Prediction of Transmembrane Proteins from Their Primary Sequence by Support Vector Machine Approach

C.Z. Cai[1,2], Q.F. Yuan[1,2], H.G. Xiao[1,2], X.H. Liu[1],
L.Y. Han[2], and Y.Z. Chen[2]

[1] Department of Applied Physics, Chongqing University, Chongqing 400044, China
caiczh@gmail.com
[2] Department of Pharmacy, National University of Singapore, Singapore 117543

Abstract. Prediction of transmembrane (TM) proteins from their sequence facilitates functional study of genomes and the search of potential membrane-associated therapeutic targets. Computational methods for predicting TM sequences have been developed. These methods achieve high prediction accuracy for many TM proteins but some of these methods are less effective for specific class of TM proteins. Moreover, their performance has been tested by using a relatively small set of TM and non-membrane (NM) proteins. Thus it is useful to evaluate TM protein prediction methods by using a more diverse set of proteins and by testing their performance on specific classes of TM proteins. This work extensively evaluated the capability of support vector machine (SVM) classification systems for the prediction of TM proteins and those of several TM classes. These SVM systems were trained and tested by using 14962 TM and 12168 NM proteins from Pfam protein families. An independent set of 3389 TM and 6063 NM proteins from curated Pfam families were used to further evaluate the performance of these SVM systems. 90.1% and 86.7% of TM and NM proteins were correctly predicted respectively, which are comparable to those from other studies. The prediction accuracies for proteins of specific TM classes are 95.6%, 90.0%, 92.7% and 73.9% for G-protein coupled receptors, envelope proteins, outer membrane proteins, and transporters/channels respectively; and 98.1%, 99.5%, 86.4%, and 98.6% for non-G-protein coupled receptors, non-envelope proteins, non-outer membrane proteins, and non-transporters/non-channels respectively. Tested by using a significantly larger number and more diverse range of proteins than in previous studies, SVM systems appear to be capable of prediction of TM proteins and proteins of specific TM classes at accuracies comparable to those from previous studies. Our SVM systems – SVMProt, can be accessed at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

1 Introduction

Transmembrane (TM) proteins play important roles for signaling, transport, recognition and interaction with extracellular molecules [1,2,3,4]. Many TM proteins, such as G-protein coupled receptors and channels, have been explored as

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 525–533, 2006.

therapeutic targets [5,6,7]. Membrane-bound transporters are responsible for absorption and excretion of drugs as well as cellular molecules [8,9]. Thus prediction of TM proteins is important for facilitating functional study of genomes, understanding molecular mechanism of diseases, and for searching new therapeutic targets.

Although TM proteins can be determined by experimental methods such as antibody-binding analysis and C-terminal fusions with indicator proteins [10,11], the number of experimentally determined TM proteins is significantly smaller than the estimated TM proteins in genomes [12,13,14]. Thus computational methods have been developed for facilitating the prediction of TM sequences [13,14,15,16,17,18]. These methods are capable of achieving high prediction accuracy for TM proteins and they can satisfactorily distinguish between TM and globular proteins and between TM and signal peptides. A study of 14 TM protein prediction methods using 270 helical TM chains, 1,418 signal peptides and 616 globular proteins showed that ∼95% TM helices are correctly predicted as TM proteins and ∼92% of globular proteins are correctly predicted as non-membrane (NM) proteins by the best methods [13]. A more recent study showed that ∼95% of the 125 TM proteins and ∼99% of the 526 soluble proteins can be correctly predicted by using a modified algorithm [18].

These methods have been developed and tested by using a few hundred to several hundred TM sequences and a slightly higher number of NM proteins. Our search of Swissprot database http://www.expasy.ch/sprot (Swissprot release 44.1, [19]) showed that there are 18358 TM protein sequences and over 134,000 NM proteins. Thus these methods may preferably need to be more adequately tested and trained by using a more diverse set of proteins. Previous studies also revealed that some TM prediction methods tend to predict proteins with more than 5 TM helices at a lower accuracy [13], which affects their prediction capability for such therapeutically and biologically relevant TM proteins as G-protein coupled receptors and certain types of channels and transporters [20]. Some methods have been found to be less capable of distinguishing between signal peptides and membrane helices [13,14]. Therefore, it is useful to evaluate the performance of TM protein prediction methods on specific therapeutically and biologically important classes of TM proteins.

The performance of a statistical learning method, support vector machine (SVM), for the prediction of TM proteins was evaluated in this work by using a diverse set of TM proteins and NM proteins. It was also tested on specific classes of TM proteins. SVM is a relatively new and promising algorithm for binary classification by means of supervised learning and it appears to be more superior than other statistical learning methods [21]. SVM has been applied to the prediction of TM proteins [15,17] and a specific TM class of G-protein coupled receptors [22] as well as other proteins [23,24,25,26,27,28,29,30,31]. These SVM TM protein prediction systems were not trained and tested by using a sufficiently diverse set of TM and NM proteins. Therefore, in this work, a large number of TM and NM proteins were used to train and test a SVM system. SVM systems were also trained and tested for the prediction of therapeutically and biologically

important individual classes of TM proteins including G-protein coupled receptors, envelope proteins, outer membrane proteins, and transporters/channels.

Many of the proteins in these four classes either contain more than 5 TM helices or have non-helix TM segments. For instance, G-protein coupled receptors contain 7 TM helices as well as intracellular and extracellular domains. Envelope proteins are located in viral lipoprotein membranes which form the outermost layer of the virion in certain viruses. Outer membrane proteins are located in the outer membrane of organelles like mitochondria, chloroplasts and some eubacteria which are surrounded by a double membrane. Almost all TM transport processes are mediated by integral TM proteins, sometimes functioning in conjunction with extracytoplasmic receptors or receptor domains as well as with cytoplasmic energy-coupling and regulatory proteins or protein domains. Thus the four classes of TM proteins have their own characteristics and they are useful for testing the performance of SVM classification.

2 Methods

2.1 Selection of Transmembrane Proteins and Non-membrane Proteins

All TM proteins used in this study were from a comprehensive search of Swissprot database (Swissprot release 44.1, TrEMBL release 27.1, [19]). A total of 18,358 TM protein sequences were obtained, which include 8,457 G-protein coupled receptors, 450 envelope proteins, 1,492 outer membrane proteins, and 980 transporters and channels. All distinct members in each group were used to construct positive samples for training, testing and independent evaluation of SVM classification system. Multiple entries for a distinct protein were evenly distributed to the training, testing, and independent evaluation set.

The negative samples, i.e. NM proteins, for training and testing our SVM classification systems were selected from seed proteins of the more than 3886 curated protein families in the Pfam database [32] that have no TM protein as a family member. Each negative set contains at least one randomly selected seed protein from each of the Pfam families. For each sub-group of non-G-protein coupled receptor, non-envelope protein, non-out membrane protein, or non-transporter/non-channel, distinct members in the other four sub-groups were added to the negative samples of each of the training, testing and independent evaluation set. For instance, distinct members of envelope proteins, out membrane proteins, transporters and channels are added into the negative samples of the G-protein coupled receptors. It is expected that the number of negative samples in each of these sub-groups may be higher than that in the group of negative samples for all TM proteins.

Training sets of both positive and negative samples were further screened so that only essential proteins that optimally represent each group are retained. The SVM training system for each group was optimized and tested by using separate testing sets of both positive and negative samples composed of all the remaining distinct proteins of a group and those outside the group respectively. The performance of

SVM classification was further evaluated by using independent sets of both positive and negative samples composed of all the remaining proteins of a group and those outside the group respectively. No duplicate protein entry was used in the training, testing and independent evaluation set for each group. For those with sufficient number of distinct members, multiple entries were assigned into each set. For those with less than three distinct members, the proteins were assigned in the order of priority of training, testing and independent evaluation set.

The number of positive and negative samples for each of the training, testing and independent evaluation set for each group of TM proteins is given in Table 1. The training set is composed of 2,105 TM and 2,563 NM proteins, 927 G-protein coupled receptors and 1,320 non-G-protein coupled receptors, 177 envelope proteins and 1,999 non-envelope proteins, 602 outer membrane proteins and 1,539 non-outer membrane proteins, and 485 transporters/channels and 3,511 non-transporters/non-channels. The testing set is comprised of 12,857 TM and 9,605 NM proteins, 4,998 G-protein coupled receptors and 13,216 non-G-protein coupled receptors, 123 envelope proteins and 7,932 non-envelope proteins, 547 outer membrane proteins and 8,385 non-outer membrane proteins, and 335 transporters/channels and 5,632 non-transporters/non-channels. The independent evaluation set is made of 3,389 TM and 6,063 NM proteins, 2,532 G-protein coupled receptors and 7,244 non-G-protein coupled receptors, 150 envelope proteins and 4,952 non-envelop proteins, 343 outer membrane proteins and 4,948 non-outer membrane proteins, and 160 transporters/channels and 3,963 non-transporters/non-channels.

2.2 Feature Vector Construction

Construction of the feature vector for a protein was based on the formula for the prediction of protein-protein interaction [33] and protein function prediction [23,24,25,26]. Details of the formula can be found in the respective publications and references therein. Each feature vector was constructed from encoded representations of tabulated residue properties including amino acids composition, hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility.

There is some level of overlap in the descriptors for hydrophobicity, polarity, and surface tension. Thus the dimensionality of the feature vectors may be reduced by principle component analysis (PCA). Our own study suggests that the use of PCA reduced feature vectors only moderately improves the accuracy. It is thus unclear to which extent this overlap affects the accuracy of SVM classification. It is noted that reasonably accurate results have been obtained using these overlapping descriptors in various protein classification studies [23,24,25,26,33,34,35,36].

2.3 Support Vector Machine

SVM is based on the structural risk minimization (SRM) principle from statistical learning theory [21]. SVM constructs a hyperplane that separates two different classes of feature vectors. A feature vector xi represents the structural

and physico-chemical properties of a protein. There are a number of hyperplanes for an identical group of training data. The classification objective of SVM is to separate the training data with maximum margin while maintaining reasonable computing efficiency. SVM maps feature vectors into a high dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ followed by the construction of OSH in the feature space [36]. Gaussian kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ was used in this work because it consistently gives better results than other kernel functions [35]. Linear support vector machine is applied to this feature space and then the decision function is given by:

$$f(\mathbf{x}) = sign[\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b], \tag{1}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

under conditions

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{3}$$

Positive or negative value from Eq.(1) indicates that the vector $\mathbf{x}$ belongs to the positive or negative class respectively. To further reduce the complexity of parameter selection, hard margin SVM with threshold was used in our own SVM program SVM* [36].

As in the case of all discriminative methods [37], the performance of SVM classification can be measured by the quantity of true positives TP, true negatives TN, false positives FP, false negatives FN, sensitivity $SE = TP/(TP + FN)$, specificity $SP = TN/(TN + FP)$, the overall accuracy (Q) and Matthews Correlation Coefficient (MCC) [25] are given below:

$$Q = (TP + TN)/(TP + FN + TN + FP), \tag{4}$$

$$MCC = \frac{TP \bullet TN - FN \bullet FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}. \tag{5}$$

3 Results and Discussion

The number of training and testing proteins and prediction results of specific class of TM proteins and the corresponding NM proteins are given in Table 1. In this Table, TP stands for true positive (correctly predicted TM protein of a specific TM class), FN stands for false negative (protein from a specific class

of TM proteins incorrectly predicted as a non-class-member), TN stands for true negative (correctly predicted non-class-member), and FP stands for false positive (non-class-member incorrectly predicted as a member of a specific class of TM proteins). The predicted accuracies for TM proteins, G-protein coupled receptors, envelope proteins, outer membrane proteins, and transporters/channel are 90.1%, 95.6%, 90.0%, 92.7% and 73.9% respectively. The predicted accuracies for NM proteins, non-G-protein coupled receptors, non-envelope proteins, non-outer membrane proteins, and non-transporters/non-channels are 86.7%, 98.1%, 99.5%, 86.4%, and 98.6% respectively.

Table 1. Prediction accuracies and the number of positive and negative samples in the training, test, and independent evaluation set of transmembrane proteins (Tr), G-protein coupled receptors (Gp), Envelope proteins (En), Outer Membrane proteins (OM), and Transporters and Channels (TC). Predicted results are given in TP, FN, TN, FP, accuracy for positive samples SE, accuracy for negative samples SP, overall accuracy Q and Matthews correlation coefficient MCC. The number of positive or negative samples in the training set is P or N respectively. The number of positive or negative samples in the test and independent evaluation sets is $TP + FN$ or $TN + FP$ respectively. PF represents Protein Family.

	Training Set		Test Set				Independent Set							
PF	P	N	TP	FN	TN	FP	TP	FN	TN	FP	SE (%)	SP (%)	Q (%)	MCC
Tr	2105	2563	11135	1722	8237	1368	3054	335	5254	809	90.1	86.7	87.9	0.749
Gp	927	1320	4993	5	13212	4	2421	111	7104	140	95.6	98.1	97.4	0.933
En	177	1999	112	11	7904	28	135	15	4927	25	90.0	99.5	99.2	0.867
OM	602	1539	547	0	8384	1	318	25	4276	672	92.7	86.4	86.8	0.499
TC	485	3511	331	4	5628	4	127	33	3909	54	73.9	98.6	97.8	0.735

A direct comparison with results from previous protein studies is inappropriate because of the differences in the specific aspects of proteins classified, dataset, descriptors and classification methods. Nonetheless, a tentative comparison may provide some crude estimate regarding the level of accuracy of our method with respect to those achieved by other studies. With the exception of and transporters/channels, the accuracies for various TM classes are comparable to those of ∼95% obtained from previous studies [13,18]. The prediction accuracy for transporters and channels is substantially lower primarily because the collected proteins in this class are not sufficiently diverse to adequately train the corresponding SVM classification system. There are 250 identified families of transporters and 115 families of channels, some of which contain substantial number of distinct proteins [20]. Thus the collected 980 transporters and channels are not enough to fully represent all of the identified families.

The prediction accuracy for the NM proteins and those of negative samples of individual TM classes is comparable to the level of 92%∼99% obtained from previous studies. Unlike that of the positive samples, the prediction accuracy

for the negative samples of transporters and channels is comparable to those of other classes and those from other studies. This is because the corresponding SVM system was trained by using a diverse set of negative samples that include all representative NM proteins and proteins from other TM classes.

4 Conclusion

SVM appears to be capable of prediction of MP proteins and proteins in specific TM classes from a large number and diverse range of proteins at accuracies comparable to those from other studies. The prediction accuracy of SVM may be further enhanced with the improvement of SVM algorithms particularly the use of multi-class prediction models, more adequate training for distantly related proteins, and the use of the expanded knowledge about specific classes of TM proteins such as transporters and channels. To assist their evaluation and exploration, our SVM classification systems – SVMProt, can be accessed at http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi.

References

1. Stack, J.H., Horazdovsky, B.,Emr, S.D.: Receptor-mediated Protein Sorting to the Vacuole in Yeast: Roles for a Protein Kinase, a Lipid Kinase and GTP-binding Proteins. Annu. Rev. Cell Dev. Biol. **11** (1995) 1–33
2. Le Borgne, R., Hoflack, B.: Protein Transport from the Secretory to the Endocytic Pathway in Mammalian Cells. Biochim. Biophys. Acta **1404** (1998) 195–209
3. Chen, X., Schnell, D.J.: Protein Import into Chloroplasts. Trends Cell Biol. **9** (1999) 222–227
4. Thanassi, D.G., Hutltgren, S.J.: Multiple Pathways Allow Protein Secretion Across the Bacterial Outer Membrane. Curr. Opin. Cell Biol. **12** (2000) 420–430
5. Heusser, C. , Jardieu, P.: Therapeutic Potential of Anti-IgE Antibodies. Curr. Opin. Immunol. **9** (1997) 805–813
6. Saragovi, H.U. , Gehring, K.: Development of Pharmacological Agents for Targeting Neurotrophins and their Receptors. Trends Pharmacol. Sci. **21** (2000) 93–98
7. Sedlacek, H.H.: Kinase Inhibitors in Cancer Therapy: A Look Ahead. Drugs **59** (2000) 435–476
8. Zhang, L., Brett, C.M., Giacommi, K.M.: Role of Organic Cation Transporters in Drug Absorption and Elimination. Annu. Rev. Pharmacol. Toxicol. **38** (1998) 431–460
9. Tamai, I. , Tsuji, A.: Transporter-mediated Permeation of Drugs Across the Blood-brain Barrier. J. Pharmaceut. Sci. **89** (2000) 1371–1388
10. McGovern, K., Ehrmann, M., Beckwith, J.: Decoding Signals for Membrane Proteins using Alkaline Phosphatase Fusions. EMBO J. **10** (1991) 2773–2782
11. Amstutz, P., Forrer, P., Zahnd, C., Pluckthun, A.: In Vitro Display Technologies: Novel Developments and Applications. Curr. Opin. Biotechnol. **12** (2001) 400–405
12. Wallin, E. , von Heijne, G.: Genome-wide Analysis of Integral Membrane Proteins from Eubacterial, Archaean, and Eukaryotic Organisms. Protein Sci. **7** (1998) 1029–1038

13. Chen, C.P., Kernytsky, A., Rost, B.: Transmembrane Helix Predictions Revisited. Protein Sci. **11** (2002) 2774–2791
14. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting Transmembrane Protein Topology with A Hidden Markov Model: Application to complete genomes. J. Mol. Biol. **305** (2001) 567–580
15. Cai, Y.D., Zhou, G.P., Chou, K.C.: Support Vector Machine for Predicting Membrane Protein Types by using Functional Domain Composition. Biophys. J. **84** (2003) 3257–3263
16. Gromiha, M.M., Ahmad, S., Suwa, M.: Neural Network-based Prediction of Transmembrane -strand Segments in Outer Membrane Proteins. J. Comput. Chem. **25** (2004) 762–767
17. Yuan, Z., Mattick, J.S., Teasdale, R.D.: SVMtm: Support Vector Machines to Predict Transmembrane Segments. J. Comput. Chem. **25** (2004) 632–636
18. Cserzo, M., Eisenhaber, F., Eisenhaber, B., Simon, I.: On Filtering False Positive Transmembrane Protein Predictions. Protein Eng. **15** (2002) 745–752
19. Bairoch, A., Apweiler, R.: The SWISS-PROT Protein Sequence Database And Its Supplement Tremble In 2000. Nucleic Acids Res. **28** (2000) 45–48
20. Saier, M.H.: A functional-phylogenetic Classification System for Transmembrane Solute Transporters. Microbiol. Mol. Biol. Rev. **64** (2000) 354–411
21. Vapnik, V.: The Nature of Statistical Learning Theory. Springer: New York (1999)
22. Karchin, R., Karplus, K., Haussler, D.: Classifying G-protein Coupled Receptors with Support Vector Machines. Bioinformatics **18** (2002) 147–159
23. Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z.: SVM-Prot: Web-Based Support Vector Machine Software for Functional Classification of a Protein from Its Primary Sequence. Nucleic Acids Res. **31** (2003) 3692–3697
24. Cai, C.Z., Han, L.Y., Chen, Y.Z.: Enzyme Family Classification by Support Vector Machines. Proteins **55** (2004) 66–76
25. Cai, C.Z., Wang, W.L., Sun, L.Z., Chen, Y.Z: Protein Function Classification via Support Vector Machine Approach. Math. Biosci. **185** (2003) 111–122
26. Cai, C.Z., Han, L.Y., Chen, X. et al.: Prediction of Functional Class of the SARS Coronavirus Proteins by a Statistical Learning Method. J. Proteome Res. **4** (2005) 1855–1862
27. Han, L.Y., Cai, C.Z., Lo, S.L., et al.: Prediction of RNA-binding Proteins from Primary Sequence by a Support Vector Machine Approach. RNA **10** (2004) 355–368
28. Han, L.Y., Cai, C.Z., Ji, Z.L., Chen, Y.Z.: Prediction of Functional Class of Novel Viral Proteins by a Statistical Learning Method Irrespective of Sequence Similarity. Virology **331** (2005) 136–143
29. Han, L.Y., Cai, C.Z., Ji, Z.L., et al.: Predicting Functional Family of Novel Enzymes Irrespective of Sequence Similarity: a Statistical Learning Approach. Nucleic Acids Res. **32** (2004) 6437–6444
30. Cui, J., Han, L.Y., Cai, C.Z., et al.: Prediction of Functional Class of Novel Bacterial Proteins without the Use of Sequence Similarity by a Statistical Learning Method. J. Mol. Microbiol. Biotechnol. **9** (2005) 86–100
31. Lin, H.H., Han, L.Y., Cai, C.Z., Ji, Z.L., Chen, Y.Z.: Prediction of Transporter Family from Protein Sequence by Support Vector Machine Approach. Proteins **62** (2006) 218–231
32. Bateman, A., Birney, E., Cerruti, L. et al.: The Pfam Protein Families Database. Nucleic Acids Res. **30** (2002) 276–280
33. Bock, J.R. and Gough, D.A.: Predicting Protein-protein Interactions from Primary Structure. Bioinformatics **17** (2001) 455–460

34. Lo, S.L., Cai, C.Z., Chen, Y.Z., Chung, M.C.M.: Effect of Training Datasets on Support Vector Machine Prediction of Protein-protein Interactions. Proteomics **5** (2005) 876–884
35. Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C.: Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein. J. Comput. Chem. **23** (2002) 267–274
36. Cai, C.Z., Wang, W.L., Chen, Y.Z.: Support Vector Machine Classification of Physical and Biological Datasets. Inter. J. Mod. Phys. C **14** (2003) 575–585
37. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. Bioinformatics **16** (2000) 412–424