Amino Acids (2010) 38:1497–1503
DOI 10.1007/s00726-009-0362-4

ORIGINAL ARTICLE

Using auto covariance method for functional discrimination of membrane proteins based on evolution information

Li Yang · Yizhou Li · Rongquan Xiao · Yuhong Zeng · Jiamin Xiao · Fuyuan Tan · Menglong Li

Received: 24 April 2009 / Accepted: 24 September 2009 / Published online: 7 October 2009

Abstract Membrane transporters are critical in living cells. Therefore, the discrimination of the types of membrane proteins based on their functions is of great importance both for helping genome annotation and providing a supplementary role to experimental researchers to gain insight into membrane proteins' function. There are a lot of computational methods to facilitate the identification of the functional types of membrane proteins. However, in these methods, the local sequence environment was not integrated into the constructed model. In this study, we described a new strategy to predict the functional types of membrane proteins using a model based on auto covariance and position-specific scoring matrix. The novelty of the presented approach is considering the distribution of different positions of functional conservation sites in protein sequences. Thereby, this model adequately takes into account the long-range correlation between such sites during sequential evolution. Fivefold cross-validation test shows that this method greatly improves the prediction accuracy and achieves an acceptable prediction accuracy of 87.51%. The result indicates that the current approach might be an effective tool for predicting the functional types of membrane proteins only using the primary sequences. The code and dataset used in this article are freely available at http://cic.scu.edu.cn/bioinformatics/predict_membrane.zip.

Keywords Membrane transporters · Sequence environment · Position-specific scoring matrix · Auto covariance · Support vector machine

L. Yang · Y. Li · R. Xiao · Y. Zeng · J. Xiao · F. Tan · M. Li (✉)
College of Chemistry, Sichuan University,
Chengdu 610064, People's Republic of China
e-mail: liml@scu.edu.cn

F. Tan
National Institute of Measurement and Testing Technology,
No. 10 Yushuang Road, Chengdu 610021,
People's Republic of China

Introduction

As the most basic structural and functional unit of all living organisms, a cell is "enveloped" by one or more membranes, in which membrane proteins play a key role in the cell membranes of both prokaryotic and eukaryotic organisms, by executing biologically important functions related to the cell membrane. These include transporting ions and molecules into and out of cells, recognizing the immune system and energy transducers and possessing a characteristic of nature such as membrane adhesion, catalytic activity and receptors of signal molecules. Membrane transporters are proteins that span the lipid bilayer and support basic biological processes in living cells by moving essential nutrients and metabolites across cell and cellular compartment boundaries, importing and exporting signaling molecules to mediate intercellular communications by maintaining physiological concentrations of metabolites and ionic species, and preventing the accumulation of toxins by virtue of transporters that function as toxin pumps (Yan 2003). Transporters represent a large and diverse group of proteins that differ in membrane topology, energy coupling mechanism and substrate specificities (Ren et al. 2007). Hence, membrane transport systems play indispensable roles in the fundamental cellular functioning and normal physiological processes of all organisms (Saier 2000).

There are a lot of technologies to screen transporters and determine their transport mechanisms. However, molecular biology experiment is a time-consuming and difficult work to determine their 3D structure with X-ray crystallography or NMR spectroscopy. Especially, among tens of thousands of proteins with known 3D structure, only a trivial proportion has yielded 3D structural information about transport systems (Kuhlbrandt and Wang 1994). Whereas the information about topology structure of transporters is of great significance in both basic research and drug discovery, so a series of automatic computational methods had been developed to identify and classify transporters effectively.

It is a fundamental view that two proteins shown homologous can be expected to exhibit strikingly similar 3D structures (Doolittle 1986). In recent years, there is a huge appearance of studies regarding the field of discriminating membrane proteins using bioinformatics technology. Several statistical prediction methods had been proposed (Chou and Elrod 1999; Gromiha and Suwa 2005; Cai and Chou 2006; Diao et al. 2008), using hidden Markov model (Martelli et al. 2002; Bagos et al. 2004), fuzzy K-NN method (Shen et al. 2006) and other machine learning techniques (Cai et al. 2003; Wang et al. 2004; Shen and Chou 2005, 2007; Gromiha and Yabuki 2008). However, the prediction of membrane proteins based on their functions has not received enough attention in the existing studies (Gromiha and Yabuki 2008). So far only a few studies in this field have used protein sequences information or evolution information. Recently, Gromiha has applied amino acid composition and occurrence into functional discrimination of membrane proteins. In their study, the neural network for discriminating the channels/pores, electrochemical potential-driven transporters and active transporters achieved the highest accuracy (64%) with amino acid composition among different machine learning algorithms. The application of amino acid occurrence improved the accuracy to 68% (Gromiha and Yabuki 2008). In addition, there are two articles emphasized on sequential evolution information. IAMPC developed by Pu was composed of five modules. Module 1–module 4 extracted a particular feature from PSSM and module 5 extracted features from protein sequence. The accuracy of each module was 88.2, 91.3, 90.5, 89.4 and 86.6%, respectively (Pu et al. 2007). MemType-2L developed by Chou was a prediction engine with which a query protein can be identified as membrane or non-membrane protein. Once a protein was identified as membrane protein, the engine would further identify its type among the membrane proteins. The evolution information was incorporated in this method by representing the protein samples with the Pse-PSSM. The overall accuracy of discriminating membrane protein types by the jackknife test was reported to be 85% (Chou and Shen 2007a, b). By virtue of the close correlation between function and structure, as reported in literature (Bartlett et al. 2002), there are a lot of conservation residues with similar function in the vicinity of the 3D structure of functional sites. These conservation residues play an important role in maintaining the structural integrity of the protein. Yet, these adjacent positions in 3D structure may span a certain distance in primary structure of a protein. Owing to the knowledge of protein fingerprinting, it is well recognized that exiguous local environment information of amino acids often plays crucial roles in identifying protein function. However, few works have tried to explore the relationship between sequential evolution information and the local sequence-order effects for predicting the functional types of membrane proteins.

In view of the above-mentioned reasons and on the purpose of further improving the prediction performance on functional discrimination of membrane proteins, we proposed a new PSIAC model based on Position-specific scoring matrix (PSSM) and auto covariance (AC). Firstly, sequential evolution information on the inquired sequences was obtained by PSI-BLAST. Secondly, the obtained PSSM was further applied AC transformation to incorporate local sequence-order effects between an amino acid and its neighbor (both in 3D structure and primary structure). Finally, Support vector machine (SVM) was applied with fivefold cross-validation. The flowchart of the experiment is described in Fig. 1. The overall accuracy is 87.51% (when using three-state classification method, the obtained accuracy is 82.21%). The sensitivities of correctly identifying channels/pores, electrochemical and active transporters are 0.897, 0.955 and 0.797, respectively, and the specificities are 0.878, 0.924 and 0.864, respectively.

Materials and methods

Data sets

Here, we manually compiled our dataset as channels/pores, electrochemical transporters and active transporters from the TCDB website (Saier et al. 2006) in January 2008. The IUBMB endorsed transporter classification database (TCDB) where the classification of families is based on their phylogeny, transport mode, energy coupling mechanism, and substrate specificity (Li et al. 2008). The TCDB contains seven groups of transporters in which three groups include few data for model construction and one is for incompletely characterized proteins (Gromiha and Yabuki 2008). Hence, we have excluded these four groups and focused on the three major transporters, namely channels/pores, electrochemical and active transporters. They contain 910, 1219 and 1697 proteins, respectively. To get rid of redundancy and

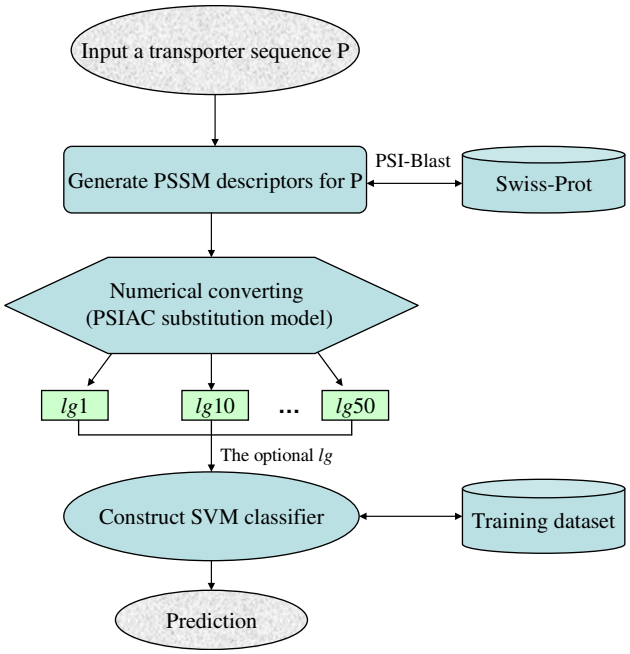

Fig. 1 Flowchart of the entire prediction system

homology bias, we have removed the redundant sequences using BLASTCLUST program (Altschul et al. 1997) so that the mutual identity in the dataset was less than 25% (Chou and Shen 2007a, b; Chou 2004). Moreover, the protein fragments with less than 50 amino acids and the sequences containing ambiguous residues such as X, B and Z were screened out. After filtering, a total of 2,373 proteins were left, which include 662 channels/pores, 670 electrochemical and 1,041 active transporters.

On the basis of the membrane protein dataset, two working datasets were constructed. The independent testing dataset contains 186 proteins by randomly extracting from our dataset proteins, of which 52 are channels/pores, 48 electrochemical transporters and 86 active transporters. The remaining 2,187 proteins were used as the training dataset (used for model construction with fivefold cross-validation), of which 610 are channels/pores, 622 electrochemical transporters and 955 active transporters. None of the 186 entries in the independent dataset occurs in the training dataset of the 2,187 entries.

Position-specific scoring matrix

Position-Specific Iterated BLAST (PSI-BLAST) is a strong measure of residue conservation in a given location and much more sensitive sequence comparison based on a gapped BLAST (Altschul et al. 1997). It is able to detect distant homologs and represent evolution information of a protein sequence. Position-specific scoring matrix is a matrix that consists of a 20-dimensional vector got by PSI-BLAST and represents the probabilities of conservation against mutations to 20 different amino acids for all residues in a given sequence. If a residue is conserved through cycles of PSI-BLAST, it is likely to be a biological function domain (Ahmad and Sarai 2005). This method has been widely used in membrane protein types prediction (Pu et al. 2007; Chou and Shen 2007a, b), protein secondary structure prediction (Jones 1999; Kaur and Raghava 2003), subcellular localization (Xie et al. 2005), and other bioinformatics problems.

To gain PSSM, each sequence in the constructed dataset was processed to search the homologous sequences against the Swiss-Prot database (version56, released on 22 July, 2008) using the PSI-BLAST with parameters e value of 0.001 and 3 iteration. The generated matrix includes $L \times 20$ elements, where L is the length of the query sequence. Each element represents the frequency of the occurrence of each 20 amino acids when it was mutated to the others at one position during the evolution process.

Auto covariance

In this study, to avoid loss of the local sequence-order information, AC variables were applied to each column of PSSM. In view of the fact that SVM requires the fixed length feature vectors as their inputs for training, AC was employed to transform these digital vectors into equal length matrices. As a powerful statistical tool published by Wold et al. (1993), it has been widely applied to the field of bioinformatics (Guo et al. 2006a, b; Wen et al. 2007; Fang et al. 2007; Doytchinova and Flower 2007). Auto covariance is a correlation factor coupling adjacent residues along a protein chain, which is a kind of variant of auto cross covariance (ACC). Compared with the other variables such as cross covariance (CC), AC variables are able to avoid producing too many variants, and have also been proved to be the most important components in ACC variables by our previous work (Guo et al. 2008). In statistics, AC is simply the covariance of the signal against a time-shifted version of itself. For protein sequence, it is the covariance of residues against those with certain distance apart along protein sequence. Here, AC variables were used to describe the average correlation between positions with a series of lag (i.e., the residue number when applied to protein sequences) apart throughout the protein sequence P. The AC can be calculated by

$$\mathrm{AC}_{\mathrm{lag},j} = \frac{1}{n-\mathrm{lag}} \sum_{i=1}^{n-\mathrm{lag}} \left(P_{i,j} - \frac{1}{n}\sum_{i=1}^{n} P_{i,j}\right) \times \left(P_{(i+\mathrm{lag}),j} - \frac{1}{n}\sum_{i=1}^{n} P_{i,j}\right) \quad (1)$$

with

$$D = \lg \times q \quad \lg = \mathrm{Max}\{\mathrm{lag} \parallel \mathrm{lag} = 1, 2, \ldots, \lg\} \qquad (2)$$

where lag denotes the distance between one residue and its neighbors, j represents one descriptor, i is the position and n is the length of the sequence. Due to the length of the shortest sequence (50 amino acids) in our dataset, the lg corresponding to the maximum lag should be smaller than 50. In terms of the AC descriptor (Eqs. 1, 2), a protein sequence P was represented by a vector of AC variables, which equals to the value of D. Here, the number of descriptors q equals to 20 (namely the dimensionality of PSSM scores), the optimal lg ($\lg = 10$) is adopted in this study by the reasons given in "Discussion". So the number of AC variables equals to 200 ($D = 20 \times 10$). Ultimately, each protein sequence was characterized by the PSIAC model, which consists of 200-D (dimensional) AC variables.

Support vector machine

Support vector machine, based on the structural risk minimization principle from statistical learning theory algorithm, had been detailedly described in the literature (Vapnik 1998). Here, we used the software libsvm 2.86 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) to construct the classification model for functional discrimination of membrane protein types. Facilitating the analysis on the capability of our model on discriminating each class, a 'one versus rest' method (Hua and Sun 2001) was adopted. In the work, the channels/pores were defined as positive samples (labeled as $+1$) and the rest two kinds were defined as negative samples (labeled as -1), etc. Besides, to facilitate comparison with other methods, we also used three-state classification method. A radial basis function (RBF) was chosen as the kernel function, and an easy search method was employed to optimize the regularization parameter C and the kernel width parameter γ.

Performance evaluation

A statistical prediction method is widely developed by jackknife test and k-fold cross-validation (Chou and Zhang 1995). The former is deemed more rigorous and objective in a series of papers (Tan et al. 2006; Guo et al. 2006a, b; Liu et al. 2005; Liu and Chou 1999). Considering the size of our dataset, the jackknife test would be time-consuming, so fivefold cross-validation was used in this work. The dataset was randomly divided into five subsets. Each was singled out in turn as a testing set and the remaining four were used for training the classifier. The process was repeated for five times and the final prediction result was the average accuracy of the five testing sets. (The accuracy for each testing sets was then averaged as the final prediction result.)

TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives, respectively. The performance of SVM classification can be measured by sensitivity, specificity, accuracy and MCC (Matthews 1975). The term 'sensitivity' shows the correct prediction of specific transporters and 'specificity' about the negative samples. Accuracy is the total accurate rate of the predictions. MCC accounts for both over- and under-predictions. These parameters can be calculated by following equations:

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad \mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \times 100\%$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}}.$$

Results and discussion

Selecting the optimal lg

As mentioned above, the maximal possible lg must be smaller than the length of the shortest sequence (50 residues) in the dataset, and the parameter lg represents the maximum distance between two considered positions in the sequence. Therefore, optimal value lg usually varies in different data sets. In our paper, in order to achieve the optimum characterization of the protein sequences, a series of lgs were investigated ($\lg = 1$, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50). The results are shown in Fig. 2. As can be observed from the curve, the overall accuracy increases when lg increases from 1 to 10, and then decreases when lg larger than 10. However, it slightly rebound when lg is 35, and then gradually reduces as lg increasing from 35 to 50. This fluctuation may be originated from the diversity and complexity of protein sequence and spatial structure.

It is obviously seen that the optimal lg is 10, corresponding to a peak with an overall accuracy of 87.51%. The observed result is in line with the viewpoint that the positions of adjacent conservation residues in 3D structure may span a certain distance in primary structure of a protein. Moreover, it proves the importance of the local environment information of amino acids.

The performance for discriminating channels/pores, electrochemical and active transporters was evaluated by fivefold cross-validation and presented in Table 1. It can be observed that electrochemical transporters are well distinguished from other two classes of transporters, with the

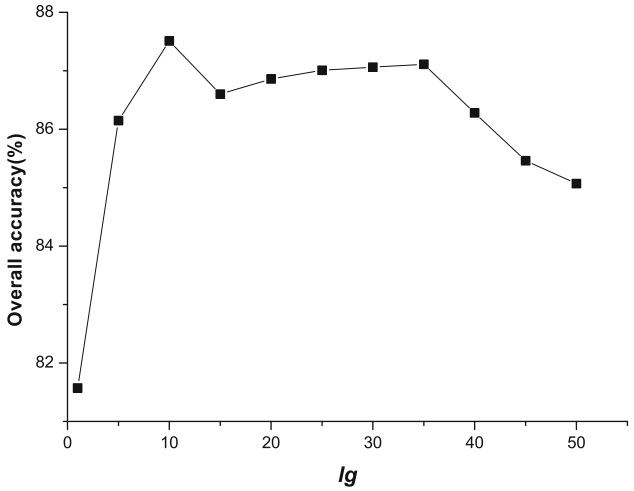

Fig. 2 The overall accuracy of the method with AC of different lgs

best sensitivity, specificity, MCC and accuracy, which are 0.955, 0.924, 0.825 and 93.08%, respectively. They are 0.897, 0.878, 0.700 and 88.19% for channels/pores, and 0.797, 0.864, 0.663 and 83.45% for active transporters. This might be due to the fact that electrochemical transporters can be better characterized using our method.

Comparison with simple PSI-BLAST search and other prediction methods

In order to highlight the advantage of our model, we analyzed the prediction capability using only PSSM. Because the proteins with different lengths will correspond to row-different matrices based on PSI-BLAST, here, Chou's method was used to generate a size-uniform matrix (Chou and Shen 2007a, b). The result is listed in Table 2. The size-uniform process roughly handled the PSSM scores without the consideration of correlation between different positions, so this method obtained the accuracy of only 60.91% in discriminating channels/pores, electrochemical and active transporters. It well proved the importance of the local sequence-order effects. In addition, we also performed the comparison of PSIAC module developed in this study with two other existing methods using amino acid composition and amino acid occurrence. For the sake of a fair comparison, the same dataset was employed to construct models. In Table 2, we observed that the obtained accuracy was 70.55% using amino acid composition and the application of amino acid occurrence improved the accuracy to 73.48%. Our method achieved the highest accuracy of 82.21%. The results demonstrated that the performance of our method is superior to the simple PSI-BLAST programmer search. It is revealed that the incorporation of evolution correlation between different positions do help to identify functional type of membrane proteins.

Performance on the independent testing dataset

In order to demonstrate and evaluate the practical prediction capability of our model, it was also tested by the aforementioned independent dataset. Here, we focus on the comparison of encoding strategies rather than algorithms, so SVM was used for all of the methods. The corresponding accuracy is listed in Table 2. It can be shown from Table 2, compared with those results gained by other methods, that the accuracy obtained by our method was 84.95% which has been significantly improved. It is consistent with our hypothesis that the model performance should be significantly improved by integrating local sequence environment.

Conclusions

Traditionally, the previous methods with only sequential information did not consider the local environment of proteins. In this study, AC was applied to PSSM. So, an inquiry protein sequence can be characterized by the sequential evolution information and the local sequence-order effects between each position and its 10 sequential vicinal positions, which adequately reflected the protein

Table 1 Performance of our method in discriminating channels/pores, electrochemical and active transporters using fivefold cross-validation test

Membrane types	No. of sequences	Correct hit	Accuracy (%)	Sensitivity	Specificity	MCC	Overall accuracy (%)
Channels/pores	610	538	88.19	0.897	0.878	0.700	–
Electrochemical transporters	622	579	93.08	0.955	0.924	0.825	–
Active transporters	955	797	83.45	0.797	0.864	0.663	–
Total	2,187	1,914	–	–	–	–	87.51

Here, we used 'one versus rest' method. For example, the channels/pores were defined as positive samples while the rest two kinds were defined as negative samples, etc.

Table 2 The overall accuracy of our method comparison with different methods by fivefold cross-validation test and independent dataset test

Protein descriptor	Fivefold cross-validation test (%)	Independent dataset test (%)
Amino acid composition[a]	70.55	72.04
Amino acid occurrence[a]	73.48	73.12
PSSM[b]	60.91	64.52
Consideration of PSSM and AC (our method)	82.21	84.95

Here, we used three-state classification method, namely all the three classes simultaneously

[a] See Gromiha and Yabuki (2008)

[b] See Chou and Shen (2007a, b)

local environment during evolution. Therefore, it is expected that this method could be effectively used to discriminate different classes of transporters in the absence of experimental structure data. Furthermore, we believe that the application sphere of our model is not limited to this field, and it can be extended to other biological problems such as the identification of enzyme functions, even the location of functional sites, which have been proved closely correlated with local environment around active sites.

Acknowledgments The authors gratefully thank Zhining Wen and the anonymous reviewers for their constructive comments on this article.

References

Ahmad S, Sarai S (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinform 6:33

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004) A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinform 5:29

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121

Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J Theor Biol 238:395–400

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

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC, Elrod DW (1999) Prediction of membrane protein types and subcellular locations. Proteins 34:137–153

Chou KC, Shen HB (2007a) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2007b) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Diao YB, Ma DC, Wen ZN, Yin JJ, Xiang J, Li ML (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel–Ziv complexity. Amino Acids 34:111–117

Doolittle RF (1986) Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley

Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 8:4

Fang YP, Guo YZ, Feng Y, Li ML (2007) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34:103–109

Gromiha MM, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 21:961–968

Gromiha MM, Yabuki Y (2008) Functional discrimination of membrane proteins using machine learning techniques. BMC Bioinform 9:135

Guo YZ, Li ML, Lu MC, Wen ZN, Huang ZT (2006a) Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins 65:55–60

Guo YZ, Li ML, Lu MC, Wen ZN, Wang KL, Li GB, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast fourier transform. Amino Acids 30:397–402

Guo YZ, Yu LZ, Wen ZN, Li ML (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030

Hua SJ, Sun ZR (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Kaur H, Raghava GP (2003) Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci 3:627–634

Kuhlbrandt W, Wang D (1994) Three-dimensional structure of plant light-harvesting complex determined by electron crystallography. Nature 350:130–134

Li HQ, Dai XB, Zhao XC (2008) A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. Bioinformatics 24:1129–1136

Liu W, Chou KC (1999) Protein secondary structural content prediction. Protein Eng 12:1041–1050

Liu H, Wang M, Chou KC (2005) Low-frequency fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739

Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 18:S46–S53

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247:259–265

Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35:D274–D279

Saier MH (2000) A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64:354–411

Saier MH, Tran CV, Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34:D181–D186

Shen HB, Chou KC (2005) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334:288–292

Shen HB, Chou KC (2007) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. J Theor Biol 240:9–13

Tan FY, Feng XY, Fang Z, Li ML, Guo YZ, Jiang L (2006) Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids 33:669–675

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Sel 17:509–516

Wen ZN, Li ML, Li YZ, Guo YZ, Wang KL (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32:277–283

Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277:239–253

Xie D, Li A, Wang M, Fan Z, Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33:W105–W110 Web Server

Yan Q (2003) Membrane transporters: methods and protocols, volume 227 of methods in molecular biology. Humana Press, Totowa