BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(9)

Identification of Penicillin-binding proteins employing support vector machines and random forest

Vinay Nair[1], Monalisa Dutta[2], Sowmya S Manian[2], Ramya Kumari S[1] & Valadi K Jayaraman[1, 3]*

[1]Center for Development of Advanced Computing, Pune, India; [2]Rajiv Gandhi Institute of Information Technology and Biotechnology, Bharati Vidyapeeth Deemed University, Pune, India; [3]Center for Informatics, Shiv Nadar University, Dadri, UP, India; Valadi K Jayaraman - Email: valadi@gmail.com; *Corresponding author

Received November 16, 2012; Accepted April 05, 2013; Published May 25, 2013

Abstract:
Penicillin-Binding Proteins are peptidases that play an important role in cell-wall biogenesis in bacteria and thus maintaining bacterial infections. A wide class of β-lactam drugs are known to act on these proteins and inhibit bacterial infections by disrupting the cell-wall biogenesis pathway. Penicillin-Binding proteins have recently gained importance with the increase in the number of multi-drug resistant bacteria. In this work, we have collected a dataset of over 700 Penicillin-Binding and non-Penicillin Binding Proteins and extracted various sequence-related features. We then created models to classify the proteins into Penicillin-Binding and non-binding using supervised machine learning algorithms such as Support Vector Machines and Random Forest. We obtain a good classification performance for both the models using both the methods.

Keywords: Penicillin-Binding Proteins, Support Vector Machines, Random Forest, Protein Classification.

Background:
Penicillin-Binding Proteins (PBPs) have been the subject of intense research ever since their discovery as the target of the β-lactam class of drugs. Studies in *E.coli* K12 have demonstrated that PBPs play a major role in cell wall biosynthesis and affect the cell shape and cell elongation and also affect cell division **[1].** By inhibiting PBPs, the β-lactam drugs imbalance the cell wall biosynthesis pathway in bacteria and inhibit cell division and lyse the cells. Research in this area has gained further importance and urgency with various important pathogens such as *Staphylococcus aureus*, *Enterococci* and *Streptococcus pneumoniae* developing resistance to various β-lactam drugs **[2].**

Peptidoglycans are important constituent of bacterial cell walls. In bacteria, PBPs affect cell wall biogenesis by functioning as transpeptidases or carboxypeptidases in the later stages of peptidoglycan metabolism **[3].** The natural substrate of PBPs is the D-Ala-D-Ala end of the stem peptides. PBPs polymerize the glycan strand (transglycosylation) and cross-link glycan chains by virtue of its transpeptidase activity. The natural substrate of PBPs is the D-Ala-D-Ala end of the stem peptides, which it hydrolyzes (DD-carboxypeptidation). The sensitivity of PBPs to penicillin is due to the similarity in structural features shared by D-Ala-D-Ala end of the stem and penicillin. This causes the PBPs to form an extremely stable acyl bond with penicillin leading to impairment of function **[4].** The penicillin-binding/transpeptidase (TP) domains in all PBPs are characterized by the presence of three conserved motifs: SXXK with the active site serine, SXN, and KT/SG. Out of the three conserved motifs, the serine of the SXXK motif is located at the catalytic center and is involved in the actual catalysis mechanism **[5, 3].**

Depending on their molecular mass, PBPs can be categorized into high molecular mass (HMM) PBPs and low molecular mass (LMM) PBPs **(Figure 1)**. HMM PBPs are multi-modular proteins

and consist of a cytoplasmic tail at the N-terminus, a transmembrane region and domains coding for transglycosylation and transpeptidase activity **[6].** HMM PBPs can be further subdivided into Class A and Class B depending on their structure and catalytic activity. At the C-terminal, both the classes have the transpeptidase domain that catalyzes the cross-linking of glycan chains. Class A PBPs have an additional transglycosylation domain whereas Class B PBPs have a N-terminal domain. LMM PBPs have a signal peptide at the N-terminal followed by the transpeptidase domain and the transmembrane region at the C-terminal. LMM PBPs are involved in the regulation of the crosslinking of the peptidoglycans **[7].**

Support Vector Machine (SVM) is a popular supervised learning algorithm employed in various classification and regression tasks. For binary classification SVM employs a maximum margin linear hyperplane to separate the data belonging to the two classes **[8].** For nonlinearly classifiable data SVM first transforms the data to a higher dimensional feature space and subsequently employs a linear hyperplane. Further, to deal with intractability problem SVM employs appropriate kernels so that all the computations can be performed in the input space itself. The concept can be easily extended to multiclass classification problems.

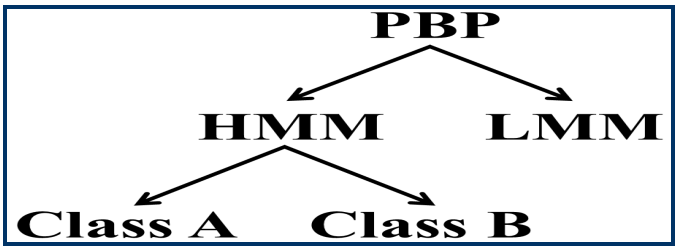

Figure 1: Classification of Penicillin Binding Proteins

Random forest (RF) **[9]** is a randomly constructed collection of independent decision trees. In the RF algorithm, randomness is introduced in two ways: while selecting the samples for making the dataset for growing the trees and while choosing the attributes to generate the subset for node splitting. Such a RF is grown in the following manner: For each tree, *'n'* Bootstrap samples (with replacement) are drawn from the original training data set to form 'In Bag' data for a particular tree, where '*n*' is the size of the training data set. In each of the Bootstrap training sets, while generating the 'In Bag' data, about one-third of the instances are unused. These are called the out of bag (OOB) data for that particular tree. Using the CART algorithm, the classification tree is then generated using the 'in bag' data. After all the trees are grown, the *k*th tree classifies the OOB instances for that tree. Subsequently, there is no need for a separate test data in RF for checking the overall accuracy of the forest. The important features of random forests are that they can handle any high dimensional and multi-class data easily. Various sequence features of proteins have been previously used to define proteins and classify them into their corresponding type/subtype. In this work, we extract a number of features from a manually-extracted and curated dataset of PBPs and pass the features to SVM and RF to generate a model that can classify proteins into non-PBPs and PBPs and its subtypes.

Methodology:

Extraction of PBP sequences and redundancy reduction: Fasta sequences of PBPs were extracted from various databases such as NCBI, UniPROT and PDB. All the PBP sequences extracted were of bacterial origin. To reduce redundancy, we used CD-HIT, a standard tool for redundancy reduction **[10].** After redundancy reduction, we had a dataset containing 377, 280, 122 and 744 sequences of Class A (HMM), Class B (HMM), LMM and non-PBP sequences respectively. We created two models, Model I for classification of proteins into PBPs and non-PBPs and Model II for sub-classification of PBPs into the respective subtypes. Model 1 consisted of 744 samples of PBP and non-PBP proteins. Model 2 consisted of 100 samples each of Class A, Class B and LMM PBP proteins.

Feature Extraction

To classify the proteins, we extracted various features from the protein sequences. The features extracted mainly involved conserved motifs/patterns, physico-chemical properties, Dipeptide / Monopeptide count and pseudo amino acid features. Conserved motifs were obtained from previously published literature. In total 19 conserved motifs/patterns were obtained. Similarly 32 physico-chemical properties were collected from various sources such as AAIndex, ExPASy etc. Chou's pseudo amino acid features **[11-13]** were also used in our set of extracted features. Additionally, we also calculated the amino-acid composition and dipeptide composition. In all, we extracted 1199 features, which included 640 physiochemical features, 19 motif-related features, 420 dipeptide and monopeptide features and 120 psuedo amino acid index features.

Support Vector Machine

To classify the dataset, we used LIBSVM **[14],** which is a popular implementation of Support Vector Machines. We used the RBF kernel for classification purposes. We scaled each feature of the dataset to a range of -1 to +1 and optimized the cost and γ parameters for the dataset using the inbuilt tools. This was followed by training SVM to generate the model and finding the accuracy for 10-fold cross-validation.

Random Forest

We also used the Weka implementation of Random Forest to classify the dataset **[15].** We generated 100 trees and 100 nodes on each tree. Using this model of Random Forest, we performed 10-fold cross-validation on the dataset

Ranking of Features

In order to select the best features and improve our model, we had ranked the features using information gain as the ranking metric. Information gain is a measure of the contribution of a particular feature to the model. Ranking using information gain was done using Weka. After ranking, the top ranked features were extracted from the feature set and passed to LIBSVM and RF for classification.

Discussion:

For binary classification of proteins into PBPs and non-PBPs, we had employed Model I. In the proposed model, after feature extraction, the individual feature sets were passed to LIBSVM and RF to create models to classify the proteins into PBPs and non-PBPs. We performed 10-fold cross validation using both

the supervised learning algorithms and saw that both the algorithms gave high classification accuracy. It was seen that some properties such as physicochemical properties and PseAA gave better classification accuracies as compared to other feature sets. In order to increase the CVA and to increase the robustness of the model, we pooled all the features and created new models. The CVA for this model was better than the highest CVA obtained using individual feature sets. Thus, we see that pooling the features increases the accuracy and robustness of the model. It is often seen that many features are noisy and do not contribute to the model. Presence of such features often leads to the classifier getting confused and giving a model which is not robust or with low accuracy. Elimination of such features leads to an increase in the CVA for the dataset. In order to eliminate such features, we ranked the features using Weka and used the top-ranked features to create a model.

We repeated the same procedure for Model II for sub-classification of PBPs into its respective subtypes. We first extracted the feature sets and created the models using LIBSVM and RF. We saw that the Dipeptide / Monopeptide count gave the highest classification. On pooling the features, we saw that there was a decrease in the classification accuracy. This might be attributed to the noisy features from the other feature sets. To remove such features, we again rank the features and create a model using the top-ranked features. We have reported the CVAs obtained in the table. We see that the use of the top-ranked features generates a model with higher prediction accuracy and increased robustness. The CVAs for the individual feature sets, the pooled and ranked features obtained using SVM and RF for Model I and Model II have been reported in **Table 1 (see supplementary material).**

Conclusion:

We extracted various features from Penicillin Binding Proteins and used these to classify proteins. Using supervised learning algorithms, we generated models using a compendium of features consisting of different types. However the models did not give the best results. We improved the models by removing the noisy features and saw that the models we obtained were robust models and gave good classification. Thus, we conclude that the features that we have extracted are features pertaining to penicillin-binding proteins and the models that we have built are robust models and can be used for classification of proteins into non-penicillin-binding proteins and penicillin binding proteins and their subtypes.

Acknowledgement:

V.K.J gratefully acknowledges financial support from Department of Science and Technology, New Delhi. V.N acknowledges Council of Scientific and Industrial Research, New Delhi for awarding a junior research Fellowsip.

References:

[1] Spratt BG, *Proc Natl Acad Sci*. 1975 **72**: 2999 [PMID: 1103132]
[2] Zapun A *et al. FEMS Microbiol Rev*. 2008 **32**: 361 [PMID: 18248419]
[3] Sauvage E *et al. FEMS Microbiol Rev.* 2008 **32**: 234 [PMID: 18266856]
[4] Goffin C & Ghuysen JM, *Microbiol Mol Biol Rev*. 1998 **62:** 1079 [PMID: 9841666]
[5] Goffin C & Ghuysen JM, *Microbiol Mol Biol Rev.* 2002 **66**: 702 [PMID: 12456788]
[6] Macheboeuf P *et al. FEMS Microbiol Rev.* 2006 **30**: 673 [PMID: 16911039]
[7] Morlot C *et al. Mol Microbiol.* 2004 **51**: 1641 [PMID: 15009891]
[8] Cortes C & Vapnik V, *Mach Learn*. 1995 **20**: 273 [DOI : 10.1007 / BF00994018]
[9] Breiman L, *Mach Learn*. 2001 **45**: 5 [DOI: 10.1023 / A:1010933404324]
[10] Huang Y *et al. Bioinformatics*. 2010 **26**: 680 [PMID: 20053844]
[11] Chou KC, *Proteins.* 2001 **43**: 246 [PMID: 11288174]
[12] Chou KC, *Bioinformatics*. 2005 **21**: 10 [PMID: 15308540]
[13] Chou KC & Cai YD, *J Chem Inf Model*. 2005 **45**: 407 [PMID: 15807506]
[14] Lin CCC & Jen C, *ACM Trans Intell Syst Technol*. 2011: 1 doi>10.1145 / 1961189.1961199
[15] Hall M *et al. SIGKDD Explorations.* 2009 **11** [doi>10.1145 / 1656274.1656278]

Edited by P Kangueane

Citation: **Nair *et al.*** Bioinformation 9(9): 481-484 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: 10-fold Cross-Validation Accuracy for Model I and Model II using SVM and RF

Feature-Set	SVM	RF
Model I : PBP/nonPBP		
Di / Mono	89.04	88.8441
Phy-Chem	90.5242	89.2473
PseAA	90.0538	88.6425
Patterns	86.3575	85.4167
Pooled	92.4731	90.5242
Best Ranked	93.75	91.378
Model II : Class A / Class B / LMM		
Di / Mono	85.667	84
Phy-Chem	80	75.333
PseAA	79	74.333
Patterns	73	75.333
Pooled	80	80
Best Ranked	86.667	81.33