Journal of
Computer Science & Systems Biology

Pillai et al. J Comput Sci Syst Biol 2011, 4:3
DOI: 10.4172/jcsb.1000075

Research Article Open Access

SVM Model for Amino Acid Composition Based Prediction of *Mycobacterium tuberculosis*

Lakshmi Pillai[1]*, Bhasker Pant[1], Usha chauhan[2] and KR Pardasani[3]

[1]*Phd Student, Department of Mathematics, Maulana Azad National Institute of Technology*
[2]*Assistant Professor, Department of Mathematics, Maulana Azad National Institute of Technology*
[3]*Professor, Department of Mathematics, Maulana Azad National Institute of Technology*

Abstract

The Tuberculosis is the classical human mycobacterial disease, caused by *Mycobacterium tuberculosis*. The disease primarily affect the lung and causes pulmonary tuberculosis, as well as affect intestine, bone, joints, meninges, lymph nodes, skin and other tissue of the body, causing extra pulmonary tuberculosis. Thus there arises the need to understand the relationships among various parameters of these proteins for prediction of their classes, structures and functionality. The computational approaches for prediction of their classes are fast and economical therefore can be used to complement the existing wet lab techniques. Realizing their importance, in this paper an attempt has been made to correlate them with their amino acid composition and predict them with fair accuracy. This is a novel method where Mycobacterium Tuberculosis has been classified on the basis of amino acid composition using Support Vector Machine. The SVM has been implemented using SVM Light package [1,2]. The method discriminates different strains of Mycobacterium Tuberculosis. The performance of the method was evaluated using 10-fold cross-validation where accuracy of 100% was obtained.

Keywords: xtra pulmonary Tuberculosis; Support vector machine; Amino acid composition; Kernel functions; Granuloma; Macrophages; Necrosis; Binary classifier; Cytotoxic T cells; Supervised machine learning; Matthews correlation coefficient

Introduction

The German scientist (**Robert Koch**) announced that he had cultured the causative agent from human TB lesions and designated as "Bacillus of Tuberculosis". Mycobacterium, the genus of Actinobacteria, given its own family of mycobacteriaceae includes certain species.

About 90% of those infected with *Mycobacterium tuberculosis* have asymptomatic, latent TB infection (sometimes called LTBI), with only a 10% lifetime chance that a latent infection will progress to TB disease. However, if untreated, the death rate for these active TB cases is more than 50%.

TB infection begins when the mycobacteria reach the pulmonary alveoli, where they invade and replicate with the endosomes of alveolar macrophages. The primary site of infection in the lungs is called the Ghon focus, and is generally located in either the upper part of the lower lobe, or the lower part of the upper lobe. Bacteria are picked up by dendritic cells, which do not allow replication, although these cells can transport the bacilli to local (mediastinal) lymph nodes. Further spread is through the bloodstream to other tissues and organs where secondary TB lesions can develop in other parts of the lung (particularly the apex of the upper lobes), peripheral lymph nodes, kidneys, brain, and bone. All parts of the body can be affected by the disease, though it rarely affects the heart, skeletal muscles, pancreas and thyroid. Tuberculosis is classified as one of the granulomatous inflammatory conditions. Macrophages, T-lymphocytes, B-lymphocytes and fibroblasts are among the cells that aggregate to form agranuloma, with lymphocytes surrounding the infected macrophages. The granuloma functions not only to prevent dissemination of the mycobacteria, but also provides a local environment for communication of cells of the immune system. Within the granuloma, T lymphocytes secrete cytokines such as interferon gamma, which activates macrophages to destroy the bacteria with which they are infected. Cytotoxic T cells can also directly kill infected cells, by secreting perforin and granulysin [3,4].

Importantly, bacteria are not always eliminated within the granuloma, but can become dormant, resulting in a latent infection. Another feature of the granulomas of human tuberculosis is the development of cell death, also called necrosis, in the center of tubercles. To the naked eye this has the texture of soft white cheese and was termed caseous necrosis.

If TB bacteria gain entry to the bloodstream from an area of damaged tissue they spread through the body and set up many foci of infection, all appearing as tiny white tubercles in the tissues. This severe form of TB disease is most common in infants and the elderly and is called miliary tuberculosis. Patients with this disseminated TB have a fatality rate of approximately 20%, even with intensive treatment.

In many patients the infection waxes and wanes. Tissue destruction and necrosis are balanced by healing and fibrosis. Affected tissue is replaced by scarring and cavities filled with cheese-like white necrotic material. During active disease, some of these cavities are joined to the air passages bronchi and this material can be coughed up. It contains living bacteria and can therefore pass on infection. Treatment with appropriate antibiotics kills bacteria and allows healing to take place. Upon cure, affected areas are eventually replaced by scar tissue.

Currently, efforts are underway to develop new therapeutic agents and elucidation of metabolic pathway associated with diseases

***Corresponding author:** Lakshmi Pillai, Department of Mathematiics, Maulana Azad National Institute of Technology, Bhopal, India, Tel: +91 (0)7828013946; E-mail: lakshmilster@gmail.com

Received May 12, 2011; **Accepted** July 30, 2011; **Published** July 31, 2011

Citation: Pillai L, Pant B, Chauhan U, Pardasani KR (2011) SVM Model for Amino Acid Composition Based Prediction of *Mycobacterium tuberculosis*. J Comput Sci Syst Biol 4: 047-049. doi:10.4172/jcsb.1000075

[5]. Moreover, the mere understanding of different strains of Mycobacterium will assist in finding novel drug target with minimum side effects. The experimental attempts are reported in the literature for functional classification of Mycobacterium Tuberculosis. But no computational technique is available in the literature for classification of Mycobacterium tuberculosis based on other parameters like dipeptide composition, amino acid composition and physiochemical properties. Since the experimental identifications of them are labor and cost-intensive task, the computational biology can provide a better alternative to develop a method for classifying different strains of Mycobacterium Tuberculosis.

In view of the above an attempt has been made in this paper to develop a computational approach for predicting and classifying two types of Tuberculosis strains i.e. Mycobacterium Tuberculosis and Non Mycobacterium Tuberculosis. This is a binary classification method where the Tuberculosis can be discriminated as Mycobacterium Tuberculosis and Non Mycobacterium Tuberculosis [6]. It has been shown in past that SVM is an elegant technique for the classification of biological data. Here SVM model has been developed for amino acid composition based prediction identification and classification of MTB and Non Mycobacterium TB.

This paper is a step in the direction where machine learning and computational biology techniques can be used to complement existing wet lab techniques [6,7].

Materials and Methods

Data set

To achieve our goal and develop our methodology we obtained the dataset from Swissprot/Uniprot databank of Expasy server (12). The following two data sets were used.

Dataset1: It consisted of all Mycobacterium Tuberculosis proteins. All the entries marked as fragments were not included in the dataset. The total instances were 28,200. The final dataset consisted of 28,200 sequences belonging to Mycobacterium TB strains i.e. H37RA 4002 sequences, CDC1551 4197 sequences, F11 3905 sequences, H37RV 8021 sequences and KZN- 1435 4026 sequences [8,9].

Dataset2: To validate our methodology proteins belonging to some other class were taken into consideration. They were treated as negative instances.

For training dataset we consider 25,000 sequences belonging to different strains while remaining 3,200 sequences were used to prepare the test dataset.

Support vector machine (Binary classification)

SVM is a supervised machine learning method which is based on the statistical learning theory. When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The SVMs were implemented using freely downloadable software, SVM light [1,2]. In this software there is a facility to define parameters and choose among various inbuilt kernals. They can be radial basis function (RBF) or a polynomial kernel (of given degree), linear, sigmoid.

SVM software; SVM light

Simulations were performed using SVM light version 6.02 (a freely available software package) [8,9]. For our study RBF Kernel was found to be the best. The SVM training was carried out by the optimization of the value of the regularization parameter and the value of RBF kernel parameter.

Amino acid composition

Previously, this parameter has been used for predicting the subcellular localization of proteins [12,13]. The amino acid composition is the fraction of each amino acid type within a protein. The fractions of all 20 natural amino acids were calculated by using Equation 1,

Fraction of amino acid i (where i can be any amino acid)

$$= \frac{TotalNumberofaminoacidi}{Totalnumberofaminoacidsinaprotein}$$

Evaluation of Performance:

The performance of our classifier was judged by 10 fold cross validation. The SVM Light provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of SVM Light tools as shown in Figure 1. Here pairs of C and Gamma are tried and the one with the best cross validation accuracy is picked. On using the values of C=0.5 and Gamma=0.5 obtained through grid search an accuracy of 100% was obtained.

Prediction system assessment

True positives (TP) and true negatives (TN) were identified as the positive and negative samples, respectively. False positives (FP) were negative samples identified as positive. False negatives (FN) were positive samples identified as negative. The prediction performance was tested with sensitivity (TP/ (TP+FN)), specificity (TN/ (TN+FP)), overall accuracy (Q2), and the Matthews correlation coefficient (MCC). The accuracy and the MCC for each subfamily of Mycobacterium tuberculosis, was calculated as described by Hua and Sun [9] and shown below in equation 2 and 3.

$$Accuracy(x) = \frac{tp + tn}{tp + tn + fp + fn}$$

$$MCC = \frac{(tp)(tn) - (fp)(fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

Results and Discussion

All strains of Mycobacterium Tuberculosis have been implicated in various diseases. Realizing their involvement in disease we have chosen these strains for our study.

The results obtained here will be helpful in differentiating between different strains of Mycobacterium Tuberculosis. A new protein discovered can be shown as belonging to any of the classified Mycobacterium strain. This model can also be an important tool to understand the differences between different strains hence a step towards assisting various wet lab techniques in devising novel drugs and therapeutic agents against these strains. The correlation of different strains with their amino acid composition explored here can be useful to obtain better insight about these strains.

The overall accuracy and MCC of the amino acid composition-based classifier [14] for classifying the different strains of Mycobacterium Tuberculosis was 100%. It proved that strains can be correlated with amino acid composition and can be easily distinguished on this basis.

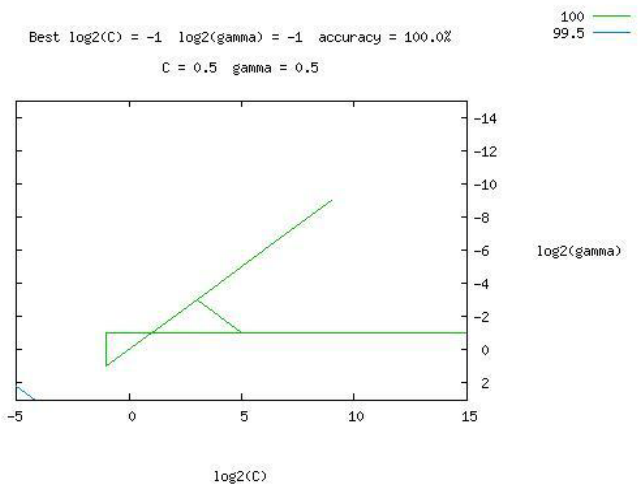

Figure 1: Coarse Grid Search on C = 2-5, 2-4 … 210 and Gamma = 25, 24 … 2 – 10.

Conclusion

The SVM model developed here is computationally efficient and effective in predicting and classifying the Mycobacterium Tuberculosis. This is evident from the accuracy (100%) in the results. Further the amino acid composition contains very significant information for discriminating the classes of above proteins.

This model can be used to analyze other strains, such as entire proteomics data. Such type of prediction systems can be very useful for understanding the above proteases in a better way so as in conclusion, a novel method for classifying Mycobacterium Tuberculosis is presented. This method will nicely complement the existing wet lab methods. It will assist in assigning the correct class to which these proteins belong. The prediction method presented here may be useful for the annotation of the piled-up proteomic data.

The author awaits discovery of more of these proteins in the future so that accuracy of the prediction model can be increased further and a server developed for public use.

Key points

- The SVM model developed here is computationally efficient and effective in predicting and classifying the Mycobacterium Tuberculosis with accuracy rate of 100%.
- This model can be used to analyze other strains, such as entire proteomics data of Mycobacterium Tuberculosis.
- All strains of Mycobacterium Tuberculosis have been implicated in various diseases. Realizing their involvement in disease we have chosen these strains for our study.
- It will assist in assigning the correct class to which the proteins belong and thus will nicely complement the existing wet lab methods.
- The correlation of different strains with their amino acid composition explored here can be useful to obtain better insight about these strains.

Acknowledgement

The authors are highly thankful to the department of Biotechnology, Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinformatics Infrastructure facility to carry out this work.

References

1. Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84: 343-348.
2. Cai YD, Pong-Wong R, Feng K, Jen JCH, Chou KC (2004) Application of SVM to predict membrane protein types. J Theor Biol 226: 373-376.
3. Cai YD, Zhou GP, Jen CH, Lin SL, Chou KC (2004) Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228: 551-557.
4. Joachims T (2002) SVM Light: a library for support vector machines, software.
5. Holmgren NB, Millman I, Youmans GP (1954) Studies on the metabolism of Mycobacterium tuberculosis. VI. The effect of Krebs' tricarboxylic acid cycle intermediates and precursors on the growth and respiration of Mycobacterium tuberculosis. J Bacteriol 68: 405-410.
6. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain *iNJ*661 and proposing alternative drug targets BMC Syst Biol 1: 26.
7. ExPASy Proteomics Server.
8. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721-728.
9. Cai YD, Zhou GP, Jen CH, Lin SL, Chou KC (2004) Identify catalytic triads of serine hydrolases by support vector machines. J Theor Biol 228: 551-557.
10. Joachims T (2001) SVM Light: a library for support vector machines.
11. Schneider E, Moore M, Castro KG (2005) Epidemiology of tuberculosis in the United States. Clin chest med 26: 183-195.
12. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765-45769.
13. Bhasin M, Raghava GP (2004) Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition. J Biol chem 279: 23262-23266.
14. Blobel CP (1999) Metalloprotease-disintegrins: modular proteins capable of promoting cell-cell interactions and triggering signals by protein-ectodomain shedding. Cell 90: 589.
15. Cosma CL, Sherman DR, Ramakrishnan L (2003) The secret lives of the pathogenic mycobacteria. Annu Rev Microbiol 57: 641-676.