Journal of Theoretical Biology 415 (2017) 13–19

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC

Muslim Khan[a], Maqsood Hayat[a,*], Sher Afzal Khan[a,b], Nadeem Iqbal[a]

[a] Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan
[b] Faculty of Computing and Information Technology in Rabigh, King Abdul Aziz University, Saudi Arabia

ARTICLE INFO

Keywords:
Mycobacterium
Oversampled features
Un-biasness
Support vector machine

ABSTRACT

This study investigates an efficient and accurate computational method for predicating mycobacterial membrane protein. Mycobacterium is a pathogenic bacterium which is the causative agent of tuberculosis and leprosy. The existing feature encoding algorithms for protein sequence representation such as composition and translation, and split amino acid composition cannot suitably express the mycobacterium membrane protein and their types due to biasness among different types. Therefore, in this study a novel un-biased dipeptide composition (Unb-DPC) method is proposed. The proposed encoding scheme has two advantages, first it avoid the biasness among the different mycobacterium membrane protein and their types. Secondly, the method is fast and preserves protein sequence structure information. The experimental results yield SVM based classification accurately of 97.1% for membrane protein types and 95.0% for discriminating mycobacterium membrane and non-membrane proteins by using jackknife cross validation test. The results exhibit that proposed model achieved significant predictive performance compared to the existing algorithms and will lead to develop a powerful tool for anti-mycobacterium drugs.

1. Introduction

Membrane protein is one of the major class of basic protein classes. Experimentally, it has been reported that there are about 20–30% genomes encode membrane proteins. There are approximately 8000 estimated membrane proteins in human body. Also, these membrane proteins perform significant roles in all cellular processes. In addition, to this as membrane proteins represent 60% of drug targets. These drug targets are critical for novel drug discovery as well as for understanding of cellular functions (Yang et al., 2012, 2015; Ji et al., 2014, 2013), It has been also proven that protein is the structural and functional unit of body, so each membrane protein has its specific function on the basis of their types. For this characterization of membrane proteins, using other existing methods are time consuming, laborious and costly. Thus there is a need of an efficient computational model to characterize membrane proteins and their classes.

As mycobacterium can cause critical diseases i.e. every year millions of people die due to tuberculosis (TB) and leprosy. Although, researchers have made numerous efforts in this regard, in order to treat these diseases through bacteria and drugs, but due to huge exploration of protein sequences the prediction of mycobacterium membrane protein characterization is difficult. For such prediction several experimental approaches were carried out, because a complicated envelop that contain of a cell wall and cytoplasmic membrane act as a major role for multidrug resistance (Niederweis et al., 2010). It addition, they perform many essential biological and physiological functions, such as receptor of many hormones and as a transporter to carry material into or out of cells.

Why this computational method is being used as compared to experimental methods to recognize membrane protein, because there are some membrane protein types which cannot be crystallize and dissolve in majority of solvent. Although the recent breakthroughs in nuclear magnetic resonance (NMR) indicate that NMR is truly a very powerful tool in determining the 3D structures of membrane proteins (Oxenoid et al., 2016; Dev et al., 2016; Schnell and Chou, 2008; Berardi et al., 2011; OuYang et al., 2013; Fu et al., 2016), however, it is costly and time-consuming. That is why the computational method is needed to predict membrane proteins types. Similarly, Cryo-electron microscopy (Cryo-EM), which is the most recently developed method also used to solve the membrane protein structure.

Earlier, many computational models have been carried out for discrimination of membrane proteins on the basis of protein sequence information (Chou and Elrod, 1999; Cai et al., 2003; Wang et al., 2004; Shen and Chou, 2005; Chou and Shen, 2007; Lin et al., 2008; Chou,

* Corresponding author.
E-mail address: m.hayat@awkum.edu.pk (M. Hayat).

http://dx.doi.org/10.1016/j.jtbi.2016.12.004
Received 2 August 2016; Received in revised form 24 October 2016; Accepted 7 December 2016
Available online 08 December 2016
0022-5193/

2001; Walzer et al., 2009). Most researchers have developed many strategies for protein encoding extraction. Some of these strategies are amino acid composition (AAC) (Yang et al., 2007), pseudo amino acid composition (PseAAC) (Lin et al., 2008; Chou, 2011), split amino acid composition (SAAC) (Afridi et al., 2012), discrete wavelet analysis (DWT) (Rezaei et al., 2008), hybrid models, translation & composition (Huang et al., 2010) and tri-peptide composition (Ung and Winkler, 2011). In order to investigate the success rates of these techniques many algorithms such as support vector machine (SVM) (Chou, 2011; Ding and Dubchak, 2001; Lin, 2008; Kumar et al., 2011; Chen et al., 2009; Du et al., 2014; Hayat and Iqbal, 2014), k-nearest neighbor (KNN) (Chou, 2011), probabilistic neural network (PNN) (Hayat and Khan, 2012; Khan et al., 2008), random forest (RF) (Breiman, 2001) and Mem-EnsSAAC were used as a protein structure and function predictors (Hayat et al., 2012). Such early works suggest that machine learning algorithms performed a vital role in discrimination of mycobacterial membrane protein and their classes, but it has been used very little for discrimination of membrane protein in mycobacterium. PROB predictor has been developed Pajon et al., for identification of beta-barrel of M. tuberculosis (Pajon et al., 2006). In this work, two membrane protein functions were found undefined. Although results of this work were very good, but no one has given concentration to the identification of mycobacterial membrane proteins. Therefore, another attempt was developed for this prediction, called identification of mycobacterial membrane protein by using over-represented tri-peptide compositions by using binomial distribution function (Chen et al., 2012).

In this work, an identification algorithm for mycobacterial membrane protein and their classes as shown in Fig. 1 is developed. The proposed algorithm utilized simultaneously both oversampling technique SMOTE to remove biasness among different types of member proteins and using dipeptide compositions to extracted features from the unbiased data. This is because although we could remove some redundant samples in the model-training process, when testing the predictor, all experiment-confirmed samples must be included, even for those removed by SMOTE approach. Only by doing so, the prediction method is really validated by all the experiment-confirmed data rather than part of them. Please see the papers (Liu et al., 2015b; Xiao et al., 2015; Jia et al., 2016a, 2016b), for a detailed analysis about this.

The SVM classification algorithm applied to deal with multi-classification. The classification performance of using jackknife test obtained an overall accuracy of 97.1% for mycobacterium membrane protein classes and 95.0% accuracy for mycobacterium membrane and non-membrane protein accordingly. Furthermore, this study extends previous research as follows.

i) In order to avoid the biasness among protein membrane types this paper used the over sampling technique SMOTE.
ii) This study considers numerous feature extraction algorithms on datasets. The empirical result shows that un-biased dipeptide composition performance is far better than other feature extraction algorithms.
iii) In order to dimensional reduction this paper used mRMR as feature selection algorithm. Also, the data are collected from two different datasets to evaluate the comparative performance with existing algorithms.

The results of this study will further improve the predication of mycobacterial membrane protein (Xiao et al., 2016; Qiu and Sun, 2016; Qiu et al., 2016; Chen et al., 2016a), and their types and will be helpful for development of anti-mycobacterium drugs.

The rest of the paper is structured as follows, Section2 represents material and methods, Section 3 shows feature selection algorithms, Section 4 and Section 5 presents classification algorithms and performance evaluation criteria respectively, Section 6 is about result and discussion, Finally, Section 7 draws conclusion.

2. Materials and methods

2.1. Dataset

In order to get precise results, an appropriate benchmark dataset is required for training and testing the computational model. In this regards, a standard dataset of mycobacterium membrane protein is used. It is constructed from universal protein Resources (uniProt) database (Magrane, 2011). For accurate and well define dataset some instructions were followed. First of all, those sequences were selected which were achieved and manually annotated by the researchers, secondly, those sequences were excluded whose type is not defined, further those sequences were knocked which were fragment of other proteins (Han et al., 2006). After this, a précise and accurate dataset was generated. The database consists of two benchmark datasets namely dataset-I and dataset-II. The dataset-I contain 274 sequences, out of which 32 sequences are single pass, 192 are multi-pass, 20 are lipids anchor and 30 are peripheral membrane protein, while in dataset-II there are 564 sequences out of which 274 are membrane protein types and 290 are non-membrane proteins.

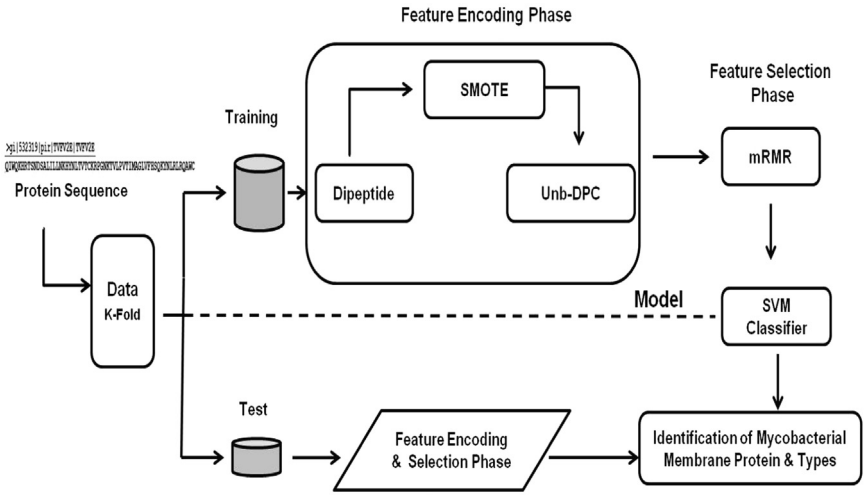

Fig. 1. Schematic diagram of the proposed algorithm.