BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Prediction Model**

Volume 6(10)

SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems

Sachin Pundhir* & Anil Kumar

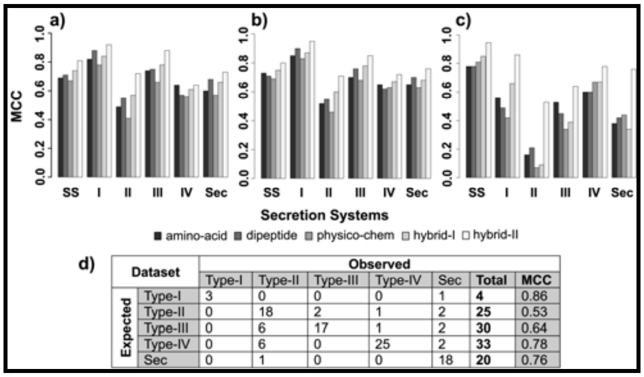
School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore (M.P) – 452001, India; Sachin Pundhir - Email: sachbinfo@gmail.com; Phone: +91-731-2470372; *Corresponding author

Received July 25, 2011; Accepted July 29, 2011; Published August 02, 2011

Abstract:
Protein secretion systems used by almost all bacteria are highly significant for the normal existence and interaction of bacteria with their host. The accumulation of genome sequence data in past few years has provided great insights into the distribution and function of these secretion systems. In this study, a support vector machine (SVM)- based method, SSPred was developed for the automated functional annotation of proteins involved in secretion systems further classifying them into five major sub-types (Type-I, Type-II, Type-III, Type-IV and Sec systems). The dataset used in this study for training and testing was obtained from KEGG and SwissProt database and was curated in order to avoid redundancy. To overcome the problem of imbalance in positive and negative dataset, an ensemble of SVM modules, each trained on a balanced subset of the training data were used. Firstly, protein sequence features like amino-acid composition (AAC), dipeptide composition (DPC) and physico-chemical composition (PCC) were used to develop the SVM-based modules that achieved an average accuracy of 84%, 85.17% and 82.59%, respectively. Secondly, a hybrid module (hybrid-I) integrating all the previously used features was developed that achieved an average accuracy of 86.12%. Another hybrid module (hybrid-II) developed using evolutionary information of a protein sequence extracted from position-specific scoring matrix and amino-acid composition achieved a maximum average accuracy of 89.73%. On unbiased evaluation using an independent data set, SSPred showed good prediction performance in identification and classification of secretion systems. SSPred is a freely available World Wide Web server at http//www.bioinformatics.org/sspred.

Background:

Recent years have witnessed a great thrust in the number of completely sequenced microbial genomes available online to the scientific community. Till date, more than 1600 microbial genomes have been completely sequenced and sequencing of ~5000 is in progress. This leads to an increase in demand of functional annotation of genomic and proteomic data through computational methods. Functional annotation allows categorization of genes in functional classes, which can be very useful to understand the physiological meaning of large amounts of genes. Bacteria on the basis of staining procedure can be classified into Gram-positive and Gram-negative bacteria. While the Gram-positive bacteria contain a single plasma membrane followed by a thick cell wall, Gram-negative bacteria comprise of double membrane layer enclosing the periplasmic space and peptidoglycan layer between the two lipid bilayers **[1]**. Bacterial organisms have evolved dedicated secretion systems that aid in the transport of polypeptides across their outer membrane. While little has been studied about the secretion system pathways in Gram-positive bacteria, various detailed studies have been performed on Gram-negative bacteria **[1, 2]**. Furthermore, genome sequencing of a variety of Gram-positive bacteria showed that many of the secretion genes, which are initially identified in *E.coli*, are also present in these organisms **[3]**. Secretion systems in gram-negative bacteria secrete a wide range of proteins across the cell membrane such as those involved in biogenesis of pili and flagella, nutrient acquisition, virulence and efflux of drugs and other toxins. On the basis of molecular nature of transport machineries and their catalyzed reactions, Secretion systems can be classified into several classes: (1) Type I Secretion (T1S); (2) Type II Secretion (T2S); (3) Type III Secretion (T3S); (4) Type IV Secretion (T4S); and (5) Sec Secretion system pathway **[1, 4]**. Being critical of the export of virulence proteins, functional annotation of proteins involved in export machinery pathways can provide novel drug targets that will be crucial in combat against rapidly evolving pathogenic microorganisms. Most of the tools developed for the identification for secretion systems are either dedicated to only one major class of secretion systems, Type-III **[5, 6, 7]**, or are not specifically meant for secretion systems **[8]**. In this context, similarity based search tools like BLAST **[9]** have aided in the functional annotation of proteomic data, but the major limitation of these tools have been in identifying novel and distantly related proteins. This work explores the use of machine learning approach, Support Vector Machine (SVM), for the identification and classification of proteins involved in secretion system pathways from their sequence. SVM is a widely used machine learning approach for biological sequence analysis due to its ability to handle high dimensional and noisy data. Further, due to its strong mathematical background, it has a great generalization capacity that makes it less susceptible to over-fitting, an important feature in learning algorithms **[10]**. SVM has been widely used for diverse range of Biological applications **[11]**. We have implemented the approach as a web-server application SSPred, available online at http://www.bioinformatics.org/sspred. SSPred predicts a protein to be involved in secretion system pathways on the basis of SVM modules developed using amino acid composition (AAC), dipeptide composition (DPC), physico-chemical composition (PCC), combination of all the three aforementioned properties and combination of PSSM profiles with AAC. A query protein sequence predicted to be part of secretion systems was further classified into one of the five major sub-types i.e. T1S, T2S, T3S, T4S and Sec secretion systems. (Supplementary Figure 1 available at http://www.bioinformatics.org/sspred**)**. The development and performance measure analysis of SSPred will be discussed in the subsequent sections of the manuscript.



Dataset		Observed						
		Type-I	Type-II	Type-III	Type-IV	Sec	**Total**	**MCC**
Expected	Type-I	3	0	0	0	1	**4**	0.86
	Type-II	0	18	2	1	2	**25**	0.53
	Type-III	0	6	17	1	2	**30**	0.64
	Type-IV	0	6	0	25	2	**33**	0.78
	Sec	0	1	0	0	18	**20**	0.76

Figure 1: Prediction performance of various SVM models. **a)** Prediction performance of models on a) training dataset evaluated using 5-fold cross-validation; **b)** Test dataset evaluated using validation test and; **c)** independent dataset evaluated using validation test. Each bar represents MCC achieved by a model based on distinct input features. Hybrid-II based models achieved highest MCC both for the classification of secretion system proteins from non-secretion system protein and at the sub-classification of secretion system proteins; **d)** Confusion matrix showing prediction results of SSPred in the sub-classification of secretion system proteins from independent dataset. A relatively low MCC was observed for Type-II secretion systems due to many false positive predictions. This can be partly explained by the fact that many of Type-II secretion systems proteins are shared among Type-III and Type-IV secretion system proteins **[21, 22]**.

Methodology:

Dataset:

Training Dataset:

1977 secretion system proteins (Positive dataset) were collected from the KEGG **[12]** and SWISS-PROT **[13]**. Similarly, 1932 non-secretion system proteins (not located in the cell wall) were obtained from PSORT-B **[14]** and SubLoc **[15]** making it as our negative dataset.

Test Dataset:

Only 70% of the sequences in the training dataset were used for training models. Remaining 30% were used as the test dataset for evaluating the performance of SVM models.

Independent Dataset:

For unbiased evaluation, an independent set of 112 secretion system and 88 non-secretion system protein sequences were retrieved from Transport Classification Database (TCDB) **[16]** and UniProtKB **[17]**, respectively**.** All the datasets were manually curated such that only validated and non-redundant protein sequences existed in the datasets (Supplementary Table 1 available at http://www.bioinformatics.org/sspred)**.**

Support Vector Machine:

SVM is a machine learning algorithm that from a set of positively and negatively labeled training vectors learns a classifier that can be used for many complex binary classification problems **[10, 11]**. Freely downloadable package, SVM_light (http://www.cs.cornell.edu/people/tj/svm_light/) was used to implement SVM. In this study, the Radial Basis Function (RBF) was adopted and all the regulatory parameters were set as default, except for C and *γ*, which were varied to get the best results. The best C and γ parameters correspond to accuracy value at which sensitivity and specificity values are nearly equal. For both C and γ parameters, a range of 0.05 to 500 was searched.

Input features:

Amino-acid composition:

Amino acid composition is a fraction of each amino acid present in the protein sequence. If L is the length of protein and Q_i is the frequency of occurrence of an amino acid i, then amino acid composition is $C_i = Q_i/L$, where, i is any of the 20 amino acids.

Dipeptide composition:

It transforms a protein into an input vector of 400 dimensions (20 by 20). Let Q_{ij} be the fraction of paired amino acids (i, j = 1, 20) and L be the total number of all possible dipeptides (L = 400) then the dipeptide composition is $C_{ij} = Q_{ij}/L$, where i, j are any of the 20 amino acid residues.

Physico-chemical properties:

Feature vectors of 36 elements corresponding to 36 physico-chemical properties for each amino acid were also used to train SVM modules. The values of each physico-chemical property were normalized between 0 and 1 for all 20 amino acids.

Fraction of phy-chem property i = (Sum total of phy-chem property i in the protein sequence) / (Sum total of all phy-chem properties in the protein sequence)

Position Specific Scoring Matrix (PSSM):

The PSSM for each query sequence was generated using three rounds of PSI-BLAST against a non-redundant protein database, with an E-value cut-off of 0.001. The PSSM provides a matrix of dimension L rows and 20 columns for a protein sequence of L amino acid residues, where, 20 columns represent occurrence/substitution of each type of 20 amino acids. This PSSM matrix was further transformed into an input vector of 400 dimensions using methodology described in earlier studies **[18]**.

SVM Models:

Different SVM models, first using amino acid composition, dipeptide composition, physico-chemical properties individually and then by combining amino acid, dipeptide and physico-chemical (hybrid-I) and amino acid and PSSM (hybrid-II) properties were developed. To overcome the imbalance in positive and negative dataset, an 'ensemble of SVM classifiers' as suggested in **[19]** were used (Supplementary Figure 2, available at http://www.bioinformatics.org/sspred)**.**

Performance evaluation:

5-fold cross validation:

The performance of various SVM models was evaluated using 5-fold cross-validation. The training dataset was randomly partitioned into five subsets of approximately equal size. The training of each module was carried using collection of four subsets as training data and the fifth subset as test data. This process was repeated five times so that each subset was used once as the test data.

Validation test:

Test and independent datasets were used to evaluate the performance of different SVM models. A confusion matrix was employed to quantify the efficiency of classification between secretion and non-secretion systems using TP (True positive – known and predicted secretion systems), TN (True negative – known and predicted non-secretion systems), FP (False positive – known non-secretion systems and predicted secretion systems), FN (known secretion systems and predicted non-secretion systems). We further defined sensitivity (TP/ (TP+FN)), specificity (TN/(TN+FP)), accuracy, Matthews correlation coefficient (MCC) and Reliability index **[20]** for evaluating model performance.

Discussion:

The aim of this study was to develop a prediction server based on SVM for the identification and classification of bacterial secretion systems. Different SVM models based on diverse set of input features were developed and their performance was evaluated based on training, testing and independent datasets. The results are shown in **Figure 1a-d** and Supplementary Table 2 available at http://www.bioinformatics.org/sspred. **Figure 1a-c** shows Mathew Correlation Coefficient (MCC) observed during 5-fold cross-validation and validation test. SVM models based on hybrid-II input features achieved highest MCC at each prediction level, for all three datasets (training, test and independent) as shown in **Figure 1a-c**. **Figure 1a** displays the MCC for various SVM models evaluated using 5-fold cross validation on training dataset. The highest MCC of 0.80 was observed for hybrid-II based SVM model in distinguishing secretion system proteins from non-secretion system proteins. Similarly, at sub-classification of secretion systems, hybrid-II based models achieved highest MCC of 0.95, 0.71, 0.85, 0.72 and 0.76 for Type-1, Type-2, Type-3, Type-4 and Sec Secretion Systems, respectively **(Figure 1a)**. This suggested that inclusion of evolutionary information using PSSM matrix significantly aided the SVM models in increasing sensitivity and specificity of predictions. It is to be noted that inclusion of additional features in the input vector for training SVM modules does not necessarily result in improvement in accuracy as observed for low MCC for some dipeptide based SVM models in comparison to amino-acid based models **(Figure 1a-c)**. **Figure 1b and 1c** displays the MCC for various SVM models evaluated using validation test and independent dataset respectively. As observed during 5-fold cross-validation, hybrid-II based models achieved the highest MCC. For the test dataset, hybrid-II based models achieved a MCC of 0.81 in the classification of secretion system proteins from non-secretion system proteins. Similarly, for the sub-classification of secretion system proteins in test dataset, hybrid-II based models performed best with the highest MCC of 0.92 and lowest MCC of 0.72 was observed for Type-I and Type-II secretion systems, respectively. On unbiased evaluation of trained SVM models using independent dataset, hybrid-II based models achieved a MCC of 0.94 for the classification of secretion system proteins from non-secretion system proteins. Similarly for the sub-classification of secretion systems, the MCC for hybrid-II based models ranged from 0.53 to 0.86 for Type-II and Type-I secretion systems respectively. The performance of other SVM models based on input features like amino-acid, dipeptide, physico-chemical and hybrid-I are also displayed in **Figure 1a-c** for performance comparison to hybrid-II based models.

Earlier studies have shown that various secretion systems share protein components among them like four of the proteins in the T2S and T4S systems, viz. the prepilin peptidase/N-methyl transferase, ATPase, the secretin and the multispanning transmembrane (TM) proteins are shown to be homologous, suggesting a common evolutionary origin **[21, 2].** Similarly, T3S protein components share their injectisome apparatus with flagellar apparatus and some proteins like type-II/III secretion proteins are common in both T2S and T3S system apparatus **[22]**. It is worth mentioning that for training SVM modules the ideal dataset should not have two identical objects with opposite labels (positive and negative) **[10]** as this may result in misclassification of shared labels. In this context, shared proteins among T2S, T3S and T4S systems may be misclassified by SSPred. In fact, the confusion matrix derived from validation test on independent dataset displays some of the protein sequences from Type-III and Type-IV secretion systems misclassified as Type-II secretion systems **(Figure 1d)**. Many of these proteins were indeed observed to be shared components of Type-II, III or IV secretion system machinery (UniProtKB Id: Q7CMH0) and this partially explains the relatively low MCC observed for Type-II in comparison to other secretion system classes. Although this may be regarded as a drawback of the current prediction tool, considering the fact that biological data does not always qualify the prerequisites of machine learning algorithms. Moreover, long recognized similarity between the T2S, T3S and T4S proteins may tend any machine learning algorithm towards a certain level of misclassification. Furthermore, if SSPred was considered as the first tool for functional annotation of huge proteomic data it may prove significant in refining the candidate proteins for further wet-lab based research. All the SVM modules trained in this study have been implemented in the form of a web-server available at http://www.bioinformatics.org/sspred. Server-side programming was implemented using Perl-CGI while for client-side programming HTML and JavaScript were used. SSPred provides a user-friendly interface where the user can type or paste the query sequence(s) in the text-area or can upload the sequence(s) as a single file. Input sequence should be in FASTA format. The server provides options to select any of the prediction approaches (AAC-, DPC-, PCC-, Hybrid-I or Hybrid-II based) for the identification and classification of secretion systems.

Conclusion:

Protein secretion plays a central role in modulating the interactions of bacteria with their environment. Despite a considerable diversity of proteins involved in various classes of secretion systems, our knowledge of the complexity of bacterial secretion systems has expanded. With the rapid accumulation of bacterial genome data, assistance of computational tools for automated functional annotation of genomic data is inevitable. We present here a prediction server, SSPred for the identification and classification of proteins involved in bacterial secretion systems. SSPred is based on a machine learning approach, SVM and is trained using fixed length input vector derived from compositional and evolutionary features of the protein sequence. SSPred has shown good prediction accuracy and authors believe that SSPred will enlighten the path of researchers in their quest in further understanding the complex machinery of bacterial secretion systems.

Acknowledgements:

This work was supported by the Grants received from the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi under the Bioinformatics Sub-Centre at DAVV, Indore.

References:

[1] Lee VT & Schneewind O. *Genes Dev*. 2001 **15**: 1725 [PMID: 11459823]
[2] Tseng TT *et al. BMC Microbiol*. 2009 **9** Suppl 1: S2 [PMID: 19278550]
[3] Paulsen IT *et al. Microbiology* 1997 **143**: 2685 [PMID: 9274022]
[4] Remaut H & Waksman G. *Curr Opin Struct Biol*. 2004 **14**: 161 [PMID: 15093830]
[5] Löwer M & Schneider G. *PLoS One*. 2009 **4**: e5917 [PMID: 19526054]
[6] Arnold R *et al. PLoS Pathog*. 2009 **5**: e1000376 [PMID: 19390696]
[7] Yang Y *et al*. *BMC Bioinformatics* 2010 **11**: S47 [PMID 20122221]
[8] Pundhir S *et al. In Silico Biol*. 2008 **8**: 223 [PMID: 19032158]
[9] Altschul SF *et al. J Mol Biol*. 1990 **215**: 403 [PMID: 2231712**]**
[10] Noble WS. *Nat Biotechnol*. 2006 **24**: 1565 [PMID: 17160063]
[11] Yang ZR. *Brief Bioinform*. 2004 **5**: 328 [PMID: 15606969]
[12] Kanehisa M *et al. Nucleic Acids Res*. 2008 **36**: D480 [PMID: 18077471]
[13] Boeckmann B *et al. Nucleic Acids Res*. 2003 **31**: 365 [PMID: 12520024]
[14] Gardy JL *et al. Nucleic Acids Res*. 2003 **31:** 3613 [PMID: 12824378]
[15] Hua S & Sun Z. *Bioinformatics* 2001 **17**: 721 [PMID: 11524373]
[16] Saier MH Jr *et al. Nucleic Acids Res*. 2006 **34**: D181 [PMID: 16381841]
[17] Magrane M & Consortium U. *Database (Oxford)* 2011 **2011**: bar009 [PMID: 21447597]
[18] Garg A & Gupta D. *BMC Bioinformatics*. 2008 **9**: 62 [PMID: 18226234]
[19] Caragea C *et al. BMC Bioinformatics*. 2007 **8**: 438 [PMID: 17996106]
[20] Hua S & Sun Z. *Bioinformatics* 2001 **17**: 721 [PMID: 11524373]
[21] Peabody CR *et al. Microbiology*. 2003 **149**: 3051 [PMID: 14600218]
[22] Nguyen L *et al. J Mol Microbiol Biotechnol*. 2000 **2**: 125 [PMID: 10939240]

Edited by P Kangueane
Citation: Pundhir & Kumar. Bioinformation 6(10): 380-382 (2011)