BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 10 2005, pages 2563–2565
doi:10.1093/bioinformatics/bti321

Databases and ontologies

MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*

Pierre-Étienne Jacques[1,2], Alain L. Gervais[1,2], Mathieu Cantin[2], Jean-François Lucier[3], Guillaume Dallaire[2], Geneviève Drouin[1], Luc Gaudreau[1], Jean Goulet[2] and Ryszard Brzezinski[1,*]

[1]Département de Biologie, [2]Département d'Informatique and [3]Département de Microbiologie et Infectiologie, Université de Sherbrooke, Sherbrooke, J1K 2R1, Quebec, Canada

Received on December 27, 2004; revised on February 4, 2005; accepted on February 9, 2005
Advance Access publication February 18, 2005

ABSTRACT

Summary: MtbRegList is a database dedicated to the analysis of gene expression and regulation data in *Mycobacterium tuberculosis*. It is designed to contain predicted and characterized regulatory DNA motifs cross-referenced with corresponding transcription factor(s), and experimentally identified transcription start sites. MtbRegList can also handle flexible and complex genomic search requests, besides having a noteworthy browsing capability.

Availability: MtbRegList is freely available at http://www.USherbrooke.ca/vers/MtbRegList

Contact: Ryszard.Brzezinski@USherbrooke.ca

Supplementary information: On the MtbRegList website.

Environmental adaptation is essential to the establishment of successful infection by bacterial pathogens. In most cases, adaptation greatly depends on transcriptional regulation. In prokaryotes, promoter recognition is effected by a sigma factor that binds the catalytic core RNA polymerase and initiates transcription (Gruber and Gross, 2003). *Mycobacterium tuberculosis* (Mtb), the causal agent of tuberculosis, has one principal and 12 alternative sigma factors (Cole *et al*., 1998), all with potentially different promoter-recognition properties. It is presumed that each sigma factor controls diverse sets of genes in response to specific environmental conditions. Furthermore, at least 200 regulatory proteins also modulate gene expression profiles in *Mtb*.

Despite the use of the BCG vaccine and antibiotic treatments, it is currently estimated that Mtb infects one-third of the human population, killing nearly 2 million people each year (Frieden *et al*., 2003). Since the publication of its entire genome sequence (Cole *et al*., 1998), data regarding gene expression regulation in Mtb has accumulated at a significant rate. The identification and understanding of gene regulatory networks could lead to the discovery of new drugs and better vaccines. Compiling transcription start sites (tss), promoters and transcription factor binding sites is the first step to rationally initiating this approach. Some compilation studies already exist for other micro-organisms (Makita *et al*., 2004; Salgado *et al*., 2004). However, no database is currently available to structure our knowledge of gene regulation in this threatening pathogen. In order to facilitate access to this data, we have organized it in a relational database complete with a web-based search form and a graphical visualization tool. MtbRegList is freely accessible at http://www.USherbrooke.ca/vers/MtbRegList.

Genome annotations found in MtbRegList have been extracted from the frequently updated TubercuList database (http://genolist.pasteur.fr/TubercuList/) (Camus *et al*., 2002). Where applicable, we provide hyperlinks to COG numbers from the GenBank annotation (Tatusov *et al*., 2003), gene information from TubercuList, protein information from the Mtb Structural Genomics Consortium and relevant references from PubMed (National Library of Medicine). The first release of MtbRegList contains 315 annotated DNA motifs (72 tss, 119 promoters, 121 transcription factor binding sites and 3 terminators) described in 56 research papers. Among annotated DNA motifs, some are predicted or deduced sites (14), and some are experimentally derived from mycobacteria other than Mtb (24). The database will be kept up-to-date in future releases with data from the literature. Researchers requiring access to the most recent features and data can access the development version following a dedicated link provided in the main page of our database.

MtbRegList has two main search pages. The first allows requests on genome information while the second focuses on annotated DNA motifs. Genome information is categorized as intergenic regions (IRs), genes and gene products (proteins and untranslated RNAs). A functional status is given to each IR based on the orientation of its flanking genes. If both genes have the same orientation, the 'IR status' is Promoter (P if genes are encoded on the positive strand, Q if encoded on the negative strand); Divergent (D) where they are in opposite orientations; and Terminator (T) in the case of opposite convergent orientations. The unique gene identifiers, and gene products, are identical to those found in TubercuList. However, the name of an IR is composed of 'ir' followed by a letter representing its status (P, Q, D or T), and by the upstream and downstream gene identifier (e.g. irD_Rv1737c_Rv1738). Typical queries regarding genome information search pages could be to list all divergent and promoter IRs >55 bp or to identify all proteins from the 'Regulatory' functional category with a molecular mass >32 kDa.

The second main search page regards annotated motifs (including tss, promoters and transcription factor binding sites). Motifs are classified as either 'Root pattern' or 'DNA motif'. Each annotated DNA motif (characterized or predicted) is derived from a Root pattern. Root patterns, akin to consensus sequences, are obtained from

*To whom correspondence should be addressed.

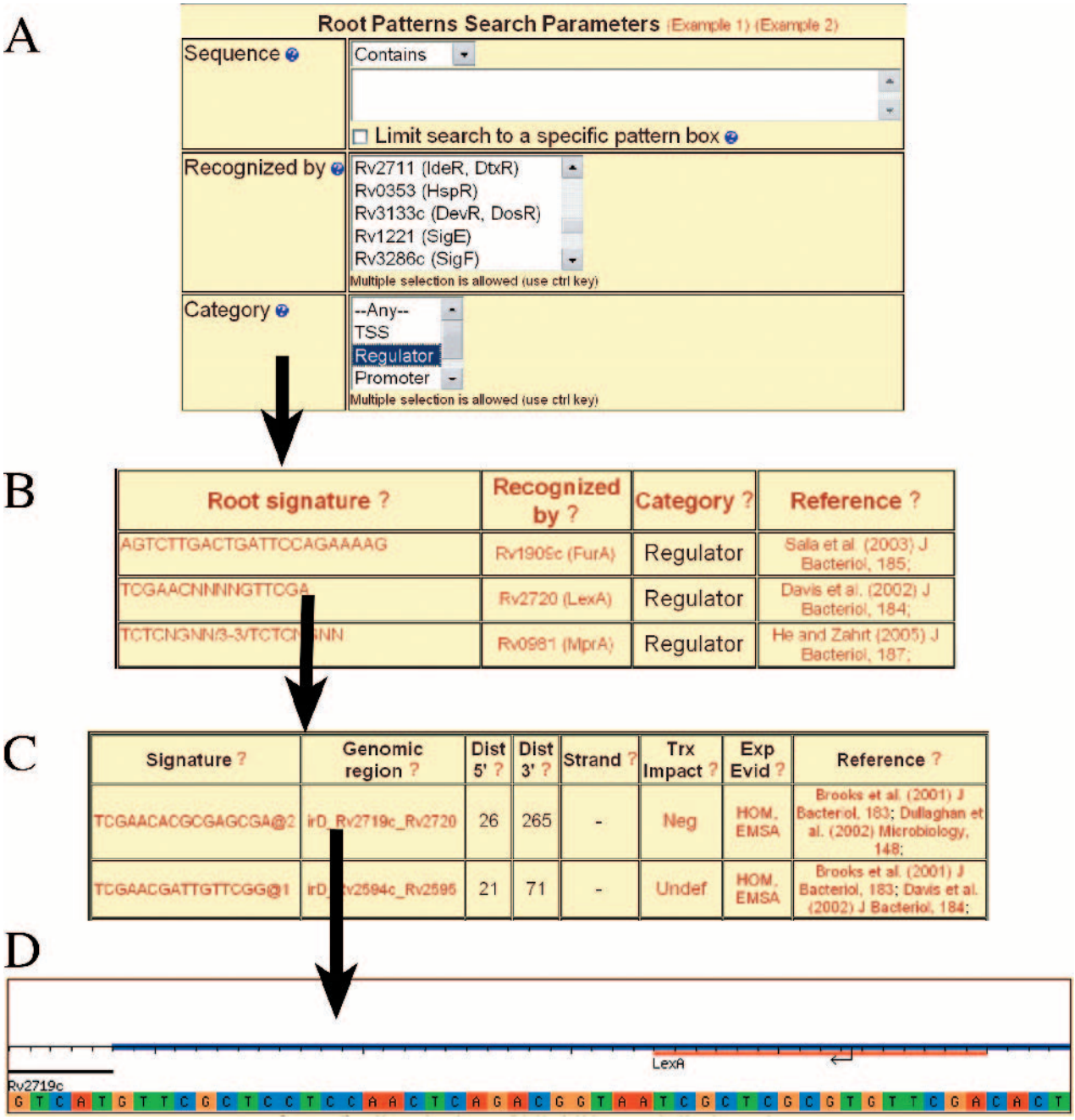

Fig. 1. An example of the browsing capability of MtbRegList. From the Root pattern search page **(A)**, list all Root of the 'Regulator' category, click on the signature of the LexA Root pattern **(B)** to access all annotated DNA motifs recognized by this regulatory protein **(C)** and visualize the IR irD_Rv2719c_Rv2720 that possesses a LexA annotated DNA motif **(D)**.

the literature or derived from experimentally identified DNA motifs. We also use the concept of 'signature', which allows users to view the information concerning several motifs at a glance. A Root pattern is represented by a signature such as S[/minSP – maxSP/S] where S stands for sequence box, SP for spacing, and what is inside brackets can be omitted or repeated depending on the number of pattern boxes. An annotated DNA motif is similarly represented by a signature such as S@m[/SP/S@m] where m stands for mismatch count relative to its respective Root pattern. For example, the Root pattern 'GGRAAC/15-19/SGTTG' (R = A or G; S = C or G) could have the following associated DNA motif instances: GGGA*T*C@1/16/CGTTG@0 and GG*T*AAC@1/18/*A*GTTG@1. A Root pattern is mainly described by its signature, a category, the regulatory protein that recognizes it and relevant references from the literature. Categories of Root patterns include promoters, regulators, terminators and tss. A DNA motif is described by features such as its

signature, the associated Root signature, the genomic region where it is found, the strand, 5′ and 3′ distances, experimental evidences and relevant references from the literature. The 'experimental evidence' field content distinguishes experimentally verified motifs from the predicted ones, as does the graphical view. A typical query could be to list all annotated DNA motifs formally identified by DNA-binding analysis or site-directed mutagenesis. The Root pattern and DNA motif search pages are dedicated exclusively to annotated motifs.

For each search page, the result columns displayed can be customized. Moreover, the results are displayed in tables that allow users to sort them according to any attribute. Hyperlinks are provided to navigate from results to definitions or associated components (for example regulatory proteins or genomic regions). Results can be saved in XML or tab delimited text format. An example of MtbRegList's browsing capability is shown in Figure 1, which simulates a sequence of operations from the Root pattern search page to the visualization of a particular IR.

In addition, it is possible to graphically navigate the genome. One can choose to view a genomic region according to its coordinates, or centered on a gene of interest. It is possible to zoom in or out of the generated view. Nucleotide sequences are shown whenever they would be readable. The view is 'clickable' and linked to relevant records of MtbRegList.

MtbRegList runs on an open source web platform (LAMP: Linux, Apache, MySQL, PHP). The design adopted for this database will allow, in subsequent versions, the support of large scale studies of DNA microarrays. For instance, it will allow users to establish links between a subset of differentially expressed genes and a particular annotated DNA motif (characterized or predicted). Moreover, the framework developed can be adapted for use with other completely sequenced bacterial genomes with minimal modifications.

ACKNOWLEDGEMENTS

The authors would like to thank Sébastien Rodrigue, Benoît Leblanc, Jean-François Jacques and Daniel Lafontaine for their valuable comments on the database and the manuscript. This work was supported by NSERC grants to R.B., L.G. and J.G. L.G. holds a Canada Research Chair on mechanisms of gene transcription, and is a Chercheur boursier Junior II of the FRSQ. P.-É.J. holds a FQRNT PhD scholarship.

REFERENCES

Camus,J.C. *et al*. (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, **148**, 2967–2973.

Cole,S.T. *et al*. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.

Frieden,T.R. *et al*. (2003) Tuberculosis. *Lancet*, **362**, 887–899.

Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space, *Ann. Rev. Microbiol.*, **57**, 441–466.

Makita,Y. *et al*. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32** (Database issue), D75–D77.

Salgado,H. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32** (Database issue), D303–D306.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.