Computers in Biology and Medicine 43 (2013) 1502–1511

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble classification

Muhammad Tayyeb Mirza [a], Asifullah Khan [a], Muhammad Tahir [a], Yeon Soo Lee [b,*]

[a] Pattern Recognition Laboratory, Department of Computer and Information Sciences, PIEAS, Nilore, Islamabad, Pakistan
[b] Department of Biomedical Engineering, College of Medical Science, Catholic University of Daegu, 330 Geumrak 1-ri, Hayang-eup, Gyeongsan-si, Gyeongbuk 712-702, Republic of Korea

ARTICLE INFO

Article history:
Received 22 April 2013
Accepted 24 July 2013

Keywords:
Plasmodium falciparum
Mitochondrial proteins
Bi-profile Bayes
PseACS
PseAAC
SAAC
SVM
Ensemble classification

ABSTRACT

Mitochondrial protein of *Plasmodium falciparum* is an important target for anti-malarial drugs. Experimental approaches for detecting mitochondrial proteins are costly and time consuming. Therefore, *MitProt-Pred* is developed that utilizes Bi-profile Bayes, Pseudo Average Chemical Shift, Split Amino Acid Composition, and Pseudo Amino Acid Composition based features of the protein sequences. Hybrid feature space is also developed by combining different individual feature spaces. These feature spaces are learned and exploited through SVM based ensemble. *MitProt-Pred* achieved significantly improved prediction performance for two standard datasets. We also developed the score level ensemble, which outperforms the feature level ensemble.

1. Introduction

Malaria is one of the most significant parasitic diseases in the human society. According to the world malaria report 2011, a total of 216 million cases of malaria were reported in the year 2010, out of which 655 thousand people died [1]. Malaria is a mosquito born infectious disease caused by the eukaryotic protists (a group of microscopic organisms) belonging to the genus *Plasmodium*. This disease is transmitted to human beings by the female mosquito belonging to the genus *Anopheles*, which acts as a vector. Examples of mosquito belonging to this genus causing malaria are *Anopheles stephensi* and *Anopheles gambiae*. *Anopheles gambiae* is one of the best-known vectors of this disease. There exists five species of the genus *Plasmodium* responsible for transmitting malaria to human beings. These species include *Plasmodium falciparum (PF)*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium knowlesi*. The most deadly reported cases are caused by *PF*.

According to the world malaria report of 2008, more than 75% cases in sub-Saharan Africa were caused by *PF* [2]. Despite the fact that we need vaccine in order to fight this disease, no cure has yet been discovered against this parasite. Drugs do exist to fight this disease but with the passage of time *PF* develops resistance against these anti-malarial drugs [3]. For example, Chloroquine drug has been using for many decades as a treatment against the malarial disease but *PF* has spread in many areas showing resistance to this cure. Similarly, the resistance against Artemisinin has also been observed. Thus, it is necessary to find new targets in *PF*, which can be used in developing novel anti-malarial drugs. The simple proposition is 'to destroy *PF*, we need to stop the functioning of power source of.

Mitochondria are the power house of a cell [4]. Therefore, targeting the mitochondria of a malaria parasite can be considered prospective. In this way, one can destroy the structure of mitochondria, which may in turn change the functionality of mitochondria. This means one can put an end to the power supply of the cell. Consequently, the cell will not be able to perform its functions and will eventually die out. Thus, mitochondrial proteins of *PF* can be considered as a likely target. Here, a concern also arises against the development of this kind of drug. The drugs that will be able to destroy the mitochondria of *PF* may also destroy the mitochondria of human beings. Fortunately, proteins of mitochondria in *PF* are different from the proteins of mitochondria in human. This in essence makes mitochondrial proteins of *PF* an important target for anti-malarial drugs [3,5]. However, the foremost query arises is that what the nature of these proteins is and in what aspects they are different from other proteins. What are

* Corresponding author. Tel.: +82 53 850 3447; fax: +82 53 850 3291.
E-mail addresses: yeonsoolee@cu.ac.kr, khan.asifullah@gmail.com (Y.S. Lee).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.07.024

their characteristics/features, and how can we identify these proteins based on these characteristics/features. In order to accomplish this, we must be able to recognize any given protein as mitochondrial/non-mitochondrial protein of malaria parasite. One way to identify the protein is the experimental approach. However, this approach is complex, time consuming and error prone due to subjective analysis. Alternatively, we can adopt bioinformatics and machine learning strategies for the prediction of mitochondrial proteins of *PF*. Consequently, annotation of mitochondrial proteins of malaria parasite becomes an important task.

2. Background

Prediction of mitochondrial proteins has largely been performed using statistical techniques and machine learning approaches. Such techniques use sequence and biological information simultaneously. Several works are already proposed for sequence based classification using ensemble [4,6–8]. In addition, there exist various techniques including Target P, Signal P3.0, WoLF PSORT, TargetLoc, MitoProt II, MITOPRED, MitPred, and Mito-GSAAC, which are all organelle specific methods [4,9–15]. In fact, they are developed to differentiate between mitochondrial and non-mitochondrial proteins. However, due to the difference between the human and *PF* mitochondria, we need the combination of organelle and organism specific methods. The combination of organelle and organism specific methods showed better performance compared to the organelle specific methods in predicting the localization of protein sequences [16–18]. Current methods consider both the organelle and organism specific features, including PlasMit [16], PFMpred [18] and the method developed by Chen et al. [19]. All of these methods were reported using a dataset developed by Bender et al. that contains 40 mitochondrial and 135 non-mitochondrial proteins of *PF* [16]. PlasMit is a neural network based system that takes the amino acid composition (AAC) of 24-N terminus amino acids of a protein. PlasMit used 20-fold cross validation test and yielded an accuracy of 90.0%. On the other hand, PFMpred uses Split Amino Acid Composition (SAAC) along with position specific scoring matrix (PSSM) based features and support vector machine (SVM) as a prediction system [16]. Employing a 5-fold cross validation test, the prediction accuracy of PFMpred was 92.0% [18]. Chen et al.'s proposed system achieved the accuracy of 92.0% [19], which used the increment of diversity to predict the mitochondrial proteins. However, the performance of this method was tested on a dataset containing limited number of positive samples. Recently, Jia et al. developed a dataset, which contains 108 mitochondrial proteins and 125 non-mitochondrial proteins [17]. They employed Bi-profile Bayes (BpB) and SAAC as input features to the SVM. Jackknife test was adopted as a cross validation technique and the reported accuracy was 90.99%. Mitochondrial prediction of Jia et al. is quite effective; however, our hypothesis in the current paper states that there is still some margin of improvement.

3. Research objective

The aim of this work is to propose an accurate and more effective prediction system for predicting mitochondria of *PF*. In this work, we adopted two different approaches. In one approach, we utilized the discriminative power of individual feature spaces including BpB, PseACS, SAAC, and PseAAC based features, as well as we constructed a hybrid feature space of BpB and PseAAC based features, which have the discrimination power of both the feature spaces. For the hybrid model, the BpB and PseAAC based features were adopted since they have higher prediction accuracies compared to PseACS and SAAC features. Then SVM is trained on this hybrid feature space. In the other approach, we trained SVM on BpB, PseACS, SAAC, and PseAAC features individually. Then the predictions of all these SVMs were combined through the majority-voting scheme. In this work, the former approach is called 'features level ensemble approach', whereas the later one is called 'scores level ensemble approach'. In addition, we have combined the highest performing features from the set of these four descriptors and obtained the final prediction using the sum rule. The parameters of SVM were tuned using non-dominated sorting genetic algorithm II (NSGA-II).

4. Materials and methods

This section provides details about the used datasets and discusses the proposed technique followed by feature extraction strategies and classification algorithm.

4.1. Datasets

The datasets, used to train and test the proposed prediction system, is adopted from a non-redundant dataset developed recently by Jia et al. [17]. Originally, Jia et al. have extracted 132 mitochondrial and 272 non-mitochondria proteins from geneDB website (http://www.genedb.org/). They have applied 25% similarity threshold using BLASTclust [20] in order to execute the homology reduction. This resulted in 109 mitochondrial and 127 non-mitochondrial proteins. In the current study, only those proteins from this dataset were selected that have length greater than or equal to 60 amino acids. Consequently, the dataset is further reduced to 108 mitochondrial and 125 non-mitochondrial proteins making a total number of 233 proteins, which is termed as DS233 in our work.

Another dataset developed by Bender et al. [16] contains 40 mitochondrial proteins as positive subset and 135 non-mitochondrial proteins as negative subset. This dataset is not balanced in terms of positive and negative instances. This dataset will be referred as DS175 in discussions onwards.

4.2. The proposed MitProt-Pred prediction system

Fig. 1 illustrates the proposed prediction system, which will be referred to as *MitProt-Pred* in the rest of the paper. In the feature extraction phase, BpB, PseACS, SAAC, and PseAAC features are extracted. BPB features are computed with two different configurations of amino acids on N and C termini that are 25, 25 and 30, 30, respectively. PseACS feature computation is based on the value of lambda for four different types of atoms. SAAC features are extracted as 20, 25, and 30 amino acids on each of the N and C termini. PseAAC features are computed with different numbers of tiers and physiochemical properties. The best performing model was then selected for the classification of *PF*. The features are forwarded to the classification phase where SVM is utilized to recognize the patterns of a particular class.

The cost and gamma parameters of SVM are optimized with NSGA-II. Since each feature vector forms a separate feature space therefore, we obtained the optimized values of cost and gamma parameters separately for each feature space. During the use of NSGA-II, we utilized all the data as training data and then with the optimized values of cost and gamma parameters we utilized LIBSVM package for testing the performance of the proposed model. The individual components of *MitProt-Pred* are discussed in the following sections.