Classification of Cell Membrane Proteins

Seyed Koosha Golmohammadi, Lukasz Kurgan, Brendan Crowley, and Marek Reformat
Department of Electrical and Computer Engineering, University of Alberta, Canada
{koosha, lkurgan, bcrowley, reform}@ece.ualberta.ca

Abstract

Membrane proteins are an important class of proteins that serve as channels, receptors, and energy transducers in a cell membrane. Knowledge of a given type of cell membrane protein is crucial for determining its function. This paper introduces an automated, in-silico method for identifying different types of membrane proteins based on their amino acid composition. Our method applies a novel, composite protein sequence representation that includes seven feature sets. The performance of the proposed method was tested on two large datasets and was compared with eight competing prediction methods. The results indicate that our method outperforms existing methods, provides improved predictions for the transmembrane protein types, and obtains 87% and 98% accuracy for the jackknife test and test on an independent dataset, respectively.

1. Introduction

Information about millions of known proteins is stored in several open-access databases, including PDB, SWISS-PROT, and NCBI. PDB (Protein Data Bank) is a manually curated database of about 45,000 tertiary protein structures. SWISS-PROT is another manually curated database that includes partial functional/structural annotation for approximately 275,000 proteins. By far the largest, the NCBI (National Center for Biotechnology Information) database currently contains over 3.3 million protein chains, but without structural and functional information. The large and widening gap between the number of annotated proteins and all known proteins serves as a motivation to develop computational models that use the knowledge of annotated proteins to predict functional/structural information for the remaining proteins.

Cell membrane proteins are very important components of a cell. They carry out many of the functions that are imperative to the cell's survival. As a consequence of their importance, they became an attractive target for both basic research and drug design [1]. Knowledge of a given type of cell membrane protein is crucial in determining its function. Determining the type of a membrane protein using traditional experimental methods is costly and time-consuming. Therefore, an automatic method of identifying thousands of uncharacterized proteins is highly desirable. Membrane proteins are classified into transmembrane proteins, which span across the membrane, and anchored proteins, which are attached to the membrane on one side. Five main sub-types of membrane proteins are usually considered [2]: Type I Transmembrane, Type II Transmembrane, Multipass Transmembrane, Lipid Chain-Anchored Membrane, and GPI-Anchored Membrane.

The classification of membrane protein chains into their corresponding types is usually accomplished in two steps. First, the sequences are converted into a feature based representation which is next fed into a classifier. The previous attempts to classify membrane protein can be divided into two categories: (1) methods that use amino acid composition to represent the input sequences [2]; (2) methods that use pseudo-amino acid composition to represent the sequences [3]. The latter representation incorporates the sequence-order, while the former is based on simple order-independent counts. The methods that use the amino acid composition apply various classifiers such as Hamming distance [4], Euclidian distance [5], ProtLock [6], and covariant discriminant analysis [2]. The methods based on pseudo amino acid composition are generally more accurate. They apply Hamming distance [3], Euclidian distance [3], ProtLock [3], covariant discriminant analysis [3], support vector machine [7,8], fuzzy K-nearest neighbor [9], optimized evidence-theoretic K-nearest neighbor [10], supervised locally linear embedding [11], and various ensembles of classifiers [1,12], to classify the protein membrane sequences.

The two most recent contributions, both of which use pseudo-amino acid composition and an ensemble of classifiers, are:

- A stacked generalization based method that combines the results of two lower level classifiers (a support vector machine and an instance-based learner) through a meta-classifier (C4.5 decision tree) to maximize the classification accuracy [1].
- An ensemble of classifiers that is formed by merging a set of nearest neighbor (NN) classifiers [12]. Each of these NN classifiers is defined in a different pseudo amino acid composition space.

In addition to cell membrane protein type classification, there is a large amount of related work with a broader focus, e.g., protein structural class prediction [13] and protein subcellular location prediction [14].

This paper introduces a novel, automated method for identifying the types of membrane proteins using their amino acid sequence as the only input. Our main goal was to improve classification accuracy when compared with existing approaches. First, each protein sequence was mapped into a novel feature-based vector. Next, the best performing classifier was selected to identify the type. Three conventional tests performed on two large benchmark datasets were used to evaluate the performance of the proposed method. The classification accuracy was compared with eight recently proposed methods in this domain. The unique characteristic of the proposed method is that the sequences are represented by seven features sets, while the existing methods usually use only one feature set.

Section 2 describes the design of the proposed method. Section 3 presents and discusses our experimental results, and section 4 concludes the paper.

2. Proposed Approach

Preparation of the input for the classifiers is a crucial and time-consuming task. The classification accuracy depends on the features that are selected to represent the protein sequence. Section 3.1 describes the raw data that was used in this work. Section 3.2 describes the features that were considered as inputs for the classification model, and section 3.3 describes the methods used to select the best performing classifier.

2.1 Data

Two datasets were used to design and test our prediction system. These datasets are widely used to evaluate the performance of cell membrane protein classification systems [1-3,7-14], allowing for fair comparison with models described in the literature. The first dataset (Dataset 1) [3] was used to design the system. It contains 2059 cell membrane proteins, including 435 type-I transmembrane proteins, 152 type-II transmembrane proteins, 1311 multipass transmembrane proteins, 51 lipid-chain anchored transmembrane proteins, and 110 GPI anchored transmembrane proteins. The second dataset (Dataset 2) [2] was used for an independent test of the developed method. It contains 2625 proteins, including 478 type-I transmembrane proteins, 180 type-II transmembrane proteins, 1867 multipass transmembrane proteins, 14 lipid-chain anchored transmembrane proteins, and 86 GPI anchored transmembrane proteins.

2.2 Feature-based Sequence Representation

There are 20 unique amino acids that are used as a protein's building blocks. All amino acids have a common basic chemical structure, but different chemical properties due to differences in their side chains. A protein can be represented by a string of amino acids. Different proteins have different sequences, in terms of the ordering of their amino acids and length of the sequence. The first step in classifying proteins is to find a common way to represent the sequences. In this work, we adopt a feature vector to represent protein chains. Any protein, regardless of the length or composition of its sequence, can be mapped to our feature vector representation. We use 7 feature sets within our feature vector. These feature sets along with the corresponding number of features in each set are shown in Table 1. The proposed feature vector contains a total of 70 features.

Table 1. Feature based sequence representation.

Vector Feature	Number of Features
Amino Acid Composition	20
Sequence Length	1
2-Gram Exchange Group Frequency	36
Hydrophobic Group	2
Electronic Group	5
Sum of Hydrophobicity	1
R-Group	5

1) ***Amino Acid Composition*** $\mathbf{CV_i\{i=1:20\}}$, which is the normalized frequency of occurrence of each of the twenty amino acids in the given protein's amino acid sequence [1]. Therefore, this feature set includes 20 features.

2) ***Sequence Length*** **L**, which is defined as the total number of amino acids in the given protein's amino acid sequence.

3) ***2-Gram Exchange Group Composition*** **$CVExG_i$ {i=1:36}** that is defined by converting the sequence into its equivalent 6-letter exchange group representation [15], (which was derived from the PAM matrix) where $e_1 \in \{H, R, K\}$, $e_2 \in \{D, E, N, Q\}$, $e_3 \in \{C\}$, $e_4 \in \{S, T, P, A, G\}$, $e_5 \in \{M, I, L, V\}$, and $e_6 \in \{F, Y, W\}$. The exchange groups are broader classes of amino acids that represent the effects of evolution. For example, all H, R, and K amino acids in the original sequence are replaced by e_1. After the amino acids are replaced, the resulting sequence consists of an alphabet of only 6 different characters. We compute the frequency of occurrence of each possible 2-gram (pair) [16] of the consecutive exchange group amino acids. Therefore, this feature set takes into account the sequence of amino acids, rather than just their composition. This set includes a total of $6^2 = 36$ features.

4) ***Hydrophobic Group*** **$CVHG_i$ {i=1:2}**. The side chains may be polarized. Non-polar side chains are hydrophobic, while polar side chains are hydrophilic [17]. The hydrophobic amino acids include {A, C, F, I, L, M, P, V, W, Y} and the hydrophilic amino acids include {D, E, G, H, K, N, Q, R, S, T} [18]. This feature set counts the number of hydrophobic and hydrophilic amino acids in the protein sequence, and thus it includes two features.

5) ***Electronic Group*** **$CVEG_i$ {i=1:5}**. The electronic group specifies whether a given amino acid is electrically neutral, donates electrons, or accepts electrons. For this feature set we again compute the frequency of amino acids in each of the electronic groups, which include donors {A, D, E, P}, weak donors {I, L, V}, acceptors {K, N, R}, weak acceptors {F, M, Q, T, Y}, and neutral {G, H, S, W}. Therefore, the electronic grouping corresponds to 5 features.

6) ***Sum of Hydrophobicity*** **Y**. Each amino acid has an associated hydrophobic affinity, which is often measured using a hydrophobic index. The Eisenberg hydrophobic index, which was used to analyze membrane-associated helices [20], is applied in this feature set. This index is normalized and ranges between -2.53 for R (the least hydrophobic) and 1.38 for I (the most hydrophobic). Simlarly to [21], we compute the sum of this hydrophobic index over all amino acids in the protein sequence, which gives one feature.

7) ***R-Group*** **$CVRG_i$ {i=1:5}**. As discussed above, each amino acid has a different side chain. However, some of these side chains have similar characteristics and can be clustered into five R Groups: non-polar aliphatic {A, I, L, V}, glycine {G}, non-polar {F, M, P, W}, polar uncharged {C, N, Q, S, T, Y}, or charged $\in \{D, E, H, K, R\}$ [21]. Compositions of amino acids in each of the above five groups is computed.

The resulting feature vector, which consists of 70 features grouped into seven features sets, constitutes the input for our classification model.

2.3 Design of the Proposed Prediction Method

To find the best performing classifier we updated our design iteratively based on a series of tests that were divided into three phases. We designed our system using the Weka[1] environment [22]. The tests were performed utilizing 10 fold cross-validation on Dataset 1.

1) ***Phase 1***

Phase one was devoted to preparing the input data for classification. We computed 70 features (described in Section 2.2) for each sequence in Datasets 1 and 2.

2) ***Phase 2***

We tested all 70 classifiers in Weka (except for certain models that required discrete input data) to compare their performance for this classification problem. These classifiers included Bayesian methods, regression, support vector machines, neural networks, instance based nearest neighbor methods, decision trees, rule based and cost based methods. The top 9 classifiers according to their overall classification accuracy over 10 fold cross-validation on Dataset 1 are shown in Table 2.

Table 2. Top 9 classifiers with the highest overall classification accuracy on Dataset 1.

Index	Classifier	Accuracy
1	Decision Tree with Naïve Bayes at the leaves	81.29
2	Bagged Decision Tree	81.78
3	Logistic Regression based metaclassifier	81.88
4	Support Vector Machine with polynomial kernel	82.31
5	Decorate based ensemble of Decision Trees	83.04
6	Random Forest	83.04
7	Neural Network with back propagation training	84.26
8	K-nearest neighbor	85.76
9	K* -nearest neighbor	86.30

The overall classification accuracy of a model is an important factor, but not sufficient to select the best classifier for this problem. The accuracies for each membrane protein type are other factors that we considered in choosing the best performing classifier. Figure 1 shows the overall accuracy and the accuracy for each protein type for the top 9 classifiers. From

[1] http://www.cs.waikato.ac.nz/ml/weka/

these 9 models we eliminated those that had the worst accuracies for individual protein types, and retained those that had the best accuracies for the different types. Among all models K* performed the best considering both overall and majority of per type accuracies.

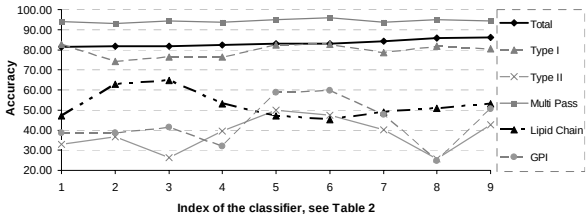

Figure 1. Overall accuracy and accuracy for each membrane protein type for 9 best performing classifiers.

The selected K* method is an instance-based classifier [23]. The basic idea behind this type of classifiers is that similar instances should have similar class labels. First, an instance-based classification algorithm uses a distance function to determine which instance is most similar to a given test instance. Second, it uses a classification function which determines the final classification of the test instance based on class labels of the similar instances. A *k*-nearest neighbor algorithm finds the *k* instances that are the most similar to the test instance. The predicted class label is equal to the most common class among the *k* nearest neighbors.

The defining characteristic of the K* classifier is that it uses an entropy-based distance function. The classification step determines the probability that a given instance (protein) is in a certain class

$$P(C \mid a) = \sum_{b \in C} P(b \mid a)$$

where P(*x*|*y*) is the conditional probability of *x* given *y*, *a* is the test instance, *C* is the current class label, and *b* is each instance from the training set known to be in class *C*. The category (class) with the highest probability is chosen as the prediction for the test instance. The "globalBlend" parameter of the algorithm specifies how many neighbors should be considered. Choosing 100% means that all neighbors have an equal weighting. Choosing 0% results in a 1-nearest neighbor algorithm.

3) ***Phase 3***

In this phase the K* classifier was tested with different parameters through 10 fold cross-validation to optimize the resulting overall accuracy. K* performed the best when the "globalBlend" parameter was equal to 38%.

3. Results and Discussion

Three test methods were used to evaluate the quality of the proposed prediction model [24]: (1) the re-substitution (self-consistency) test; (2) the jackknife (leave-one-out) test; and (3) the independent dataset test. The self-consistency test involves training the model with Dataset 1, and then testing the model with the same Dataset 1. During the jackknife test, we designed and tested the model through *n*-fold cross validation on Dataset 1, where *n* is the size of the dataset. The independent dataset test involves training the model on Dataset 1, and then testing it on Dataset 2. Among the three tests the jackknife test is the most objective [14]. This type of test is widely used to evaluate related prediction methods [2, 24-28].

In addition to reporting overall accuracy, we also report the accuracy, specificity, and Matthew's Correlation Coefficient for each membrane protein type for each test type. The Matthew's Correlation Coefficient (MCC) is a number between -1 and 1. A value of 1 means the classifier is perfect, and always classifies correctly. A value of -1 means the classifier always classifies incorrectly. The results are given in Table 3.

Table 3. Classification results for optimized K* classifier

		Test method		
		Self-consistency	**Jackknife**	**Independent**
Accuracy [%]	Overall	99.9	86.9	97.1
	Type I	100	83.5	96.4
	Type II	100	52.6	80.6
	Multipass	100	95.8	99.0
	Lipid	100	45.1	78.6
	GPI	99.1	61.5	96.5
Specificity [%]	Type I	100	94.7	99.2
	Type II	100	98.3	99.8
	Multipass	99.9	83.4	93.9
	Lipid	100	99.9	99.9
	GPI	100	98.7	99.8
MCC	Type I	1.00	0.77	0.95
	Type II	1.00	0.59	0.87
	Multipass	1.00	0.81	0.94
	Lipid	1.00	0.64	0.78
	GPI	0.99	0.65	0.96

The worst accuracies of about 50% are obtained for type II transmembrane and lipid-chain anchored membrane proteins. At the same time, type I and multipass types are predicted with 87% and 96% accuracy, respectively. The specificity values range between 95% and 100%, which shows that the proposed method is selective. The results show that the weakness of our model is in classifying lipid-chain-anchored membrane proteins, while the model

performs relatively well for transmembrane proteins. It is possible that the model performed the best for multipass transmembrane proteins because they constitute the majority of the samples in the two datasets. On the other hand, the number of samples for the lipid-chain anchored membrane proteins is the lowest, which could lead to the poorer quality of our method for this type.

The proposed method was also compared with eight competing methods that were published after 2001; see Table 4.

Table 4. Comparison of the overall accuracy with eight competing methods.

Classifier	Ref.	Test method		
		Self-consistency	Jack-knife	Inde-pendent
K*	this paper	99.9	86.9	97.7
Ensemble of NNs	[12]	not available	85.8	96.8
Fuzzy KNN	[9]	not available	85.6	95.7
Stacking	[1]	98.7	85.4	94.3
OET-KNN	[10]	99.5	84.7	94.2
Weighted SVM	[8]	99.9	82.4	90.3
SLLE	[11]	not available	82.3	95.7
Augmented covariant discriminant	[3]	90.9	80.9	87.5
SVM	[7]	not available	80.4	85.4

Table 4 shows that prediction methods based on nearest neighbor (NN) and *k*-nearest neighbor (KNN) classifiers, including the proposed method, perform quite well, suggesting that this type of the classifier is an appropriate choice for the membrane type prediction problem. Our method produced the highest accuracies for both the jackknife and the independent dataset tests. The proposed method improved the error rate of the jackknife and independent dataset test by 8% (1.1/14.2) and 28% (0.9/3.2), respectively, when compared with the second best ensemble classifier [12]. Table 5 provides a detailed breakdown of differences in prediction quality between the proposed and the second best methods.

Table 5. Comparison of results.

		Jackknife test		Independent set test	
		[12]	This paper	[12]	This paper
Accuracy [%]	Type I	81.2	83.5	96.0	96.4
	Type II	44.7	52.6	79.4	80.6
	Multipass	95.8	95.8	99.0	99.0
	Lipid	47.1	45.1	57.1	78.6
	GPI	60.0	61.5	90.7	96.5
MCC	Type I	0.737	0.772	0.950	0.955
	Type II	0.527	0.587	0.862	0.875
	Multipass	0.800	0.810	0.934	0.940
	Lipid	0.654	0.638	0.675	0.784
	GPI	0.640	0.652	0.915	0.958

When compared with the ensemble classifier published in [12], the proposed method improves predictions for type I and type II transmembrane proteins, while providing comparable quality for the anchored and multipass proteins.

4. Conclusion

Empirical tests on two benchmark datasets indicate that the proposed method outperforms state-of-the-art existing methods, as it achieves the highest accuracies for both the jackknife and the independent dataset tests. Results show that our method improves the quality of prediction for transmembrane proteins when compared with the second best ensemble-based method. At the same time, there is still some room for further improvement, as the jackknife test accuracy of the proposed method equals 86.9%. The quality of the prediction model is highly dependent on the features used to represent the sequences. In contrast to existing methods that use either composition or pseudo composition to represent sequences, the proposed method uses seven feature sets for the same task. We believe that this more comprehensive representation resulted in the reported improvements.

Although the proposed feature representation is a step in the right direction, there may be other features that were not considered in this work that might provide further improvements. This challenging protein classification problem can also be improved by using related classification problems such as a method proposed by Koza, which aims at classifying protein segments (segments of protein sequence) into transmembrane / non-transmembrane classes [29]. This method could help in designing new features that would improve the contrast between anchored and transmembrane proteins.

5. References

[1] S. Q. Wang, J. Yang, K.-C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo amino acid composition", *Journal of Theoretical Biology*, (2006), Vol. 242, pp. 941-46.

[2] K.-C. Chou and D. W. Elrod, "Prediction of membrane protein types and subcellular locations", *Proteins: Proteins: Structure, Function, and Genetics*, (1999), Vol. 34, 137-53

[3] K.-C. Chou. "Prediction of protein cellular attributes using pseudo amino acid composition", *Proteins: Structure, Function, and Genetics*, (Erratum: *Proteins: Structure, Function, and Genetics*, Vol. 44, 60) (2001), Vol. 43, 246-55.

[4] PY. Chou, "Prediction of protein structural classes from amino acid composition". In: GD. Fasman, editor. "*Prediction of protein structure and the principles of protein conformation*". New York: Plenum Press, (1989), pp. 549–586.

[5] H. Nakashima, K. Nishikawa and T. Ooi. „The folding type of a protein is relevant to the amino acid composition", *Journal of Biochemistry*, (1986), Vol. 99, pp.152–162.

[6] J. Cedano, P. Aloy, J.A. Perezpons and E. Querol. "Relation between amino acid composition and cellular location of proteins", Journal of Molecular Biology, (1997), Vol. 266, pp. 594–600.

[7] Y. D. Cai, P. W. Ricardo, C. H. Jen, K.-C. Chaou, "Application of SVM to predict membrane protein types" *Journal of Theoretical Biology*, (2004), Vol. 226, pp. 373-6

[8] Wang M., Yang J., Liu G. P., K.-C. Chou, "Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition", *Journal of Protein Engineering Design and Selection*, (2004), Vol. 17, pp. 509-16

[9] H.-B. Shen, J. Yang and K.-C. Chou, "Fuzzy KNN for predicting membrane protein types from pseoudo-amino acid composition", *Journal of Theoretical Biology*, (2006), Vol. 240, pp. 9-13

[10] H.-B. Shen, K.-C. Chou, "Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types." *Biochemical and Biophysical Research Communications*, (2005), Vol. 334, pp. 288-92.

[11] M. Wang, J. Yang, Z. J. Xu, K.-C. Chou, "SLLE for predicting membrane protein types", *Journal of Theoretical Biology*, (2005), Vol. 232, pp. 7-15.

[12] H.-B. Shen and K.-C. Chou, "Using ensemble classifier to identify membrane protein types", *Amino Acids*, (2007), Vol. 32, pp. 483-8.

[13] Y. Cai, X. J. Liu, X. B. Xu, and K.-C. Chou, "Support vector machines for prediction of protein domain structural class", *Journal of Theoretical Biology*, (2003), Vol. 221, pp. 115–20.

[14] Q.-B. Gao, Z.-Z. Wang, C. Yan, and Y.-H. Du, "Prediction of protein subcellular location using a combined feature of sequence", *Federation of European Biochemical Societies Letters*, (2005), Vol. 579, pp. 3444-8.

[15] C. H. Wu, M. Berry, Y. S. Fung, and J. McLarty, "Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition", *Machine Learning*, (1995), Vol. 21, pp 177–93.

[16] C. H. Wu, G. Whitson, J. McLarty, A. Ermongkonchai, and T .C. Chang, "Protein Classification Artificial Neural System", *Protein Science*, (1992), Vol. 1, 667-77.

[17] S. Zumdahl and S. Zumdahl, "Chemistry", Fifth edition, Houghton Mifflin Company, (2000).

[18] D. F.Waugh, "Protein-protein interactions", *Advances in Protein Chemistry*, (1954), Vol. 9, 325-437.

[19] M. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, and R. Reddy, "Characterization of protein secondary structure", *IEEE Signal Processing Magazine*, (2004), Vol. 21, pp. 78–87.

[20] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, "Analysis of membrane and surface protein sequences with the hydrophobic moment plot", *Journal of Molecular Biology*, (1984), Vol. 179, 125-42.

[21] K. Kedarisetti, L. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology", *Biochemical and Biophysical Research Communication*, (2006), Vol. 348, pp. 981-8.

[22] I.H. Witten and E. Frank, "*Data Mining: Practical Machine Learning Tools and Techniques*", 2nd edition, Morgan Kaufmann, San Francisco, (2005).

[23] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure", *Proceedings of the 12th International Conference on Machine Learning*, (1995), pp. 108–14.

[24] K.-C. Chou, C.T. Zhang, "Review: prediction of protein structural classes", *Critical Reviews in Biochemistry and Molecular Biology*, (1995), Vol. 30, pp. 275–349.

[25] Y.-D. Cai, K.-Y. Feng, W.-C. Lu, K.-C. Chou, "Using LogitBoost classifier to predict protein structural classes", *Journal of Theoretical Biology*, (2006), Vol. 238, pp. 172–6.

[26] G.-P. Zhou, K. Doctor, "Subcellular location prediction of apoptosis proteins", *Proteins: Structure, Function, and Genetics*, (2003), Vol. 50, pp. 44–8.

[27] G.-P. Zhou, N. Assa-Munt, "Some insights into protein structural class prediction", *Proteins: Structure, Function, and Genetics*, (2001), Vol. 44, 57–9.

[28] G.-P. Zhou, "An intriguing controversy over protein structural class prediction", *Journal of Protein Chemistry*, (1998), Vol. 17 pp. 729–38.

[29] J.R. Koza and D. Andre. "Classifying protein segments as transmembrane domains using architecture-altering operations in genetic programming". In J.A. Peter and K.E. Kinnear, editors, *Advances in Genetic Programming II*, MIT Press, Cambridge, MA, USA, (1996), pp. 155-176.