Original Article

Predictions of Protein-Protein Interfaces within Membrane Protein Complexes

Ebrahim Barzegari Asadabadi and Parviz Abdolmaleki *

Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran

Abstract

Background: Prediction of interaction sites within the membrane protein complexes using the sequence data is of a great importance, because it would find applications in modification of molecules transport through membrane, signaling pathways and drug targets of many diseases. Nevertheless, it has gained little attention from the protein structural bioinformatics community.

Methods: In this study, a wide variety of prediction and classification tools were applied to distinguish the residues at the interfaces of membrane proteins from those not in the interfaces.

Results: The tuned SVM model achieved the high accuracy of 86.95% and the AUC of 0.812 which outperforms the results of the only previous similar study. Nevertheless, prediction performances obtained using most employed models cannot be used in applied fields and needs more effort to improve.

Conclusion: Considering the variety of the applied tools in this study, the present investigation could be a good starting point to develop more efficient tools to predict the membrane protein interaction site residues.

*** Corresponding author:**
Parviz Abdolmaleki, Ph.D.,
Department of Biophysics,
Faculty of Biological Sciences,
Tarbiat Modares University,
Tehran, Iran
Tel: +98 21 82883404
E-mail:
parviz@modares.ac.ir
Received: 12 Dec 2012
Accepted: 10 Apr 2013

Avicenna J Med Biotech 2013; 5(3): 148-157

Keywords: Interaction sites, Membrane proteins, Support vector machines

Downloaded from http://www.ajmb.org

Introduction

A wide range of essential cellular functions are mediated by membrane proteins. For example, the exchange of membrane-impermeable molecules between organelles and between a cell and its extracellular environment are facilitated by channels and pumps. In addition, transmembrane receptors sense changes in the environment and commence specific cellular responses typically via their associated proteins. Membrane proteins are also of great diagnostic and therapeutic importance, so that they are targets of >50% of all current drugs [1].

For the diverse biological functions in a living cell, various interactions among proteins are inevitable. Knowledge about these interactions will improve our understanding of the general principles underlying biological systems function [2]. Structural details of protein-protein interactions will also help in posing experimentable mechanistic hypotheses for protein complexes. In addition, it provides a basis for structure-based discovery of therapeutic compounds to manipulate these interactions.

Although for the cytosolic proteins these interactions have been the subject of intense research, but less is known for the membrane proteins [3]. In fact, traditional protein chemistry techniques have not been helpful to study the membrane proteins, because they are typically hydrophobic macromolecules [4]. It is well known that the largest part of the binding free energy of protein interaction is contributed by a few key residues [5]. Experimental methods for detecting the key residues

Copyright © 2013, Avicenna Journal of Medical Biotechnology. All rights reserved.

of interaction domains, such as alanine scanning mutagenesis are also not applicable on large scales because they are expensive and time-consuming [6]. Therefore, efficient and reliable computational methods for identifying these residues from sequences are urgently required.

To our knowledge, research on developing computational approaches for predicting the interaction sites in membrane proteins is limited to only one study by Bordner *et al* [7]. In the present investigation, we have attempted to address this problem with the aim of achieving a more efficient and accurate method for prediction of key residues of membrane protein interaction sites. For this reason, we have constructed many predictive models to classify the surface lipid-facing residues of membrane proteins based on their existence in the interaction interface within membrane protein complexes.

Materials and Methods

Dataset

The only dataset of membrane proteins to date collected and used for the only membrane protein interaction site prediction study, was taken from Bordner [7]. In addition to the taken dataset which was used for performance comparison purposes, another dataset was collected and used as an independent test set. Again for comparison purposes, we utilized the PDBTM database [8,9] in order to construct this dataset in the same way as the Bordner's was collected. PDBTM updates since 2010 including 502 added complexes plus 9 modified structures was taken and culled using PISCES web server [10] to form a non-redundant set of membrane protein complexes in which no pair of complexes had all proteins differing by more than 30% sequence identity. The set contains alpha-helical as well as beta-barrel complexes in different oligomeric states including monomeric, homo- and hetero-multimeric complexes.

From the non-redundant set of membrane proteins obtained this way, the independent test data were extracted. Only surface residues, with relative solvent accessible surface area (SASA) ≥0.2, which are also within the hydrophobic core of the membrane, were included in this set. In order to attain relative values of accessible surface area, solvent accessibility is defined as the ratio of the solvent accessible surface area of a residue in the folded state and that of the residue in an extended tripeptide (Gly-X-Gly) conformation. The residue solvent accessibility value was computed by means of ASAView program, a web server available online at http://gibk26.bio.kyutech.ac.jp/jouhou/shandar/netasa/asaview [11].

The membrane boundaries were predicted using PDBTM-TMDET server [12] and used for determining the membrane core residues. If a Z-axis is hypothesized perpendicular to the plane of the membrane predicted by TMDET, with the origin being in the center of the membrane, then residues in the membrane core have Z-coordinates with $|Z| \leq 15$ *Å*. In other words, the membrane core is assumed to be 30 *Å* thick. Each surface residue located at the membrane core was labeled either as a binding site or interface residue (denoted with I), if it had <4 *Å* non-H atom separation from another protein chain in the complex structure, otherwise as a non-binding site or non-interface residue (denoted with N).

Training data for each individual residue included the frequencies of each of the 20 standard amino acids in a multiple sequence alignment of similar sequences and the evolutionary rate. To create the sequence alignments, similar protein sequences were first searched for in the NCBI nr database with BLAST [13] at an E-value cutoff of 10^{-2}; then redundant sequences at the 90% sequence identity level were removed using the CD-HIT program [14], and the multiple alignments of the remaining sequences were generated with MUSCLE [15]. The residue frequency for a particular residue type was simply computed as the fraction of residues of that type in the corresponding multiple sequence alignment column. Only proteins with at least 20 sequences in the final alignment were included

Downloaded from http://www.ajmb.org

in the training and testing sets.

The evolutionary rate was calculated using the REVCOM method [16], which in evolutionary conservation values are more robust to the particular set of sequences and local alignment errors than other methods. Evolutionary rates obtained this way vary inversely with conservation.

In classification problems, the training data will considerably affect the classification accuracy. However, the data in real applications have often imbalanced class distribution, *i.e.* most of the data are in majority class and little data are in minority class. In this case, if all the data are used as the training data, the classifier would tend to predict that most of the incoming data belongs to the majority class. Therefore, it is important to adopt suitable methods for classification in the imbalanced data problems [17].

Of the most common approaches to deal with the class imbalance problem are over-sampling and under-sampling techniques. The over-sampling approach increases the number of minority class samples to reduce the degree of imbalanced distribution. Under-sampling approach is supposed to reduce the number of samples with the majority class [17]. Generally, the performances of over-sampling approaches are worse than those of under-sampling ones [18]; so, we applied the under-sampling approach to address the data imbalance. The whole dataset contains 122 proteins which include the whole sample size of 8,365 residues. 2,391 out of 8,365 residues are interfacial residues and the remaining 5,974 are non-interfacial. To implement the under-sampling method, all class N samples with NaN values for the evolutionary parameter (1,477 N samples) were removed, and an additional 2,106 samples were also randomly removed to reach equal sample size between the classes I and N (both classes containing 2,391 samples). Missing values of evolutionary parameter in class I (332 I samples) were replaced by intraclass mean.

Prediction using weka classifiers

Weka [19] is an open source data mining and machine learning package implemented in Java. Many researchers and practitioners in the data mining and machine learning community commonly use Weka. As a comprehensive tool, Weka provides an interface for implementing many modeling algorithms in a user-friendly manner. Classification tools are also included in this software environment. The total 110 classifiers are grouped in seven categories.

We employed Weka version 3.6.8 to classify the membrane protein residues in I and N classes. Due to different limitations in data and/or algorithms we could apply 71 of these classifiers. All the parameters of classifier models were set to their default values.

Prediction using ℓ_1/ℓ_q-regularized logistic regression

The ℓ_1/ℓ_q-regularized logistic regression (RLR) model used in this study is a generalization of the ℓ_1-regularization logistic regression. This model has strong theoretical guarantee, and has exhibited great empirical success arisen from recent studies in areas such as machine learning, statistics, and applied mathematics [20-24]. Regarding this fact, we adopted this model to deal with the problem of predicting the interaction class of residues at the interfaces of membrane proteins.

The ℓ_1/ℓ_q-regularized logistic regression is an expression of the form:

$$\min_{x} \sum_{l=1}^{k} \sum_{i=1}^{m} w_{il} \log\left(1 + \exp\left(-y_{il}(x_l^T a_{il} + c_l)\right)\right) + \lambda x_{\iota_1/\iota_q}$$

where $a_{i\ell}$ indicates vector of size $1\times n$; n is the number of features for i-th residue of the ℓ-th interaction class; $w_{i\ell}$ is the weight for $a_{i\ell}$; $y_{i\ell}$ is the response of $a_{i\ell}$; and c_ℓ is the intercept for the ℓ-th interaction class. Since the RLR model assigns weights to dependent variables, it gives a measure of preference and avoidance of different residues in the interaction sites and thus it can be used as a feature selection tool [25]. To construct the ℓ_1/ℓ_q-regularized logistic regression we used mcLogisticR function of SLEP package version 4.0 [26] which is written in Matlab. In this function, the elements in y are required to be an $m\times k$

Downloaded from http://www.ajmb.org

matrix including elements of 1 or −1 (m is the number of residues and k is the number of interaction classes).

Prediction using tuned support vector machine

Kinds of learning machines based on statistical learning theory are called Support Vector Machines (SVMs). They have three remarkable characteristics: the absence of minima, the sparseness of the solution, and the implementation using the kernel Adatron algorithm. The kernel Adatron maps inputs to a high-dimensional feature space, and then optimally separates data into their respective classes by isolating those inputs which fall close to the data boundaries. Therefore, the kernel Adatron is especially effective in separating sets of data which share complex boundaries. SVMs seek a global optimized solution and avoid over-fitting in the training process; they can only be used for classification, not for function approximation. The theory and algorithms of SVMs can be found in Vapnik (1995, 1998) [27,28].

In this study, we applied the tune function of e1071 package [29] in the R environment version 2.12-0 [30] to develop our SVM-based method. The tune function uses Grid Search to find the best functions. Using the tune function through cross-validation procedure, it provides as many simulations as the number of cross-validation folds in databases to select optimum model structure each time.

Evaluation criteria

In the two-class scenarios, one class with high identification importance is referred to as the positive class and the other as the negative class. After a classification process, samples are categorized into four groups, including TP (true positives: number of correctly classified interface residues), TN (true negatives: number of correctly classified non-interface residues), FP (false positives: number of non-interface residues incorrectly classified as interface) and FN (false negatives: number of interface residues incorrectly classified as non -interface). Several measures for the evaluation of model's performance can be derived using these scalar indices. Prediction accuracy (PA) is the best-known and most common of these measures, which is defined as:

$$PA = (TP + TN)/(TP + FP + FN + TN)$$

ROC curve provides a good summary of the performance of a classification model. It measures the classifier performance over the whole range of thresholds from 0 to 1 from the plots of Sensitivity (TP/(TP+FN)) and Specificity (TN/(TN+FP)). The area under a ROC curve (AUC) gives a single measure of classifiers' performance for evaluating which model is better on average. The ROC curve was plotted using the 10-fold cross validation results.

Results

Many predictor models were applied in this investigation to achieve the best possible classification accuracy in prediction of interaction site residues. In order to obtain an accurate estimation of the prediction performance for novel data, the data was divided so that all residue data for a particular protein was contained entirely within one of the training or testing sets. And the predictions were made for a distinct set of proteins from those used to train the predictor model. The classifier models were trained and tested using 10-fold cross-validation technique, whereby the whole set is divided into ten sets, each containing equal number of samples. The method was trained on nine sets and the performance was measured on the remaining tenth set. This procedure was then repeated ten times in order to trust that all members of the dataset had been selected in the testing procedure. The performance of the model was evaluated by averaging the mentioned measures over ten sets. This way, we could hopefully expect to get a global conclusion on the whole of dataset.

Weka classifiers

Table 1 indicates the prediction accuracies obtained by the total 71 predictor models of Weka, on the independent test set. As can be seen, the accuracy of most models is less than the acceptable value of 75%, and the best classifiers show accuracies around 76 or 77%.

Downloaded from http://www.ajmb.org

Table 1. The Weka classifier models and their accuracy of prediction on the independent test set

Classifier	Accuracy (%)	Classifier	Accuracy (%)	Classifier	Accuracy (%)
NBTree	77.1435	Random forest	71.6437	Voted perceptron	60.9159
SMO	76.9305	Classification via regression	71.1627	IB1	60.7905
Decision table	76.2861	Bayes Net	71.1209	IBk	60.7905
Attribute selected classifier	76.2233	Rotation forest	70.4308	Multilayer perceptron	59.7449
Filtered classifier	76.1188	LADTree	70.2217	Naïve bayes multinomial	58.9921
Bagging	75.366	LogitBoost	69.6989	Complement naïve bayse	58.9293
Decorate	75.0314	ADTree	68.4442	Naïve bayes simple	58.7829
JRip	74.4458	FT	68.2978	Naïve bayes	58.5529
END	74.3622	AdaBoostM1	67.2313	Naïve bayes multinom updateable	58.5529
Nested Dichotomies Class Balanced ND	74.3622	RandomTree	66.1857	Naïve bayes updateable	58.5529
Nested Dichotoies Data Near Balanced ND	74.3622	Raced incremental logitBoost	65.0774	DMNBtext	57.7373
Nested Dichotomies ND	74.3622	OneR	63.279	Threshold selector	56.3363
Ordinal class classifier	74.3622	Conjunctive rule	62.4843	RBF network	56.0435
J48	74.3622	KStar	61.9824	VFI	51.7984
PART	74.3413	LWL	61.857	Classification via clustering	50.2928
J48graft	74.3413	Decision stump	61.857	CV parameter selection	49.9791
Random sub space	74.2158	SPegasos	61.7942	Grading	49.9791
Simple cart	73.923	Multi boost AB	61.7315	Multi scheme	49.9791
LMT	73.6303	NNge	61.7315	Stacking	49.9791
Ridor	73.4421	Bayesian logistic regression	61.606	StackingC	49.9791
BFTree	73.3166	Logistic	61.5851	Vote	49.9791
DTNB	73.1493	Multi class classifier	61.5851	ZeroR	49.9791
REPTree	72.3756	Simple logistic	61.376	Hyper pipes	49.9164
Random committee	71.9155	Dagging	61.2505		

Such a weak performance would be too inadequate to be applied in the prediction of membrane protein interaction sites. However, the parameters of classifier models in Weka can be modified so that a higher performance is achieved. Regarding this fact, we chose to modify the best-performing classifiers to reach a higher prediction accuracy. The highest accuracy (77.14%) was obtained by the NBTree, a decision tree-based algorithm, but this classifier does not present additional parameters to be modified. The second best classifier was SMO, which is the Weka implementation of support vector machines (SVM). We relegated the modification of the SMO hyper-parameters to the tune function of another SVM implementation (Tuned SVM), whose promising results have been reported in subsection 3.3.

ℓ_1/ℓ_q-Regularized logistic regression

We ran the Regularized Logistic Regression (RLR) method on the dataset in two subclasses using 10-fold cross-validation. The performance of this method was not satisfactory, and the model did not give a prediction accuracy higher than 66.28% on Bordner's dataset and 59.08% on the independent test set. Thus, despite the strong theoretical guarantee and the great empirical success of the model, we can only state that the regularized logistic regression model is not applicable in the membrane protein interaction site prediction problem using the available features.

Nevertheless, we used the capability of feature selection of this model as a criterion to estimate the relative contribution of each fea-

Downloaded from http://www.ajmb.org

ture in the interaction interfaces of membrane proteins. Table 2 lists the features in the order of importance suggested by RLR model. The contribution order is also compared with that obtained in the only previous similar study by Bordner [7]. The contribution orders are in relative accordance with each other. Evolutionary rate and frequency of Alanine, Leucine, Glycine and Valine are proposed by both studies to mostly contribute in interaction interfaces and frequency of Glutamine, Glutamic acid, Asparagine, Tyrosine, Aspartic acid and Threonine are proposed by both to be the least contributing factors.

Tuned support vector machine

We used the tune function to select optimized structure of SVM through 10-fold cross-validation test. The most important hyper-parameter of tuned SVM topology is kernel function and search for the best one among four different kernel functions, *i.e.*, linear, polynomial, radial, and sigmoid was carried out. The best kernel was found to be the radial basis function. In addition, the kernel-related parameter gamma was searched for the best value and gamma=0.01 was reported as the optimum value by the model. Cost of constraints violation, which is the "C" constant of the regularization term in the Lagrange formulation of the SVM model was also searched and the optimum value of cost=32 was obtained. Then, the SVM model using the optimal set of hyper-parameters was constructed and employed for classification of membrane protein residues.

This tuned support vector machine model could classify the samples of Bordner's dataset by 86.95% accuracy and showed a prediction accuracy of 82.17% on the independent test set. The ROC curve was plotted using the 10-fold cross validation results, and has been illustrated in figure 1. AUC value for the tuned SVM model is obtained as 0.812 which shows a remarkable improvement in comparison with the value AUC=0.75 obtained in Bordner's study which similarly has used all lipid-facing residues [7]. Prediction of the interaction site residues in the independent test set also gives acceptable AUC value of 0.786. Therefore, the applied tuned SVM model outperforms the more complicated Random for-

Table 2. Feature selection by RLR and its comparison with the previous study

Features/Weights	Order of importance by RF model
Evol. Rate/1.480945	Ala
His/0.989602	Leu
Ala/0.255016	Gly
Cys/0.251367	Val
Ile/0.239659	Evol. Rate
Lys/0.231607	Met
Val/0.229424	Phe
Leu/0.215648	Ile
Gly/0.173995	Trp
Pro/0.103435	Ser
Trp/-0.05232	Arg
Gln/-0.08103	Lys
Glu/-0.14467	Thr
Ser/-0.17687	Asn
Asn/-0.23034	Cys
Phe/-0.24444	Pro
Tyr/-0.27445	His
Asp/-0.35331	Tyr
Arg/-0.42569	Gln
Thr/-0.80005	Asp
Met/-0.83732	Glu

All predictor variables are sorted by their RLR-assigned weights and the order of importance is compared with that of previously reported RF model (Bordner, 2009). Positive values show preference and negative values show avoidance of the features in interaction sites

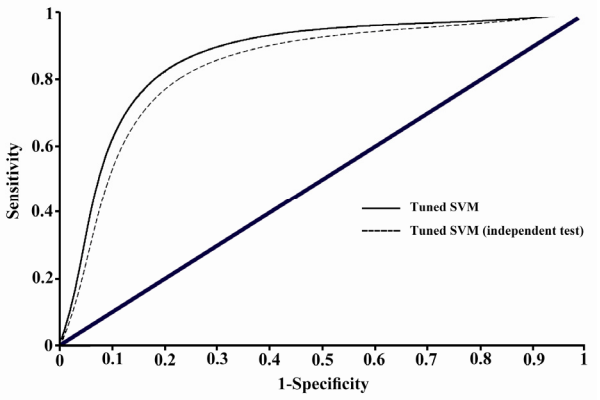

Figure 1. ROC plot illustrating the classification performance of tuned SVM model with the reference dataset and the collected independent test set. Related AUC values are 0.812 and 0.786, respectively

Downloaded from http://www.ajmb.org

ests model employed in the only previous similar investigation [7].

Discussion

In this study, more than 70 predictive and classifier models were applied to classify the surface lipid-facing residues of membrane proteins based on their existence in the interaction interface within membrane protein complexes. Among experimented models, the tuned support vector machine classifier could show a high performance in distinguishing the interacting residues from non-interacting ones, at the interaction interfaces of membrane proteins. This model outperforms the result obtained by the only previous study which has been devoted to prediction of membrane protein interaction sites [7].

Though the obtained performance by tuned SVM is a satisfactory result, the performances of the most models applied may still be interpreted as weak. In this regard, it should be emphasized that achieving high performance of predictive models in the prediction of protein-protein interaction sites has been a difficult task [6]. Considering this issue, it is necessary to discover new theories, algorithms and features in order to further improve the performance of such prediction tasks, especially about membrane-associated proteins. As may be expected, separate predictors trained on membrane or non-membrane protein classes are required in order to achieve good prediction accuracy, because each of these classes experience different physiochemical environments, resulting in different frequencies of surface residue types for each class [7].

Computational studies of membrane protein interaction sites have been previously performed with different aims, including improvement of membrane protein crystallization [31], recognizing the membrane protein types [32], identifying the hub proteins within complicated membrane protein network systems [33], and discrimination of outer membrane proteins [34].

Another research has considered the protein-protein interfaces in transmembrane domains of outer membrane proteins, with the purpose of determining their oligomerization states. Their predictions using only sequence information has obtained the accuracy and specificity of 96% and 94%, respectively [35]. This indicates that high amount of information about the protein-protein interaction is hidden and can be extracted from the protein sequence data. Prediction of protein interaction sites has also found its applications in the form of web servers, like MEMSAT [36], MEMSATSVM [36], MEMPACK [36], PPI-Pred [37], cons-PPISP [38], meta- PPISP [39], PINUP [40], ProMate [41], SPPIDER [42], WHISCY [43], ConSurf [44], InterProSurf [45], ProteMot [46] and PrISE [47].

Weka classifiers did not give a satisfactory result, however construction of many models by using this tool provided a basis to choose the most powerful model (the tuned SVM) to perform the purposed classification task. Regularized logistic regression also could not classify the samples with desirable accuracy, but its feature selection capability allowed to obtain a measure for the degree of contribution of each predictor variable in the interaction class of residues. As proposed by this model, evolutionary rate and frequency of Alanine, Leucine, Glycine and Valine are the factors which mostly contribute in interaction interfaces and frequency of Glutamine, Glutamic acid, Asparagine, Tyrosine, Aspartic acid and Threonine are the least contributing factors. Our result, to a great extent, confirms the findings of the previous similar study [7]. Thus, the preferred and avoided features proposed by these models may provide a basic knowledge which could be helpful in mechanistic, protein function and even the protein design studies.

From structural point of view, our results are also consistent with findings of previous researches. According to these investigations, membrane proteins prefer a wide range of moderately stabilizing interactions instead of strong ones, which lends them a greater degree of flexibility in terms of conformation

Downloaded from http://www.ajmb.org

and stability [48]. Furthermore, it has been found that the membrane protein-protein interface is enriched with weakly stable strands [35]. In accordance with these findings, our results of preference and avoidance of amino acids in membrane protein interfaces show that non-polar amino acids which form weak hydrophobic interactions are preferred in these sites.

Conclusion

Prediction of interaction sites could be a good starting point which helps to identify pharmacological targets, thereby helping drug design studies. The prediction methods could also find applications in guiding experimental investigations of membrane protein interactions, and also, in the prediction of protein complex structures using computational methods such as docking or threading [7]. Identifying the binding site residues is also crucial for understanding the function of proteins.

Regarding the few investigations concerning the computational prediction of membrane protein binding sites, it would be recommended to explore new ideas, methods and features to further improve the performance of such predictions. Such a task could be performed by considering the results obtained by the numerous classifiers in this study. Modifications of the algorithms, tuning the parameters of better-performing models, adding more features to the available feature set and/or changing the data structures are of ways through which the prediction performance could be improved.

Acknowledgement

We give our greatest thanks to all coworkers in Bioinformatics lab of Dr. Abdolmaleki, especially Mrs. Naddafi as the lab expert, for her kind helps in providing facility.

References

1. Stagljar I, Fields S. Analysis of membrane protein interactions using yeast-based technologies. Trends Biochem Sci 2002;27(11):559-563.
2. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 2003;19(10):551-560.
3. Maurel D, Kniazeff J, Mathis G, Trinquet E, Pin JP, Ansanay H. Cell surface detection of membrane protein interaction with homogeneous time-resolved fluorescence resonance energy transfer technology. Anal Biochem 2004;329(2):253-262.
4. Njus D. Fundamental principles of membrane biophysics. Department of Biological Sciences, Wayne State University: Wayne State University; 2000.
5. Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics 2010;11:174.
6. Liu B, Wang X, Lin L, Tang B, Dong Q. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics 2009;10:381.
7. Bordner AJ. Predicting protein-protein binding sites in membrane proteins. BMC Bioinformatics 2009;10:312.
8. Tusnady GE, Dosztanyi Z, Simon I. Transmembrane proteins in the protein data bank: identification and classification. Bioinformatics 2004;20 (17):2964-2972.
9. Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 2005;33(Database issue):D275-278.
10. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12): 1589-1591.
11. Ahmad S, Gromiha M, Fawareh H, Sarai A. ASA view: database and tool for solvent accessibility representation in proteins. BMC Bioinformatics 2004;5:51.
12. Tusnady GE, Dosztanyi Z, Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 2005;21(7):1276-1277.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17): 3389-3402.
14. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13): 1658-1659.

Downloaded from http://www.ajmb.org

15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32(5):1792-1797.

16. Bordner AJ, Abagyan R. REVCOM: a robust Bayesian method for evolutionary rate estimation. Bioinformatics 2005;21(10):2315-2321.

17. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 2009;36(3):5718-5727.

18. Drummond C, Holte RC. C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. Workshop on Learning from Imbalanced Datasets II; 2003: Citeseer.

19. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Soft ware: An Update. SIGKDD Explorations 2009; 11(1).

20. Bach FR. Consistency of the group Lasso and multiple kernel learning. JMLR 2008;9:1179-1225.

21. Duchi J, Singer Y. Efficient online and batch learning using forward backward splitting. JMLR 2009; 10:2873-2898.

22. Kowalski M. Sparse regression using mixed norms. Appl Comput Harmon 2009;27(3):303-324.

23. Negahban SN, Ravikumar P, Wainwright MJ, Yu B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist Sci 2012;27(4):538-557.

24. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Royal Statistical Society: Series B (Statistical Methodology) 2006;68(1):49-67.

25. Jahandideh S, Srinivasasainagendra V, Zhi D. Comprehensive comparative analysis and identification of RNA-binding protein domains: Multiclass classification and feature selection. J Theor Biol 2012;312C:65-75.

26. Liu J, Ji S, Ye J. SLEP: Sparse learning with efficient projections. 4.1 Version. Arizona State University; 2009.

27. Vapnik V. The nature of statistical learning theory. New York: Springer; 1999.

28. Vapnik V. The nature of statistical learning theory. New York: John Wiley and Sons; 1998.

29. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. Misc functions of the Department of Statistics (e1071), TU Wien. Available at cranr-projectorg/web/packages/e1071/indexhtml (verified 6 Apr 2009) R Foundation for Statistical Computing, Vienna, Austria; 2008.

30. Team RDC. R: A language and environment for statistical computing. Vienna, Austria 2009.

31. Grabe M, Neu J, Oster G, Nollert P. Protein interactions and membrane geometry. Biophys J 2003; 84(2 Pt 1):854-868.

32. Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 2003;84(5):3257-3263.

33. Shen YZ, Ding YS, Gu Q, Chou KC. Identifying the hub proteins from complicated membrane protein network systems. Med Chem 2010;6(3):165-173.

34. Park KJ, Gromiha MM, Horton P, Suwa M. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 2005;21(23): 4223-4229.

35. Naveed H, Jackups R Jr, Liang J. Predicting weakly stable regions, oligomerization state, and protein-protein interfaces in transmembrane domains of outer membrane proteins. Proc Natl Acad Sci USA 2009;106(31):12735-12740.

36. Buchan DW, Ward SM, Lobley AE, Nugent TC, Bryson K, Jones DT. Protein annotation and modelling servers at University College London. Nucleic Acids Res 2010;38(Web Server issue):W563-568.

37. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 2005;21(8): 1487-1494.

38. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins 2005;61(1):21-35.

39. Qin S, Zhou HX. meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 2007;23(24):3386-3387.

40. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic Acids Res 2006;34(13):3698-3707.

41. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 2004;338(1):181-199.

42. Porollo A, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins 2007;66 (3):630-645.

43. de Vries SJ, van Dijk AD, Bonvin AM. WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins 2006; 63(3):479-489.

Downloaded from http://www.ajmb.org

44. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, et al. ConSurf2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 2005;33(Web Server issue):W299-302.

45. Negi SS, Schein CH, Oezguen N, Power TD, Braun W. InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 2007;23(24):3397-3399.

46. Chang DT, Weng YZ, Lin JH, Hwang MJ, Oyang YJ. Protemot: prediction of protein binding sites with automatically extracted geometrical templates. Nucleic Acids Res 2006;34(Web Server issue): W303-309.

47. Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V. Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics 2012;13:41.

48. Joh NH, Min A, Faham S, Whitelegge JP, Yang D, Woods VL, et al. Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. Nature 2008;453(7199):1266-1270.

Downloaded from http://www.ajmb.org