



Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology

Mohammad Reza Bakhtiarizadeh ^{a,*}, Mohammad Moradi-Shahrbabak ^b,
Mansour Ebrahimi ^c, Esmaeil Ebrahimie ^{d,e,**}

^a Department of Animal and Poultry Science, College of Aburayhan, University of Tehran, Tehran, Iran

^b Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

^c Department of Biology, School of Basic Sciences, University of Qom, Qom, Iran

^d Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran

^e School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia

HIGHLIGHTS

- Performance of SVM and neural network were compared to classify different lipid binding proteins (LBPs) classes from non-LBPs.
- SVM was more successful at discriminating between LBPs and non-LBPs than neural network.
- Degree of diversity in different datasets of LBPs classes was an important factor for prediction.
- SVM can work quite well and better in a diverse set of datasets than neural network.
- Our study improved the predictive accuracy of the LBPs than previous studies.

ARTICLE INFO

Article history:

Received 15 January 2014

Received in revised form

3 April 2014

Accepted 29 April 2014

Available online 10 May 2014

Keywords:

Support vector machine

Protein features

Machine learning

Lipid metabolism

ABSTRACT

Due to the central roles of lipid binding proteins (LBPs) in many biological processes, sequence based identification of LBPs is of great interest. The major challenge is that LBPs are diverse in sequence, structure, and function which results in low accuracy of sequence homology based methods. Therefore, there is a need for developing alternative functional prediction methods irrespective of sequence similarity. To identify LBPs from non-LBPs, the performances of support vector machine (SVM) and neural network were compared in this study. Comprehensive protein features and various techniques were employed to create datasets. Five-fold cross-validation (CV) and independent evaluation (IE) tests were used to assess the validity of the two methods. The results indicated that SVM outperforms neural network. SVM achieved 89.28% (CV) and 89.55% (IE) overall accuracy in identification of LBPs from non-LBPs and 92.06% (CV) and 92.90% (IE) (in average) for classification of different LBPs classes. Increasing the number and the range of extracted protein features as well as optimization of the SVM parameters significantly increased the efficiency of LBPs class prediction in comparison to the only previous report in this field. Altogether, the results showed that the SVM algorithm can be run on broad, computationally calculated protein features and offers a promising tool in detection of LBPs classes. The proposed approach has the potential to integrate and improve the common sequence alignment based methods.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Lipids play important roles in biological processes such as energy homeostasis, formation of cellular membranes and cell

signaling. Numerous diseases are related to lipid metabolic disorder (Gross et al., 2005; Lelliott et al., 2007; van Meer et al., 2008; Xiong et al., 2010). Coordinated function of enzymes, receptors and lipid binding proteins (LBPs) produce, identify, and transport lipids (Fahy et al., 2010).

LBPs are functional proteins with central roles in many biological processes such as cellular lipid uptake, lipid metabolism, lipid transport, cell growth, regulation of gene expression, cell signaling, membrane trafficking, and innate immune response to bacterial infections (Glatz et al., 2002; Levy-Favatier et al., 2004;

* Corresponding author.

** Corresponding author at: School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, Australia.

E-mail addresses: mrbakhtiarizadeh@ut.ac.ir (M.R. Bakhtiarizadeh), esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie).

Tang et al., 2009). Consequently, the study of LBPs may lead to the development of new therapeutic strategies and better understanding of biological processes (Glatz et al., 2002).

In an exponentially growing rate, recent genome sequencing projects have produced an enormous amount of protein sequences with unknown function and structure (Friedberg, 2006; Galagan et al., 2002). Increasing efforts have been made to develop computational tools for prediction of protein function (Friedberg, 2006). Recently, many studies have considered LBPs genetics, functional classes of LBPs, lipid binding site prediction, and mechanisms of interaction between lipid and protein (Hunte and Richers, 2008; Lin et al., 2006; Xiong et al., 2010). These methods are based on (1) similarity search and clustering techniques, (2) predicting signal sequences and motifs, and (3) machine learning methods irrespective sequence similarity (Cui et al., 2007; Eisenhaber et al., 2004; Lin et al., 2006; Tang et al., 2009; Xiong et al., 2010).

There are some shortcomings in similarity based methods. Similarity search based methods will be difficult to identify the function of the query protein in the absence of experimentally annotated homologous proteins in the database. Furthermore, proteins that contain a particular domain do not always have a similar function (Bhardwaj et al., 2006; Lin et al., 2006; Tang et al., 2009; Xiong et al., 2010). As mentioned above, LBPs are diverse in sequence, structure, and function (Bhardwaj et al., 2006; Lin et al., 2006). Additionally, many LBPs sharing common domains are known to have different functions (Bhardwaj et al., 2006; Blatner et al., 2004). Hence, there is a clear need to develop alternative functional prediction methods that are not based solely on sequence similarity.

Recently, machine learning methods implementing amino acid composition and physicochemical properties have been widely used in predicting protein function (Ebrahimi et al., 2010, 2011; Gromiha et al., 2008; Hosseinzadeh et al., 2012; Kumar et al., 2011; Xiong et al., 2010; Yuan et al., 2010). These methods have two main steps. The first step is extracting a n -dimensional features vector (which is composed of descriptors derived from the protein sequences to reflect different aspects of structural and physicochemical properties of protein) with a class label attached. Various sets of protein features including amino acid compositions, dipeptide compositions, pseudo amino acid compositions, normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, and distribution of various structural and physicochemical properties have been frequently used (Ebrahimi et al., 2011; Hosseinzadeh et al., 2012; Lin et al., 2006; Ong et al., 2007; Tang et al., 2009). Ong et al. (2007) investigated the impact of the different protein features in predicting protein functional families. They showed that combination of protein features tends gives a better prediction performance than the use of individual protein features (Ong et al., 2007).

The second step is application of machine learning method (or classifier) for prediction of the class label of the input features (Ashrafi et al., 2011; Ebrahimi et al., 2010, 2011; Ebrahimie et al., 2011; Liu et al., 2010; Tang et al., 2009). Currently, many machine learning methods, such as neural networks, support vector machine (SVM), and decision trees have been successfully employed for prediction of protein function (Ashrafi et al., 2011; Ding and Dubchak, 2001; Ebrahimi et al., 2010, 2011; Ebrahimie et al., 2011; Gromiha et al., 2008; Krishnan and Westhead, 2003; Kumar et al., 2011; Xiong et al., 2010).

Neural network is a mathematical structure able to process information through many connected neurons that respond to inputs through modifiable weights, thresholds, and mathematical transfer functions (Wen et al., 2012). Neural networks are widely used in various fields such as engineering, chemistry and biology (Cartwright, 2008). Also, neural network models have been

applied in a number of protein studies including protein secondary structure prediction (Tsilo, 2009), protein–nucleotide interactions (Patel et al., 2012), protein fold class prediction (Ding and Dubchak, 2001) and protein localization prediction (Westerlund et al., 2009).

SVM is a binary classification method, proposed by Vapnik (1995), which originally designed for classification and regression tasks. The basic theory of SVM and its applications in prediction of protein functions has been alternatively been described in the previous studies (Cai et al., 2003; Noble, 2004). The SVM method has been employed for pattern recognition problems in computational biology, including gene expression analysis (Brown et al., 2000), protein–protein interactions (Cui et al., 2012), protein fold class prediction (Ding and Dubchak, 2001; Markowitz et al., 2003), and protein–nucleotide interactions (Kumar et al., 2011). This method has recently been used for predicting the LBPs (Tang et al., 2009), lipid interacting amino acid residues (Wang et al., 2008) and lipid binding sites (Xiong et al., 2010).

To our knowledge, Lin et al. (2006) study was the first report which applied SVM to predict different LBPs classes (Lin et al., 2006). In the mentioned study, LBPs were divided into nine major classes, including lipid degradation (LD), lipid metabolism (LM), lipid synthesis (LS), lipid transport (LT), lipid binding (LB), lipopolysaccharide biosynthesis (LPB), lipoprotein (LP), lipoyl and all of LBPs group (Lin et al., 2006). They used nine feature properties to describe physicochemical characteristics of each protein. Accuracy of their results in predicting different LBPs classes and non-LBPs ranged from 76.6% to 90.6% and 97% to 99.9%, respectively. In fact, the prediction accuracy of non-LBPs was higher than that of LBPs in each of the classes. In this situation, the SVM method tends to overfit and perform poorly in the minority class (LBPs) which is a significant shortcoming in prediction efficiency (Zhao et al., 2008).

Altogether, neural network and SVM have been found to be the most effective machine learning methods in protein science. To our knowledge, there is no report to evaluate the performance of neural network in predicting LBPs classes and compares its performance against SVM. Also, there is only one report on classification of different LBPs classes using the SVM method (Lin et al., 2006) which still need to be improved in accuracy of the classifiers as well as range of features. Constructing a proper feature vector of proteins is vital for a successful prediction/classification with high accuracy (Ebrahimi and Ebrahimie, 2010; Ebrahimie et al., 2011; Tahrokh et al., 2011; Yuan et al., 2010; Zhang et al., 2009).

As demonstrated by a series of recent publications (Chen et al., 2013, 2012; Min et al., 2013; Qiu et al., 2014; Xu et al., 2013a, 2013b) and summarized in a comprehensive review (Chou, 2011), to develop a really useful predictor for a protein system, one needs to go through the following five steps: (i) select or construct a valid benchmark dataset to train and test the predictor; (ii) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm to conduct the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; and (v) establish a user-friendly web-server or software for the predictor that is accessible to the public.

Feature extraction and constructing a comprehensive and proper feature vector of proteins is vital for success of machine learning classifiers. Recently, we showed that increasing feature number and adding features such as dipeptides significantly contribute in achieving high prediction accuracy (Ebrahimi and Ebrahimie, 2010; Ebrahimi et al., 2010; Ebrahimie et al., 2011). To avoid completely losing the sequence-order information, the pseudo amino acid composition (Chou, 2001, 2005) or Chou's PseAAC (pseudo amino acid composition) (Lin and Lapointe, 2013)

was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein. Since the concept of PseAAC was proposed in 2001, it has been widely used to study various attributes of proteins (Esmaili and Mohabatkar, 2010; Mohabatkar et al., 2011, 2013; Mohammad Beigi et al., 2011; Nanni and Lumini, 2008; Nanni et al., 2012; Sahu and Panda, 2010; Zhang et al., 2008). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides (Chen et al., 2013, 2012). Because it has been widely and increasingly used, recently two powerful soft-wares, called “PseAAC-Builder” (Du et al., 2012) and “propy” (Cao et al., 2013), were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server “PseAAC” (Shen and Chou, 2008) built in 2008.

In the present study, we used two supervised machine learning methods, neural network and SVM. The main aims of this study were to compare the performances of these methods in predicting LBPs classes from their protein features and improve the prediction accuracy of LBPs classes using more appropriate and comprehensive protein features. Obtaining a comprehensive view on protein structure through calculation of a large number of protein attributes for each protein sequence provided the opportunity of discovery of the key features distributing different LBP classes and increasing prediction accuracy.

2. Materials and methods

2.1. Datasets

Our method relied on structural and physicochemical properties of protein sequences and contained a two-step approach to annotate LBPs, (1) prediction whether the query sequence belongs to the LBPs family or not, and (2) prediction the class of LBPs.

At first, two datasets of positive (LBPs) and negative (non-LBPs) were created to train and test the machine learning classifiers.

2.1.1. Positive dataset

Positive dataset was a dataset which contained LBP proteins. To create this dataset, a comprehensive search of the Swiss-Prot database (release 2012, at <http://www.expasy.uniprot.org>) was carried out to retrieve different classes of LBPs by using the following keywords: “lipid degradation (LD)”, “lipid metabolism (LM)”, “lipid synthesis (LS)”, “lipid transport (LT)”, “lipid binding (LB)”, “lipopolysaccharide biosynthesis (LPB)”, “lipoprotein (LP)” and “lipoyl (LiP)”.

Protein sequences with length > 6000 amino acids or < 60 amino acids and also unreviewed protein sequences were removed. Furthermore, in order to avoid overtraining, CD-HIT program was employed to remove the homologous sequences that have > 90% sequence similarity. CD-HIT is a widely used program for clustering a large number of biological sequences with high sequence identity thresholds to reduce sequence similarity (Huang et al., 2010). Our method resulted in 706, 616, 3355, 235, 777, 553, 4026 and 335 protein sequences in LD, LM, LS, LT, LB, LPB, LP and LiP classes, respectively (Supplementary material S1). In total, 10603 protein sequences belonging to the different LBPs classes were retrieved.

2.1.2. Negative dataset

Negative dataset was defined as a dataset containing all protein classes except LBP. At first, Swiss-Prot database was depleted by searching with a list of keywords suspicious implying lipid binding functionality, using the “or” logic. In order to obtain high quality data, we exclude unreviewed protein sequences and protein sequences with length > 6000 amino acid or < 60 amino acid.

To increase confidence that a negative dataset is solely constructed by non-LBP interacting protein, we compared protein domains of positive and negative datasets. To this end, firstly all of domains related to the proteins in the positive and negative datasets were extracted from Pfam database. Then, the proteins at the negative dataset were removed if they had any common domain with domains in the positive dataset. Finally, aforementioned refinements (LBP-domain based trimming), we obtained 185,628 protein sequences in the negative dataset.

2.1.3. Independent evaluation (IE) datasets

As mentioned above, all distinct protein sequences in each LBPs class were used to construct a positive dataset. To achieve a balance between the positive and negative training data in classifier training methods, we randomly selected proteins (with a similarity threshold of < 90%) from the negative dataset at the same number of the positive dataset. Furthermore, to remove the probable bias in selecting negative subsets, five negative subsets were randomly selected from the negative dataset for each of the LBPs classes. After that, positive and negative subsets were combined together to construct the dataset related to a distinct LBPs class. In other words, for each of the LBPs classes, five distinct datasets were created. Consequently, nine function specific distinct datasets including All of LBPs group, LD, LM, LS, LT, LB, LPB, LP and LiP were created, as each of them had five repetitions (a total of 45 datasets). Before using the classifier methods, to create training and IE (for validating the classifier methods) datasets, the datasets were randomly divided into two subsets: validation set (or IE, including 20% of the whole dataset) and training set (including 80% of the whole dataset). Both the classifier methods were trained and validated on the same dataset. Data distribution for training and IE subsets in each LBPs classes are given in Table 1.

2.2. Protein features calculation

Machine learning-based methods such as neural network and SVM require feature vectors of fixed dimension as their inputs for training. Constructing a proper feature vector of proteins is a key step for a successful prediction/classification (Das Roy and Dash, 2014; Ebrahimi and Ebrahimie, 2010; Ebrahimie et al., 2011; Tahrokh et al., 2011; Yuan et al., 2010; Zhang et al., 2009). Based on Ong et al. (2007) in order to achieve a higher prediction performance, a combination of protein features was used in this study.

To provide vector of protein features, protein features were calculated using Protein Feature Server (PROFEAT) program (Li et al., 2006) to compute structural and physicochemical features of proteins and peptides based on amino acid sequence.

Table 1

Data distribution for training and independent evaluation subsets in each LBPs classes.

LBPs class	Total	Training datasets	Independent evaluation datasets
All of LBPs	21206	16965	4241
Lipid degradation	1412	1130	282
Lipid metabolism	1232	986	246
Lipid synthesis	6710	5368	1342
Lipid transport	470	376	94
Lipid binding	1554	1243	311
Lipopolysaccharide biosynthesis	1106	885	221
Lipoprotein	8052	6442	1610
Lipoyl	670	536	134

Note: LBPs, lipid binding proteins; each subset has the same number of positive and negative proteins.

In this study, seven feature groups composing diverse structural and physicochemical features of proteins were used. These protein features were: amino acid composition, dipeptide composition, normalized moreau–broto autocorrelation, moran autocorrelation, geary autocorrelation, composition, transition, distribution, quasi-sequence-order (schneider–wrede distance), quasi-sequence-order (normalized grantham chemical distance), amphiphilic Pseudo-amino acid composition and total amino acid properties as described in previous studies (Chou, 2005; Ebrahimi et al., 2011; Hosseinzadeh et al., 2012, 2013; Liu et al., 2013, 2014; Ong et al., 2007; Qiu et al., 2014; Yuan et al., 2010). As a result, the dimensionality of protein feature vector had 1080 features for each protein sequence. Also, for each protein sequence in training and IE datasets, 1080 protein features were transformed into the normalized values varying from 0 to 1. The detailed information of these protein features is provided in [Supplementary material S2](#).

2.3. Machine learning methods

Two classifier methods (SVM and neural network) were trained to identify the different classes of LBPs. For each of the two machine learning methods nine binary classifiers were constructed separately. One binary classifier investigated whether a protein belongs to the LBPs family or not. Then, if the protein accepted as LBPs family, another eight binary classifiers identified its probable LBPs class (or classes).

2.3.1. Neural network

The multilayer perceptron (a feed-forward back-propagation) neural network structure is one of the most commonly used classes of neural networks that have been developed over the years (Attarzadeh and Ow, 2010; Bishop, 1995). A feed forward multilayer perceptron network comprised of an input layer (receives the values of the input variables), an output layer (provides the model output) and one or more hidden layers. In the prediction stage, protein feature vector of each protein was presented to the input layer which subsequently feed forward into the hidden layer. Activations from the hidden layer feed forward to produce the output layer activations. Details on the theoretical aspects of neural network is reviewed by Cartwright's, (2008) study. The network topology is one of the parameters that has a significant effect on the performance of a neural network (Zhao et al., 1998). It has been reported that a neural network with three layers (one hidden layer) is capable of approximating any finite nonlinear function with high accuracy as systems with more than one hidden layer can lead to unnecessary computational overload (Wen et al., 2012; Zhao et al., 1998). In the present study, neural network was employed as a binary classifier and a feed forward back propagation network (multilayer perceptron) with a single hidden layer was used for each of the datasets using Neural Net operator in RapidMiner (Version 5.2, www.rapidminer.com) data mining tool as described previously (Ebrahimi et al., 2011). The input was the protein feature vector of LBPs and non-LBPs and the output was either 1 or 0 depending on whether the protein was predicted to be involved in the desired class or not, respectively. For each of the datasets, the number of nodes in the hidden layer varied from 3 to 10 in order to find a network with the highest performance in distinguishing LBPs classes from non-LBPs. The sigmoid function was applied as activation function in the hidden layer and also in the output layer. Furthermore, the learning rate and momentum rate were set to 0.3 and 0.2. The accepted average squared error was 0.00001 and the training epochs were 2000.

2.3.2. SVM

SVM is a margin classifier that is trained by a group of labeled data (here, LBPs and non-LBPs) formulated as feature vectors. SVM attempts to find an optimal boundary that separates the two different classes of feature vectors with a maximum margin (distance between the optimal hyperplane and a vector which lies closest to it). To classify non-separable dataset, a nonlinear SVM project feature vectors into a high-dimensional feature space using a kernel function such as the radial basis function (RBF) kernel function, $K(\vec{x}, \vec{y}) = \exp(-\gamma \|\vec{x} - \vec{y}\|^2)$ where \vec{x} and \vec{y} are the two input vectors and γ is training parameter. The linear SVM procedure was then applied to the feature vectors in this feature space. The class of a new feature vector could be identified based on the side of the hyperplane on which it is located. In this study, SVM was constructed by using LIBSVM operator (Chang and Lin, 2011) in RapidMiner (Version 5.2) data mining tool. The RBF was employed as kernel function to discover the optimal solution for all the SVM classifiers as it has widely used in general cases and most often gives the better results (Lin et al., 2008, 2010). Parameter optimization is critical in accuracy of the established SVM models (Zhong et al., 2011). Here, the regularization parameter C (controls the trade-off between maximizing the margin and minimizing the errors) and the kernel width parameter γ was optimized based on five-fold cross-validation (CV) using a grid search strategy, as the optimal C and γ parameters that maximized accuracy were chosen for each of the training subsets separately.

2.4. Evaluation of the performance

To investigate the validity of both the neural network and SVM, similar measures have to be employed. To this end, two methods, including five-fold cross-validation and IE tests were used. In five-fold cross-validation method, each of the training subsets randomly divided into five equally sized groups. Then four of them were used for training and the remaining was applied for testing the method. The same process was repeated for five times until each set was used for testing once and the average accuracy of the five testing set was computed.

The performances of the two methods were evaluated by the following measures. Sensitivity = $TP / (TP + FN)$ which shows the correct prediction of LBPs classes, specificity = $TN / (TN + FP)$ which is the ratio of correctly predict to be non-LBPs, and the overall accuracy = $(TP + TN) / (TP + TN + FP + FN)$ which correspond to the total accurate rate of the predictions. Here, TP, TN, FP, FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

In addition threshold independent measure receiver operating characteristic (ROC) was used to evaluate the performance of neural network and SVM methods. The ROC curve is a plot of the sensitivity against the false positive rate. A perfect prediction would have sensitivity and specificity both equal to 1. The area under the ROC curve (AUC) can be applied as a reliable measure of classifier performance, because it provides a single measure of overall accuracy that is not dependent on a particular threshold. The maximum value of AUC is 1 that denotes a perfect prediction (Bradley, 1997). In this study, the ROC curves were generated based on IE subsets for two classifiers methods in different LBPs classes. Also for simple comparison of the ROC curves, just AUCs were compared. AUCs were computed and reported based on a subset with highest overall accuracy in each of the LBPs classes.

3. Results

Neural network and SVM classifiers were assessed to distinguish different classes of LBPs and non-LBPs. To select the best parameters, the performance of different neural network

topologies and also different kernel parameters of SVM were compared in prediction of each of the LBP classes. For both neural network and SVM classifiers, the same partitioning of the data into training and IE subsets were used. To build nine corresponding models, both classifiers were trained with eight LBPs classes as well as “All of LBPs” group training subsets.

In statistical prediction, the following three cross-validation methods are often used to examine the effectiveness of predictor in practical application: independent dataset test, subsampling or k-fold crossover test, and jackknife test. The jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset (Chou, 2011; Chou and Shen, 2008). Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the quality of various predictors (Chen et al., 2013, 2012; Chou et al., 2012; Esmaeili and Mohabatkar, 2010; Feng et al., 2013; Hajisharifi et al., 2014; Qiu et al., 2014). However, to reduce the computational time, we adopted the k-fold cross-validation and independent dataset test in this study as extensively employed in many machine learning studies (Beiki et al., 2012; Ebrahimi et al., 2011; Hosseinzadeh et al., 2012, 2013). The prediction accuracies of the machine learning methods were primarily evaluated by five-fold CV. N-fold CV is a standard and commonly used method for evaluating classifier methods, as set of proteins used for training and testing are mutually exclusive. But it has been reported that performance of the N-fold CV method is biased with optimization (Bhasin and Raghava, 2004). Moreover, an appropriate classifier method does not only predict the training set of proteins correctly, but also is capable to classify the external proteins of the training set properly. Therefore, it is ideal to evaluate classifiers on an IE dataset (proteins not used for training the method) to demonstrate their true or unbiased performance. Here, for each LBPs classe, in addition of five-fold CV, an IE subset was created and used for unbiased evaluation of neural network and SVM methods.

3.1. Neural network

A three-layer feed forward back propagation network was used in this study. To find the optimal network that allows most accurate classification system for each of LBPs classes in the training subsets, different number of nodes in the hidden layer (3–10) were tested. To test a proper number of hidden layer nodes, five-fold CV of training subsets was applied and the number of hidden layer nodes that maximized accuracy was recorded for each of the training subsets separately. Tables 2 and 3 compares the performances of neural networks using five-fold CV and IE test, respectively. Each row in these tables represents an independent assessment of neural network as a binary classifier over different training and IE subsets. For each binary classifier, the

performances on the training and IE subsets were approximately same (Tables 2 and 3).

The highest overall accuracy based on five-fold CV (97.91%) and IE (99.40%) tests were belonged to lipoyl class of LBPs. Also, the lowest overall accuracy was occurred in all of LBPs group on five-fold CV (75.12%) and IE (74.88%) tests. As it can be inferred from Tables 2 and 3, the variation of overall accuracies in classifying different LBPs classes is high. As example, “All of LBPs group” was classified with an overall accuracy of 75% whereas lipoyl class achieved 98% accuracy.

3.2. SVM

In this study, SVM was used with RBF as the kernel function. Grid search is an appropriate way to determine the optimal values for the two major parameters of the RBF (parameters C and γ). We set the parameters values to achieve the highest classification overall accuracy rate. When grid search strategy carried out, the regularization parameter C was fixed at 10 for all of the datasets and parameter γ for different datasets ranged from 0.02 to 0.17. Performance results for the five-fold CV and IE tests for each of the LBPs classes using SVM are shown in Tables 4 and 5, respectively.

The results showed that the SVM is able to discriminate between all of the LBPs classes with more than 88% overall accuracy, confirmed by both of five-fold CV and IE test methods. The highest (98.58%) and lowest (88.84%) overall accuracy in five-fold CV test belonged to lipoyl and LT classes, respectively (Table 4). However, in IE test, the highest (99.55%) and lowest (89.55%) overall accuracy belonged to lipoyl and “All of LBPs group”, respectively. Results of SVM (Tables 4 and 5) shows that the performances of predicting different LBPs classes are approximately the same and high. In average, lipoyl obtained the highest performance in comparison to the rest of seven classes.

3.3. Comparison of neural network with SVM

Here, we compared the performance of neural network and SVM in terms of overall accuracy and AUC. From Tables 2–5, it can be seen that performances of both five-fold CV and IE test in SVM approach are higher than those of neural network. SVM outperformed neural network by 14.15~0.67% and 14.7~0.15% in term of overall accuracy based on five-fold CV and IE test methods, respectively.

In order to have a threshold IE of the methods, we created ROC curves and computed AUCs for the two classifiers in different datasets. These curves and AUCs showed the continuous relationship between sensitivity and specificity that are two competing measures of quality for any classifier method. The goal is to achieve the highest possible value of 1 for both of them. However, practically, it is very difficult to achieve this goal, consequently, a

Table 2
Five-fold cross-validation test results of the neural network method.

Neural networks	Node number	Accuracy (%)			Sensitivity (%)			Specificity (%)		
		Low	High	Average	Low	High	Average	Low	High	Average
All of LBPs	10	74.79	75.64	75.12	69.36	75.97	72.99	73.50	79.63	77.26
Lipid degradation	7	87.52	89.38	88.17	88.32	91.50	89.48	86.37	87.61	86.87
Lipid metabolism	3	75.55	83.82	81.12	69.78	91.14	83.22	80.32	82.15	81.05
Lipid synthesis	3	85.39	87.07	86.11	83.27	85.62	84.95	87.52	88.52	87.32
Lipid transport	9	84.58	88.83	86.81	85.11	90.96	88.94	81.91	88.30	84.68
Lipid binding	9	82.07	84.49	83.57	83.76	86.01	85.34	80.39	82.96	81.80
Lipopolysaccharide biosynthesis	7	89.25	90.95	90.20	88.91	92.08	90.59	89.37	90.27	89.82
Lipoprotein	7	79.96	80.81	80.52	77.93	81.22	80.42	79.26	81.99	80.63
Lipoyl	3	97.01	98.69	97.91	97.01	99.25	98.36	96.64	98.51	97.46

Note: LBPs, lipid binding proteins.

Table 3
Independent evaluation (IE) test results of the neural network method.

Neural networks	Node number	Accuracy (%)			Sensitivity (%)			Specificity (%)		
		Low	High	Average	Low	High	Average	Low	High	Average
All of LBPs	10	72.18	76.94	74.88	63.08	82.23	76.63	64.50	83.59	73.14
Lipid degradation	7	86.88	91.13	89.93	88.65	91.49	89.77	88.65	93.62	90.07
Lipid metabolism	3	81.30	84.96	82.44	86.99	92.68	89.29	77.24	82.11	75.61
Lipid synthesis	3	85.25	89.05	86.69	85.25	89.72	87.39	80.77	89.57	85.99
Lipid transport	9	86.17	87.23	86.59	82.98	87.23	85.11	87.23	89.36	88.08
Lipid binding	9	82.90	86.77	84.76	84.52	90.97	88.00	81.29	84.52	82.32
Lipopolysaccharide biosynthesis	7	89.54	92.79	91.42	89.19	93.69	91.53	89.19	93.69	91.35
Lipoprotein	7	81.18	83.04	81.99	78.76	84.84	82.95	79.38	83.73	81.02
Lipoyl	3	98.51	100	99.40	98.51	100	99.40	98.51	100	99.40

Note: LBPs, lipid binding proteins.

Table 4
Five-fold cross-validation test results of the SVM method.

SVMs	Accuracy (%)			Sensitivity (%)			Specificity (%)		
	Low	High	Average	Low	High	Average	Low	High	Average
All of LBPs	89.04	89.49	89.28	88.96	89.55	89.20	89.04	89.35	89.22
Lipid degradation	91.95	92.74	92.25	92.04	93.45	92.85	90.44	92.39	91.65
Lipid metabolism	88.44	89.68	89.15	89.05	91.68	90.55	87.22	88.24	88.75
Lipid synthesis	94.45	94.91	94.74	93.33	94.63	94.08	94.60	95.98	95.41
Lipid transport	86.70	89.89	88.84	86.17	90.43	87.55	90.96	93.09	90.11
Lipid binding	89.87	91.00	90.61	89.87	91.32	90.55	89.23	91.96	90.68
Lipopolysaccharide biosynthesis	92.99	93.78	93.26	91.18	93.89	92.58	92.99	94.80	93.94
Lipoprotein	88.92	89.27	89.06	88.26	89.10	88.85	88.57	89.44	89.06
Lipoyl	98.32	98.69	98.58	97.39	98.51	97.76	98.88	99.63	99.40

Note: LBPs, lipid binding proteins.

Table 5
Independent evaluation test results of the SVM method.

SVMs	Accuracy (%)			Sensitivity (%)			Specificity (%)		
	Low	High	Average	Low	High	Average	Low	High	Average
All of LBPs	89.01	90.05	89.55	89.16	89.63	89.31	88.78	90.81	89.79
Lipid degradation	89.72	92.55	90.85	87.94	90.78	89.80	90.78	94.33	92.20
Lipid metabolism	88.21	91.87	90.00	92.68	93.50	93.00	83.74	90.24	86.99
Lipid synthesis	94.78	96.27	95.36	93.44	94.78	94.22	95.68	98.06	96.51
Lipid transport	89.36	92.55	91.06	85.11	87.23	86.38	91.49	97.87	95.74
Lipid binding	92.26	93.55	93.23	91.61	94.19	93.03	92.90	94.19	93.42
Lipopolysaccharide biosynthesis	92.79	93.24	93.15	90.99	94.59	92.61	91.89	94.59	93.69
Lipoprotein	89.38	90.50	90.10	88.44	90.05	89.51	88.52	92.17	90.42
Lipoyl	98.51	100	99.55	98.51	100	99.40	98.51	100	99.70

Note: LBPs, lipid binding proteins.

reasonable trade-off point should be found based on the goal of the search. An AUC of 0.50 is equivalent to chance discrimination, and an AUC of 1 is equivalent to perfect discrimination.

Table 6 shows AUCs values for different LBPs classes according to neural network and SVM methods. Comparison of the AUCs values shows that SVM has a balanced specificity and sensitivity at all thresholds, and performed better than neural network. For example, when SVM was used to classify the LBPs classes, AUC achieved 0.958, 0.972, and 0.961 on the “All of LBPs group”, LD and LM datasets, respectively. In contrast, neural network only attained AUC of 0.830, 0.948, and 0.913 for the above datasets (“All of LBPs group”, LD and LM), respectively.

3.4. Comparison of the presented SVM in this study with previous application of SVM in LBPs classification

In this study, an attempt was made to develop a better SVM method for discriminating LBPs from non-LBPs against previous

Table 6
Area under the ROC curve for independent evaluation test of SVM and NN.

LBPs classes	SVM (AUC)	NN (AUC)
All of LBPs	0.95	0.83
Lipid degradation	0.970	0.95
Lipid metabolism	0.96	0.91
Lipid synthesis	0.98	0.931
Lipid transport	0.96	0.93
Lipid binding	0.96	0.93
Lipopolysaccharide biosynthesis	0.97	0.96
Lipoprotein	0.96	0.87
Lipoyl	1	0.99

Note: LBPs, lipid binding proteins; SVM, Support vector machine; NN, neural network; AUC, area under the ROC curve.

study by extending the number of protein features and optimizing of SVM parameters. The method developed by Lin et al. for predicting LBPs is available at the SVMProt server

Table 7

Performance comparison between SVMprot and the presented SVM method in this (based on our independent evaluation datasets).

LBP's class	Total	SVMprot accuracy (%)	Our SVM accuracy (%)
Non-LBPs	615	67.80	88.62
All of LBPs	2120	72.92	89.31
Lipid degradation	141	68.79	89.80
Lipid metabolism	123	62.60	93.00
Lipid synthesis	671	86.58	94.22
Lipid transport	47	38.29	86.38
Lipid binding	155	57.41	93.03
Lipopolysaccharide biosynthesis	111	40.54	92.61
Lipoprotein	805	00.86	89.51
Lipoyl	67	82.08	99.40

Note: LBPs, lipid binding proteins; SVM, Support vector machine.

(<http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi>). Here, to make an appropriate comparison, we benchmarked our SVM method against SVMProt using IE subsets. To this end, IE subsets of different LBPs classes and also non-LBPs dataset (used in our study) were run against the SVMProt server (accession number of these proteins are in [Supplementary material S3](#)). The results obtained from this section are presented in [Table 7](#). As can be seen, the results clearly documents that the performance of our “SVM method in junction with increasing feature number” is very higher than that of the “SVMProt server”.

4. Discussions

We have presented a detailed study of the performance of the two machine learning methods to build nine binary classifiers for prediction of different LBPs classes. As mentioned previously, the effectiveness of machine learning methods depends mainly on the protein features that extracted from protein sequences ([Ashrafi et al., 2011](#); [Das Roy and Dash, 2014](#); [Tahrokh et al., 2011](#); [Yuan et al., 2010](#); [Zhang et al., 2009](#)). Currently, there are a large number of protein features suggested for representing structural and physicochemical features of proteins ([Ebrahimi et al., 2011](#); [Hosseinzadeh et al., 2012](#); [Tang et al., 2009](#)). These features serve to represent sequence properties of functionally similar proteins. Therefore, for a successful prediction, a proper coding of proteins is necessary. It has been demonstrated that using a full and proper protein feature set gives the best result ([Ong et al., 2007](#)). We used a comprehensive protein feature set which has been used frequently for the prediction of different proteins classes ([Lin et al., 2006](#); [Ong et al., 2007](#); [Patel et al., 2012](#); [Tang et al., 2009](#); [Xiong et al., 2010](#)). Our results confirmed the effectiveness of these protein features in classification of LBPs classes. Obtained results from neural network and SVM for classifying LBPs classes using identical training and IE subsets, clearly showed that the SVM outperformed neural network in all of the LBPs classes. Our results were confirmed by the use of two appropriate evaluation tests which reinforced the reliability of this conclusion.

The results of this study showed that neural network ([Tables 2 and 3](#)) has a weaker potential than SVM ([Tables 4 and 5](#)) to classify different LBPs classes. Previous studies have reported that there are different degrees of diversity in sequence, structure and function in LBPs classes ([Bhardwaj et al., 2006](#); [Blatner et al., 2004](#); [Lin et al., 2006](#)). As a result, the LBPs group in our study is highly diverse, because this group contains all LBPs classes. This finding suggests that SVM can find important classifying factors in a diverse set of positive and negative training data (like the LBPs group) better

than neural network and use these factors to find the optimal classification. Moreover, the results of AUCs (ROC curves) indicated that the balance between sensitivity and specificity in SVM is better than neural network. In contrast to the LBPs group, lipoyl proteins are more specialized, as there is less diversity in their sequences. In total, both of the machine learning methods achieved the best performance in the lipoyl proteins prediction.

We found that neural network has a similar performance to SVM when the degree of diversity in the datasets is very low. However, SVM performance is much better than neural network even with the high degree of diversity in datasets. This finding may be due to the fact that SVM classifier basically depends on the support vectors, and the classifier function is not influenced by the entire dataset. Moreover, we applied a very large number of protein features to utilize machine learning methods for LBPs classification. It has been demonstrated that SVM is able to efficiently deal with a very large number of features due to the exploitation of kernel functions. The results confirmed previous findings that SVM performs better than neural network when large numbers of protein features are used ([Byvatov et al., 2003](#)). In addition to this advantage, convergence of training with SVM (which is a deterministic quadratic optimization procedure) is much faster than neural network (which is a randomized procedure). It should be noted that we do not claim that SVM outperforms neural network in general. Thus, our results suggested that in terms of LBPs classes prediction and training speed, SVM is more efficient than neural network.

To the best of our knowledge, no previous study has compared the performance of these machine learning methods in classification of LBPs classes. However, these methods were compared in order to classify different classes of proteins. In some cases, the better performance of neural network than SVM has been reported ([Hayat and Khan, 2010](#); [Kakumani et al., 2008](#)). However, in many reports, in agreement with our study, SVM outperformed neural network ([Byvatov et al., 2003](#); [Ding and Dubchak, 2001](#); [Markowetz et al., 2003](#)).

Through a tentative comparison, we can compare our methods and results to previous studies even though, protein features, classification methods and parameters are different. The reported overall accuracies in previous studies are in the range of 70–92% for the prediction of lipid binding sites ([Bhardwaj et al., 2006](#); [Irausquin and Wang, 2007](#); [Xiong et al., 2010](#)). Our dataset is much larger than those studies and considers different aspects of protein. As mentioned above, the Lin et al. study is the only report that used the SVM method (with gaussian kernel function) to classify LBPs classes from non-LBPs ([Lin et al., 2006](#)). However, we applied RBF as a kernel function in SVM and a more complete protein feature set in comparison with their study. Results from [Lin et al., 2006](#), showed a high imbalanced sensitivity and specificity among different LBPs classes. They reported that this is directly caused by using imbalanced datasets as numbers of non-LBPs were a lot higher than LBPs in each dataset. It has been demonstrated that the existence of a high imbalanced sensitivity and specificity in a prediction system is inappropriate as it can bias the classifier during training process ([Dai et al., 2012](#); [Han et al., 2004](#)).

In the present study, we used balanced datasets with the novelty of application of several different negative subsets along with a positive subset. Using this method, the diversity of all non-LBPs fully investigates as each of the LBPs classes can be assessed against different non-LBPs subsets. Here, by using SVM as the classifier, an appropriate balance between sensitivity and specificity was achieved in all datasets. All together, the presented method in this study which is based on using the different negative subsets is much more accurate than methods based on a single negative subset.

5. Conclusion

In this study, the performance of the two powerful machine learning methods including SVM and neural network were compared in classification of different LBPs classes from non-LBPs. More importantly, the prediction accuracy of different LBPs classes was significantly improved by using the appropriate protein features. The results of this investigation showed that SVM is more successful at discriminating between LBPs and non-LBPs in all of the classes than the neural network. These results were confirmed by two commonly used evaluation tests (CV and IE) and showed that SVM is more successful at discriminating between LBPs and non-LBPs in all of the classes compared to neural network. SVM achieved 89.55% overall accuracy for the identification of LBPs from non-LBPs and 92.90% (in average) for classification of different LBPs classes (based on IE test results). However, obtained overall accuracies of neural network were 74.88% and 87.90%, respectively. In addition, we demonstrated that the degree of diversity among different datasets of LBPs classes is an important factor for prediction. This is because neural network has a poor performance in datasets with high degree of diversity but SVM performance will be high in this situation. The optimized SVM method in this study reinforced with a comprehensive protein feature set outperformed results of the only previous study in application of SVM in LBPs classification (Lin et al., 2006). This study illustrates how combination of literature mining and machine learning classifiers can be exploited to address a long standing problem of distinguishing lipid metabolism involved proteins. The SVM method established here can be applied as a screening tool for predicting of LBPs from non-LBPs with high potential. The presented approach in this study can also integrate and improve sequence alignment methods.

Recent progress in machine learning packages such as Rapid-Miner (<http://rapidminer.com/>), Dortmund, Germany) and SPSS Clementine (<http://spss-clementine.software.informer.com/>, USA), which offer a user friendly environment, provides this opportunity for the general biologist (without the knowledge of software programming) to easily run and employ the selected data mining models without any difficulty. Additionally, since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou and Shen, 2009; Min et al., 2013; Xiao et al., 2013a, 2013b), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Contributors

M.R.B, E.E, MM, M.E: Conceived and designed the experiments. M.R.B: performed the experiments and analyzed the data: M.R.B, E.E, M.M: Wrote the paper.

Acknowledgments

This work was supported by Tehran University for PhD thesis of first author. We thank Bioinformatics Research Group of Qom University for providing the server hardware and other logistics support.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jtbi.2014.04.040>.

References

- Ashrafi, E., Alemzadeh, A., Ebrahimi, M., Ebrahimi, E., Dadkhodaei, N., Ebrahimi, M., 2011. Amino acid features of P1B-ATPase heavy metal transporters enabling small numbers of organisms to cope with heavy metal pollution. *Bioinform. Biol. Insights* 5, 59–82. <http://dx.doi.org/10.4137/BBI.S6206>.
- Attarzadeh, I., Ow, S.H., 2010. A novel soft computing model to increase the accuracy of software development cost estimation. In: *Proceedings of the 2nd International Conference on Computer and Automation Engineering (ICCAE)*. IEEE, vol. 3, pp. 603–607.
- Beiki, A.H., Saboor, S., Ebrahimi, M., 2012. A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. *PLoS One* 7, e44164. <http://dx.doi.org/10.1371/journal.pone.0044164>.
- Bhardwaj, N., Stahelin, R.V., Langlois, R.E., Cho, W., Lu, H., 2006. Structural bioinformatics prediction of membrane-binding proteins. *J. Mol. Biol.* 359, 486–495 [pii] 10.1016/j.jmb.2006.03.039 [http://dx.doi.org/S0022-2836\(06\)00375-5](http://dx.doi.org/S0022-2836(06)00375-5).
- Bhasin, M., Raghava, G.P., 2004. ESIPred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32, W414–W419 (32/suppl_2/W414 [pii]) <http://dx.doi.org/10.1093/nar/gkh350>.
- Bishop, C.M., 1995. *Neural networks for pattern recognition*. Oxford university press.
- Blatner, N.R., Stahelin, R.V., Diraviyam, K., Hawkins, P.T., Hong, W., Murray, D., Cho, W., 2004. The molecular basis of the differential subcellular localization of FYVE domains. *J. Biol. Chem.* 279, 53818–53827 (M408408200 [pii]) <http://dx.doi.org/10.1074/jbc.M408408200>.
- Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159.
- Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares Jr., M., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.* 97, 262–267.
- Byvatov, E., Fechner, U., Sadowski, J., Schneider, G., 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* 43, 1882–1889 <http://dx.doi.org/10.1021/ci0341161>.
- Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697.
- Cao, D.-S., Xu, Q.-S., Liang, Y.-Z., 2013. Propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29, 960–962 <http://dx.doi.org/10.1093/bioinformatics/btt072>.
- Cartwright, H.M., 2008. Artificial neural networks in biology and chemistry: the evolution of a new analytical tool. *Methods Mol. Biol.* 458, 1–13.
- Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* 2, 27.
- Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* 41, e68.
- Chen, W., Lin, H., Feng, P.-M., Ding, C., Zuo, Y.-C., Chou, K.-C., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* 7, e47843.
- Chou, K.-C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct. Funct. Bioinform* 43, 246–255 <http://dx.doi.org/10.1002/prot.1035>.
- Chou, K.-C., Shen, H.-B., 2009. Recent advances in developing web-servers for predicting protein attributes. *Nat. Sci.* 1, 2.
- Chou, K.-C., Wu, Z.-C., Xiao, X., 2012. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* 8, 629–641.
- Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21, 10–19.
- Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* 236–247.
- Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162 <http://dx.doi.org/10.1038/nprot.2007.494>.
- Cui, G., Fang, C., Han, K., 2012. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinform.* 13 (Suppl. 7), S5. <http://dx.doi.org/10.1186/1471-2105-13-S7-S5> (1471-2105-13-S7-S5 [pii]).
- Cui, J., Han, L., Lin, H., Tang, Z., Ji, Z., Cao, Z., Li, Y., Chen, Y., 2007. Advances in exploration of machine learning methods for predicting functional class and interaction profiles of proteins and peptides irrespective of sequence homology. *Curr. Bioinform.* 2, 95–112 <http://dx.doi.org/10.2174/157489307780618222>.
- Dai, D., Wang, J., Hua, J., He, H., 2012. Classification of ADHD children through multimodal magnetic resonance imaging. *Front. Syst. Neurosci.* 6, 63.
- Das Roy, R., Dash, D., 2014. Selection of relevant features from amino acids enables development of robust classifiers. *Amino Acids* 46, 1343–1351. <http://dx.doi.org/10.1007/s00726-014-1697-z>.
- Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358.
- Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* 425, 117–119 <http://dx.doi.org/10.1016/j.ab.2012.03.015>.
- Ebrahimi, M., Ebrahimi, E., 2010. Sequence-based prediction of enzyme thermostability through bioinformatics algorithms. *Curr. Bioinform.* 5, 195–203.

- Ebrahimi, M., Ebrahimi, E., Shamabadi, N., 2010. Are there any differences between features of proteins expressed in malignant and benign breast cancers? *J. Res. Med. Sci.* 15, 299–309.
- Ebrahimi, M., Lakizadeh, A., Agha-Golzadeh, P., Ebrahimi, E., 2011. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. *PLoS One* 6, e23146 (PONE-D-11-06772 [pii]) <http://dx.doi.org/10.1371/journal.pone.0023146>.
- Ebrahimi, E., Ebrahimi, M., Sarvestani, N.R., Ebrahimi, M., 2011. Protein attributes contribute to halo-stability, bioinformatics approach. *Saline Syst.* 7, 1 <http://dx.doi.org/10.1186/1746-1448-7-1>.
- Eisenhaber, B., Eisenhaber, F., Maurer-Stroh, S., Neuberger, G., 2004. Prediction of sequence signals for lipid post-translational modifications: insights from case studies. *Proteomics* 4, 1614–1625 <http://dx.doi.org/10.1002/pmic.200300781>.
- Esmaili, M., Mohabatkar, H., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol.* 203–209.
- Fahy, E., Subramaniam, S., Brown, H.A., Glass, C.K., Merrill, A.J., Murphy, R.C., Raetz, C.R., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L., Dennis, E.A., 2010. A comprehensive classification system for lipids. *J. Lipid Res.* 51, 1618.
- Feng, P.-M., Chen, W., Lin, H., Chou, K.-C., 2013. iHSP-PseRAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442, 118–125.
- Friedberg, I., 2006. Automated protein function prediction—the genomic challenge. *Brief Bioinform.* 7, 225–242 ([pii] 0.1093/bib/bbl004) <http://dx.doi.org/bbl004>.
- Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R.D., Cann, I., Graham, D.E., Grahame, D.A., Guss, A.M., Hedderich, R., Ingram-Smith, C., Kuettner, H.C., Krzycki, J.A., Leigh, J.A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J.N., Smith, K., Springer, T.A., Umayam, L.A., White, O., White, R.H., Conway de Macario, E., Ferry, J.G., Jarrell, K.F., Jing, H., Macario, A.J., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W., Birren, B., 2002. The genome of *M. aceticivorans* reveals extensive metabolic and physiological diversity. *Genome Res.* 12, 532–542 <http://dx.doi.org/10.1101/gr.223902>.
- Glatz, J.F.C., Luiken, J.J.F.P., van Bilsen, M., van der Vusse, G.J., 2002. Cellular lipid binding proteins as facilitators and regulators of lipid metabolism. *Mol. Cell. Biochem.* 239, 3–7.
- Gromiha, M.M., Ahmad, S., Suwa, M., 2008. Neural network based prediction of protein structure and Function: Comparison with other machine learning methods. In: Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN 2008 (IEEE World Congress on Computational Intelligence), pp. 1739–1744.
- Gross, R.W., Jenkins, C.M., Yang, J., Mancuso, D.J., Han, X., 2005. Functional lipidomics: the roles of specialized lipids and lipid-protein interactions in modulating neuronal function. *Prostaglandins Other Lipid Mediat.* 77, 52–64, ([pii] 10.1016/j.prostaglandins.2004.09.005) [http://dx.doi.org/S1098-8823\(04\)00093-0](http://dx.doi.org/S1098-8823(04)00093-0).
- Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., Mohabatkar, H., 2014. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* 341, 34–40 <http://dx.doi.org/10.1016/j.jtbi.2013.08.037>.
- Han, L.Y., Cai, C.Z., Ji, Z.L., Cao, Z.W., Cui, J., Chen, Y.Z., 2004. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* 32, 6437–6444 ([pii] 10.1093/nar/gkh984) <http://dx.doi.org/32/21/6437>.
- Hayat, M., Khan, A., 2010. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J. Theor. Biol.* ([pii] 10.1016/j.jtbi.2010.11.017) [http://dx.doi.org/S0022-5193\(10\)00602-8](http://dx.doi.org/S0022-5193(10)00602-8).
- Hosseinizadeh, F., Ebrahimi, M., Goliaei, B., Shamabadi, N., 2012. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PLoS One* 7, e40017 (PONE-D-12-08806 [pii]) <http://dx.doi.org/10.1371/journal.pone.0040017>.
- Hosseinizadeh, F., Kayvanjoo, A., Ebrahimi, M., Goliaei, B., 2013. Prediction of lung tumor types based on protein attributes by machine learning algorithms. *SpringerPlus* 2, 1–14 <http://dx.doi.org/10.1186/2193-1801-2-238>.
- Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682 (btq003 [pii]) <http://dx.doi.org/10.1093/bioinformatics/btq003>.
- Hunte, C., Richers, S., 2008. Lipids and membrane protein structures. *Curr. Opin. Struct. Biol.* 18, 406–411 ([pii] 10.1016/j.sbi.2008.03.008) [http://dx.doi.org/S0959-440X\(08\)00055-9](http://dx.doi.org/S0959-440X(08)00055-9).
- Irausquin, S., Wang, L., 2007. A machine learning approach for prediction of lipid-interacting residues in amino acid sequences. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007, pp. 315–319.
- Kakumani, R., Devabhaktuni, V., Ahmad, M., 2008. A two-stage neural network based technique for protein secondary structure prediction. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2008, 1355–1358, <http://dx.doi.org/10.1109/IEMBS.2008.4649416>.
- Krishnan, V.G., Westhead, D.R., 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19, 2199–2209.
- Kumar, M., Gromiha, M.M., Raghava, G.P., 2011. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.* 24, 303–313, <http://dx.doi.org/10.1002/jmr.1061>.
- Lelliott, C.J., Ljungberg, A., Ahnmark, A., William-Olsson, L., Ekroos, K., Elmgren, A., Arnerup, G., Shoulders, C.C., Oscarsson, J., Linden, D., 2007. Hepatic PGC-1beta overexpression induces combined hyperlipidemia and modulates the response to PPARalpha activation. *Arterioscler. Thromb. Vasc. Biol.* 27, 2707–2713 ([pii] 10.1161/ATVBAHA.107.155739) <http://dx.doi.org/ATVBAHA.107.155739>.
- Levy-Favatier, F., Leroux, A., Antoine, B., Nedelec, B., Delpech, M., 2004. Upregulation of rat P23 (a member of the YjgF protein family) by fasting, glucose diet and fatty acid feeding. *Cell. Mol. Life Sci.* 61, 2886–2892, <http://dx.doi.org/10.1007/s00018-004-4231-8>.
- Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., Chen, Y.Z., 2006. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 34, W32–W37 ([pii] 10.1093/nar/gkl305) http://dx.doi.org/34/suppl_2/W32.
- Lin, H.H., Han, L.Y., Zhang, H.L., Zheng, C.J., Xie, B., Chen, Y.Z., 2006. Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J. Lipid Res.* 47, 824–831 ([pii] 10.1194/jlr.M500530-JLR200) <http://dx.doi.org/M500530-JLR200>.
- Lin, S.-X., Lapointe, J., 2013. Theoretical and experimental biology in one. *J. Biomed. Sci. Eng.* 6, 4.
- Lin, S.W., Lee, Z.J., Chen, S.C., Tseng, T.Y., 2008. Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl. Soft Computing* 8, 1505–1512.
- Liu, B., Wang, X., Zou, Q., Dong, Q., Chen, Q., 2013. Protein Remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol. Inform.* 32, 775–782.
- Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.C., 2014. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 30, 472–479, <http://dx.doi.org/10.1093/bioinformatics/btt709>.
- Liu, T., Zheng, X., Wang, J., 2010. Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile. *Biochimie* 92, 1330–1334, <http://dx.doi.org/10.1016/j.biochi.2010.06.013> (S0300-9084(10)00227-0 [pii]).
- Markowitz, F., Edler, L., Vingron, M., 2003. Support vector machines for protein fold class prediction. *Biom. J.* 45, 377–389.
- Min, J.-L., Xiao, X., Chou, K.-C., 2013. iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Res. Int.* 2013, 13, <http://dx.doi.org/10.1155/2013/701317>.
- Mohabatkar, H., Mohammad Beigi, M., Esmaili, A., 2011. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* 281, 18–23, <http://dx.doi.org/10.1016/j.jtbi.2011.04.017>.
- Mohabatkar, H., Mohammad Beigi, M., Abdolahi, K., Mohsenzadeh, S., 2013. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem.* 9, 133–137, <http://dx.doi.org/10.2174/157340613804488341>.
- Mohammad Beigi, M., Behjati, M., Mohabatkar, H., 2011. Prediction of metallo-proteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics* 12, 191–197, <http://dx.doi.org/10.1007/s10969-011-9120-4>.
- Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondrial localization. *Amino Acids* 34, 653–660, <http://dx.doi.org/10.1007/s00726-007-0018-1>.
- Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 467–475, <http://dx.doi.org/10.1109/tcbb.2011.117>.
- Noble, W.S., 2004. Support vector machine applications in computational biology. *Kernel Methods Computational Biol.* 71–92.
- Ong, S.A., Lin, H.H., Chen, Y.Z., Li, Z.R., Cao, Z., 2007. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinform.* 8, 300 (1471-2105-8-300 [pii] 10.1186/1471-2105-8-300).
- Patel, A.K., Patel, S., Naik, P.K., 2012. Prediction and classification of DNA binding proteins into four major classes based on simple sequence derived features using Ann. Digest J. Nanomater. Biostruct. 5, 191–200.
- Qiu, W.-R., Xiao, X., Chou, K.-C., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* 15, 1746–1766.
- Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* 34, 320–327, <http://dx.doi.org/10.1016/j.compbiolchem.2010.09.002>.
- Shen, H.-B., Chou, K.-C., 2008. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* 373, 386–388, <http://dx.doi.org/10.1016/j.ab.2007.10.012>.
- Tahroki, E., Ebrahimi, M., Ebrahimi, M., Zamansani, F., Sarvestani, N., Mohammadi-Dehcheshmeh, M., Ghaemi, M., Ebrahimi, E., 2011. Comparative study of ammonium transporters in different organisms by study of a large number of structural protein features via data mining algorithms. *Genes Genomics* 33, 565–575, <http://dx.doi.org/10.1007/s13258-011-0057-6>.
- Tang, Z.Q., Lin, H.H., Zhang, H.L., Han, L.Y., Chen, X., Chen, Y.Z., 2009. Prediction of functional class of proteins and peptides irrespective of sequence homology by support vector machines. *Bioinform. Biol. Insights* 1, 19–47.

- Tsilo, L.C., 2009. Protein Secondary Structure Prediction Using Neural Networks and Support Vector Machines. Doctoral dissertation, Rhodes University.
- van Meer, G., Voelker, D.R., Feigenson, G.W., 2008. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell. Biol.* 9, 112–124 (pii) 10.1038/nrm2330) <http://dx.doi.org/nrm2330>.
- Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.
- Wang, L., Irausquin, S.J., Yang, J.Y., 2008. Prediction of lipid-interacting amino acid residues from sequence features. *Int. J. Comput. Biol. Drug Des.* 1, 14–25.
- Wen, X., Fang, J., Diao, M., Zhang, C., 2012. Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. *Environ. Monit. Assess.* <http://dx.doi.org/10.1007/s10661-012-2874-8>.
- Westerlund, I., Von Heijne, G., Emanuelsson, O., 2009. LumenP—a neural network predictor for protein localization in the thylakoid lumen. *Protein Sci.* 12, 2360–2366.
- Xiao, X., Lin, W.-Z., Chou, K.-C., 2013a. Recent advances in predicting protein classification and their applications to drug development. *Curr. Top. Med. Chem.* 13, 1622–1635.
- Xiao, X., Min, J.-L., Wang, P., Chou, K.-C., 2013b. Predict drug–protein interaction in cellular networking. *Curr. Top. Med. Chem.* 13, 1707–1712.
- Xiong, W., Guo, Y., Li, M., 2010. Prediction of lipid-binding sites based on support vector machine and position specific scoring matrix. *Protein J.* 29, 427–431, <http://dx.doi.org/10.1007/s10930-010-9269-x>.
- Xu, Y., Ding, J., Wu, L.-Y., Chou, K.-C., 2013a. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* 8, e55844, <http://dx.doi.org/10.1371/journal.pone.0055844>.
- Xu, Y., Shao, X.-J., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2013b. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* 1, e171, <http://dx.doi.org/10.7717/peerj.171>.
- Yuan, Y., Shi, X., Li, X., Lu, W., Cai, Y., Gu, L., Liu, L., Li, M., Kong, X., Xing, M., 2010. Prediction of interactiveness of proteins and nucleic acids based on feature selections. *Mol. Divers* 14, 627–633, <http://dx.doi.org/10.1007/s11030-009-9198-9>.
- Zhang, G., Li, H., Fang, B., 2009. Discriminating acidic and alkaline enzymes using a random forest model with secondary structure amino acid composition. *Process Biochem.* 44, 654–660.
- Zhang, S.-W., Zhang, Y.-L., Yang, H.-F., Zhao, C.-H., Pan, Q., 2008. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* 34, 565–572, <http://dx.doi.org/10.1007/s00726-007-0010-9>.
- Zhao, R., Xu, G., Yue, B., Liebich, H.M., Zhang, Y., 1998. Artificial neural network classification based on capillary electrophoresis of urinary nucleosides for the clinical diagnosis of tumors. *J. Chromatogr. A* 828, 489–496.
- Zhao, X.M., Li, X., Chen, L., Aihara, K., 2008. Protein classification with imbalanced data. *Proteins: Struct., Funct. Bioinform.* 70, 1125–1132.
- Zhong, L., Ma, C.Y., Zhang, H., Yang, L.J., Wan, H.L., Xie, Q.Q., Li, L.L., Yang, S.Y., 2011. A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG-SVM method. *Comput. Biol. Med.* 41, 1006–1013.