Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

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

Q1 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*. *A. 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 *P. falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium knowlesi*. The most deadly reported cases are caused by the *P. falciparum* (*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 *PF*'.

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 Published by Elsevier Ltd.
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.

Please cite this article as: M.T. Mirza, et al., MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble..., Comput. Biol. Med. (2013), http://dx.doi.org/10.1016/j.compbiomed.2013.07.024

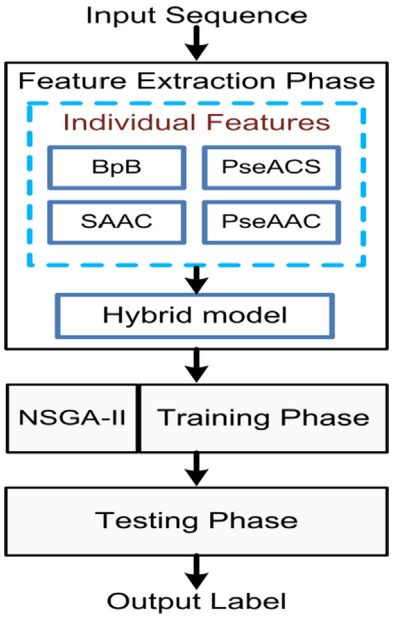

Fig. 1. Block-diagram of the proposed MitProt-Pred.

4.2.1. Feature extraction techniques

In the field of machine learning and pattern recognition, particular patterns are first extracted from objects to be classified. For this purpose, researchers have developed numerous feature extraction strategies [21–24]. In this work, we utilized BpB, PseACS, SAAC, and PseAAC based features, which are explained next.

(1) Bi-profile Bayes features.
Bi-profile Bayes (BpB) technique provides a noteworthy improvement in the performance of a predictor. These features have been used in the annotation of protein methylation sites [25], caspase cleavage sites [26] and mitochondrial protein prediction of *PF* [17]. In BpB, $S=s_1, s_2, \ldots, s_n$ represents the protein sequence, where s_i $(i=1, 2, 3, \ldots, n)$ represents single amino acid, and n is the total number of amino acids selected on both the N-terminus and the C-terminus of a protein sequence. S belongs either to C_1 or to C_{-1} where C_1 represents the mitochondrial proteins and C_{-1} represents the non-mitochondrial proteins. A protein sample S corresponds to the feature vector $\boldsymbol{p}=(p_1, p_2, p_3, \ldots, p_n, p_{n+1}, p_{n+2}, p_{n+3}, \ldots, p_{2n})$. Each of $p_1, p_2, p_3, \ldots, p_n$ represents the posterior probability of each amino acid at each position in the set of mitochondrial proteins belonging to C_1. Similarly, each of $p_n, p_{n+1}, p_{n+2}, p_{n+3}, \ldots, p_{2n}$ represents the posterior probability of each amino acid at each position in the set of non-mitochondrial proteins belonging to C_{-1}. According to Bayes' rule, the posterior probabilities of S for these two categories are given by the following:

$$P(C_1|S)=\frac{P(S|C_1)P(C_1)}{P(S)} \quad (1)$$

$$P(C_{-1}|S)=\frac{P(S|C_{-1})P(C_{-1})}{P(S)} \quad (2)$$

where $P(C_1)$ and $P(C_{-1})$ are the prior probabilities of mitochondrial and non-mitochondrial proteins, respectively. $P(S|C_1)$ and $P(S|C_{-1})$ are the likelihoods. If we assume that s_j are mutually independent for $j=1, 2, 3, \ldots, n$, then Eqs. (1) and (2) can be rewritten as follows:

$$P(S|C_1)=\prod_{j=1}^{n} P(s_j|C_1) \quad (3)$$

$$P(S|C_{-1})=\prod_{j=1}^{n} P(s_j|C_{-1}) \quad (4)$$

While computing BpB features, the number of selected amino acid residues, in this work, on each of N and C-termini of a protein sequence are given as follows: (30, 30), (30, 0), (25, 25), (25, 0), (20, 20), (20, 0), (0, 30), (0, 25), and (0, 20).

(2) Pseudo Average Chemical Shift features.
In a molecule, the chemical shift of an atom is the measure of relative nuclear energy level, which is due to the surrounding environment. Protons are sensitive to their chemical environment; therefore, in different chemical environments, protons undergo slightly different magnetic fields. Due to the difference in magnetic fields, the protons are absorbed at different frequencies [27]. This variation in nuclear magnetic resonance frequency, due to the variation in electron cloud distribution, is called the chemical shift.
Chemical shift is measured by the Nuclear Magnetic Resonance (NMR) spectroscopy. They are sensitive to local environment, so chemical shift can be used to indicate the information in its surroundings. The Average Chemical Shift (ACS) correlates well to the protein's secondary structure or backbone dihedral angles [28,29]. In order to compute Pseudo Average Chemical Shift (PseACS) features, we first obtain the protein's secondary structure using Porter [30], a server for protein's secondary structure prediction. It can be accessed online using the link: http://distill.ucd.ie/porter/.
Let us represent a protein sequence A of L amino acids by P, on the availability of the secondary structure of protein sequence A, the ACS can replace every amino acid in A. Accordingly, P can be expressed as follows:

$$P=[A_1^i, A_2^i, A_3^i, \ldots, A_L^i] \quad (i={}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}) \quad (5)$$

where N stands for nitrogen, C_α for alpha carbon, H_α for alpha hydrogen, and H_N for hydrogen linked with nitrogen. After replacing each amino acid by its ACS, the protein sequence is represented as follows:

$$P_{PseACS}=[\varphi_i^0, \varphi_i^1, \varphi_i^2, \ldots, \varphi_i^\lambda] \quad (i={}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_\mathrm{N}; \lambda<L) \quad (6)$$

$$\varphi_i^\lambda=\frac{1}{L-\lambda}\sum_{k=1}^{L-\lambda}[A_k^i-A_{k+\lambda}^i]^2 \quad (i=1,2; \lambda<L) \quad (7)$$

where i indicates the atom selected with chemical shift and it can be any one of the four or their combination. After obtaining the secondary structure of protein sequence, we forward it to PseACS web server along with the original protein sequence to obtain the PseACS features [27,31]. PseACS web server is accessible at http://wlxy.imu.edu.cn/college/biostation/fuwu/PseACS/index.asp. In this work, we have computed PseACS for all the mentioned atoms with different values of λ as shown in Table 2.

(3) Split Amino Acid Composition
Split Amino Acid Composition (SAAC) based features have been used successfully in many works for the prediction of protein functions [4,17,18]. SAAC is the successor of Amino Acid Composition (AAC) based features [24]. In SAAC, a protein sequence is first decomposed into its constituent parts and then the composition of each fragment is figured out separately. The given protein sequence is usually split into three parts: N-terminus (amino-terminus), C-terminus (carboxyl-terminus), and the internal segment [4]. The AAC of each of these three parts is calculated separately resulting in the 60D feature vector. In fact, the AAC of each part gives 20D vectors; the three parts collectively form 3-by-20D=60D vector.
The 20D feature vector can be represented as follows:

$$P=[f_1, f_2, f_3, \ldots, f_{20}] \quad (8)$$

where f represents the occurrence frequency of each amino acid. In SAAC, Eq. (8) can be written as follows:

$$P = [f_1^N, \ldots, f_{20}^N, f_1^{\text{int}}, \ldots, f_{20}^{\text{int}}, f_1^C, \ldots, f_{20}^C,] \quad (9)$$

where N indicates N-terminus, int represents the internal segment and C stands for C-terminus. In this work, the length of N or C termini is selected variably as 20, 25, and 30 amino acids.

(4) Pseudo Amino Acid Composition

Pseudo Amino Acid Composition (PseAAC) based features were proposed in order to replace simple amino acid composition based features [32]. PseAAC has successfully been used by many researchers [6,19,31,33,34]. In simple amino acid composition, we take the occurrence frequency of all the 20 amino acids, and hence it includes 20 components. In PseAAC, we take into account not only the 20 features of simple amino acid composition but also some physiochemical properties of protein sequences, which add up more components in the feature space. Due to these additional components, the order and length information of protein sequences are preserved [33]. The type of PseAAC, which we have considered in this work, is called *series-correlation type* [22,23]. A mitochondrial protein sequence P with total number of L amino acids can be represented as given in Eq. (10).

$$P = A_1, A_2, A_3, \ldots, A_L \quad (10)$$

where A_i represents the amino acid at position i ($i=1, 2, \ldots, L$). Its respective PseAAC feature vector will be as follows:

$$\text{PseAAC} = P_1, P_2, \ldots, P_{20}, \ldots, P_{\Lambda} \quad (11)$$

where $\Lambda = 20 + n \times \lambda$ (λ is the total number of tiers used in PseAAC and n is the number of physiochemical properties used for mitochondrial protein sequence). The classification performance may be affected by the number of selected tiers λ and physiochemical properties n. The first 20 elements P_1, P_2, …, P_{20} are the amino acid frequency components. The next elements P_{21}, P_{22}, …, P_{Λ} are the first-tier to λ-tier correlation factors of amino acid sequence, which are determined on the basis of physiochemical properties. In this work, we consider five physiochemical properties including hydrophobicity, hydrophilicity, mass, electronic and bulk properties. We have chosen hydrophobicity and hydrophilicity properties because the functioning of proteins in a cell is dependent on the arrangement of these hydrophobic and hydrophilic amino acids in a protein sequence. In PseAAC, we look for the tier-correlation of hydrophobic and hydrophilic amino acids in order to see the arrangement in which these amino acids are linked with each other. Likewise, the mass distribution in a protein sequence is demonstrated by the Mass property of amino acids. Similarly, the bulk property exhibits the composition, polarity and molecular volume of amino acids. The polarity and volume have direct effect on the folding property of protein sequences. In the same way, the electron ion interaction pseudo-potential (EIIP) is utilized as electronic property. Besides, the literature reveals that the aforementioned properties exist among the important and discriminative properties expressing the PseAAC features. Therefore, we adopted these properties for building the proposed model.

These five properties have been assessed in different combinations as given below:

- Hydrophobicity and hydrophilicity (HH).
- Hydrophobicity, hydrophilicity and mass (HHM).
- Bulk and electronic (BE).
- Hydrophobicity, hydrophilicity, mass, bulk and electronic (HHMBE).

(5) The hybrid model

Besides, we constructed a hybrid model of BpB features and PseAAC features. The hybrid model may lead to a better performance because they utilize the discriminative power of both the feature spaces simultaneously.

4.2.2. Classification

In a classification task, a classification system distributes data into predetermined groups. Each group has its own characteristics based on the attributes of available data. In this work, we utilize SVM for the classification of mitochondrial proteins of *PF*. We now provide detailed discussion about SVM as follows.

4.2.2.1. Support vector machine. SVM is a machine learning technique, which is based on the statistical learning theory [35]. It has been successfully used in many fields of machine learning [36], pattern recognition, and bioinformatics such as GPCRs hierarchical classification [33,34], mitochondrial protein prediction [4,17], protein methylation sites [25], and prediction of membrane protein types [24]. SVM is inherently a binary classifier; therefore, it was utilized efficiently for the prediction of mitochondrial proteins of malaria parasite [17,18].

The SVM algorithm builds a decision hyper-plane that has a maximum distance to the closest points in the training dataset. The principle is to find the hyper-plane such that the training classification error is minimized. Therefore, the classification problem is solved as a quadratic optimization problem. The package used for our SVM training and prediction was LIBSVM [37]. For example, we have N training pairs (x_i, y_i) where $x_i \in R^N$ and $y_i \in \{-1, 1\}$ then the decision surface can be formulated as follows:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i^T . x + bias, \quad \alpha_i > 0 \quad (12)$$

where α is the Lagrange multiplier. In case of linearly separable classification problem, SVM adopts the notion of dot product of two points in the input space as a kernel function. In case of non-separable classification problem, the hyper-plane is computed as follows:

$$\varphi(W, \zeta) = 1/2 W^T W + C \sum_{i=1}^{N} \zeta_i \quad (13)$$

Provided that the following condition is $\text{mety}_i(W^T \varphi(x_i) + b) \geq 1 - \zeta_i$, $\zeta_i > 0$ where $C > 0$ is the penalty term of $\sum_{i=1}^{N} \zeta_i$ and $\varphi(x)$ is the nonlinear mapping. Weight vector w is used to minimize the cost function term $W^T W$. SVM transforms nonlinear data from the low dimensional space N to higher dimensional space M using $\varphi(x)$ such that $\varphi : R \rightarrow F^M$, $M > N$. In this work, radial basis kernel function (RBF) was utilized for SVM training. RBF is defined as follows:

$$K(x_i, x_j) = \exp\{-\gamma ||x_i - x_j||^2\} \quad (14)$$

where the parameter gamma 'γ' indicates the width of Gaussian function. These parameters were optimized using NSGA-II algorithm. The objective was to minimize the total support vector count of the model. Support vector count has a direct relation with the space and time complexity of SVM machine. Removing a non-support vector from the training dataset does not affect the model built by the machine [38]. If *Nsv* represents the number of support vectors, then the support vector count T_{nSV} is formulated as given in [38]:

$$T_{nSV} = \frac{Nsv}{l} \quad (15)$$

where l is the total number of samples in the training dataset. The parameter optimization package NSGA-II algorithm is described briefly in the following section.

4.2.3. Non-dominated sorting genetic algorithms-II

The non-dominated sorting genetic algorithm II (NSGA-II) is one of the efficient optimization multi-objective evolutionary algorithms (MOEAs). The main advantage of this algorithm is that it has very low computational complexity [39]. In addition, it is simple to use and more efficient than other genetic algorithms (GA), therefore, we have employed it in this work. The main procedure of NSGA-II is given below:

Step 1 Initialize chromosomes.
Step 2 Multiple objectives evaluation.
Step 3 Apply operator for non-dominated sorting.
Step 4 While (condition is true).
  (a) Selection operator.
  (b) Mutations and crossovers.
  (c) Offspring chromosome is evaluated.
  (d) Non-dominated sorting on intermediate chromosome (initial and offspring chromosomes).
  (e) Initial chromosome is selected and changed.
  (f) Go to Step 4 until criterion is met.
Step 5 End.
Step 6 Most optimum solution is chosen.

In our work, the optimization has only one objective that is the minimization of total support vector count. In optimizing the T_{nSV}, the population size was set to 50 and total number of generations to 100. The values of c and γ were optimized and used in training the SVM model.

5. Results and discussion

In classification problems, cross validation methods are often used to assess the performance of a classifier. Among various cross validation methods, the jackknife test is considered the most objective one that always generates discernment results. Hence, researchers are increasingly using it for assessing the success of different predictors. The current study also adopts this strategy for DS233 and DS175 datasets. In jackknife test, model is trained on $N-1$ samples and tested on the remaining one sample where N is the total number of instances in the dataset. The procedure is repeated N times for each sample. Further, for comparison purpose we have also adopted 5-fold cross validation test for DS175.

Performance metrics, used to measure the performance of the propose model, are accuracy, sensitivity, specificity, Mathew's Correlation Coefficient (MCC) and Area under the ROC curve.

5.1. Performance analysis on DS233

In this section, we analyze the performance of *MitProt-Pred* for *DS*233 using Bi-profile Bayes, SAAC, PseAAC, and PseACS based feature extraction strategies. For the analysis of each feature extraction strategy, we devote a separate section to discuss their performance on pattern extraction from mitochondrial protein sequences.

In all the tables presenting the results, Acc, Sen, Spe, and MCC represent percent accuracy, percent sensitivity, percent specificity and Mathew's Correlation Coefficient, respectively.

5.1.1. Performance analysis of MitProt-Pred using individual feature spaces

Table 1 presents the classification results of *MitProt-Pred* using Bi-profile Bayes (BpB) features for DS233. First two columns show the number of selected amino acids on each of the N and C termini during feature extraction, respectively.

MitProt-Pred achieved the highest accuracy of 89.2% using BpB features selecting 30, 30 amino acids on both the N-terminus and the C-terminus. BpB discriminates the classes with a good accuracy because it considers the peptide sequence features, which are found in both the mitochondrial proteins and the non-mitochondrial proteins. Other performance measures have also indicated good performance of *MitProt-Pred* using these features. For example, MCC value of 0.78 shows a reasonable balance among the true positive and true negative rates of the proposed system as can also be observed from the fourth and fifth columns and the second row of Table 1.

Performance predictions of *MitProt-Pred* using *Pseudo Average Chemical Shift* features are shown in Table 2. The first column is devoted to show the value of λ, whereas the second column shows the atom name for which PseACS features are extracted. The highest accuracy of 57.6% is achieved for the value of $\lambda=25$ and atom $^{1}H_{N}$. The obtained results revealed that the discriminative power of PseACS features for DS233 is less compared to other utilized features. However, these features might be diverse in nature as can be proven by the results of score level ensemble presented at the end of this section. The sensitivity and specificity values show that the true positives and true negatives predicted by *MitProt-Pred* are not promising. This is due to the fact that these features do not possess discriminating power of differentiation between mitochondrial and non-mitochondrial proteins of *PF*.

Table 3 demonstrates the output predictions of *MitProt-Pred* using SAAC features. First column shows the number of amino acids selected on each of the N and C termini during the feature extraction. The highest accuracy of 70.3% has been obtained for SAAC features with 25 amino acids on each of the N and C termini.

Table 1
Performance of *MitProt-Pred* using Bi-profile Bayes features for DS233.

N-terminus amino acids	C-terminus amino acids	Acc	Sen	Spe	MCC
30	30	**89.2**	88.6	89.8	0.78
30	0	82.4	81.9	82.8	0.64
0	30	82.4	81.9	82.8	0.69
25	25	84.9	85.0	84.8	0.74
25	0	87.1	86.1	87.9	0.63
0	25	81.1	80.8	81.3	0.66
20	20	81.5	81.9	81.1	0.64
20	0	83.2	83.2	83.2	0.62
0	20	82.8	81.2	84.2	0.65

Table 2
Performance of *MitProt-Pred* using PseACS features for DS233.

λ	Atom	Acc	Sen	Spe	MCC
12	$^{1}H_{\alpha}$	52.1	49.1	54.7	0.03
	$^{1}H_{N}$	53.8	50.2	56.8	0.07
	$^{13}C_{\alpha}$	52.9	49.9	55.6	0.05
	^{15}N	51.6	49.8	53.4	0.03
16	$^{1}H_{\alpha}$	51.6	49.8	53.4	0.03
	$^{1}H_{N}$	53.8	50.6	56.4	0.07
	$^{13}C_{\alpha}$	50.0	46.9	52.5	−0.005
	^{15}N	51.2	49.8	52.4	0.02
20	$^{1}H_{\alpha}$	54.2	52.0	56.2	0.08
	$^{1}H_{N}$	55.5	54.4	56.4	0.10
	$^{13}C_{\alpha}$	50.4	48.2	52.3	0.006
	^{15}N	47.8	44.9	50.3	−0.04
25	$^{1}H_{\alpha}$	50.4	48.9	51.5	0.00
	$^{1}H_{N}$	57.6	55.5	59.5	0.15
	$^{13}C_{\alpha}$	51.2	47.3	54.7	0.02
	^{15}N	50.8	50.3	51.4	0.01

Please cite this article as: M.T. Mirza, et al., MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble..., Comput. Biol. Med. (2013), http://dx.doi.org/10.1016/j.compbiomed.2013.07.024

Table 3
Performance of *MitProt-Pred* using SAAC features for DS233.

N/C terminus amino acids	Acc	Sen	Spe	MCC
20	69.4	67.2	71.3	0.38
25	**70.3**	69.0	71.3	0.40
30	68.6	67.9	69.3	0.37

Table 4
Performance of *MitProt-Pred* using PseAAC features for DS233.

Physiochemical properties	Tier	Acc	Sen	Spe	MCC
BE	1	**74.1**	71.8	76.0	0.48
	5	73.3	71.4	74.8	0.46
	10	72.4	70.5	74.0	0.44
	15	72.4	71.0	73.6	0.44
	20	72.8	70.7	74.6	0.45
HH	1	66.1	64.1	67.6	0.31
	5	69.9	68.3	71.1	0.39
	10	64.8	62.8	66.6	0.29
	15	69.4	67.2	71.3	0.38
	20	65.6	62.7	68.2	0.31
HHM	1	71.1	68.1	73.7	0.42
	5	69.0	66.1	71.5	0.37
	10	67.7	66.1	69.3	0.35
	15	66.1	63.2	68.4	0.31
	20	66.9	64.1	69.3	0.33
HHMBE	1	73.7	71.4	75.8	0.47
	5	73.7	71.6	75.4	0.47
	10	70.7	69.2	71.9	0.41
	15	71.1	69.0	72.9	0.42
	20	66.5	63.4	69.0	0.32

The capital letters in the first column represent bulk (B), hydrophobicity (H), hydrophilicity (H), mass (M), and electronic (E).

The MCC value of 0.40 confirms the reasonably good performance of these features. The true positive and negative rates are also good for the other two SAAC features.

The performance predictions of *MitProt-Pred* using PseAAC based features for different combinations of physiochemical properties are shown in Table 4. The first column shows various combinations of physiochemical properties of PseAAC, and the second column demonstrates the tier value for that specific combination.

The tier values are selected to be 1, 5, 10, 15, and 20. *MitProt-Pred* achieved the highest accuracy of 74.1% for the combination of bulk and electronic properties with tier value of one. The MCC value of 0.480, the true positive value of 71.8%, and the true negative value of 76.0% also confirm the discriminative power of combined Bulk and Electronic properties compared to other combinations. PseAAC based features have contained the knowledge of length and order of amino acid sequences in order to classify the *PF* proteins with good accuracy.

This signifies these characteristics of PseAAC since it takes into account not only the physiochemical properties of the amino acid sequence but also preserves the sequence order and sequence length information. Similarly, combining the physiochemical properties of the same sequence extracts more information.

5.1.2. Performance analysis of MitProt-Pred through the feature level ensemble

Table 5 demonstrates the prediction performance of *MitProt-Pred* using the hybrid model of *PseAAC* with different combinations of physiochemical properties and Bi-profile Bayes features with 30 amino acids selected on each of the N and C termini. The first column of Table 5 shows various combinations of physiochemical properties in PseAAC, and the second column represents the values of tier in PseAAC.

Table 5
Performance of *MitProt-Pred* using the hybrid model of PseAAC and Bi-profile Bayes features having 30 amino acids on each of N and C termini for DS233.

Physiochemical properties	Tiers	Acc	Sen	Spe	MCC
HH	1	89.2	89.5	89.0	0.78
	5	87.1	86.7	87.4	0.74
	10	87.9	87.7	88.1	0.76
	15	87.5	87.6	87.5	0.75
BE	1	90.1	90.4	89.8	0.80
	5	89.2	90.2	88.4	0.78
	10	88.4	88.5	88.2	0.77
	15	88.4	90.2	87.7	0.78
HHMBE	1	89.2	89.5	89.0	0.78
	5	89.2	89.5	89.0	0.78
	10	89.7	89.6	89.7	0.79
	15	91.4	92.3	90.7	0.83
HHM	1	90.1	91.2	89.2	0.80
	5	**93.1**	91.0	**94.9**	**0.86**
	10	88.4	87.1	89.5	0.77
	15	88.4	87.8	88.8	0.77

The capital letters in the first column represent bulk (B), hydrophobicity (H), hydrophilicity (H), mass (M), and electronic (E) in PseAAC.

MitProt-Pred has achieved the highest accuracy of 93.1% using the combination of hydrophobicity, hydrophilicity and mass properties of PseAAC in conjunction with Bi-profile Bayes features. The accuracy is 3.90% higher than the highest accuracy achieved by BpB features having 30 amino acids on each of the N and C termini among the individual feature spaces as shown in Table 1. This signifies the characteristics of PseAAC since it takes into account not only the physiochemical properties of the amino acid sequence but also preserves the sequence order and sequence length information. Similarly, the hybrid model possesses the peptide sequence information of both the classes as well as their information about physiochemical properties. The combined feature space performed well compared to the individual constituents of this combined feature space. The prediction performance is better compared to each of the Bi-profile Bayes and PseAAC features as shown in Tables 1 and 4, respectively.

Through the analysis of principal components 1 and 2, we observe that the two components are well separated and easily be distinguishable from each other as shown in Fig. 2.

Similarly, simulation results are obtained for the hybrid model of *PseAAC* with different combinations of physiochemical properties and Bi-profile Bayes features with 25 amino acids selected on each of the N and C termini as shown in Supplementary Table 1. The performance of this hybrid model is also better than that of the individual constituents. However, less impressive as compared to the predictions presented in Table 5. Figs. 3 and 4 visualize the results of Table 5 and Supplementary Table 1, respectively, for different tier values used in PseAAC features.

Area under the ROC curve (AUC) is also calculated for the best performing individual and hybrid feature spaces as shown in Fig. 5. This strengthens our findings regarding the best performance of our proposed MitProt-Pred system. The highest AUC value of 0.97 is obtained using the hybrid model of BpB and PseAAC features. The AUC values achieved by PseAAC, SAAC, PseACS and BpB features are 0.78, 0.76, 0.60 and 0.95, respectively.

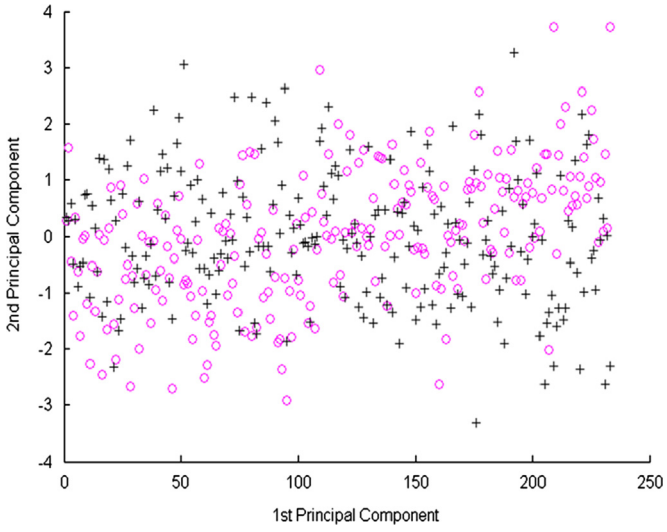

Fig. 2. The First and second principal components for the combination of hydrophobicity, hydrophilicity and mass properties of PseAAC in conjunction with Bi-profile Bayes features with 30 amino acids on each of N and C termini.

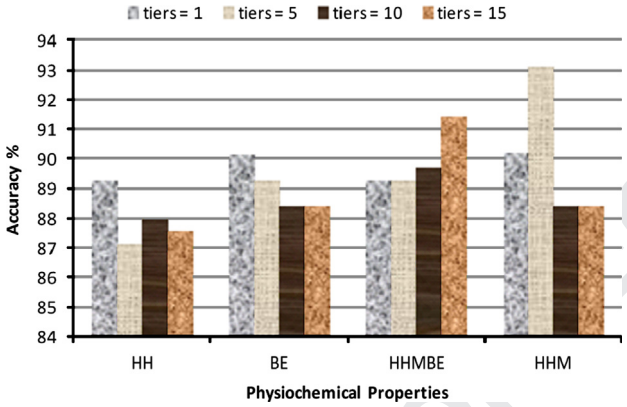

Fig. 3. Visual representation of the results presented in Table 5.

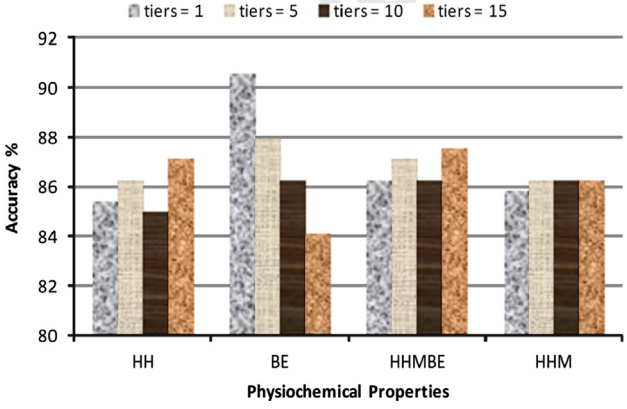

Fig. 4. Visual representation of the results presented in Supplementary Table 1.

5.1.3. *Performance analysis of MitProt-Pred through the score level ensemble*

Table 6 highlights the performance predictions of the ensemble system at score level. The ensemble approach employed here is based on the majority-voting scheme. First *MitProt-Pred* is trained using the individual features presented in Tables 1–4. There are 48 different features in these tables. To form the ensemble, we remove anyone to make them odd in number so that in decision

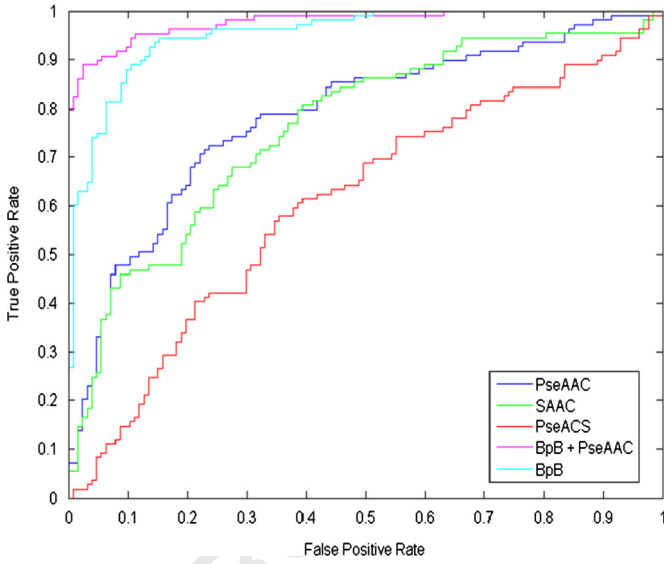

Fig. 5. ROC analysis for the best performing features using DS233.

Table 6
Performance of the scores level ensemble for DS233.

Scheme	Number of voters	Acc	Sen	Spe	MCC
Majority voting	47	**99.1**	98.6	99.5	0.98

making there is no scope of a tie among the individuals. In this work, we remove SAAC features computed using 20 amino acids on each of the N and C termini, as shown in Table 3. Consequently, we have 47 individuals in the majority voting ensemble formation.

The accuracy achieved using the score level ensemble is 6.00% higher than the accuracy yielded using the feature level ensemble and 9.90% higher than the BpB features having 30 amino acids on the N and C termini among the individual feature spaces as shown in Tables 5 and 1, respectively. This shows the significance of combining classifiers at the score level rather than at feature level. The obtained results reveal that different feature extraction strategies dig out diverse type of information from the protein sequences. Any feature extraction strategy that shows poor performance on some protein sequence cannot be declared as not being able to extract discriminative information. They may have extracted some important information that might be missed by other powerful feature extraction techniques. It is evident from Table 6 that combining these weak and powerful feature extraction techniques at score level enhanced the performance predictions.

We have also adopted the notion of sum rule. For this purpose, we have selected BpB with 30 amino acids on each of the N and C termini, PseACS with $\lambda=25$ and atom ^1H_N, PseAAC with physiochemical properties of bulk and electronic along with tier value = 1 and SAAC with 25 amino acids on each of the N and C termini. The values of accuracy, sensitivity, specificity, and MCC are 68.6%, 82.01%, 57.13%, and 0.41, respectively. These measures reveal that the performance of sum rule based ensemble is not promising, for this particular dataset, compared to majority voting based ensemble.

5.2. *Performance analysis on DS175*

In this section, the performance of MitProt-Pred is assessed on DS175 dataset using the same features as discussed previously. The

Please cite this article as: M.T. Mirza, et al., MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble..., Comput. Biol. Med. (2013), http://dx.doi.org/10.1016/j.compbiomed.2013.07.024

Table 7
Prediction performance of *MitProt-Pred* for DS175 using jackknife test.

Feature	N-terminus amino acids	C-terminus amino acids	λ	Atom	Physiochemical properties	Tier	Acc	Sen	Spe	MCC
BpB	25	25	–	–	–	–	**97.6**	95.0	98.4	0.94
PseACS	–	–	20	$^{13}C_{\alpha}$	–	–	80.5	75.2	82.1	0.55
SAAC	30	30	–	–	–	–	94.2	90.8	95.2	0.86
PseAAC	–	–	–	–	BE	20	84.0	77.0	86.0	0.62
BpB+PseAAC	30	30	–	–	HH	1	**99.4**	98.4	99.7	0.98
BpB+PseAAC	25	25	–	–	HHMBE	1	98.2	96.6	98.7	0.95

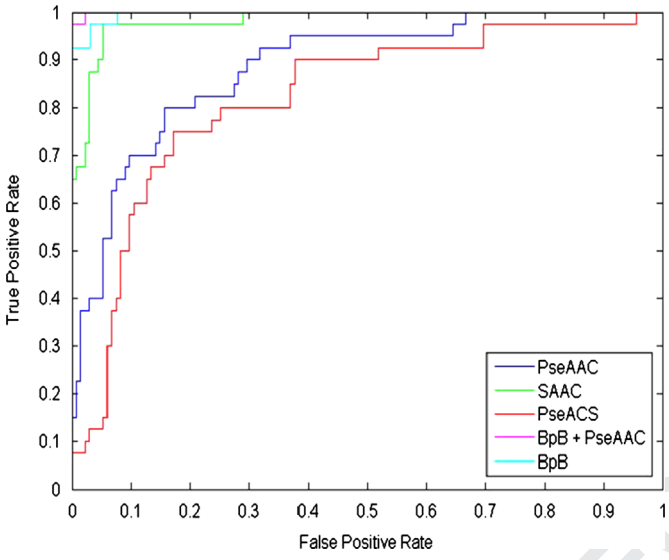

Fig. 6. ROC analysis for the best performing features using DS175.

best performing features are presented in Table 7. The detailed results and analysis are provided in the Supplementary material file.

MitProt-Pred has achieved the highest success rates among individual features, using BpB features with 25 amino acids on each of the N and C termini. The true positive and negative rates also agree with the achieved accuracy using BpB features. Similarly, the features' level ensemble enhances the discriminative power of MitProt-Pred in identifying mitochondrial and non-mitochondrial proteins. The highest accuracy achieved by BpB and PseAAC is 99.4%. AUCs under the ROC curves are shown in Fig. 6.

The highest AUC value of 0.99 is achieved using BpB as well as the hybrid model of BpB and PseAAC features. Similarly, PseAAC, SAAC and PseACS obtained ACU values of 0.88, 0.98 and 0.82, respectively. This shows the significance of the discriminative power of BpB features over other adopted features.

For comparison purpose, we have also provided the results through 5-fold cross validation protocol for DS175 dataset. Table 8 highlights the prediction performance of the best performing features through MitProt-Pred through 5-fold cross validation. Detailed results can be found in the Supplementary material file.

5.2.1. *Performance analysis of MitProt-Pred using the score level ensemble*

Results obtained by utilizing the score level ensemble approach are presented in Table 9. We have utilized the majority voting technique to combine the predictions of various classifications. The number of individual voters is 47.

The analysis of the results, presented in Table 9, reveals that the majority voting technique here did not improve the performance of *MitProt-Pred* ensemble system. The obtained accuracy through the score level ensemble is 1.20% higher than the accuracy achieved using the BpB features with 25 amino acids selected on each of the N and C termini among the individual feature spaces. However, this approach could not outperform the accuracy achieved by the feature level ensemble as shown in Supplementary Table 6 where BpB and PseAAC based features were combined.

Table 10 presents the prediction performance of score level ensemble of the predictions obtained through 5-fold cross-validation technique. The highest performance of 100% accuracy is achieved, which outperformed other techniques.

This ensemble is composed of the predictions presented in Supplementary Tables 8–11 excluding the SAAC based predictions with 20 amino acids on each of the C and N termini.

We have also obtained the predictions using the sum rule. In order to form the ensemble based on the sum rule, we have chosen BpB with 25 amino acids on each of the N and C termini, PseACS with $\lambda=20$ and atom $^{13}C_{\alpha}$, PseAAC with physiochemical properties of bulk and electronic along with tier value=20 and SAAC with 30 amino acids on each of the N and C termini. The values of accuracy, sensitivity, specificity, and MCC are 80.3%, 71.4%, 83.0%, and 0.53, respectively. These measures indicate that sum rule based ensemble is less effective compared to majority voting based ensemble in this particular case.

5.3. *Comparison with the existing approaches*

We also provide the comparative analysis with already reported approaches, which use the same datasets as we have utilized to assess the performance of our proposed *MitProt-Pred* system. The results achieved by *MitProt-Pred*, as shown in Table 11, clearly outperforms the approach proposed by Jia et al. [17], which were best reported before. MitProt-Pred, for DS175, has achieved the accuracy of 99.4% using the feature level ensemble, which is 0.60% higher than Jia et al.'s method. However, in case of score level ensemble, it does not show good performance compared to our feature level ensemble approach for DS175 still it is better compared to the approaches reported by other researchers. Employing 5-fold cross validation for DS175, *MitProt-Pred* has achieved 100% accuracy and outperformed all the previous approaches. The AUC under the ROC curve is 0.99 using the hybrid of BpB and PseAAC features for DS175, which agrees with the accuracy values yielded by *MitProt-Pred* system.

Likewise, *MitProt-Pred* has achieved the accuracy of 93.1% using the features level ensemble approach for DS233 that is 2.2% higher than that of Jia et al.'s method. Additionally, MitProt-Pred has obtained the accuracy of 99.1% using the score level ensemble approach, which is 8.2% higher than that of Jia et al.'s method and 6.00% higher than that of our feature level ensemble approach. The AUC under the ROC curve is 0.97 using the hybrid of BpB and PseAAC features for DS233, which ensures the reliability of MitProt-Pred system.

Please cite this article as: M.T. Mirza, et al., MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble..., Comput. Biol. Med. (2013), http://dx.doi.org/10.1016/j.compbiomed.2013.07.024

Table 8
Prediction performance of MitProt-Pred for DS175 using 5-fold cross-validation test.

Feature	N-terminus amino acids	C-terminus amino acids	λ	Atom	Physiochemical properties	Tier	Acc	Sen	Spe	MCC
BpB	25	25	–	–	–	–	98.8	96.9	99.4	0.97
PseACS	–	–	20	$^{13}C_{\alpha}$	–	–	82.2	73.6	84.8	0.57
SAAC	20	20	–	–	–	–	96.0	93.0	96.8	0.90
PseAAC	–	–	–	–	BE	20	85.1	78.8	87.0	0.64
BpB+PseAAC	30	30	–	–	HH	15	**100**	100	100	1.00
BpB+PseAAC	25	25	–	–	BE	5	99.4	98.4	99.7	0.98

Table 9
Performance of the scores level ensemble for DS175 through jackknife test.

Scheme	Number of voters	Acc	Sen	Spe	MCC
Majority voting	47	**98.8**	96.9	99.4	0.97

Table 10
Performance of the scores level ensemble for DS175 through 5-fold cross-validation test.

Scheme	Number of voters	Acc	Sen	Spe	MCC
Majority voting	47	**100**	100	100	1.00

Table 11
Performance comparison with other methods.

Method	Accuracy		
	DS233	DS175	
TargetP [9]	–	86.9 (unknown)	–
MITOPRED [14]	–	80.0 (5-fold)	–
Mitpred [15]	–	82.9 (5-fold)	–
PlasMit [16]	–	90.0 (20-fold)	–
PFMpred [18]	–	92.0 (5-fold)	–
Mito-GSAAC [4]	–	93.2 (5-fold)	–
Jia et al. [17]	90.9 (*N*-fold)	98.8 (*N*-fold)	–
Chen et al. [19]	–	92.0 (*N*-fold)	–
MitProt-Pred (score level ensemble)	**99.1** (*N*-fold)	**98.8** (*N*-fold)	**100** (5-fold)
MitProt-Pred (feature level ensemble)	**93.1** (*N*-fold)	**99.4** (*N*-fold)	**100** (5-fold)

N-fold represents jackknife test.

6. Conclusions

The current study presents an accurate and effective prediction model for mitochondrial protein sequences of *P. falciparum* (*PF*). This study utilizes BpB, SAAC, PseAAC, and PseACS based individual features and a step ahead combining BpB and PseAAC to form a highly discriminative hybrid feature space. Ensemble method at feature level as well as at score level for the prediction of *PF* mitochondrial proteins has been developed. At the feature level ensemble, Bi-profile Bayes features in combination with PseAAC features enhanced the classification capability of *MitProt-Pred*, which achieved the highest prediction rate of 93.1% for DS233. The ensemble at score level has shown significant improvement over other approaches including our proposed features level ensemble. *MitProt-Pred* has achieved 99.1% accuracy through the majority-voting scheme for scores level ensemble on DS233. Parameter optimization of SVM using NSGA-II is fast and efficient in developing accurate and reliable models for protein sequence prediction.

It is observed that ensemble methods either at feature level or at score level are able to lead towards enhanced performance of the prediction model. Comparative analysis confirms the significance of our proposed *MitProt-Pred* approach. *MitProt-Pred* approach can assist in the identification of mitochondrial proteins of *PF*, which are considered potential targets for anti-malarial drugs.

7. Summary

This study utilizes BpB, SAAC, PseAAC, and PseACS based features to form highly discriminative feature spaces. Ensemble method at feature level as well as at score level for the prediction of *PF* mitochondrial proteins has been developed. At the feature level ensemble, Bi-profile Bayes features are combined with PseAAC features in order to enhance the discrimination capability of the two feature spaces. The ensemble at score level has been constructed by combining the decisions through the majority-voting scheme. Parameter optimization for SVM is performed using the NSGA-II package.

MitProt-Pred as decision support system would help the medical experts in recognizing the mitochondria of *PF*, which can be consequently targeted to destruction. Additionally, during the experiments the effectiveness of drugs, developed for the destruction of *PF*, can be assessed.

Conflict of interest statement

We, the authors of this paper, state that there is no conflict of interest among authors and to any commercial or noncommercial personnel or units.

Acknowledgment

The project was supported by Higher Education Commission of Pakistan under the indigenous Ph.D. scholarship program 17-5-4 (Ps4-124)/HEC/Sch/2008/, National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2012-0001655), and National Agenda Project (NAP) funded by Korea Research Council of Fundamental Science and Technology (P-09-JC-LU63-C01).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2013.07.024.

References

[1] World Malaria Report Fact Sheet, World Health Organization, 2011.
[2] World Malaria Report, World Health Organization, 2008, p. 10.

[3] A.B. Vaidya, M.W. Mather, Mitochondrial evolution and functions in malaria parasites, Annu. Rev. Microbiol. 63 (2009) 249–267.
[4] T.H. Afridi, A. Khan, Y.S. Lee, Mito-GSAAC: mitochondria prediction using genetic ensemble classifier and split amino acid composition, Amino Acids 42 (2012) 1443–1454.
[5] A. Rao, S.J. Yeleswarapu, R. Srinivasan, G. Bulusu, An integrated rule-set for protein localization in *Plasmodium falciparum*, Curr. Bioinformatics 3 (2008) 66–73.
[6] M. Naveed, A. Khan, GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble, Amino Acids (2011).
[7] L. Nanni, A. Lumini, S. Brahnam, An empirical study on the matrix based protein representations and their combination with sequence based approaches, Amino Acids 44 (2013) 887–901.
[8] L. Nanni, S. Brahnam, A. Lumini, High performance set of PseAAC and sequence based descriptors for protein classification, J. Theor. Biol. 266 (2010) 1–10.
[9] O. Emanuelsson, H. Nielsen, S. Brunak, G.v. Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300 (2000) 1005–1016.
[10] J.D. Bendtsen, H. Nielsen, G.v. Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (2004) 783–795.
[11] P. Horton, K.-J. Park, T. Obayashi, K. Nakai, Protein subcellular localization prediction with WoLF PSORT, in: Proceedings of the Fourth Annual Asia Pacific Bioinformatics Conference APBC06, Taipei, Taiwan, 2006, pp. 39–48.
[12] A. Hoglund, P. Donnes, T. Blum, H.-W. Adolph, O. Kohlbacher, MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition, Bioinformatics 22 (2006) 1158–1165.
[13] M. Claros, P. Vincens, Computational method to predict mitochondrial proteins and their targeting sequences, Eur. J. Biochem. FEBS 241 (1996) 779–786.
[14] C. Guda, E. FahyS. Subramaniam, MITOPRED: a genome scale method for prediction of nucleus-encoded mitochondrial proteins, Bioinformatics 20 (2004) 1785–1794.
[15] M. Kumar, R. Verma, G.P.S. Raghava, Prediction of mitochondrial proteins using support vector machine and hidden Markov model, J. Biol. Chem. 281 (2006) 5357–5363.
[16] A. Bender, G.G.v. Doren, S.A. Ralph, G.I. McFadden, G. Schneider, Properties and prediction of mitochondrial transit peptides from *Plasmodium falciparum*, Mol. Biochem. Parasitol. 132 (2003) 59–66.
[17] C. Jia, T. Liu, A.K. Chang, Y. Zhai, Prediction of mitochondrial proteins of malaria parasite using Bi-profile Bayes feature extraction, Biochimie 93 (2011).
[18] R. Verma, G.C. Varshney, G.P.S. Raghava, Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile, J. Amino Acids 39 (2009) 101–110.
[19] Y.L. Chen, Q.Z. Li, L.Q. Zhang, Using increment of diversity to predict mitochondrial proteins of malaria parasite: integrating pseudo-amino acid composition and structural alphabet, Amino Acids 42 (2012) 1309–1316.
[20] C.N. Hayes, D. Diez, N. Joannin, W. Honda, M. Kanehisa, M. Wahlgren, C.E. Wheelock, S. Goto, varDB: a pathogen-specific sequence database of protein families involved in antigenic variation, Bioinformatics 24 (2008) 2564–2565.
[21] Q.-B. Gao, Z.-C. Jin, C. Wu, Y.-L. Sun, J. He, X. He, Feature extraction techniques for protein subcellular localization prediction, Curr. Bioinformatics 7 (2012) 120–128.
[22] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[23] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model. 45 (2005) 407–413.
[24] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.
[25] J. Shao, D. Xu, S.N. Tsai, Y. Wang, S.M. Ngai, Computational identification of protein methylation sites through Bi-profile Bayes feature extraction, J. PLoS ONE 4 (2009).
[26] J. Song, H. Tan, H. Shen, K. Mahmood, S.E. Boyd, G.I. Webb, T. Akutsu, J.C. Whisstock, Cascleave: towards more accurate prediction of caspase substrate cleavage sites, Bioinformatics 26 (2010) 752–760.
[27] G.-l. Fan, Q. Li, Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition, J. Theor. Biol. 304 (2012) 88–95.
[28] A.B. Sibley, M. Cosman, V.V. Krishnan, An empirical correlation between secondary structure content and averaged chemical shifts in proteins, Biophys. J. 84 (2003) 1223–1227.
[29] S.P. Mielke, V.V. Krishnan, Protein structural class identification directly from NMR spectra using averaged chemical shifts, Bioinformatics 19 (2003) 2054–2064.
[30] G. Pollastri, A. McLysaght, Porter: a new, accurate server for protein secondary structure prediction, Bioinformatics 21 (2005) 1719–1720.
[31] G.-l. Fan, Q. Li, Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition, Amino Acids 43 (2012) 545–555.
[32] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255.
[33] Z.U. Rehman, A. Khan, G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties, Anal. Biochem. 412 (2011) 173–182.
[34] Z.U. Rehman, M.T. Mirza, A. Khan, H. Xhaard, Chapter four—predicting G-protein-coupled receptors families using different physiochemical properties and pseudo amino acid composition, Methods Enzymol. 522 (2013) 61–79.
[35] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[36] M. Hassan, A. Chaudhry, A. Khan, J.Y. Kim, Carotid artery image segmentation using modified spatial fuzzy c-means and ensemble clustering, Comput. Methods Programs Biomed. 108 (2012) 1261–1276.
[37] C. Chang, C. Lin, LIBSVM: A Library for Support Vector Machines, .
[38] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Mach. Learn. 46 (2002) 131–159.
[39] K. Deb, A. Ptatap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm NSGA-II, IEEE Trans. Evol. Comput. 6 (2002) 182–197.

Please cite this article as: M.T. Mirza, et al., MitProt-Pred: Predicting mitochondrial proteins of *Plasmodium falciparum* parasite using diverse physiochemical properties and ensemble..., Comput. Biol. Med. (2013), http://dx.doi.org/10.1016/j.compbiomed.2013.07.024