Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Method

Mining Genomic Patterns in *Mycobacterium tuberculosis* H37Rv Using a Web Server Tuber-Gene

Lavanya Rishishwar[1*], Bhasker Pant[2], Kumud Pant[3], and Kamal R. Pardasani[4]

[1]*School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332, USA;*
[2]*Department of Information Technology, Graphic Era University, Dehradun, Uttrakhand 248002, India;*
[3]*Department of Bioinformatics, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh 462051, India;*
[4]*Department of Mathematics, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh 462051, India.*

Genomics Proteomics Bioinformatics 2011 Oct; 9(4-5): 171-178 DOI: 10.1016/S1672-0229(11)60020-X

Received: Feb 03, 2011; Accepted: Sep 01, 2011

Abstract

Mycobacterium tuberculosis (MTB), causative agent of tuberculosis, is one of the most dreaded diseases of the century. It has long been studied by researchers throughout the world using various wet-lab and dry-lab techniques. In this study, we focus on mining useful patterns at genomic level that can be applied for *in silico* functional characterization of genes from the MTB complex. The model developed on the basis of the patterns found in this study can correctly identify 99.77% of the input genes from the genome of MTB strain H37Rv. The model was tested against four other MTB strains and the homologue *M. bovis* to further evaluate its generalization capability. The mean prediction accuracy was 85.76%. It was also observed that the GC content remained fairly constant throughout the genome, implicating the absence of any pathogenicity island transferred from other organisms. This study reveals that dinucleotide composition is an efficient functional class discriminator for MTB complex. To facilitate the application of this model, a web server Tuber-Gene has been developed, which can be freely accessed at http://www.bifmanit.org/tb2/.

Key words: dinucleotide composition, *Mycobacterium tuberculosis*, support vector machine, data mining

Introduction

Today, an estimated one-third of the world's population is infected with *Mycobacterium tuberculosis* (MTB), the causative agent of tuberculosis, which causes nearly 2 million deaths annually (*1, 2*). WHO estimates that over 13 million people are suffering from tuberculosis (*3*). However, to date, many aspects of the interactions between MTB and its human host remain occult. The bacterium not only evades the defences of the host's immune system, but also stays in the host's body for years and may reactivate, causing the disease decades after infection. Hence it is not hard to imagine the significance of understanding the pathogen at genomic level in order to improve or develop treatment strategies.

The MTB complex comprises of four species, including *M. tuberculosis*, *M. bovis*, *M. microti*, and *M. africanum*. These four species share high similarity in DNA sequence, which are even completely conserved in several gene regions (16S rDNA, the 16S-to-23S rDNA internal transcribed spacer, and DNAJ family)

*Corresponding author.
E-mail: lavanya.rishishwar@gatech.edu

(*4-6*). Moreover, numerous strains of MTB are known to exist. This has limited the use of sequence analysis for strain differentiation. Out of these strains, H37 variants are widely used as reference strains in mycobacteriology and molecular biology studies. The genotyping of MTB, primarily for outbreak identification, has become a model for the application of strain typing in the field of molecular epidemiology (*7*).

Functionally characterizing genes through experimental techniques roughly takes a year. The development of computational models for characterizing genes will complement the existing experimental technologies and provide a much quicker, less expensive way to molecular biologists and other scientists working towards alleviating the sufferings of mankind.

Proteomic analysis has been conventionally used for the classification and understanding of organisms and their physiology (*8-10*). A number of attempts are reported in the literature to classify protein sequences in various organisms based on amino acid composition (*11, 12*). However, to our knowledge, no attempt has been reported to classify nucleotide sequences based on genomic information of MTB. Support vector machine (SVM) is widely used in pattern recognition tasks over other classification algorithms. The seminal work pertaining to the application of SVM in solving biological problem was reported by Guyon *et al* (*13*).

In view of above, here an effort has been made to recognize patterns correlating the gene with the function of its protein product, employing SVM algorithm for classification and prediction. The genome information contained in the nucleotides and the higher-order composition has been used for evaluation of the organism. A web-based tool has also been developed, which can identify the function of 99.77% of the total genes of H37Rv, since H37Rv is extensively employed in biomedical research (*14*).

Methods

Sequence dataset

In the current study, 3,906 gene sequences of the MTB H37Rv strain were retrieved from the Tuberculosis Database (*15*) and arranged according to their functional classes (derived from www.sanger.ac.uk/Projects/M_tuberculosis/Gene_list). A complete list of all the gene accession numbers used in this study can be found in Table S1.

Functional hierarchy

The adopted classification is strictly hierarchical, constituting three levels of branching. The top two levels of the hierarchy with the gene distribution are shown in **Table 1**. According to the function that their products perform, genes are broadly divided into six categories: I. small-molecule metabolism; II. macromolecule metabolism; III. cell processes; IV. other; V. conserved hypotheticals and VI. unknowns.

Support vector machine

SVM is widely used for learning separating functions in pattern recognition tasks and in performing functional estimation in regression problems. Classification tasks usually involve training and testing data that consist of some data instances. Every training dataset constitutes of a class label and a set of attributes. The aim of SVM is to produce a model based on these training datasets, which are capable of correctly predicting the target label of an instance, given its attributes. To achieve this, the instances are first mapped to a higher (maybe infinite) dimension and then a linear separating hyper-plane with the maximal margin in this higher dimensional space is computed (*16*).

The function $K(x_i, x_j)$ represented by Equation 1 is called the kernel function:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (1)$$

It describes the nature of the separating hyper-plane that will be used for classification purpose. There are four types of basic kernels (*17-19*):

Linear: $K(x_i, x_j) = x_i^T x_j$ (2)

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (3)

Radial Basis Function (RBF):

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0 \quad (4)$$

Sigmoid or Hyperbolic:

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \quad (5)$$

Here, γ, *r* and *d* are kernel parameters.

Table 1 Top two levels of gene functional class hierarchy

Hierarchy order	Class name	No. of genes	Percentage distribution (%)
I.	**Small-molecule metabolism**	**1,066**	**27.29**
I.A.	Degradation	163	4.17
I.B.	Energy metabolism	292	7.48
I.C.	Central intermediary metabolism	45	1.15
I.D.	Amino acid biosynthesis	95	2.43
I.E.	Polyamine synthesis	1	0.03
I.F.	Purines, pyrimidines, nucleosides and nucleotides	60	1.54
I.G.	Biosynthesis of cofactors, prosthetic groups and carriers	117	3.00
I.H.	Lipid biosynthesis	65	1.66
I.I.	Polyketide and non-ribosomal peptide synthesis	41	1.05
I.J.	Broad regulatory functions	187	4.79
II.	**Macromolecule metabolism**	**653**	**16.72**
II.A.	Synthesis and modification of macromolecules	215	5.5
II.B.	Degradation of macromolecules	87	2.23
II.C.	Cell envelope	351	8.99
III.	**Cell processes**	**206**	**5.27**
III.A.	Transport/binding proteins	123	3.15
III.B.	Chaperones/heat shock	16	0.41
III.C.	Cell division	19	0.49
III.D.	Protein and peptide secretion	14	0.36
III.E.	Adaptations and atypical conditions	12	0.31
III.F.	Detoxification	22	0.56
IV.	**Other**	**463**	**11.85**
IV.A.	Virulence	38	0.97
IV.B.	IS elements, repeated sequences and phage	132	3.38
IV.C.	PE and PPE families	164	4.20
IV.D.	Antibiotic production and resistance	14	0.36
IV.E.	Bacteriocin-like proteins	3	0.08
IV.F.	Cytochrome P450 enzymes	22	0.56
IV.G.	Coenzyme F420-dependent enzymes	3	0.08
IV.H.	Miscellaneous transferases	61	1.56
IV.I.	Miscellaneous phosphatases, lyases, and hydrolases	18	0.46
IV.J.	Cyclases	6	0.15
IV.K.	Chelatases	2	0.05
V.	**Conserved hypotheticals**	**914**	**23.40**
VI.	**Unknowns**	**604**	**15.46**
Total		**3,906**	**100**

RBF kernel was selected for the training purpose since it can handle both linear and non-linear data efficiently with fewer numerical difficulties in contrast to polynomial kernel, in which kernel values may go to infinity or zero at large degrees (*20*). The parameter selection for RBF was done by running a grid analysis on each dataset using LIBSVM software (*19*).

Design of the prediction systems

In this study, correlations are mined by three different approaches to generate a system that is most competent of predicting the correct gene function for a given input gene sequence. The basic difference between the developed systems is the feature vectors used (**Table 2**). Since MTB belongs to the Kingdom Bacteria and bacterial genes lack intron regions, the compositions would, hence, be characterizing the operon (*21*) as well as coding region only.

In each case, the addressed problem becomes a multi-level multi-class classification problem. One strategy to handle this problem is to reduce it to a series of binary-class classification problems, *i.e.*, for the i^{th} sub-problem, all the instances belonging to the class i are labelled as positive samples and the instances belonging to other classes are treated as negative samples.

Each system is a model consisting of 119 model files representing each node of the hierarchy. **Figure 1** depicts how each system was developed.

To resolve the conflicts arising at different levels due to the presence of hierarchy, a systematic approach of decision making is adopted at each level. Rather than testing the input desultorily with the whole set of model files, the subset of model with which the input is to be tested is decided by the precedent level.

Assessment of the prediction systems

To assess the absolute prediction quality of the three systems, each gene of H37Rv was provided as an input to each system and the output was recorded. Afterwards, the predicted class was compared with the actual class to evaluate the accuracies of the prediction.

To estimate the generalization capacity, the best performing system was tested against four other strains of MTB (H37Ra, F11, C1, and CDC1551) and a homologue, *M. bovis*. The gene functional hierarchies for these organisms were first generated by running BLASTn for all the sequences from each strain

Table 2 The parameters used to form the vectors in the three systems in the current study

System	Composition employed	Vector length	Feature attributes
A	Mononucleotide	4	Composition of A, T, G, C
B	A+T and C+G	2	Composition of A+T and G+C
C	Dinucleotide	16	Composition of AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GG, GC, CA, CT, CG, CC

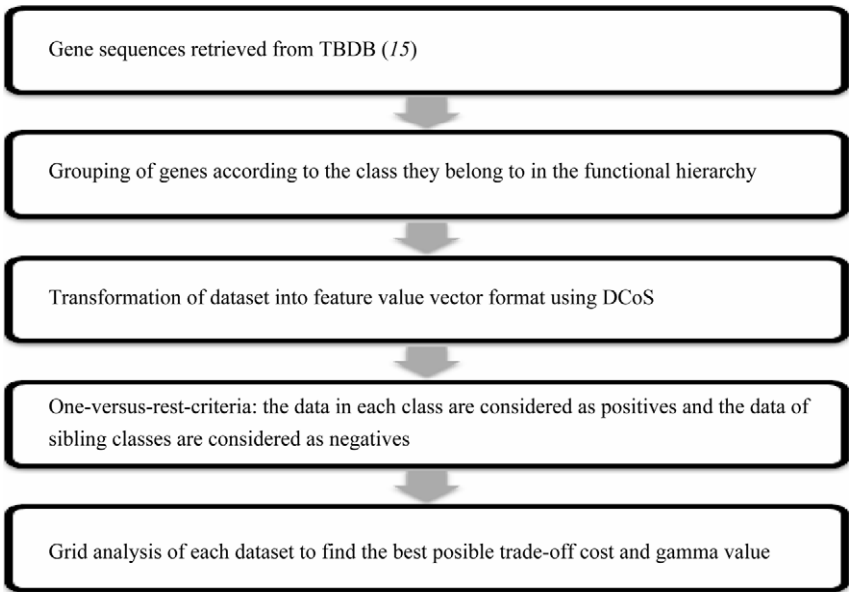

Figure 1 Flow chart describing the whole process. The steps involved in designing of the system A, B and C are depicted.

against the database of H37Rv. Thus, a comparison between BLAST (sequence-order dependent) and the concerned system (sequence-order independent) was also performed.

Results

Prediction accuracy

Misclassification rate was evaluated in two terms: (a) incorrect classification rate – percent of false positive instances and (b) failed classification rate – percent of instances that the system failed to predict completely at the concerned level.

The resulting accuracies are summarized in **Table 3**. SVM algorithm failed to find any separating hyper-planes in all the kernels at any level in System B, implicating that the dataset was inseparable. System A correctly predicted 97.49% of the input sequences at the top level of the hierarchy and with 97.19% and 75.49% accuracy at the following levels. Incorrect classification was 0.67%, and failed classification of 1.84% of the input genes was observed at level 1. The performance of System C was found to be most efficient, with an accuracy of 99.77% at level 1, 99.96% and 78.77% accuracy at subsequent levels. In addition, no incorrect classifications were present and only 0.23% (9 instances) failed classification was noticed at level 1.

The compression ratio of information in the output models was calculated as:

$$Compression\ Ratio = \frac{No.\ of\ SVs\ in\ the\ model}{No.\ of\ vectors\ in\ the\ input} \quad (6)$$

where SVs stand for support vectors, *i.e.*, the vectors in the output model. The compression ratio for system A and C was observed to be 0.991 and 0.876, respectively.

Based on these results, system C was preferred over the other two for further analysis. In the test with homologues, system C performed fairly well in predicting the function of their genes (Table 3).

Software tools and access

In order to carry out the study, we developed a web-based tool, DCoS (DNA COmposition Server), to assist calculating the composition of a given input nucleotide sequence. DCoS can be accessed for free at http://www.bifmanit.org/dcos/. Based on the current study, another free web-based server, Tuber-Gene, has been developed (Figure S1), which can be accessed at http://www.bifmanit.org/tb2/. Sequences of length >50 bp can be input in Fasta/Pearson format in the text area provided. Depending on the level of accuracy, the output will display the putative function of the input gene with the computed degree of confidence or will display "No function predicted" if it fails to do so. The degree of confidence is determined as the product of output class value at each level as computed by the SVM. The core SVM algorithm software is SVM Light developed by Joachims (*22*).

Table 3 Prediction accuracies of gene functions

Strain		No. of genes	Accuracy (%)			Not classified (%)	Misclassified (%)
			Level 1	Level 2	Level 3		
H37Rv	System A	3,906	97.49	97.19	75.49	1.84	0.67
	System B	3,906	0.00	0.00	0.00	100.00	0.00
	System C	3,906	99.77	99.96	78.77	0.23	0.00
H37Ra		3,960	95.48	95.83	75.59	4.52	0.00
F11		3,898	87.04	87.49	68.27	12.96	0.00
M. bovis BCG		3,910	73.81	67.52	50.86	26.19	0.00
C1		3,841	73.57	73.66	56.59	26.43	0.00
CDC1551		3,893	73.21	75.93	60.19	26.79	0.00

Note: Gene functions were predicted using various systems and the resulting accuracies at each level are shown. The model built by system C was further tested against other members of MTB complex and the homologue *M. bovis* to evaluate the capability of System C to generalize its prediction power.

Discussion

The purpose of the study was to discover the relationship between the composition of nucleotides within a gene and the gene function (System A). The results clearly depict that the nucleotide composition varies notably among the genes of H37Rv (**Figure 2**). System B was developed to verify any dependency that the gene function has with its CG content. This system highlighted the high and fairly constant percentage of CG as compared to AT, with a mean of 65.44% and 34.56%, respectively, supported by a standard deviation of 3.36 in both cases (**Figure 3**). These data suggest that no pathogenicity island of unusual composition exists (*23*), which are consistent with the findings of Cole *et al* (*14*).

Dinucleotides have long been used in practice as potential discriminators and for generating various probabilistic models such as the Markov Chains (*24*). Probably beginning with the discovery of nearest neighbor patterns in both prokaryotic and eukaryotic DNA sequences (*25*), dinucleotide composition (DNC) has found numerous applications. Nakashima *et al* (*26*) effectively separated genes from nine different genomes based on the DNC space. They noted that the DNC varied significantly from organism to organism and postulated that the distinct feature in the DNC may reflect the phylogenetic relationship of organisms. Later on, they succeeded in establishing a linear relationship between optimal growth temperature and DNC on the basis of regression analysis of the sequence data for thermophilic, mesophilic and psychrophilic bacteria (*27*).

A Markov chain for DNA is shown in **Figure 4** as explained previously (*28*). Each edge in the graph represents the probability of occurrence of a nucleotide following another nucleotide. DNC is a version of these probabilities with 100 times scaled up. Exercising DNC as a class differentiation factor is synonymous to comparing the characteristic of probabilistic models between classes. System C models these differences and produces the gene function as the output using a given gene sequence as an input. This model was able to correctly classify all except 9 cases, for which no functional class could be characterized. These cases were found to belong to class VI (Unknowns) and their biological function is yet not available/known. Apparently, DNC provides the highest degree of differentiation and thus can be used as signatures for each class. Furthermore, the top occurrence of transitions was C to G, followed by G to C and G to G (Figure S2).

Additionally, the decrease in the accuracy from level 1 to 3 was due to the less availability of data at lower levels. Based on the basic performance of classical techniques of modelling, it has been observed

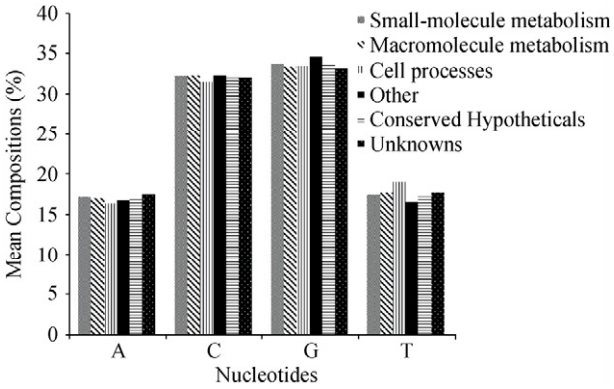

Figure 2 Compositional differences between the classes using System A. The clustered column graph indicates the mean composition of the four nucleotides in each of the six functional classes. Although the variation of each nucleotide among the classes is not significant, all the classes have markedly higher percentage (in the range of 30%-35% for each) of cytosine and guanine in their sequences. This CG-richness in the genome is the basis for system B.

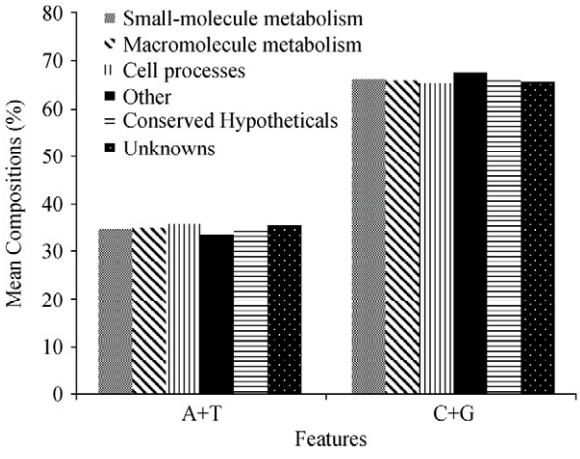

Figure 3 Compositional differences between the classes using System B. The clustered column graph indicates the mean composition of AT and CG for each of the six functional classes. A more stable composition can be observed as compared to that of System A, implying the absence of pathogenicity islands in the genome (*23*).

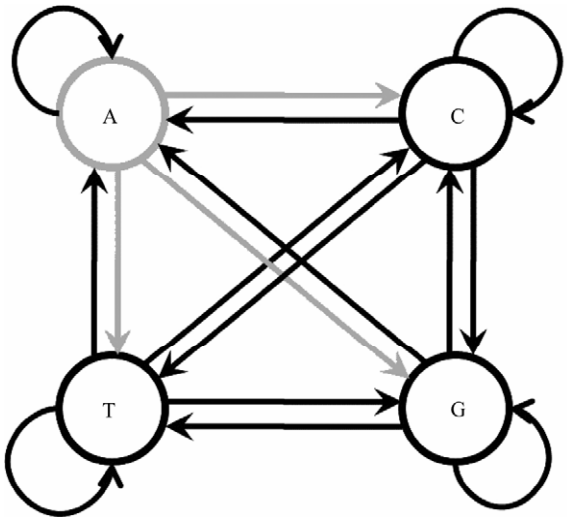

Figure 4 Markov chain model for DNA. Each edge in the graph represents the probability of occurrence of a nucleotide following another nucleotide. All the possible four transitions starting from A are shown in grey.

that the modelling error increases as the input data size decreases (*15*).

Conclusion

This study was aimed in analyzing any "sequence pattern – gene function" relationship in the genome of *M. tuberculosis* H37Rv strain. To perform the analysis, three systems (System A, B and C) were developed, differing by the type of sequence patterns used. System A, which was based on simple mononucleotide composition, clearly depicted that the nucleotide composition differed notably among the genes. System B was developed to verify any dependency of the gene function to its CG content. However, the results showed a nearly static composition throughout the genome. System C was based on the influence of a nucleotide base on the base following it, depicted by the DNC of genes. The results showed a noticeable variation in the pattern of DNC among the functional classes, which makes System C more efficient in differentiating genes belonging to different classes. It was established that the success rate of System A and C was 97.49% and 99.77%, respectively, while System B had failed completely concurring to the stable CG content throughout. The higher success rate in System C leads to a higher degree of confidence in the results obtained by System C. Thus, the study suggests that the relationship between DNC and gene function is more potent and can be safely used in the practical application of predicting the gene function given its DNC.

Acknowledgements

We are highly thankful to the Department of Biotechnology, New Delhi and Madhya Pradesh Council of Science and Technology, Bhopal for providing bioinformatics infrastructure facility at Maulana Azad National Institute of Technology, Bhopal for carrying out this work.

Authors' contributions

LR conceived the original idea and carried out data collection, computational processing and interpretation of results. BP, KP and KRP guided the study. LR and KP jointly drafted the manuscript. All the authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Dye, C., *et al*. 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *JAMA* 282: 677-686.
2 Murray, C.J. and Salomon, J.A. 1998. Modeling the impact of global tuberculosis control strategies. *Proc. Natl. Acad. Sci. USA* 95: 13881-13886.
3 Alteri, C.J., *et al*. 2007. *Mycobacterium tuberculosis* produces pili during human infection. *Proc. Natl. Acad. Sci. USA* 104: 5145-5150.
4 Frothingham, R., *et al*. 1994. Extensive DNA sequence conservation throughout the *Mycobacterium tuberculosis* complex. *J. Clin. Microbiol.* 32: 1639-1643.
5 Kirschner, P., *et al*. 1993. Genotypic identification of mycobacteria by nucleic acid sequence determination: report of a 2-year experience in a clinical laboratory. *J. Clin. Microbiol.* 31: 2882-2889.
6 Takewaki, S.L., *et al*. 1994. Nucleotide sequence comparison of the mycobacterial dnaJ gene and

PCR-restriction fragment length polymorphism analysis for identification of mycobacterial species. *Int. J. Syst. Bacteriol.* 44: 159-166.

7 Bifani, P., *et al*. 2000. Molecular characterization of *Mycobacterium tuberculosis* H37Rv/Ra variants: distinguishing the mycobacterial laboratory strain. *J. Clin. Microbiol*. 38: 3200-3204.

8 Carlsson, A., *et al*. 2010. Plasma proteome profiling reveals biomarker patterns associated with prognosis and therapy selection in glioblastoma multiforme patients. *Proteomics Clin. Appl*. 4: 591-602.

9 Anderson, D.C., *et al*. 2010. Extensive and varied modifications in histone H2B of wild-type and histone deacetylase 1 mutant *Neurospora crassa*. *Biochemistry* 49: 5244-5257.

10 Mao, Y., *et al*. 2005. Constructing support vector machine ensembles for cancer classification based on proteomic profiling. *Genomics Proteomics Bioinformatics* 3: 238-241.

11 Rishishwar, L., *et al*. 2011. Support vector machine classification and prediction of lyases. *Online J. Bioinformatics* 12: 1-8.

12 Rishishwar, L., *et al*. 2010. Support vector machine approach for isomerases prediction problem. *CiiT Int. J. Data Min. Knowl. Eng*. doi: DMKE052010005.

13 Guyon, I., *et al*. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn*. 46: 389-422.

14 Cole, S.T., *et al.* 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537-544.

15 Reddy, T.B., *et al*. 2009. TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res*. 37: D499-508.

16 Kecman, V. 2001. Learning and soft computing: support vector machines, neutral networks and fuzzy logic models. In *Support Vector Machines*, pp.121-189. MIT Press, Cambridge, MA, USA.

17 Han, J. and Kamber, M. 2006. Data mining: concepts and techniques. In *Classification and Prediction*, pp.285-344. Morgan Kaufmann, Waltham, MA, USA.

18 Cristianini, N. and Taylor, J.S. 2000. Support vector machines. In *Support Vector Machines and Other Kernel-based Learning Methods*, pp.93-112. Cambridge University Press, Cambridge, UK.

19 Chang, C.C. and Lin, C.J. 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 27.

20 Vapnik, V.N. 1995. *The Nature of Statistical Learning Theory.* Springer Verlag, Heidelberg, Germany.

21 Roback, P., *et al.* 2007. A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res*. 35: 5085-5095.

22 Joachims, T. 1999. Advances in kernel methods—support vector learning. In *Making Large-Scale SVM Learning Practical* (eds. Schölkopf, B., *et al*.), pp.169-184. MIT Press, Cambridge, MA, USA.

23 Schmidt, H. and Hensel, M. 2004. Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.* 17: 14-56.

24 Churchill, G.A. 1992. Hidden Markov chains and the analysis of genome structure. *Comput. Chem.* 16: 107-115.

25 Nussinov, R. 1981. Nearest neighbour nucleotide patterns. Structural and biological implications. *J. Biol. Chem*. 256: 8458-8462.

26 Nakashima, H., *et al*. 1998. Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res*. 5: 251-259.

27 Nakashima, H., *et al*. 2003. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem.* 133: 507-513.

28 Durbin, R., *et al*. 2002. Biological sequence analysis. In *Markov Chains and Hidden Markov Models*, pp.48-49. Cambridge University Press, Cambridge, UK.

Supplementary Material

Figures S1 and S2; Table S1

DOI: 10.1016/S1672-0229(11)60020-X