Published in final edited form as:

J Theor Biol. 2011 May 7; 276(1): 132–137. doi:10.1016/j.jtbi.2011.01.048.

Application of density similarities to predict membrane protein types based on pseudo-amino acid composition

Abbas Mahdavi^{a,1} and Samad Jahandideh^{b,*}

^aDepartment of Statistics, Faculty of Science, Shiraz University, Shiraz, Iran

^bDepartment of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Abstract

Cell membranes provide integrity of living cells. Although the stability of biological membrane is maintained by the lipid bilayer, membrane proteins perform most of the specific functions such as signal transduction, transmembrane transport, etc. Then it is plausible membrane proteins being attractive drug targets. In this article, based on the concept of using the pseudo-amino acid composition to define a protein, three different density similarities are developed for predicting the membrane protein type. The predicted results showed that the proposed approach can remarkably improve the accuracy, and might become a useful tool for predicting the other attributes of proteins as well.

Keywords

Algorithmic predictor model; Non-algorithmic predictor model; Protein structure

1. Introduction

Whole of the cell and compartments of the cell, such as the endoplasmic reticulum, Golgi apparatus, mitochondria, and other membrane-bound organelles, enclosed by membranes which in turn reach specific structures and functions for each of cell compartments. Membranes are composed mainly of lipids and proteins. Membrane proteins are embedded in the lipid bilayer, which creates a suitable environment for their actions. Although the biophysical aspect of protein–lipid interactions is inadequately understood, it is believed that most of the specific functions are carried out by the membrane proteins and their associations with the lipid bilayer (Thomas and Hidalgo, 1978; Lodish et al., 1995). Accordingly, membrane proteins are attractive targets for drug design (Feng et al., 2005, 2006; Sanders et al., 2006).

Many studies from various research laboratories around the world have indicated that mathematical analysis, computational modeling, and introducing novel physical concept to solve important problems in biology and medicine, such as protein structural class prediction (Zhou and Assa-Munt, 2001; Chou and Zhang, 1995), modeling 3D structures of targeted

*Corresponding author. Tel.: +1 205 975 9208. sjahandideh@uab.edu (S. Jahandideh).

¹Authors contributed equally to this work.

proteins for drug design (Zhou and Troy, 2005; Chou, 2004; Sharma et al., 2008; Zhou et al., 2004), diffusion-controlled reaction simulation (Zhou and Zhong, 1982; Chou and Zhou, 1982; Zhou et al., 1981, 1983), cellular responding kinetics (Qi et al., 2007), bio-macromolecular internal collective motion simulation (Chou, 1987, 1988; Zhou, 1989), identification of proteases and their types (Zhou and Cai, 2006), membrane protein type prediction (Cai et al., 2003), protein cleavage site prediction (Chou, 1993, 1996), and signal peptide prediction (Chou, 2002), can timely provide very useful information and insights for both basic research and hence are widely welcome by science community. The present study is related to the fundamental problems in system biology, network biology, and structural biology of proteins. The relationship between these systems will be of use for the global research.

Conventionally membrane proteins classified into five or eight main types. In classification system of membrane proteins to five types, these main types are including: (1) type I membrane protein, which are transmembrane proteins with C-terminal region in cytoplasmic side; (2) type II membrane protein, which are transmembrane proteins with N-terminal region in cytoplasmic side; (3) multipass transmembrane proteins; (4) lipid-chain-anchored membrane proteins; and (5) glycosylphosphatidylinositol (GPI) anchored membrane proteins (Fig. 1). In second one classification system membrane proteins classified into eight main types which is including: (1) type I transmembrane, (2) type II, (3) type III, (4) type IV, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane (Fig. 2). The five and eight mentioned types of membrane protein are classified by their modes associated with the lipid bilayer and by their biological functions.

Determination of types of membrane proteins using traditional experiments is restricted to limited number of proteins due to the inherent complications in their technical limitations in addition to time-consuming and costly disadvantages. These practical limitations encouraged the development of theoretical methods to predict the type of membrane proteins based on the available data. In addition, as a result of large-scale sequencing projects, the gap between the number of known protein sequences and number of known structures is increasing. This gap highlights the demand to theoretical methods of protein structure determination.

From historical point of view, in a pioneer work the covariant-discriminant algorithm were applied by Chou and Elrod (1999) to predict the types of membrane proteins based on the amino acid composition. Conventionally, the amino acid composition of a protein consists of 20 components, have been defined as the occurrence frequency of each of its 20 native amino acids (Chou, 1989; Nakashima et al., 1986). Obviously, if using the conventional amino acid composition as the representation for a protein, all the sequence-order and sequence-length effects would be missed. To improve this situation, pseudo-amino acid composition, as a novel concept, was proposed recently by Chou (2001). In practical applications, particularly in developing high throughput tools for predicting various biological attributes, many different pseudo-amino acid compositions for biological sequences feature representation have been developed and widely used. For example, cellular automata image (Xiao et al., 2006, 2008, 2009a; Lin et al., 2009), complexity

measure factor (Xiao et al., 2005, 2006); Grey dynamic model (Xiao et al., 2008, 2009a); and functional domain composition (Xiao et al., 2009b).

Based on the concept, so many investigations have been reported high prediction accuracy using pseudo-amino acid composition in the field of bioinformatics. Motivated by the concept of pseudo-amino acid composition, the present study was initiated in an attempt to introduce a simple and fast novel approach based on different density similarities to predict the types of membrane proteins.

According to a recent comprehensive review (Chou, 2011), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (5) establish a user-friendly web-server for the predictor that is accessible to the public. In the following sections, we describe how to deal with these steps one by one.

2. Materials and methods

2.1. Database

Two different databases were used to evaluate performance of our methods. The first database is constructed by Chou and Elrod (1999) were used to demonstrate the current method (Chou, 2001). This database contains is including two groups of proteins; 2059 membrane protein sequences, of which 435 are type I transmembrane proteins, 152 type II transmembrane proteins, 1311 multipass transmembrane proteins, 51 lipid-chain-anchored membrane proteins, and 110 GPI-anchored membrane proteins and the 2625 proteins group were also taken from Chou and Elrod (1999), in which, the number of proteins belonging to types (1)–(5) is 478, 180, 1867, 14, and 86, respectively (Fig. 1). The names of membrane proteins of this database, classified into five groups, were given in Table 1 of Chou and Elrod (1999). Moreover, as a demonstration of practical application, predictions were also conducted for the second database including two membrane proteins groups including 3249 and 4333 protein sequences. Second groups of these databases tested based on the rule parameters derived from the first groups in the training step.

2.2. Pseudo-amino acid composition

According to the classical definition, the amino acid composition contains 20 components, each reflecting the occurrence frequency of one of the 20 native amino acids in a protein. The pseudo-amino acid composition (parallel correlation type) consists of $20 + \lambda$ discrete numbers (Chou, 2001). Thus, a protein is expressed by

$$P = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix},$$

where the first 20 components are the same as those in the classic amino acid composition and the components $p_{20+1}, \dots, p_{20+\lambda}$ are related to λ different ranks of sequence-order correlation factors. For a detailed explanation of pseudo-amino acid composition, see Chou (2001). In this paper, the hydrophilic values for the 20 native amino acids were taken from Hope and Woods (1981), and the weight factor is set as 0.3 and $\lambda=40$. Generation of pseudo AA composition, easy attainable through this website: <http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/>.

2.3. Model development

Briefly after building up the method on databases we used jackknife test to verify the performance of the methods. Subsequently the method was tested on independent databases.

2.3.1. Measures of disparity between probability distributions—Discriminant analysis is a technique for classifying an observed vector into predefined classes. There are a number of different discriminant measures that may distinguish between two or more populations. In this study we will use three measures to predict membrane protein types. These three measures are as follows.

2.3.1.1. Bhattacharyya distance: In statistics, one of the methods used to measure the similarity of two probability distributions and separability of classes in classification is the Bhattacharyya distance (Bhattacharyya, 1943). The Bhattacharyya coefficient for probability distributions P and Q of a discrete random variable is defined by

$$\rho(P, Q) = \sum_{x \in S} \sqrt{p(x)q(x)}, \quad (1)$$

when $p(x) > 0$ and $q(x) > 0$ for all values of $x \in S$, $\sum_{x \in S} p(x) = \sum_{x \in S} q(x) = 1$ and S is a support for the discrete random variable (the definition for these values is the same for the three measures). The Bhattacharyya distance is defined as

$$D_B(P, Q) = -\ln(\rho). \quad (2)$$

Clearly $0 \leq \rho(P, Q) \leq 1$ and $0 \leq D_B(P, Q) < \infty$. It can be shown that D_B does not obey the triangle inequality.

2.3.1.2. Kullback–Leibler divergence: Kullback–Leibler information (Kullback and Leibler, 1951; Kullback, 1978) is a non-symmetric measure of the difference between two probability distributions P and Q , and is defined for probability distributions P and Q of a discrete random variable by

$$I(P, Q) = \sum_{x \in S} p(x) \log \left(\frac{p(x)}{q(x)} \right) \quad (3)$$

Since the Kullback–Leibler information is a non-symmetric measure of disparity, i.e. in general $I(P, Q) \neq I(Q, P)$. Then it is not a distance metric. A symmetric measure of disparity (the J -divergence) is defined as

$$J(P, Q) = I(P, Q) + I(Q, P). \quad (4)$$

The Kullback–Leibler divergence involves all the properties of distance except the triangle inequality and therefore is called quasi-distance, furthermore this measure is nonnegative, and is zero if and only if $P=Q$.

2.3.1.3. Jensen–Shannon divergence: Another symmetrized and smoothed version of the Kullback–Leibler information is the Jensen–Shannon divergence (Lin, 1991) that for two probability distributions P and Q of a random variable is defined as follows:

$$JS(P, Q) = \frac{1}{2} \left(I \left(P, \frac{P+Q}{2} \right) + I \left(Q, \frac{P+Q}{2} \right) \right) \quad (5)$$

Therefore the Jensen–Shannon divergence for two probability distributions P and Q of a discrete random variable is obtained as

$$JS(P, Q) = \frac{1}{2} \left(\sum_{x \in S} p(x) \log \left(\frac{2p(x)}{p(x)+q(x)} \right) + \sum_{x \in S} q(x) \log \left(\frac{2q(x)}{p(x)+q(x)} \right) \right) \quad (6)$$

In other words, the Jensen–Shannon divergence is the average of the two possible Kullback–Leibler information between two probability distributions.

2.3.2. Constructing densities—Let g be a nonnegative function defined on a countable set S and let $c = \sum_{x \in S} g(x)$ be total functions of set S , if c is finite, then a discrete density function on S is defined by

$$f(x) = g(x)/c \quad (7)$$

Since the values of this type of pseudo-amino acid composition are nonnegative and density similarities defined for probability distributions with equal support, we take a weight of at least 10^{-4} for the modified densities.

2.3.3. Classification method—Suppose that the density associated with the query protein is denoted by f and let f_i^j , denote the density of i th protein in j class, $j = 1, 2, 3, \dots, k$, $i = 1, 2, 3, \dots, n_j$, where k is the number of classes and n_j denote the size of j th class. Since the low similarity between proteins in each class and unbalanced database, we cannot use of mean class for prediction. Therefore we use of maximum similarity between density of query protein and density of proteins in training dataset for prediction. Thus, the prediction rule can be expressed by

$$J(f, f^m) = \min_j \{ \min_i (J(f, f_i^j)) \}, \quad (8)$$

where m can be $1, 2, 3, \dots, k$. And the operator *min* means taking minimum value in the brackets.

2.4. Evaluation of the method

The training and testing of the method was conducted by three typical statistical tests, that is, jackknife test, and testing on a different dataset, or independent dataset, not being used in the training step, as reported below. The jackknife test, in which all cases were used in both the training and testing processes, was applied to train and test models on the database. The procedure is as follows: given a training set of N proteins, the first protein in the training set, t_1 , is set aside (left out). Then the model is trained on the remaining $N-1$ proteins and tested on the left out sample. Then sample t_1 is inserted back into the database and the next protein, t_2 , is left out. This procedure is repeated until every protein in the database had the opportunity to be a left out sample. It therefore provided as many simulations as the number of samples in each database. Although this method is time-consuming, it is the most rigorous and objective one test. Moreover, as a demonstration of practical application, predictions were also conducted for the 2625 and 4333 independent membrane proteins based on the rule parameters derived from the 2059 and 3249 proteins in the training database, respectively.

3. Results and discussion

The programming and implementation of the presented algorithm was performed in R environment version 2.11.0.

3.1. Comparing results of our methods on two different databases

As shown in Table 1, the obtained results of the overall accuracies in jackknife test, and independent dataset test on the first database were 85.9%, and 96%, respectively. Comparing these results with similar works (Chou and Elrod, 1999; Chou, 2001; Feng et al., 2006) indicates the efficiency of the presented model.

The jackknife test is reported as the most effective and rigorous one; see, for example, Mardia et al. (1979) for the mathematical principle and Chou and Zhang (1995) for a comprehensive discussion about this. During jackknifing, the type of each membrane protein is identified by the rule parameters derived from using all the other membrane proteins

except the one which is being identified. The results of jackknife test on the first database are also given in Table 1.

In addition to jackknife tests, we applied trained model to test it on the independent group of first database including 2625 proteins. Obtained results of this test are also given in Table 1. High accuracy of this test once again highlights performance of this method.

In addition to first database we used a second one to discuss redundancy effect on obtained results. Membrane proteins in this database are classified into eight different classes. Obtained results of our methods presented in Table 2 show obtained accuracies in different classes and overall accuracies of three other different methods in comparison to our methods. From obtained results it is obvious that although our methods are traditional simple models overcome two of compared methods, Least Euclidean distance (Cedano et al., 1997) and protLoc (Nakashima et al., 1986), which are simple methods like ours. However in comparison to MemType-2L (Chou and Shen, 2007) our methods show lower accuracies and the major reasons of difference is using complicate process to generate input parameters and also using KNN classifier, which is a well known accurate machine learning method. On the other hand, in comparison to MemType-2L, advantages of our method are simplicity of methods and effortlessness of generating input parameters, which in turn result in reduction of time consumption.

Performance measure values of non-algorithmic models such as SVMs and ANNs in comparison with algorithmic, or statistical models, such as logistic regression model (LRM), Linear Discriminant Analysis (LDA) and density similarities, presented here, is truly high in situations that we are dealing with a large number of features. Then we could not be sure non-algorithmic models provide higher prediction accuracy in more simple problems. On the other hand, performance of predictor models is dependent on characteristics of database. For example, a database with high sequence identity results in high performance measures values. In these cases obtained higher prediction accuracy is more probable using algorithmic models.

In the procedure of applying the density similarities for prediction of membrane protein types, we concluded that these methods can extended to be applied for other biological or even more general problems suffering from data samples imbalance in different groups to be classified. This would in turn serve us with simple, fast and more accurate tools to be used in related disciplines such as rational drug design and drug discovery studies.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors (Chou and Shen, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgments

S.J. is supported by an NIH Grant Number R00RR024163.

References

- Bhattacharyya A. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*. 1943; 35:99–110.
- Cai YD, Zhou GP, Chou KC. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys J*. 2003; 84:3257–3263. [PubMed: 12719255]
- Cedano J, Aloy P, P'erez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. *J Mol Biol*. 1997; 266:594–600. [PubMed: 9067612]
- Chou KC. The biological functions of low-frequency phonons: 6. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers*. 1987; 26:285–295. [PubMed: 3828475]
- Chou KC. Review: low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem*. 1988; 30:3–48. [PubMed: 3046672]
- Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem*. 1993; 268:16938–16948. [PubMed: 8349584]
- Chou KC. Review: prediction of HIV protease cleavage sites in proteins. *Anal Biochem*. 1996; 233:1–14. [PubMed: 8789141]
- Chou KC. Prediction of protein cellular attributes using pseudoamino-acid-composition. *Protein Struct Funct Genet*. 2001; 43:246–255.
- Chou KC. Review: prediction of protein signal sequences. *Curr Protein Pept Sci*. 2002; 3:615–622. [PubMed: 12470215]
- Chou KC. Review: structural bioinformatics and its impact to biomedical science. *Curr Med Chem*. 2004; 11:2105–2134. [PubMed: 15279552]
- Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J Theor Biol*. 2011; 10.1016/j.jtbi.2010.12.024
- Chou KC, Shen HB. MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun*. 2007; 360:339–345. [PubMed: 17586467]
- Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. *Proteins Struct Funct Genet*. 1999; 34:137–153. [PubMed: 10336379]
- Chou KC, Shen HB. Review: recent advances in developing web-servers for predicting protein attributes. *Natural Science*. 2009; 2:63–92. (openly accessible at <<http://www.scirp.org/journal/NS/>>).
- Chou KC, Zhang CT. Prediction of protein structural classes. *Crit Rev Biochem Mol Biol*. 1995; 30:275–349. [PubMed: 7587280]
- Chou KC, Zhou GP. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J Am Chem Soc*. 1982; 104:1409–1413.
- Chou, PY. Amino acid composition of four classes of proteins. Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent; Las Vegas.. 1980.
- Chou, PY. Chou. In: Fasman, GD., editor. Prediction of protein structure and the principles of protein conformation. Plenum Press; New York: 1989. p. 549-586.
- Feng ZP, Keizer DW, Stevenson RA, Yao S, Babon JJ, Murphy VJ, Anders RF, Norton RS. Structure and inter-domain interactions of domain II from the blood-stage malarial protein, apical membrane antigen 1. *J Mol Biol*. 2005; 350:641–656. [PubMed: 15964019]
- Feng ZP, Zhang X, Han P, Arora N, Anders RF, Norton RS. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol Biochem Parasitol*. 2006; 150:256–267. [PubMed: 17010454]
- Hope TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci USA*. 1981; 78:3824–3828. [PubMed: 6167991]
- Kullback, S. Information theory and statistics. Peter Smith; Gloucester: 1978.
- Kullback SR, Leibler A. On information and sufficiency. *The Ann Math Stat*. 1951; 22:79–86.
- Lin WZ, Xiao X, Chou KC. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng Des Sel*. 2009; 22(11):699–705. [PubMed: 19776029]

- Lin J. Divergence measures based on the shannon entropy. *IEEE J Info Theor.* 1991; 37:145–151.
- Liu H, Wang M, Chou KC. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem Biophys Res Commun.* 2005; 336:737–739. [PubMed: 16140260]
- Lodish, HD.; Baltimore, A.; Berk, SL.; Zipursky, P.; Matsudaira, J.; Darnell. *Molecular cell biology.* Vol. Chapter 3. Scientific American Books; New York: 1995.
- Mardia, KV.; Kent, JT.; Bibby, JM. *Multivariate analysis.* Vol. 322. Academic Press; London: 1979. p. 381
- Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. *J Biochem.* 1986; 99:152–162.
- Qi JP, Shao SH, Li DD, Zhou GP. A dynamic model for the p53 stress response networks under ion radiation. *Amino Acids.* 2007; 33:75–83. [PubMed: 17072789]
- Sanders PR, Kats LM, Drew DR, O'Donnell RA, O'Neill M, Maier AG, Coppel RL, Crabb BS. A set of glycosylphosphatidyl inositol-anchored membrane proteins of *Plasmodium falciparum* is refractory to genetic deletion. *Infect Immun.* 2006; 74:4330–4338. [PubMed: 16790807]
- Spiess M. Heads or tails—what determines the orientation of proteins in the membrane. *FEBS Lett.* 1995; 369:76–79. [PubMed: 7641889]
- Sharma AK, Zhou GP, Kupferman J, Surks HK, Christensen EN, Chou JJ, Mendelsohn ME, Rigby AC. Probing the interaction between the coiled coil leucine zipper of cGMP-dependent protein kinase I α and the C terminus of the myosin binding subunit of the myosin light chain phosphatase. *J Biol Chem.* 2008; 283(47):32860–32869. [PubMed: 18782776]
- Thomas DD, Hidalgo C. Rotational motion of the sarcoplasmic reticulum Ca²⁺-ATPase. *Proc Natl Acad Sci USA.* 1978; 75:5488–5492. [PubMed: 214789]
- Xiao X, Lin WZ, Chou KC. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *J Comput Chem.* 2008; 29:2018–2024. [PubMed: 18381630]
- Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC. Using complexity measure factor to predict protein subcellular location. *Amino Acids.* 2005; 28:57–61. [PubMed: 15611847]
- Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids.* 2006; 30:49–54. [PubMed: 16044193]
- Xiao X, Shao SH, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J Comput Chem.* 2006; 27:478–482. [PubMed: 16429410]
- Xiao X, Wang P, Chou KC. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. *J Theor Biol.* 2008; 254:691–696. [PubMed: 18634802]
- Xiao X, Wang P, Chou KC. Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. *J Appl Crystallogr.* 2009a; 42:169–173.
- Xiao X, Wang P, Chou KC. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J Comput Chem.* 2009b; 30(9):1414–1423. [PubMed: 19037861]
- Zhou GP. Biological functions of soliton and extra electron motion in DNA structure. *Phys Scr.* 1989; 40:698–701.
- Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. *Protein Struct Funct Genet.* 2001; 44:57–59.
- Zhou GP, Cai YD. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. *Protein Struct Funct Genet.* 2006; 63:681–684.
- Zhou GP, Li TT, Chou KC. The flexibility during the juxtaposition of reacting groups and the upper limits of enzyme reactions. *Biophys Chem.* 1981; 14:277–281. [PubMed: 7326350]
- Zhou GP, Surks HK, Schnell JR, Chou JJ, Michael E, Mendelsohn ME, Rigby AC. The three-dimensional structure of the cGMP-dependent protein kinase I α leucine zipper domain and its interaction with the myosin binding subunit. *Blood.* 2004; 104:963a.

- Zhou GP, Troy FA. NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. *Curr Protein Pept Sci.* 2005; 6:399–411. [PubMed: 16248792]
- Zhou GQ, Zhong WZ. Diffusion-controlled reactions of enzymes. A comparison between Chou's model and Alberty–Hammes–Eigen's model. *Eur J Biochem.* 1982; 128:383–387. [PubMed: 7151785]
- Zhou GZ, Wong MT, Zhou GQ. Diffusion-controlled reactions of enzymes. An approximate analytic solution of Chou's model. *Biophys Chem.* 1983; 18:125–132. [PubMed: 6626685]

Fig. 1.

Schematic drawing showing the following five types of membrane proteins: (1) type I transmembrane, (2) type II transmembrane, (3) multipass transmembrane, (4) lipid-chain-anchored membrane, and (5) GPI-anchored membrane. Reproduced from Chou (2001) with permission.

Fig. 2.

Schematic drawing showing the following five types of membrane proteins: (1) type I transmembrane, (2) type II, (3) type III, (4) type IV, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane. Reproduced from Chou and Shen (2007) with permission. see Spiess (1995) for a detailed description about their difference.

Table 1

Comparing obtained results of different methods on the first database.

Classifier	Test method	
	Jackknife (%) ^a	Independent database (%) ^b
Least Hamming distance (Chou, 1980)	62.10	66.70
Least Euclidean distance (Nakashima et al., 1986)	62.80	69.20
ProtLock (Cedano et al., 1997)	65.50	63.80
Covariant-discriminant algorithm (Chou and Elrod, 1999)	76.40	79.40
Covariant-discriminant algorithm and pseudo-amino acid composition (Chou, 2001)	80.90	87.50
Low-frequency Fourier spectrum (Liu et al., 2005)	78.00	87.00
Bhattacharyya	86.01	96.50
Kullback–Leibler	85.96	96.50
Jensen–Shannon	85.96	96.53

^aConducted for the 2059 membrane proteins in the training database classified into five different types as described in the text and Fig. 1.

^bConducted for the 2625 proteins in the independent testing database based on the rule parameters derived from the 2059 membrane proteins in the training dataset.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Comparing obtained results of different methods.

Type of membrane protein	Least Euclidean distance (%)	ProtLoc (%)	MemType-2L (%)	Bhattacharyya (%)	Kullback–Leibler (%)	Jensen–Shannon(%)
Jackknife test on the second database						
Single-pass type I	$\frac{329}{610}=53.90$	$\frac{329}{610}=53.90$	$\frac{532}{610}=87.20$	$\frac{510}{610}=83.61$	$\frac{511}{610}=83.77$	$\frac{512}{610}=83.93$
Single-pass type II	$\frac{227}{312}=72.76$	$\frac{125}{312}=40.10$	$\frac{227}{312}=72.80$	$\frac{166}{312}=53.21$	$\frac{166}{312}=53.21$	$\frac{166}{312}=53.21$
Single-pass type II	$\frac{9}{24}=37.50$	$\frac{8}{24}=33.3$	$\frac{10}{24}=41.70$	$\frac{7}{24}=29.17$	$\frac{7}{24}=29.17$	$\frac{7}{24}=29.17$
Single-pass type IV	$\frac{32}{44}=72.70$	$\frac{29}{44}=65.90$	$\frac{33}{44}=75.00$	$\frac{21}{44}=47.73$	$\frac{22}{44}=50.00$	$\frac{20}{44}=45.45$
Multipass	$\frac{948}{1316}=72.00$	$\frac{834}{1316}=63.40$	$\frac{1260}{1316}=95.70$	$\frac{1210}{1316}=91.95$	$\frac{1213}{1316}=92.17$	$\frac{1210}{1316}=91.95$
Lipid-chain-anchor	$\frac{67}{151}=44.40$	$\frac{69}{151}=45.70$	$\frac{85}{151}=56.30$	$\frac{66}{151}=43.71$	$\frac{68}{151}=45.03$	$\frac{65}{151}=43.05$
GPI-anchor	$\frac{125}{182}=42.30$	$\frac{78}{182}=42.90$	$\frac{372}{444}=83.78$	$\frac{111}{182}=60.99$	$\frac{111}{182}=60.99$	$\frac{111}{182}=60.99$
Peripheral	$\frac{126}{610}=20.70$	$\frac{78}{610}=12.80$	$\frac{491}{610}=80.50$	$\frac{393}{610}=64.43$	$\frac{397}{610}=65.08$	$\frac{395}{610}=64.75$
Overall	$\frac{1679}{3249}=51.70$	$\frac{1688}{3249}=52.00$	$\frac{2763}{3249}=85.00$	$\frac{2484}{3249}=76.45$	$\frac{2495}{3249}=76.79$	$\frac{2486}{3249}=76.52$
Independent database test on the second database						
Single-pass type I	$\frac{229}{444}=51.60$	$\frac{241}{444}=54.30$	$\frac{386}{444}=86.90$	$\frac{372}{444}=83.78$	$\frac{370}{444}=83.33$	$\frac{372}{444}=83.78$
Single-pass type II	$\frac{27}{78}=34.60$	$\frac{35}{78}=44.90$	$\frac{55}{78}=70.50$	$\frac{43}{78}=55.13$	$\frac{43}{78}=55.13$	$\frac{43}{78}=55.13$
Single-pass type II	$\frac{1}{6}=16.67$	$\frac{2}{6}=33.33$	$\frac{1}{6}=16.67$	$\frac{1}{6}=16.67$	$\frac{1}{6}=16.67$	$\frac{1}{6}=16.67$
Single-pass type IV	$\frac{7}{12}=58.33$	$\frac{8}{12}=66.67$	$\frac{8}{12}=66.67$	$\frac{7}{12}=58.33$	$\frac{8}{12}=66.67$	$\frac{7}{12}=58.33$
Multipass	$\frac{2282}{3265}=69.90$	$\frac{1148}{3265}=35.20$	$\frac{3103}{3265}=95.00$	$\frac{2957}{3265}=90.57$	$\frac{2955}{3265}=90.51$	$\frac{2956}{3265}=90.54$
Lipid-chain-anchor	$\frac{14}{38}=36.80$	$\frac{15}{38}=39.50$	$\frac{16}{38}=42.10$	$\frac{13}{38}=34.21$	$\frac{13}{38}=34.21$	$\frac{13}{38}=34.21$
GPI-anchor	$\frac{18}{46}=39.10$	$\frac{22}{46}=47.80$	$\frac{35}{46}=76.10$	$\frac{29}{46}=63.04$	$\frac{29}{46}=63.04$	$\frac{29}{46}=63.04$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Type of membrane protein	Least Euclidean distance (%)	ProtLoc (%)	MemType-2L (%)	Bhattacharyya (%)	Kullback-Leibler (%)	Jensen-Shannon(%)
Peripheral	81 = 18.20 444	139 = 31.30 444	365 = 82.20 444	299 = 67.34 444	296 = 66.67 444	296 = 66.67 444
Overall	2659 = 61.40 4333	1610 = 37.20 4333	3970 = 91.60 4333	3721 = 85.88 4333	3715 = 85.74 4333	3717 = 85.78 4333