Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Theoretical Biology 226 (2004) 373–376

Journal of Theoretical Biology

www.elsevier.com/locate/jtbi

Application of SVM to predict membrane protein types

Yu-Dong Cai[a,*], Pong-Wong Ricardo[b], Chih-Hung Jen[c], Kuo-Chen Chou[d,e]

[a] *Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai 200233, China*
[b] *Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*
[c] *Bioinformatics Group, School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK*
[d] *Gordon Life Science Institute, 13784 Tomey Del Mar Drive, San Diego, CA 92130, USA*
[e] *Tianjin Institute of Bioinformatics & Drug Discovery (TIBDD), Tianjin, China*

Received 29 May 2003; received in revised form 22 August 2003; accepted 28 August 2003

Abstract

As a continuous effort to develop automated methods for predicting membrane protein types that was initiated by Chou and Elrod (PROTEINS: Structure, Function, and Genetics, 1999, 34, 137–153), the support vector machine (SVM) is introduced. Results obtained through re-substitution, jackknife, and independent data set tests, respectively, have indicated that the SVM approach is quite a promising one, suggesting that the covariant discriminant algorithm (Chou and Elrod, Protein Eng. 12 (1999) 107) and SVM, if effectively complemented with each other, will become a powerful tool for predicting membrane protein types and the other protein attributes as well.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Type I membrane protein; Type II membrane protein; Multipass transmembrane proteins; Lipid chain-anchored membrane proteins; GPI-anchored membrane proteins; Chou's invariance theorem

1. Introduction

A cell is enclosed by the plasma membrane (cell envelope). Inside the cell there are various organelles such as the endoplasmic reticulum, Golgi apparatus, mitochondria, and other membrane-bound organelles. Although the basic structure of biological membranes is provided by the lipid bilayer, most of the specific functions are carried out by the membrane proteins. Among membrane proteins, some of them are transmembrane proteins. They contain one or more transmembrane segments with one or more hydrophobic segments to ensure stable association with the hydrophobic interior of the membrane, and hence is relatively easily discriminated from non-membrane proteins (Rost et al., 1995). The other membrane proteins are anchored membrane proteins. They do not have the hydrophobic membrane spanning portions, but they have a consensus sequence motif at either the N- or C-terminus. So they also can be relatively easily discriminated from non-membrane proteins (Casey, 1995; Resh, 1994). In this paper, the discrimination is confined within the scope of membrane proteins only. This is because membrane proteins can be reliably distinguished by using existing methods, as elaborated by many previous investigators (Chou, 2000; Chou and Elrod, 1999a; Reinhardt and Hubbard, 1998).

The way that a membrane-bound protein is associated with the lipid bilayer usually reflects its function. For example, the transmembrane proteins can function on both sides of membrane and transport molecules from one side to the other; whereas the proteins that associated with one side of the lipid monolayer can only function on that side. Accordingly, it would greatly expedite the process of determining the function of new proteins if an automated method is available to identify the types of membrane proteins. For a detailed discussion about this, see a recent review (Chou, 2000). The first automated method for identifying membrane protein type and location was developed by Chou and Elrod (1999a). In that pioneer study, membrane proteins were classified into (1) type I membrane protein, (2) type II membrane protein, (3) multipass transmembrane proteins, (4) lipid chain-anchored membrane proteins, and (5) GPI-anchored membrane proteins (see Figs. 1–3

*Corresponding author. Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1 QD, UK. Tel.: +44-161-200-8936; fax: +44-161-236-0409.

E-mail address: y.cai@umist.ac.uk (Y.-D. Cai).

0022-5193/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2003.08.015

of Chou and Elrod, 1999a). A much more brief and clear illustration about the five membrane protein types can also be found in Fig. 3 of Chou (2001) or Fig. 3 of Chou (2002). Based on such a classification scheme, the elegant covariant discriminant algorithm, which is a combination of the Mahalanobis distance (Mahalanobis, 1936) and Chou's invariance theorem (Chou, 1995), was introduced for identifying the type for a given membrane protein Chou and Elrod (1999a).

In this paper, we would like to propose a different approach, i.e. the SVM to deal with this problem.

2. Support vector machine

Support vector machines are a kind of learning machine based on statistical learning theory. The most remarkable characteristics of SVMs are the absence of local minima, the sparseness of the solution, and the use of the kernel-induced feature spaces. The basic idea of applying SVMs to pattern classification can be outlined as follows. First, map the input vectors into a feature space (possible with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Then, within the feature space, seek an optimized linear division; i.e. construct a hyperplane which can separate two classes (this can be extended to multi-classes) with the least error and maximal margin. The SVMs training process always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is given in the book by Vapnik (1998).

Support vector machines have been used to deal with protein fold recognition (Ding and Dubchak, 2001), protein–protein interaction prediction (Bock and Gough, 2001) and protein secondary structure prediction (Hua and Sun, 2001).

In this paper, the Vapnik's SVM (Vapnik, 1995) was introduced to predict protein sub-cellular location. Specifically, the SVMlight, which is an implementation (in C Language) of SVM for the problems of pattern recognition, was used for computations. The optimization algorithm used in SVMlight can be found in Joachims (1999). Given a set of N samples, i.e. a series of input vectors

$$\mathbf{X}_k \in \Re^{\tau} \quad (k = 1, \ldots, N), \tag{1}$$

where $\mathbf{X}_k$ can be regarded as the kth protein or vector defined in the 20-D space according to the amino acid composition (Chou, 1995), and $\Re^{\tau}$ is a Euclidean space with τ dimensions. Thus, all the relevant mathematical principles can be found in the aforementioned references, and hence there is no need to repeat here.

However, there is one point that needs to be explicitly addressed here. Although SVM only deal with two-class, the multi-class identification problem can always be converted into a two-class identification problem. In this paper which actually involves a five-class problem, we used the "all-versus-all" method to transfer it into a two-class problem (Ding and Dubchak, 2001).

3. Results and discussion

The same training data set originally constructed by Chou and Elrod (1999a) was used for the current study. It contains 2059 membrane protein sequences, of which 435 are type I transmembrane proteins (Fig. 1a), 152 type II transmembrane proteins (Fig. 1b), 1311 multipass transmembrane proteins (Fig. 1c), 51 lipid-chain anchored membrane proteins (Fig. 1d), and 110 GPI anchored membrane proteins (Fig. 1e). The names of

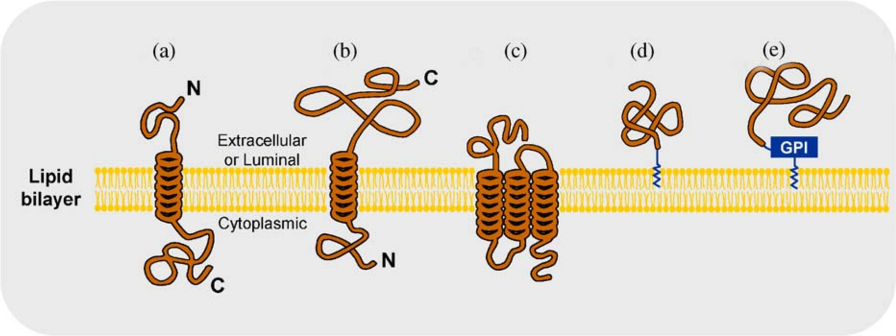

Fig. 1. Schematic drawing showing the following five types of membrane proteins: (a) type I transmembrane, (b) type II transmembrane, (c) multipass transmembrane, (d) lipid-chain anchored membrane, and (e) GPI-anchored membrane. As shown from the figure, although both type I and type II membrane proteins are of single-pass transmembrane, type I has a cytoplasmic C-terminus and an extracellular or luminal N-terminus for plasma membrane or organelle membrane, respectively, while the arrangement of N- and C-termini in type II membrane proteins is just reverse. No such distinction was drawn between the extracellular (or luminal) and cytoplasmic sides for the other three types in the current classification scheme. Reproduced from Fig. 3 of Chou (2001) with permission.

the 2059 membrane proteins, classified into five groups, were given in Table 1 of Chou and Elrod (1999a).

Following (Chou and Elrod, 1999a), a protein is represented by a point or a vector in a 20-D space according to its amino acid composition. In other words, the amino acid composition was taken as the input for the SVM operation. The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000). The width of the Gaussian RBFs selected was such that it minimized an estimate of the VC-dimension. The parameter C that controlled the error-margin trade-off was set at 100. After being trained, the hyperplane output by the SVM was obtained, indicating that the trained model had the function to identify the membrane protein types.

The demonstration was conducted by three most typical approaches in statistical prediction (Chou and Zhang, 1995); i.e. the re-substitution test, jackknife test, and independent data set test, as reported below.

3.1. Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of an identification method. When the re-substitution test is performed for the current study, the sub-cellular location of each protein in the data set is in turn identified using the rule parameters derived from the same data set, the so-called training data set. The success rate thus obtained for the five membrane protein types (Fig. 1) were successively 417/435 = 95.9%, 131/152 = 86.2%, 1292/1311 = 98.6%, 50/51 = 98.0%, and 90/110 = 81.8%, with an overall rate of 1980/2059 = 96.2%, indicating that after being trained, the SVMs model has grasped the complicated relationship between the amino acid composition and the types of membrane proteins. However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Accordingly, the success rate thus obtained represents some sort of optimistic estimation (Cai, 2001; Chou, 1995; Chou and Elrod, 1999b; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003). Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. An identification algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating an identification method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of an identification method in practical application. This is important especially for checking the validity of a training database: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

3.2. Jackknife test

As is well known, the independent data set test, sub-sampling test and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one; see, e.g. Chou and Zhang (1995) for a comprehensive discussion about this, and Mardia et al. (1979) for the mathematical principle. During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule-parameters are calculated based on the remaining proteins. In other words, the sub-cellular location of each protein is identified by the rule parameters derived using all the other proteins except the one that is being identified. During the process of jackknifing both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. The overall success rate thus obtained for the 2059 membrane proteins was 1655/2059 = 80.4%, which is 3.6% higher than the corresponding success rate obtained by the covariant discriminant algorithm (Chou and Elrod, 1999a).

3.3. Independent data set test

Moreover, as a demonstration of practical application, predictions were also conducted for an independent data set based on the rule-parameters derived from the 2059 proteins in the training data set. The independent data set was also adopted from Chou and Elrod (1999a). It consists of 2625 membrane proteins, of which 478 are type I transmembrane proteins, 180 type II transmembrane proteins, 1867 multi-pass transmembrane proteins, 14 lipid-chain anchored membrane proteins, and 86 GPI anchored membrane proteins. The overall success rate was 2243/2625 = 85.4%, which is 4.5% higher than the corresponding success rate obtained by the covariant discriminant algorithm (Chou and Elrod, 1999a).

Finally, it is instructive to conduct an analysis of the sequence identity for the membrane proteins studied here. The sequence identity percentage between two protein sequences is defined as follows. Suppose the maximum number of residues matched by sliding one sequence along the other is M, and the alignment length is L, the sequence identity between the two sequences is defined as M/L. The treatment for gaps is according to CLUSTALW (Thompson et al., 1994). The average sequence identities obtained by the sequence match operation for the 5 membrane protein subsets (Fig. 1) in the training data set are 0.079581, 0.079658, 0.077192, 0.101482, and 0.079054, respectively. From these data we can see that the majority of sequences in a same subset have very low sequence identity, a clear indication of exclusion of redundant and homologous sequences, which is fully consistent with the redundancy-excluding procedures described in Chou and Elrod (1999a) during constructing the working data sets.

4. Conclusion

The results obtained from the current study, together with those by the covariant discriminant prediction algorithm (Chou and Elrod, 1999a), have indicated that the types of membrane proteins is considerably correlated with their amino acid composition. For the case studied here, the SVM yields better results than the covariant discriminant algorithm; but for some others, e.g. in the case of predicting the G-protein coupled receptor (GPCR) type (Chou and Elrod, 2002), the outcomes were just reverse. Therefore, in practical application, it would be wise to complement the two prediction algorithms with each other. To further improve the prediction quality, it is necessary to take into account the sequence-order effect. How to develop a statistical prediction algorithm that can effectively reflect the sequence-order effect is a critical challenge in this area.

References

Bock, J.R., Gough, D.A., 2001. Predicting protein-protein interactions from primary structure. Bioinformatics 17, 455–460.

Cai, Y.D., 2001. Is it a paradox or misinterpretation. Proteins: Struct. Funct. Genet. 43, 336–338.

Casey, P.J., 1995. Protein lipidation in cell signalling. Science 268, 221–225.

Chou, K.C., 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Struct. Funct. Genet. 21, 319–344.

Chou, K.C., 2000. Review: prediction of protein structural classes and subcellular locations. Curr. Protein Peptide Sci. 1, 171–208.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins: Struct. Funct. Genet. 43, 246–255 (Erratum: Proteins: Struct. Funct. Genet. 2001, Vol. 44, 60).

Chou, K.C., 2002. A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer, P.W., Lu, Q. (Eds.), Gene Cloning & Expression Technologies. Eaton Publishing, Westborough, MA, pp. 57–70 (Chapter 4).

Chou, K.C., Elrod, D.W., 1999a. Prediction of membrane protein types and subcellular locations. Proteins: Struct. Funct. Genet. 34, 137–153.

Chou, K.C., Elrod, D.W., 1999b. Protein subcellular location prediction. Protein Eng. 12, 107–118.

Chou, K.C., Elrod, D.W., 2002. Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1, 429–433.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Ding, C.H., Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17, 349–358.

Hua, S.J., Sun, Z.R., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge, pp. 169–184.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proc. Natl Inst. Sci. India 2, 49–55.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. In: Multivariate Analysis. Academic Press, London, pp. 322, 381.

Reinhardt, A., Hubbard, T., 1998. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. 26, 2230–2236.

Resh, M.D., 1994. Myristylation and palmitylation of Src family members: the fats of the matter. Cell 76, 411–413.

Rost, B., Casadio, R., Fariselli, P., Sander, C., 1995. Transmembrane helices predicted at 95% accuracy. Protein Sci. 4, 521–533.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, Berlin.

Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.

Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct. Funct. Genet. 44, 57–59.

Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct. Funct. Genet. 50, 44–48.