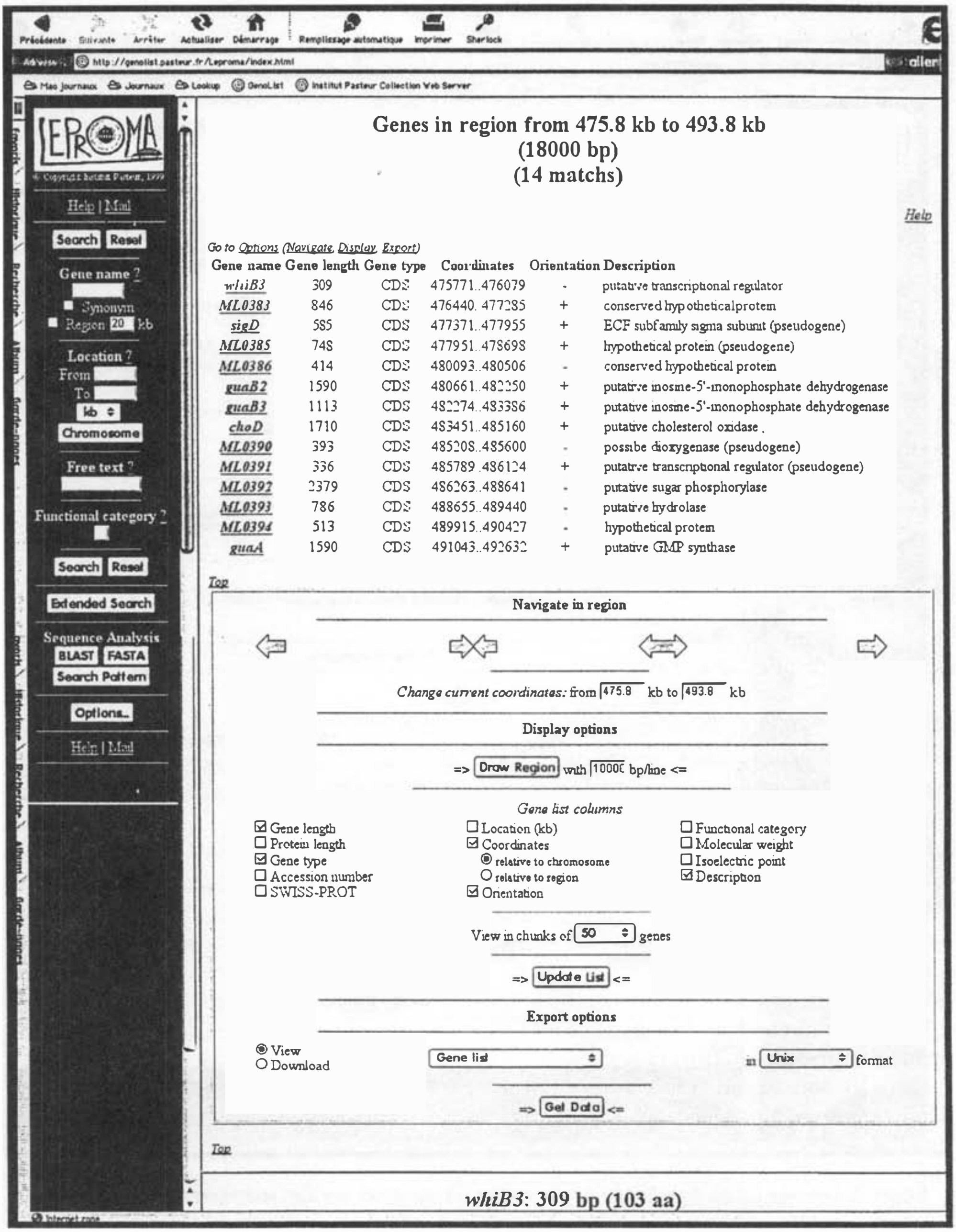
For a gene list of a region, the **Navigate in region** options are displayed. The user can shift the region to the left (lower genome coordinates) by clicking the yellow left arrow, expand the region, contract the region, or shift the region to the right. The region coordinates may be changed directly with the **From** and **To** fields and clicking the **Update List** button. Also, the list display can be changed to a drawing (Figure 3; see also accompanying colour poster) by clicking the **Draw Region** button.

The list columns may be changed by selecting the appropriate gene list columns and clicking the **Update List** button.

The user may export information from the list including the list of genes and selected columns, the protein sequence for each CDS in the list, the DNA sequence for all entries in the list, and for a region list, the DNA sequence for the region in direct or reverse complement form. The data may be displayed on the user's screen or it may be downloaded to a local file on the user's system.

Leproma drawing of a region

The region drawing (Figure 3; see also accompanying colour poster) displays a region in the Leproma genome. Each genomic element is represented by a graphic symbol on the drawing.



Genes in region from 475.8 kb to 493.8 kb
(18000 bp)
(14 matchs)

Help

Go to Options (Navigate, Display, Export)

Gene name	Gene length	Gene type	Coordinates	Orientation	Description
whiB3	309	CDS	475771..476079	-	putative transcriptional regulator
ML0383	846	CDS	476440..477285	+	conserved hypotheticalprotein
sigD	585	CDS	477371..477955	+	ECF subfamily sigma subunit (pseudogene)
ML0385	748	CDS	477951..478698	+	hypothetical protein (pseudogene)
ML0386	414	CDS	480093..480506	-	conserved hypothetical protein
guaB2	1590	CDS	480661..482250	+	putative inosine-5'-monophosphate dehydrogenase
guaB3	1113	CDS	482274..483386	+	putative inosine-5'-monophosphate dehydrogenase
choD	1710	CDS	483451..485160	+	putative cholesterol oxidase .
ML0390	393	CDS	485208..485600	-	possibe dioxygenase (pseudogene)
ML0391	336	CDS	485789..486124	+	putative transcriptional regulator (pseudogene)
ML0392	2379	CDS	486263..488641	-	putative sugar phosphorylase
ML0393	786	CDS	488655..489440	-	putative hydrolase
ML0394	513	CDS	489915..490427	-	hypothetical protein
guaA	1590	CDS	491043..492632	+	putative GMP synthase

Top

Navigate in region

Change current coordinates: from 475.8 kb to 493.8 kb

Display options

=> Draw Region with 10000 bp/line <=

Gene list columns

☑ Gene length
□ Protein length
☑ Gene type
□ Accession number
□ SWISS-PROT

□ Location (kb)
☑ Coordinates
● relative to chromosome
○ relative to region
☑ Orientation

□ Functional category
□ Molecular weight
□ Isoelectric point
☑ Description

View in chunks of 50 genes

=> Update List <=

Export options

● View
○ Download

Gene list in Unix format

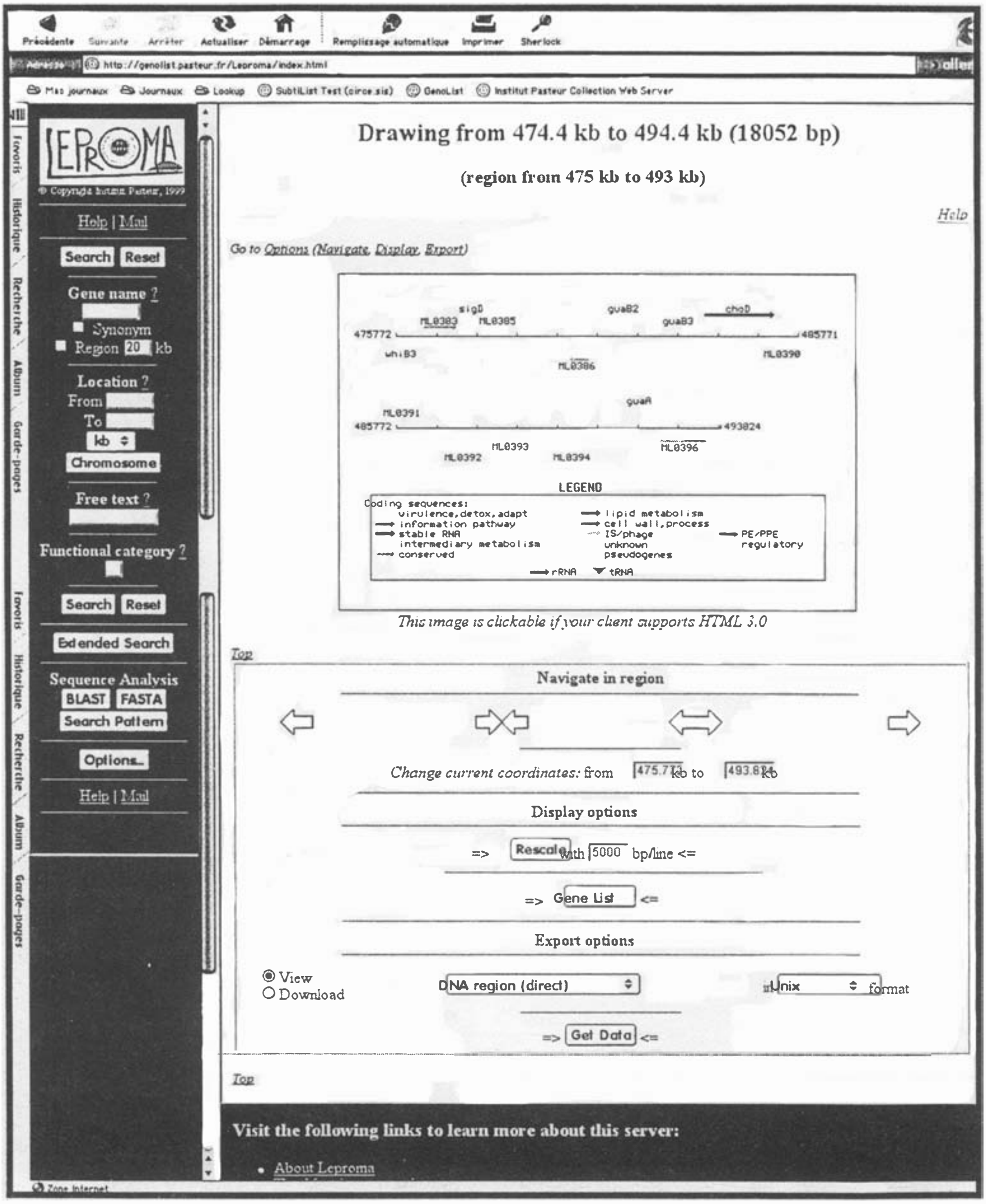
=> Get Data <=

Top

***whiB3*: 309 bp (103 aa)**

Figure 2. *Leproma* region list (for colour version of this figure see also accompanying colour poster).

A gene or CDS is represented by a coloured, horizontal line with the start-end points representing the beginning and end of the CDS and the direction of transcription. The colour represents the functional classification of the CDS that has been used for both the *M. leprae* and the *M. tuberculosis* genome projects.[5] There is one exception, as pink is used for the



http://genolist.pasteur.fr/Leproma/index.html

© Copyright Institut Pasteur, 1999

Region 20 kb

From kb

Drawing from 474.4 kb to 494.4 kb (18052 bp)

(region from 475 kb to 493 kb)

Go to Options (Navigate, Display, Export)

Coding sequences:
virulence, detox, adapt
information pathway
stable RNA
intermediary metabolism
conserved
lipid metabolism
cell wall, process
IS/phage
unknown
pseudogenes
PE/PPE
regulatory
rRNA
▼ tRNA

This image is clickable if your client supports HTML 3.0

Change current coordinates: from 475.773 kb to 493.824 kb

=> Rescale with 5000 bp/line <=

=> Gene List <=

● View
○ Download

DNA region (direct)

in Unix format

=> Get Data <=

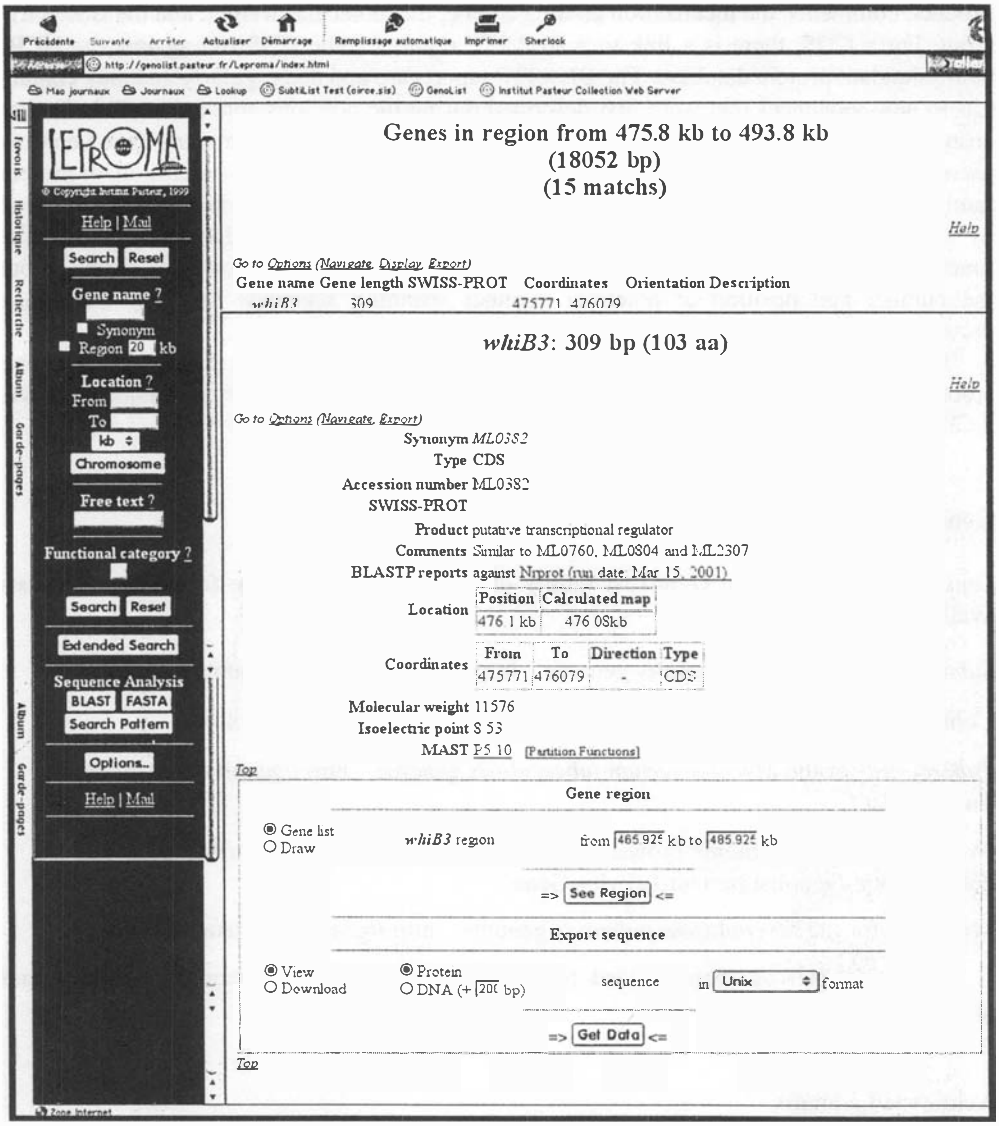
Visit the following links to learn more about this server:

- About Leproma

Figure 3. *Leproma* region drawing (for colour version of this figure see also accompanying colour poster).

many pseudogenes in *M. leprae*. The caption at the bottom of the drawing gives the colour code for each category of classification. RNA is represented by a black vertical arrow. Each graphic symbol and its name is clickable and displays detailed annotation in the lower right frame of the window.

As with the region gene list (Figure 2), the user may shift the region to the left, expand the region, contract the region, or shift the region to the right. The user may change the base-pair



Genes in region from 475.8 kb to 493.8 kb
(18052 bp)
(15 matchs)

Help

Go to Options (Navigate, Display, Export)

Gene name	Gene length	SWISS-PROT	Coordinates	Orientation	Description
whiB3	309		475771 476079		

whiB3: 309 bp (103 aa)

Help

Go to Options (Navigate, Export)

Synonym *ML0382*
Type CDS
Accession number ML0382
SWISS-PROT
Product putative transcriptional regulator
Comments Similar to ML0760, ML0804 and ML2307
BLASTP reports against Nrprot (run date: Mar 15, 2001)

Location

Position	Calculated map
476.1 kb	476.08kb

Coordinates

From	To	Direction	Type
475771	476079	-	CDS

Molecular weight 11576
Isoelectric point 8.53
MAST P5 10 [Partition Functions]

Top

Gene region

● Gene list
○ Draw
whiB3 region from 465.925 kb to 485.925 kb

=> See Region <=

Export sequence

● View
○ Download
● Protein
○ DNA (+ 200 bp)
sequence in Unix format

=> Get Data <=

Top

Figure 4. *Leproma* detailed gene annotation (for colour version of this figure see also accompanying colour poster).

density per line and change to a list display of the region. The user may export the genes or the DNA sequence in the region as for the gene list of a region.

Leproma detailed gene annotation

The gene detail display (Figure 4; see also accompanying colour poster) shows all annotations available for the selected CDS or RNA. These annotations include the size in base-pairs and in amino acids, for a CDS, the synonyms (if any), the accession number, the product, comments, the localization on the genome, the molecular weight, and the isoelectric point. For a CDS, there is a link to a BLASTP report for the CDS run against the NCBI non-redundant protein database. The BLASTP reports are updated regularly and may contain hits to new sequences that were first described during the previous annotation. This facility ensures that new clues to function are made available and we recommend that Leproma users check them regularly. Information about membership of a CDS to a partition, or protein family, can be found by clicking the MAST link as this provides graphic output generated by the MEME[6] and MAST[7] programs. A table of partition functions summarizes the functional information about all the families and the TMHMM[8] link provides details about the number and position of potential member spanning segments in likely membrane proteins.

The user has the option to display in a list (Figure 2) or a drawing (Figure 3) the region around the gene or export the gene sequence as a DNA sequence or a protein sequence if it is a CDS.

GenoList genome browsers

Leproma is a part of the **GenoList** family of genome browsers. The following table lists available **GenoList** browsers:

SubtiList for the *Bacillus subtilis* genome http://genolist.pasteur.fr.subtilList

Colibri for the *Escherichia coli* genome http://genolist.pasteur.fr/Colibri

TubercuList for the *Mycobacterium tuberculosis* genome http://genolist.pasteur.fr/TubercuList

Pylorigene, a multi-genome browser for the two *Helicobacter pylori strains* J99 and 26695 http://genolist.pasteur.fr/PyloriGene

MypuList, for the *Mycoplasma pulmonis* genome http://genolist.pasteur.fr/MypuList

Other genomes are being added regularly. An updated status can be seen at http://genolist.pasteur.fr

Acknowledgements

We are grateful to Fredj Tekaia for his help with data analysis. This work was supported by the Association Française Raoul Follereau, the Wellcome Trust, and the Institut Pasteur.

References

[1] Cole ST, Eiglemeier K, Parkhill J *et al*. Massive gene decay in the leprosy bacillus. *Nature*, 2001; **409:** 1007–1011.
[2] Altschul S, Gish W, Miller W *et al*. A basic local alignment search tool. *J Mol Biol*, 1990; **215:** 403–410.
[3] Gish W, States D. Identification of protein coding regions by database similarity search. *Nature Genet*, 1993; **3:** 266–272.
[4] Pearson W, Lipman D. Improved tools for biological sequence comparisons. *Proc Natl Acad Sci USA*, 1988; **85:** 2444–2448.

[5] Cole ST, Brosch R, Parkhill J *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature (Lond)*, 1998; **393:** 537–544.

[6] Bailey T, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Second International Conference on Intelligent Systems for Molecular Biology*, 1994, 28–36.

[7] Bailey T, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 1998; **14:** 45–48.

[8] Sonnhammer ELL, von Heijne C, Kmogh A. A hidden markov model for predicting transmembrane helices in protein sequences. *Sixth International Conference on Intelligent Systems.*