VICMpred: An SVM-based Method for the Prediction of Functional Proteins of Gram-negative Bacteria Using Amino Acid Patterns and Composition

Sudipto Saha and G.P.S. Raghava*

Institute of Microbial Technology, Chandigarh, India.

In this study, an attempt has been made to predict the major functions of gram-negative bacterial proteins from their amino acid sequences. The dataset used for training and testing consists of 670 non-redundant gram-negative bacterial proteins (255 of cellular process, 60 of information molecules, 285 of metabolism, and 70 of virulence factors). First we developed an SVM-based method using amino acid and dipeptide composition and achieved the overall accuracy of 52.39% and 47.01%, respectively. We introduced a new concept for the classification of proteins based on tetrapeptides, in which we identified the unique tetrapeptides significantly found in a class of proteins. These tetrapeptides were used as the input feature for predicting the function of a protein and achieved the overall accuracy of 68.66%. We also developed a hybrid method in which the tetrapeptide information was used with amino acid composition and achieved the overall accuracy of 70.75%. A five-fold cross validation was used to evaluate the performance of these methods. The web server VICMpred has been developed for predicting the function of gram-negative bacterial proteins (http://www.imtech.res.in/raghava/vicmpred/).

Key words: virulence factor, cellular process, information molecule, tetrapeptide, VICMpred, gram-negative bacteria

Introduction

Though there has been an exponential growth in sequence databases of proteins in the last decade, the experimental assessment of the function of every protein in each newly sequenced genome is beyond foreseeable. Our knowledge of most of the new proteins will be from prediction. Function prediction is a major challenge in the field of bioinformatics (*1*). In the past, a number of methods have been developed to predict the function of proteins (*2–4*), but the results were obtained by analyzing a significant number of true sequence similarities, pointing to the complexity of function prediction. Most of the methods are indirect ones that make attempts to predict the subcellular localization of proteins rather than the function. The subcellular localization prediction methods are based on the observation that proteins belonging to the same compartment have similar amino acid composition (*5*, *6*) and functions.

*** Corresponding author.**
E-mail: raghava@imtech.res.in

In this study, an attempt has been made to develop a direct method for predicting the major functions (virulence factors, information molecules, cellular process, and metabolism) of gram-negative bacterial proteins. Most of the proteins in an organism involve in the cellular process, metabolism, and information storage, and the remaining can be classified into virulence factors, which allow the germs to establish themselves in the host. Virulence factors include adhesions (*7*), toxins (*8*), and hemolytic molecules (*9*). The identification of virulence factors is crucial for the drug development. Therefore, we classified the bacterial proteins into four broad functional classes. The other three classes were taken from the functional annotation of the COGs (Clusters of Orthologous Groups of proteins) database (*10*). They are (1) cellular process, which includes cell division, cell envelope biogenesis, cell motility, and signal transduction molecules; (2) information storage and processing, in which transcription, translation, and DNA replication and repair molecules are included; (3) metabolic process, including energy production and the trans-

port and metabolism of carbohydrate, amino acid, nucleotide, and lipid.

The similarity search tools like BLAST (*11*), FASTA (*11*), and PSI-BLAST (*12*) are commonly used for the annotation of genomes. Besides, machine learning tools are also used for the classification of proteins, where amino acid, pseudo, dipeptide, and property composition are used as protein features. The function prediction of proteins is much more complex than other classifications because the sequence similarity is very poor in the proteins that have the same function, thus most of the methods based on similarity search fail to predict the function of proteins (*13*, *14*).

In this study, we made a systematic attempt to develop a better method for predicting the function of proteins. First, we tried traditional strategies for the classification of proteins that include (1) similarity search using PSI-BLAST; (2) support vector machine (SVM)-based method using amino acid composition; and (3) SVM-based method using dipeptide composition, which also considers the local order of amino acids. It was observed that the performance of traditional approaches was very poor in the functional classification of proteins. In order to improve the performance, we used tetrapeptides as features of protein similar to the deterministic pattern of Class A as defined by Brazma *et al* (*15*). The approach relies on identifying short signaling patterns and the group of patterns of each four broad functional classes present in a higher number (*16*). The performance of our method based on tetrapeptides was much better than that of traditional methods based on residue composition. It was further improved when the new and traditional approaches were combined. In this study, we classified the gram-negative bacterial proteins obtained from PSORTdb v.20 (http://www.psort.org/dataset; ref. *17*), which were used in the development of SubLoc (*18*). Based on our study, we have made a web server, VICMpred (http://www.imtech.res.in/raghava/vicmpred/) for predicting the function of proteins from their amino acid sequences.

Results

The performance of all the modules developed in this study is shown in Table 1, which was evaluated through a five-fold cross-validation. The composition-based module [kernel = RBF (radial basis function), Υ = 80, C = 2, and j = 4] was able to predict with accuracy of 52.39%. In the case of the dipeptide-based module, the performance of the RBF kernel (Υ = 100, C = 50, and j = 1) was 5% lower than that of the amino acid composition. The PSI-BLAST module predicted cellular, information, metabolism, and virulence protein sequences with accuracy of 23.13%, 8.33%, 28.77%, and 25.71%, respectively. During the five-fold cross-validation, only 172 hits were obtained out of the total 670 proteins. Therefore, the performance of this module was poorer in comparison to that of the SVM modules based on amino acid and dipeptide composition.

It is interesting to note that the performance for the dipeptide-based module was lower than the simple amino acid composition based module, despite dipeptide provides composition as well as the order of local amino acids. This is because in the case of dipeptide, the total number of features are 400 (20×20), which is too high to occur in a small number of proteins. Thus SVM is unable to learn properly on too many features.

Table 1 The Performance of Various Modules Including SVM Modules Based on Various Features of Protein Sequences and PSI-BLAST

Approach	Cellular		Information		Metabolism		Virulence		Overall
	ACC*	MCC*	ACC	MCC	ACC	MCC	ACC	MCC	ACC
Composition-based (A)	47.06	0.12	0.12	0.41	0.41	0.31	27.14	0.32	52.39
Dipeptide-based (B)	45.10	0.11	15.00	0.21	60.35	0.23	27.14	0.20	47.01
Pattern-based (C)	70.20	0.46	48.33	0.57	72.98	0.51	62.86	0.61	68.66
PSI-BLAST	23.13	/	8.33	/	28.77	/	25.71	/	/
Hybrid 1 (A+C)	69.41	0.48	50.00	0.59	77.19	0.54	62.86	0.65	70.30
Hybrid 2 (B+C)	69.02	0.54	48.33	0.52	74.04	0.53	58.57	0.54	68.21
Hybrid 3 (A+B+C)	69.80	0.51	53.33	0.58	77.54	0.56	61.43	0.59	70.75

*ACC: Accuracy (%); MCC: Matthew's correlation coefficient.

In order to avoid this problem, we introduced a new concept for prediction, where we consider peptides that occur in each class of proteins in a significant amount. Here we used the frequency of significant tetrapeptides found in a class of proteins. This *ab initio* pattern based module was able to predict the function of proteins with accuracy of 68.66% (kernel = RBF, $\Upsilon = 0.001$, C = 50, and j = 5), which was higher than the modules based on amino acid and dipeptide composition.

To further improve the prediction accuracy, hybrid modules on the basis of various features of proteins were constructed. The first hybrid (Hybrid 1) was developed on the basis of pattern information and amino acid composition. The prediction accuracy of the Hybrid 1 module was 70.30%, which was better than that of any individual feature-based module. Another module (Hybrid 2) was developed on the basis of pattern information and dipeptide composition; its performance was similar to that of the Hybrid 1 module. Then, a hybrid module (Hybrid 3) based on pattern information, amino acid, and dipeptide composition was developed. This hybrid used an input vector of 424 dimensions, comprising 4 for pattern information, 20 for amino acid composition, and 400 for dipeptide composition. As shown in Table 1, the performance of this module was better than that of any individual feature-based or other hybrid modules (Hybrids 1 and 2). Finally, the Hybrid 3 module with the RBF kernel ($\Upsilon = 0.001$, C = 100,000, and j = 1) was able to achieve the overall accuracy of 70.75%.

VICMpred server

Based on our study, we have developed a web server, VICMpred, which allows users to predict the function of a protein (virulence factors, information molecules, cellular process, and metabolism) from its amino acid sequences. VICMpred is freely available at http://www.imtech.res.in/raghava/vicmpred/. The common gateway interface (CGI) script for VICMpred was written using PERL version 5.03. This server is installed on a Sun Server (420E) under a UNIX (Solaris 7) environment. Users can enter the primary amino acid sequence for prediction using file uploading or cut-and-paste options.

Discussion

The functional annotation of proteins is one of the major challenges in the post-genomic era. The most widely used methods for predicting the function of a new protein involve sequence alignment, similarity search, or profile search, like FASTA, BLAST, and PSI-BLAST (*12*, *13*). These methods fail in the absence of significant similarity between queried and annotated proteins. One of the reasons of the failure of the similarity-based methods is that the variation in the size of proteins either belongs to the same or different classes.

The problem with profiles is that they are complicated models with many free parameters. There are a number of difficult problems like the best ways to set the position-specific residue scores, to score gaps and insertions, and to combine structural and multiple sequence information. An alternative way for predicting the function of a protein is to predict its location in the cell, which is based on the assumption that proteins residing in the same location also have the same functions. Most of these subcellular localization methods are based on the composition (amino acid or dipeptide) of proteins.

In this study, an attempt has been made to develop a direct method for predicting the function of proteins. First we tried traditional approaches that are commonly used in the prediction of the subcellular localization. It was observed that the performance of PSI-BLAST was poorer compared to that of composition-based methods (Table 1). This demonstrates that similarity search based methods are not very effective in function prediction. It was also observed that the dipeptide-based method performed poorer than the amino acid composition based method. This fact was unexpected as dipeptide provides more information (composition with local order) than simple amino acid. In the past we had observed that dipeptide performed better than amino acid composition in the subcellular localization of proteins. We examined our data and observed that the number of dipeptides was either rare or completely absent due to the small number of proteins used for classification. This demonstrates that the higher order composition is not successful on the small dataset. We tried a new approach in order to overcome this problem. In this approach, we used tetrapeptides that provide more local orders than dipeptide and tripeptide. Instead of using the composition of all tetrapeptides, we identified the tetrapeptides found in a significant number in each class of proteins, and only used significant tetrapeptides for classification. We calculated the number of tetrapeptides of each class present in a

query sequence. This information was used to classify the proteins by using SVM and obtained very high accuracy. One may compare this approach with pattern searching approaches (like PROSITE) where one needs to detect known patterns in a sequence. Here the patterns are tetrapeptides instead of PROSITE patterns. There is a limited number of PROSITE, so the number of proteins does not have any PROSITE pattern. Whereas in our case, we used all tetrapeptides found in a significant amount in each class of proteins, so the number of patterns in our case was very high (1,248, 381, 1,443, and 1,168 for cellular process, information, metabolism, and virulence, respectively). Thus there is a chance that each query protein will have a large number of tetrapeptides in each class. Though the specificity of our tetrapeptides is lower than that of the PROSITE patterns, the number is 100 times more.

We also developed hybrid modules, which combined our composition-based modules and pattern-based approach, in order to further improve the performance of the method. The performance of hybrid methods was better than that of any individual.

In summary, we have developed an effective method for predicting the function of bacterial proteins. This method will be very useful in the development of drug and vaccine as it allows predicting virulence proteins. Though we tried our best to improve the accuracy of prediction, still it is not very high. Another limitation of this method is that it just predicts the single function of a protein, whereas in realistic situation it is observed that a protein may have multiple functions.

Materials and Methods

Datasets

We obtained 1,572 proteins from Hua and Sun's work (*18*), examined the functions of these proteins using SWISS-PROT (*19*) version 33.0, and kept 1,048 proteins for further processing, whose functions were already known. We used the PROSET software to create a dataset of non-redundant proteins where no two proteins have more than 90% sequence identity. The final dataset consists of 670 non-redundant gram-negative bacterial proteins (255 of cellular process, 60 of information molecules, 285 of metabolism, and 70 of virulence factors).

Evaluation of the predictive performance

The performance of the modules constructed in this study was evaluated using a five-fold cross-validation technique. In the five-fold cross-validation, the relevant dataset was randomly divided into five sets. The training and testing were carried out for five times, each time using one distinct set for testing and the remaining four sets for training. For evaluating the performance of various modules, accuracy and Matthew's correlation coefficient (MCC) were calculated using the following equations:

$$Accuracy : (x) = \frac{p(x)}{Exp(x)}$$

MCC:

$$(x) = \frac{p(x)n(x) - u(x)o(x)}{\sqrt{[p(x) + u(x)][p(x) + o(x)][n(x) + u(x)][n(x) + o(x)]}}$$

where x can be any functional class (cellular process, information, metabolism, and virulence), $Exp(x)$ is the number of sequences observed in function x, $p(x)$ is the number of correctly predicted sequences of function x, $n(x)$ is the number of correctly predicted sequences not of function x, $u(x)$ is the number of under-predicted sequences, and $o(x)$ is the number of over-predicted sequences.

Support vector machine

SVM was implemented using the freely downloadable software package SVM_light written by Joachims (*20*). The software enables the user to define a number of parameters as well as to select from a choice of inbuilt kernel functions, including an RBF and a polynomial kernel. Preliminary tests show that the RBF kernel gives results better than other kernels. Therefore, in this work we used the RBF kernel for all the experiments. The prediction of functional classes is a multi-class classification problem. We developed a series of binary classifiers to handle this problem. We constructed N SVMs for the N-class classification using 1 vs r (one against the rest) strategy. Here, the class number was equal to four for bacterial protein sequences. The i^{th} SVM was trained with all samples in the i^{th} class with positive labels and all other samples with negative labels. In this way, four SVMs were constructed for the functional classes of bacterial proteins to cellular process, information, metabolism, and virulence.

Protein features

Amino acid composition

Amino acid composition is the fraction of each amino acid in a protein. The fraction of each of the 20 natural amino acids was calculated using the following equation:

$$\text{Fraction of amino acid } i = \frac{\text{Total number of amino acid } i}{\text{Total number of amino acids in protein}}$$

where i can be any amino acid.

Dipeptide composition

Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20×20). This representation encompassed the information about amino acid composition along the local order of amino acids. The fraction of each dipeptide was calculated using the following equation:

$$\text{Fraction of } dipep(i) = \frac{\text{Total number of } dipep\,(i)}{\text{Total number all possible dipeptides}}$$

where $dipep(i)$ is one out of 400 dipeptides.

Ab initio patterns

We have calculated the frequency of all possible tetrapeptides (20×20×20×20=160,000) in each class of proteins. Then we identified the significant tetrapeptides for that class, which are generally found more than a threshold for a class of proteins. In our case, we considered a tetrapeptide as significant if it is found ≥6 times in the case of cellular proteins; ≥3 times in the case of information molecules; ≥6 times in the case of metabolic proteins; and ≥4 times in the case of virulence proteins. In the next step, we computed the number of significant tetrapeptides of each class in a protein. Thus, four features represent a protein, where each feature represents the significant number of tetrapeptides of a class of proteins. Finally we used SVM for the classification of proteins based on these four features. In our study, significant tetrapeptides were only calculated from proteins in training datasets in order to avoid any biasness in prediction. An outline of this method is shown in Figure 1.

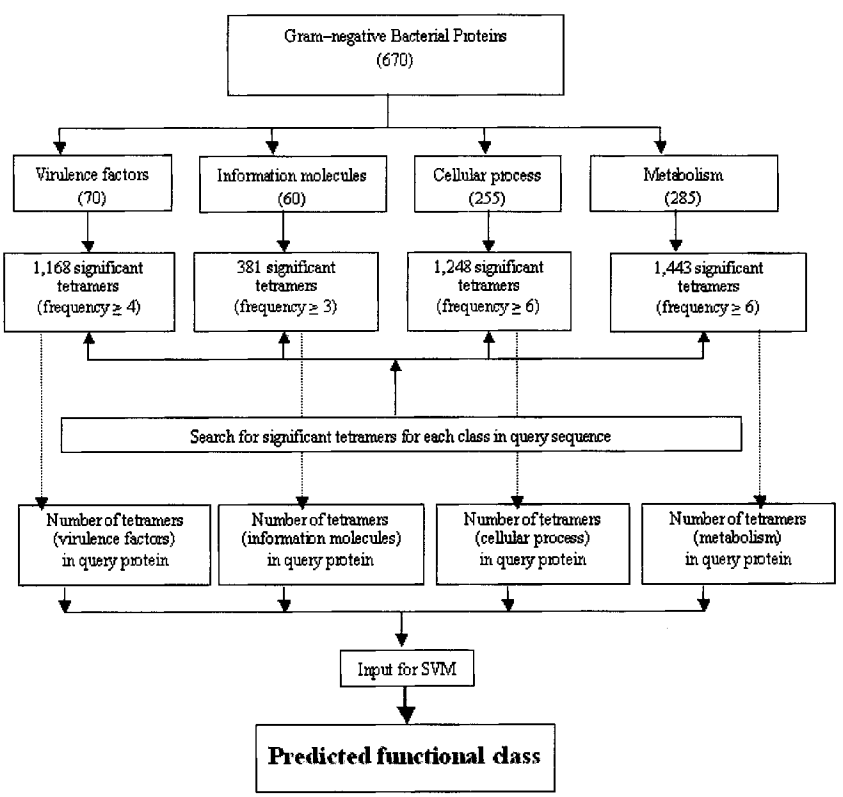

Fig. 1 An outline of the *ab initio* pattern prediction method.

PSI-BLAST

A module of PSI-BLAST was designed, in which query sequences in testing datasets were searched against proteins in training datasets using PSI-BLAST. Three iterations of PSI-BLAST were carried out at a cut-off E-value of 0.001. PSI-BLAST was used instead of normal standard BLAST because PSI-BLAST has the capability to detect remote homologies. The module could predict any of the four functions (cellular process, information, metabolism, and virulence) depending upon the similarity of the query protein to the protein in the dataset.

Acknowledgements

This work was supported by the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India (Grant No. CMM-17).

References

1. Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins* 41: 98-107.
2. Rost, B., *et al.* 2003. Automatic prediction of protein function. *Cell. Mol. Life Sci.* 60: 2637-2650.
3. Panchenko, A.R., *et al.* 2004. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* 13: 884-892.
4. Cai, Y.D. and Doig, A.J. 2004. Prediction of *Saccharomyces cerevisiae* protein functional class from functional domain composition. *Bioinformatics* 20: 1292-1300.
5. Bhasin, M. and Raghava, G.P. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32: W414-419.
6. Garg, A., *et al.* 2005. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* 280: 14427-14432.
7. Irie, Y., *et al.* 2004. The Bvg virulence control system regulates biofilm formation in *Bordetella bronchiseptica*. *J. Bacteriol.* 186: 5692-5698.
8. Geric, B., *et al.* 2004. Distribution of Clostridium difficile variant toxinotypes and strains with binary toxin genes among clinical isolates in an American hospital. *J. Med. Microbiol.* 53: 887-894.
9. Ethelberg, S., *et al.* 2004. Virulence factors for hemolytic uremic syndrome, Denmark. *Emerg. Infect. Dis.* 10: 842-847.
10. Tatusov, R.L., *et al.* 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28: 33-36.
11. Altschul, S.F., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
12. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
13. Li, L., *et al.* 2003. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci. USA* 100: 4463-4468.
14. Hannenhalli, S.S. and Russell, R.B. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* 303: 61-76.
15. Brazma, A., *et al.* 1998. Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.* 5: 279-305.
16. Perez, A.J., *et al.* 2002. A computational strayegy for protein function assigment which addresses the multidomain problem. *Comp. Funct. Genomics* 3: 423-440.
17. Gardy, J.L., *et al.* 2003. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* 31: 3613-3617.
18. Hua, S. and Sun, Z. 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17: 721-728.
19. Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28: 45-48.
20. Joachims, T. 1999. Making large-scale SVM learning particle. In *Advances in Kernal Methods: Support Vector Learning* (eds. Scholkopf, B., *et al.*), pp.42-56. MIT Press, Cambridge, USA.