MycoRRdb: A Database of Computationally Identified Regulatory Regions within Intergenic Sequences in Mycobacterial Genomes

Mohit Midha, Nirmal K. Prasad, Vaibhav Vindal*
Department of Biotechnology, School of Life Sciences, University of Hyderabad, Hyderabad, India

Abstract

The identification of regulatory regions for a gene is an important step towards deciphering the gene regulation. Regulatory regions tend to be conserved under evolution that facilitates the application of comparative genomics to identify such regions. The present study is an attempt to make use of this attribute to identify regulatory regions in the Mycobacterium species followed by the development of a database, MycoRRdb. It consist the regulatory regions identified within the intergenic distances of 25 mycobacterial species. MycoRRdb allows to retrieve the identified intergenic regulatory elements in the mycobacterial genomes. In addition to the predicted motifs, it also allows user to retrieve the Reciprocal Best BLAST Hits across the mycobacterial genomes. It is a useful resource to understand the transcriptional regulatory mechanism of mycobacterial species. This database is first of its kind which specifically addresses cis-regulatory regions and also comprehensive to the mycobacterial species. Database URL: http://mycorrdb.uohbif.in.

Introduction

Over the past few years the genomic sequence repertoire of mycobacterial sequences has increased tremendously. The availability of complete genome sequences makes it possible to efficiently employ computational approaches to understand the genome function and its complexity [1]. One of the important aspects to compare genome sequences is to find orthologous proteins among the existing species [2,3]. The identification of orthology is important not only to assist the functional annotation of a gene but also to identify its regulatory region. These regions are known to evolve at a slower rate than non-functional elements, and therefore finding the conserved DNA motifs within non-coding region is an efficient method to predict these regions [4,5]. Different approaches have been used to find the regulatory regions [6–8]. Generally, identification of these DNA elements relies on an extensive set of known target genes [4,9]. Therefore, identification of regulatory region for a novel transcriptional regulator remains a challenging task.

Extensive research on mycobacteria has produced a number of online resources, providing information on pathogenicity, cellular physiology, operon arrangement, microarray, etc. [10–15]. These resources also include a database, MtbRegList, which contains the reported regulatory regions in Mycobacterium tuberculosis [16]. Nevertheless, there is still need to document the putative regulatory regions for all the mycobacterial genomes. Our present study addresses this issue, as it identifies the putative cis-regulatory sequences within the intergenic regions of mycobacterial species and also the similar DNA motif in a genome. In addition to the predicted regulatory regions, the database includes list of Reciprocal Best BLAST Hits (RBBHs) for all 25 mycobacterial species. The database also has a search feature to identify the sequences similar to a query DNA motif. This database can assist in the characterization of gene regulation in all the mycobacterial species.

Methods

Retrieval and filtering the genome sequences

The complete genome sequences of 25 Mycobacterium species were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Some of the proteins were found to be present in more than one copy, identical in sequence, in certain species. In present study, such Multiple Identical Proteins (MIPs) were identified and replaced with only one representative protein sequence for further analysis.

Identification of orthologs

Reciprocal Best BLAST Hit (RBBH) method was used to predict orthologous proteins in mycobacterial proteomes. Pairs of proteins, from two mycobacterial species, covering the at least 50% sequence length of both the proteins in alignment and E-values lower than of 10−20 for both directions using BLASTP program with all other parameters at default values were selected as RBBHs [17–19].

Retrieval of operons and the intergenic sequences

Information for all mycobacterial operons in genomes of all 25 species were retrieved from the DOOR database (version2) [20,21] Intergenic sequence upstream of the first gene of each
Figure 1. Flowchart of the methodology.  
doi:10.1371/journal.pone.0036094.g001

Figure 2. Schematic architecture of MycoRRdb.  
doi:10.1371/journal.pone.0036094.g002
operon was retrieved using perl script. Sets of intergenic sequences were compiled for each orthologous gene. These sets of sequences were further subjected to the identification of a regulatory region.

Identification of regulatory regions

MEME suite was used to identify the conserved regulatory DNA elements from the set of sequences described earlier [4]. The DNA motif length, from minimum of 20 bases to maximum length of 30, was optimized using known DNA targets from *M. tuberculosis* [16]. DNA search was carried out to look for palindromes within the given strand as well as its complementary strand. Additionally, the top predicted DNA motifs observed associated with three or more orthologous sequences were selected as potential regulatory DNA element. All other parameters were kept on its default values. These DNA motifs were further searched in their respective genomes to identify the significantly similar motifs with minimum aligned length (L) of 16 bases (allowing N mismatch where N <= 0.2L; L-N > 14).

Database development

Subsequent to the identification of regulatory regions from all mycobacterial genomes, a web resource, MycoRRdb was developed. This database has been developed using MySQL. It is constructed to allow user to browse the outcome of study in an easy accessible mode. Web interface of the database is designed using PHP, HTML and Javascripts. Flow chart of the methodology followed in the study is depicted in Figure 1.

Results and Discussion

RBBHs across the mycobacterial species

The ortholog prediction is not only important to identify the regulatory region but also helps in functional annotation of a
sequenced genome. Our study also began with the identification of the RBBHs which serves as potential ortholog. All the RBBHs from the mycobacterial species were identified using the methodology discussed. The identified lists of RBBHs for any mycobacterial gene across all 25 mycobacterial genome were used as a data source for the MycoRRdb.

Mycobacterial regulatory regions

Subsequently, DNA regulatory regions were identified across the all 25 Mycobacterium species. The total predicted regulatory motifs were 37101 in number for all 25 mycobacterial genomes. Further, the motifs predicted across the Mycobacterial species were compared with the known DNA motifs reported in the literature [3,5,16,22–41]. It was observed that 116 DNA motifs, out of 181 retrieved, were mapped in MycoRRdb and notified through the link given in the database. The comparative list of the predicted and the reported DNA motifs is given in Table S1. The maximum number of motifs was predicted from Mycobacterium tuberculosis H37Ra while the minimum number was from Mycobacterium abscessus ATCC 19977. These predicted DNA motifs are the putative Transcription Factor Binding Sites (TFBS). The TFBS identified, positioned at more than 400 nucleotide upstream to the translational start site, are highlighted with red colour font. Further in view of over representation, similar DNA motifs were searched to find the similar motifs within the predicted list of intergenic regulatory region. All the identified motifs are displayed with the strand information and the position from translational start site.

Database access

MycoRRdb can be accessed through the database web interface at http://mycorrdb.uohbif.in. There are two kind of data that has been stored in MycoRRdb: (i) Reciprocal Best BLAST Hits (RBBHs), and (ii) Predicted Regulatory Region for each transcription unit (Figure 2). This information for any mycobacterial gene can be retrieved from MycoRRdb in either browsable or searchable fashion. Homepage of the database provides links for the mycobacterial genome which further leads to complete list of genes/protein id/ORF id of a particular species. From the list one can proceed to find the RBBHs of any gene across other mycobacterial species and associated regulatory DNA motifs along with its occurrence in the orthologous intergenic sequences. It also gives link, to facilitate user, to the retrieve the known motif reported in literature. In addition to this list of similar DNA motifs in a genome is also available (Figure 3).
Besides browsing data from complete genes list, separate links have also been made available on web interface to quickly retrieve RBBHs or regulatory DNA motifs by gene name/protein id/ORF id. A searchable interface to retrieve RBBHs is shown in Figure 4A. The predicted regulatory regions and the similar sequences present in that genome can be also be retrieved by searchable interface using gene name/protein id/ORF id (Figure 4B). Moreover, user can scan the availability of its desired DNA sequence, if it exists in any Mycobacterial species, in the identified DNA motifs set of the Database (Figure 4C).

This database is under constant development to gather the experimentally validated DNA motifs to incorporate in the database. It also provides link for biologist to put forward the experimentally validated mycobacterial regulatory regions, if any.

Conclusions

The availability of whole genome sequences makes Mycobacterium one of the highly sequenced genera. This wealth of sequence data provides unique opportunity to extract the genome information in order to address cellular physiology and to develop better intervention strategies for pathogenic species. This study is a systematic approach to reveal the putative regulatory regions and RBBHs across the mycobacterial species. On the one hand, the identified regulatory regions will help to understand the transcriptional regulation of the mycobacterial genes, and on the other hand, the identified RBBHs will assist to impart the functional knowledge of one gene to another. The availability of all the identified regulatory regions and RBBHs from the mycobacterial species at a websource, MycoRRdb, will help to access the data and will have potential implications to unravel the genomic complexity of the mycobacteria.

Supporting Information

Table S1 DNA motifs in MycoRRdb mapped with regulatory regions reported in literature. (Available at: http://mycorrdb.uohbif.in/links.php).

(XLS)

Acknowledgments

Research in VV’s laboratory is supported under RGYI scheme from Department of Biotechnology, Government of India. VV acknowledges the Bioinformatics Infrastructure Facility, University of Hyderabad, to host the database.

Author Contributions

Conceived and designed the experiments: VV. Performed the experiments: MM. Analyzed the data: MM NKP VV. Wrote the paper: VV. Database design: MM.

References

1. Young DB (2001) A post-genomic perspective. Nature medicine 7: 11–13.
2. Lee SA, Chao CH, Tsai CH, Lai JM, Wang JS, et al. (2000) Orthology-based protein–protein interaction prediction and its application to inter-species interactions. BMC bioinformatics 9 Suppl 12: S11.
3. Vindal V, Ashwantha Kumar E, Ranjan A (2008) Identification of operator sites within the upstream region of the putative μrrR gene from mycobacteria. FEMS letters 582: 1117–1122.
4. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research 34: W369–W373.
5. Vindal V, Suma K, Ranjan A (2007) GntR family of regulators in Mycobacterium tuberculosis. Bioinformatics 23: e67.
6. Thiery D, Salgado H, Huerta AM, Collado-Vides J (1998) Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K12. Bioinformatics 14: 391–400.
7. Bulyk ML, McGuire AM, Masuda N, Church GM (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. Genome research 14: 201–208.
8. Siddharthan R, Siggia ED, van Nimwegen E (2003) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. PLoS computational biology 1: e1.
9. Ranjan S, Seshadri J, Vindal V, Yellaboina S, Ranjan A (2006) iCR: a web tool to identify conserved targets of a regulatory protein across the multiple related prokaryotic species. Nucleic acids research 34: W350–W357.
10. Bergh S, Cole ST (1994) MyDB: an integrated mycobacterial database. Molecular microbiology 12: 517–534.
11. Catanho M, Mascarenhas D, Degrave W, Miranda AB (2006) GenoMycDB: a database for computational biology and molecular research: GMR 5: 115–126.
12. Rand L, Hinds J, Springer B, Sander P, Buxton RS, et al. (2003) The majority of inducible DNA repair genes in Mycobacterium tuberculosis are induced independently of RecA. Mol Microbiol 50: 1031–1042.
13. Davis EO, Duggalhan EM, Rand L (2002) Definition of the mycobacterial SOS box and use to identify LexA-regulated genes in Mycobacterium tuberculosis. J Bacteriol 184: 3267–3275.
14. Brooks PC, Movahedzadeh F, Davis EO (2001) Identification of some DNA damage-inducible genes of Mycobacterium tuberculosis: apparent lack of correlation with LexA binding. J Bacteriol 183: 4659–4667.
15. Movahedzadeh F, Colston MJ, Davis EO (1997) Determination of DNA sequences required for regulated Mycobacterium tuberculosis RecA expression in response to DNA-damaging agents suggests that two modes of regulation exist. J Bacteriol 179: 3309–3318.
16. Palaksh P, Yellaboina S, Ranjan A, Hasnain SE (2005) Computational prediction and experimental verification of novel IdDR binding sites in the upstream sequences of Mycobacterium tuberculosis open reading frames. Bioinformatics 21: 2161–2166.
17. Rodriguez GM, Gold B, Gomez M, Dusserge O, Smith I (1999) Identification and characterization of two divergently transcribed iron regulated genes in Mycobacterium tuberculosis. Tuberc Lung Dis 79: 287–298.
18. Gold B, Rodriguez GM, Marras NA, Prentecost M, Smith I (2001) The Mycobacterium tuberculosis IdrR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. Mol Microbiol 42: 851–865.
19. Rodriguez GM, Vonkil ML, Gold B, Schoenick GR, Smith I (2002) idrR, an essential gene in Mycobacterium tuberculosis: role of IdrR in iron-dependent gene expression, iron metabolism, and oxidative stress response. Infect Immun 70: 3571–3581.
20. Hoyles LB, Benjamin WH, Jr., Dunlap NE, Clark-Carrino JE (2002) Expression, autoregulation, and DNA binding properties of the Mycobacterium tuberculosis TreR response regulator. J Bacteriol 184: 2192–2202.
21. Cavet JS, Meng W, Femmer MA, Appelhoff RJ, Giedroc DP, et al. (2002) A nickel-cobalt-sensing ArcR/SnbR family repressor. Contributions of cytotoxic and effector binding sites to metal selectivity. J Biol Chem 277: 38414–38448.
22. Stewart GR, Wernisch L, Stabler R, Mangan JA, Hinds J, et al. (2002) Dissection of the heat-shock response in Mycobacterium tuberculosis using mutants and microarrays. Microbiology 148: 3129–3138.
23. Duggalhan EM, Brooks PC, Davis EO (2002) The role of multiple SOS boxes upstream of the Mycobacterium tuberculosis lexA gene identification of a novel DNA-damage-inducible gene. Microbiology 148: 3609–3615.
34. Recchi C, Selvi R, Rauzier J, Gicquel B, Reyrat JM (2003) Mycobacterium tuberculosis Rv1395 is a class III transcriptional regulator of the AraC family involved in cytochrome P450 regulation. J Biol Chem 278: 33763–33773.

35. Sala C, Fori F, Di Florio E, Canneva F, Milano A, et al. (2003) Mycobacterium tuberculosis FurA autoregulates its own expression. J Bacteriol 185: 5357–5362.

36. Engohang-Ndong J, Baillet D, Aumercier M, Bellefontaine F, Bera G, et al. (2004) EthR, a repressor of the TetR/CamR family implicated in ethionamide resistance in mycobacteria, octamerizes cooperatively on its operator. Mol Microbiol 51: 173–188.

37. He H, Zahri TC (2005) Identification and characterization of a regulatory sequence recognized by Mycobacterium tuberculosis persistence regulator MprA. J Bacteriol 187: 202–212.

38. Rickman L, Scott C, Hunt DM, Hutchinson T, Menendez MC, et al. (2005) A member of the cAMP receptor protein family of transcription regulators in Mycobacterium tuberculosis is required for virulence in mice and controls transcription of the rpfA gene coding for a resuscitation promoting factor. Mol Microbiol 56: 1274–1286.

39. Kendall SL, Burgess P, Balhana R, Withers M, Ten Bokum A, et al. (2010) Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: katR and katR2. Microbiology 156: 1362–1371.

40. Kendall SL, Withers M, Sofiär CN, Moreland NJ, Gursha S, et al. (2007) A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in Mycobacterium smegmatis and Mycobacterium tuberculosis. Mol Microbiol 65: 684–699.

41. Festa RA, Jones MB, Butler-Wu S, Sinsimer D, Gerads R, et al. (2011) A novel copper-responsive regulon in Mycobacterium tuberculosis. Mol Microbiol 79: 133–148.