Computational Biology and Chemistry 58 (2015) 199–204

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

MATEPRED-A-SVM-Based Prediction Method for Multidrug And Toxin Extrusion (MATE) Proteins

Tamanna, Jayashree Ramana*

Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form 20 July 2015
Accepted 25 July 2015
Available online 29 July 2015

Keywords:
Antibiotics
Drug resistance
MATE
PSSM
SVM
Diarrheal pathogens

ABSTRACT

The growth and spread of drug resistance in bacteria have been well established in both mankind and beasts and thus is a serious public health concern. Due to the increasing problem of drug resistance, control of infectious diseases like diarrhea, pneumonia etc. is becoming more difficult. Hence, it is crucial to understand the underlying mechanism of drug resistance mechanism and devising novel solution to address this problem. Multidrug And Toxin Extrusion (MATE) proteins, first characterized as bacterial drug transporters, are present in almost all species. It plays a very important function in the secretion of cationic drugs across the cell membrane. In this work, we propose SVM based method for prediction of MATE proteins. The data set employed for training consists of 189 non-redundant protein sequences, that are further classified as positive (63 sequences) set comprising of sequences from MATE family, and negative (126 sequences) set having protein sequences from other transporters families proteins and random protein sequences taken from NCBI while in the test set, there are 120 protein sequences in all (8 in positive and 112 in negative set). The model was derived using Position Specific Scoring Matrix (PSSM) composition and achieved an overall accuracy 92.06%. The five-fold cross validation was used to optimize SVM parameter and select the best model. The prediction algorithm presented here is implemented as a freely available web server MATEPred, which will assist in rapid identification of MATE proteins.

1. Introduction

The discovery of the antibiotics has been the most important step towards the control of infectious disease. However, with the emergence of drug resistant pathogens, currently available drugs are becoming ineffective (Putman et al., 2000). Multidrug efflux is an important mechanism of biocide and antimicrobial agent resistance in bacteria. They have been divided into various groups, which include the Major Facilitator Super (MFS) family, the Small Multidrug Resistance (SMR) family, the Resistance Nodulation and Cell Division (RND) family, the ATP binding cassette (ABC) family, and the Multidrug And Toxin Extrusion (MATE) family (Otsuka et al., 2005). Multidrug and Toxin Extrusion (MATE) proteins form a class of proteins that function as drug and proton antiporters. Initially, due to the presence of 12 transmembrane helices, they were designated as the member of MFS family. Shortly afterwards, it was reported that they showed no sequence identity to other known multidrug transporters, therefore, categorized as a new family of multidrug transporters, and are widely propagated in all realms of living beings (Omote et al., 2006). MATE proteins have been characterized as important transporters that mediate the final excretion of cationic drugs into bile and urine (Nies et al., 2012). In plants, transporter proteins from the MATE family are essential in metabolite transport, which directly changes crop yields. In bacteria and mammals, these MATE transporters facilitate multiple-drug resistance (MDR), thus regulating the efficacy of many pharmaceutical drugs used in curing a variety of diseases (He et al., 2010). MATE family transporters are conserved in the three pinion domains of life (Archaea, Bacteria and Eukarya), and export xenobiotics using an electrochemical exchange of H+ or Na+ across the tissue layer. MATE transporters confer resistance to bacterial pathogens and cancer cells, thus causing critical reductions in the curative efficacies of antibiotics and anti-cancer drugs, respectively (Tanaka et al., 2013). An example of one such protein is NorM, of *Vibrio parahaemolyticus* which is a multidrug Na +-antiporter, and was found to confer resistance to dyes, fluoroquinolones and aminoglycosides (Mohanty et al., 2012; Li and Nikaido, 2004).

As reported, MATE family efflux pumps depend upon Na+/H+ gradient for transport and have three major branches: the NorM

* Corresponding author.
E-mail address: jayashree.ramana@juit.ac.in (J. Ramana).

http://dx.doi.org/10.1016/j.compbiolchem.2015.07.011
1476-9271/

branch, a branch containing several eukaryotic proteins and a branch containing *E. coli* DinF (Mohanty et al., 2012). MATE protein length varies from 400 to 700 residues comprising of 12 transmembrane helices. In MATE proteins, there is no conserved consensus sequence; however, they share ~40% sequence similarity (Omote et al., 2006). Granting to the studies, it has been reported that sequence information for very few MATE proteins is available till date. Also, due to its primary structure heterogeneity, it is hard to recognize these proteins based on sequence. The alignment based tools like Pfam and BLAST are not sufficient to identify all the MATE proteins (as described in Section 3.1).

To combat the problem of drug resistance, it is all important to extensively understand and identify multidrug resistance proteins at a faster pace. Owing to the time limit and cost of experiments, there is a demand to have computational methods to rapidly examine and interpret relevant data (Ramana and Gupta, 2010a). In this study, we attempted to develop a prediction tool for identification of MATE proteins on the basis of Position Specific Scoring Matrix (PSSM) using Support Vector Machine (SVM). First, we used different features for generating SVMs (i) Amino Acid composition (ii) Dipeptide composition (iii) Hydrophobicity (iv) Charge composition (v) Mutliplets composition (vi) PSSM composition. However, it was observed that the performance of these SVMs was poor when compared to PSSM based SVM.

2. Methodology

2.1. Datasets generated for training

MATE proteins (assigned as positive set) and all other types of proteins (assigned as negative set) were collected through a broad and critical study of research articles from PubMed. Using CD-Hit (http://weizhong-lab.ucsd.edu/cd-hit/) (Li and Godzik, 2006) program the redundancy in both the sets was scaled down to 40%. So we had two datasets positive and negative, each comprising of 63 and 126 sequences, respectively.

2.2. Benchmark datasets for testing

For checking the efficiency of the SVM model generated, its performance was tested on independent datasets consisting of 8 positive sequences and 112 negative sequences, obtained after scaling down its redundancy to 40% against NR database.

2.3. SVM Algorithm

Support Vector Machine (SVM) is a supervised machine learning method first introduced by Vapnik in 1995 (Vapnik, 1998). SVM in combination with kernel functions is used to map input data to some vector space. In order to avoid over fitting, SVM then finds a hyperplane separating the positive data from the negative ones in high dimensional space (Ben-Hur et al., 2008).

SVM in this approach was implemented using LibSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) (Chang and Lin, 2001) which allows us to optimize a number of parameters (Ramana and Gupta, 2009) and to use kernels (e.g., linear, polynomial, radial basis function, sigmoid) for obtaining the best hyperplane (Ramana and Gupta, 2010a). In this study Radial Basis Function (RBF) kernel was used.

2.4. Five-fold cross validation

For evaluating the performance of modules generated in this study, we used five-fold cross validation in which the data is first partitioned into 5 equal sized datasets. Later, five iterations of training and validation are done such that within each iteration, a different fold of the data is held-out for validation while the remaining four-folds are used for learning (Refaeilzadeh et al., 2009). Several performance measures were then applied to evaluate the best parameters (γ and C) and then averaged to bring forth an overall assessment of the model (Ramana and Gupta, 2010b).

2.5. Performance Measures

Applying the following equations accuracy, sensitivity, specificity, Matthew correlation coefficient (MCC) and *F*-score were calculated for evaluating the accuracy of SVM classifiers:

1) **Sensitivity**: It is determined as the percentage of MATE that is correctly predicted as MATE.

Sensitivity = TP/TP + FN × 100

2) **Specificity**: It is the percentage of non-MATE that is correctly predicted as non-MATE.

Specificity = TN/TN + FP × 100

3) **Accuracy**: It is the percentage of correct predictions out of the total number of predictions.

Accuracy = TP + TN/TP + FP + TN + FN × 100

4) **Matthews correlation coefficient (MCC)**: It is a measure of both sensitivity and specificity. MCC = 0 is the indication of completely random prediction, while MCC = 1 indicates perfect prediction.

MCC = (TP × TN) − (FN × FP)/(sqrt (TP + FN) × (TN + FP) × (TP + FP) × (TN + FN)

5) **F-score**: It is the harmonic mean of precision and recall. The best value for *F*-score is 1 and worst score is 0.

F-score = 2 × Precision × Recall/(Precision + Recall)

2.6. Feature Selection

2.6.1. Composition based SVM classifiers

a) **Amino acid composition (AAC)**: It is the fraction of each of the 20 amino acids present in a protein sequence and generates an input vector of 20 dimensions.

b) **Dipeptide composition (DPC)**: It is the fraction of a dipeptide divided by the total number of possible dipeptides and gives information in the form of 400 dimensions (20 × 20).

c) **Charge composition (CC)**: It is the fraction of charged amino acids divided by the total length of the protein. The fractions of positively and negatively charged amino acids yields a fixed length input vector of 20 dimensions.

d) **Hydrophobicity composition (HC)**: Based on their hydrophobicity properties, the amino acids may be classified into five groups (Brendel et al., 1992). Moments of the positions of the five groups were calculated using the formula as below with *r* varying from 2 to 5. This yields a fixed length input vector of 25 dimensions.

Table 1
Performance of different SVM classifiers in five-fold CV (where SN: sensitivity, SP: specificity, MCC: Matthews correlation coefficient and *F*-score).

Model	C	γ	SN (%)	SP (%)	Accuracy (%)	MCC	*F*-score
AAC	5	0.06	73.02	99.21	90.47	0.78765	0.83632
DPC	4	0.01	68.25	94.44	85.71	0.67006	0.76111
CC	30	0.1	48.43	94.4	78.84	0.50581	0.60784
MPC	20	0.25	47.62	73.81	65.08	0.21428	0.25806
CH	25	0.9	27.34	96	72.75	0.34112	0.40462
ACP	10	5	76.8	89.09	74.60	0.65548	0.76042
DCP	2	6	73.68	86.36	82.53	0.59204	0.717794
PSSM	13	0.01	100	89.42	92.06	0.82436	0.86301