Prediction of Integral Membrane Protein Type by Collocated Hydrophobic Amino Acid Pairs

KE CHEN, YINGFU JIANG, LI DU, LUKASZ KURGAN
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

Received 16 February 2008; Revised 28 April 2008; Accepted 9 May 2008
DOI 10.1002/jcc.21053
Published online 19 June 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: A computational model, IMP-TYPE, is proposed for the classification of five types of integral membrane proteins from protein sequence. The proposed model aims not only at providing accurate predictions but most importantly it incorporates interesting and transparent biological patterns. When contrasted with the best-performing existing models, IMP-TYPE reduces the error rates of these methods by 19 and 34% for two out-of-sample tests performed on benchmark datasets. Our empirical evaluations also show that the proposed method provides even bigger improvements, i.e., 29 and 45% error rate reductions, when predictions are performed for sequences that share low (40%) identity with sequences from the training dataset. We also show that IMP-TYPE can be used in a standalone mode, i.e., it duplicates significant majority of correct predictions provided by other leading methods, while providing additional correct predictions which are incorrectly classified by the other methods. Our method computes predictions using a Support Vector Machine classifier that takes feature-based encoded sequence as its input. The input feature set includes hydrophobic AA pairs, which were selected by utilizing a consensus of three feature selection algorithms. The hydrophobic residues that build up the AA pairs used by our method are shown to be associated with the formation of transmembrane helices in a few recent studies concerning integral membrane proteins. Our study also indicates that Met and Phe display a certain degree of hydrophobicity, which may be more crucial than their polarity or aromaticity when they occur in the transmembrane segments. This conclusion is supported by a recent study on potential of mean force for membrane protein folding and a study of scales for membrane propensity of amino acids.

© 2008 Wiley Periodicals, Inc. J Comput Chem 30: 163–172, 2009

Key words: type of integral membrane protein; transmembrane protein; hydrophobic AA pairs; PSI-BLAST profile; support vector machine

Introduction

Integral membrane protein (IMP) refers to a protein that is permanently attached to a biological membrane. Although IMPs play a crucial role in numerous cellular functions by acting as receptors and transporters, significant amount of information concerning their function is still unknown due to the lack of high-resolution structures. Despite the estimates show that 20–30% of the human genome encodes membrane proteins, only 60 IMP structures were solved in the last 6 years.[1,2] The limited number of available structures is a result of substantial difficulties with overexpression and crystallization of membrane proteins.[3] As a result, in the last decade, the structures of new IMPs are often determined using NMR.[4–6] IMPs are generally categorized into five types. The first two types concern single-pass transmembrane (TM) proteins; type I IMP has an extracellular N-terminus and cytoplasmic C-terminus, whereas type II IMP has an extracellular C-terminus and cytoplasmic N-terminus. Type III IMP concerns multipass TM proteins that cross the membrane at least twice. The last two types concern proteins anchored to membrane and include lipid chain-anchored membrane protein and GPI-anchored membrane protein. The lipid chain-anchored membrane is associated with the bilayer only, whereas GPI-anchored membrane protein is bound to the membrane by a glycophosphatidylinositol (GPI) anchor. A recent study proposed a more detailed classification of IMPs, which includes eight types.[7] The new categories subdivide type I and type II IMPs into two types based on their orientation in the bilayer and based on the topogenic sequences that direct their insertion into the endoplasmic reticulum (ER) membrane, and a new category named peripheral membrane was added. At the same time, in this study, we concentrate on categorizing the five types of IMPs.

Correspondence to: K. Chen; e-mail: kchen1@ece.ualberta.ca

Prediction of the type of IMP is usually performed in two steps. First, sequences of different length are represented by a fixed length feature vector and next the feature values are fed into a classification algorithm. Early attempts in computational prediction of IMP types were observed in 1990s. One of the first studies was based on component-coupled algorithm and the protein sequence was represented by a conventional composition vector.[8] Recent studies applied more complex features, including pseudo amino acid composition, which considers position of amino acids (AAs) in the sequence,[9–11] functional domain composition,[12] amphiphilic effect features,[13] and GO-PseAA features.[14] Different classification algorithms, including support vector machine (SVM),[15,16] weighted SVM,[10] optimized evidence-theoretic K-nearest neighbor,[17] Fuzzy K-nearest neighbor,[18] and ensemble models,[19,20] were used in recent studies. The earlier methods classify a membrane protein into one of the five types assuming that another method is used to predict whether a given sequence is a membrane protein. The latter can be accomplished using a number of high-quality predictions methods, which include MemType-2L,[21] SOSUI,[22] DAS,[23] and SVMtm,[24] and that can find membrane proteins with accuracy of over 90%. Therefore, our study focuses on prediction of the five types of membrane proteins while assuming that the user would apply one of the aforementioned servers to predict whether a given sequence is a membrane protein. Most importantly, motivation for the development of the proposed method stems from the fact that although the prediction accuracy of state-of-the-art membrane protein type prediction algorithms is improving in the last few years, the corresponding prediction models are not interpretable and could not be used to disclose the underlying biological patterns.

Although many recent works concentrate on improving the prediction accuracy of the membrane type prediction by applying various, modern classification algorithms, we concentrate on development of a novel sequence representation which provides better accuracy and which can be analyzed and explained based on associated biological patterns. To this end, we propose a novel representation called PSI-BLAST profile-based collocation of AA pairs. This representation is based on a combination of PSI-BLAST profile[25] and a concept of collocation of AAs in a given sequence.[26] Three feature selection methods were applied to reduce the dimensionality of the original feature space based on the PSI-BLAST profile-based collocation of AA pairs. After feature selection, the selected profile-based pairs mainly include hydrophobic residues, which are regarded as the driving force for the formation of TM helices. We also note that amino acid Met (M) and Phe (F), which are commonly classified as polar and aromatic residue, occur frequently in the selected AA pairs. We discuss these findings and show that they are consistent with other published results that concern membrane proteins.

Materials and Methods

Datasets

Training set 1 and test set 1 were originally generated by Chou based on release 35.0 of SWISS-PROT.[8] The training and test set contain 2059 and 2625 membrane protein sequences, respectively. In the training set, 434 proteins belong to type-I IMP, 152 proteins belong to type-II IMP, 1311 proteins belong to type-III IMP, 51 proteins belong to type IV IMP, and 111 proteins belong to type-V IMP. In the test set, 477 proteins belong to type-I IMP, 180 proteins belong to type-II IMP, 1867 proteins belong to type-III IMP, 14 proteins belong to type IV IMP, and 87 proteins belong to type-V IMP.

Additionally, we selected subsets of training set 1 and test set 1 that are characterized by a sequence identity at 40%, i.e., any pair of the sequences in these subsets share no more than 40% similarity. The training set 2 was prepared by running CD-hit with 40% identity threshold[27] on the training set 1. As a result, training set 2 includes 1384 sequences. Test set 2 was generated in two steps. In the first step, we run CD-hit on test set 1 with 40% threshold to generate a subset of sequences that share sequence identity below 40%. In the second step, sequences from this subset that share a sequence identity of above 40% with sequences in training set 2 were removed. As a result, test set 2 contains 458 sequences. This dataset is used to analyze and contrast the prediction accuracy of the proposed method for sequence with hold identity.

Test set 3 contains membrane protein sequences that were deposited into SWISS-PROT database in 2007. Sequences annotated with ambiguous or uncertain classifications, i.e., "potential," "by similarity," "probable," were excluded. Additionally, proteins that share sequence identity of above 40% were removed to assure that the remaining set is nonredundant. As a result, test set 3 includes 165 IMP sequences, which allow for an independent evaluation of the proposed method on a nonredundant set of recently published proteins. The training set 2 and test sets 2 and 3 are freely available from the authors upon request.

Feature Representation

The proposed feature representation includes two feature sets: (1) the PSI-BLAST profile based collocation of AA pairs and (2) the autocorrelation function based on hydrophobicity indices.

PSI-BLAST Profile-Based Collocation of AA Pairs

The proposed representation combines PSI-BLAST profile and the concept of frequency of collocation of AA pairs[26,28] in the sequence. The original motivation to introduce the collocation of AA pairs comes from an insufficient sequence representation that is offered by the commonly used AA composition vector, which merely counts the frequencies of individual AAs in the sequence. On the other hand, the frequencies of AA pairs (dipeptides) provide more information since they may reflect local interaction and spatial arrangement between AA pairs. Based on this argument, we would count all dipeptides in the sequence. Since there are 400 possible AA pairs (*AA*, *AC*, *AD*, . . ., *YY*), a feature vector of that size is used to represent occurrence of these pairs in the sequence. At the same time, short-range interactions happen not only between immediately adjacent AAs, but also between residues a few positions away. For instance, the hydrogen bond in α-helix is established between the i^{th} AA and the $(i+4)^{th}$ AA. As a result, the proposed representation should consider collocated pairs of AAs, i.e., pairs that are separated by p other AAs. These pairs could be understood as the dipeptides with gaps. Collocated pairs for $p = 0, 1, \ldots, 4$ are considered, where for $p = 0$ they reduce to dipeptides. There are 400 feature values for each value of p.

Meanwhile, numerous successful applications of PSI-BLAST profile imply that the evolutionary information is more informative than the sequence itself.[29–32] PSI-BLAST aligns a given query sequence to a database of sequences and searches for these sequences that are similar to the query sequence. The alignment performed by PSI-BLAST produces so called PSI-BLAST profile, which is an $N \times 20$ matrix of frequencies of each AA at each position in the query sequence. The PSI-BLAST profile can be used to identify key conserved positions and positions in which residues undergo mutations.

The proposed approach combines the frequency of collocation of AA pairs and the PSI-BLAST profile into so called PSI-BLAST profile-based collocation of AA pairs. The PSI-BLAST profile, the $N \times 20$ matrix, can be denoted as $[a_{i,j}]$, where $i = 1, 2, \ldots, N$ denotes position in the query sequence and $j = 1, 2, \ldots, 20$ denotes 20 types of AAs. After applying the substitution matrix and log function, a_{ij} values range between -9 and 11. The proposed representation is just a generalized form of counting frequency of AA pairs based on binary coding. The binary coding uses a 20-dimensional vector to encode each AA. If the 20 AAs are represented as $AA_1, AA_2, \ldots, AA_{19}$, and AA_{20}, AA_i is hence encoded as (0, 0, ... , 0, 1, 0, ... , 0, 0), where only the i^{th} value is 1. The binary coding matrix is denoted as $[b_{i,j}]$ and it has the same dimensionality ($N \times 20$) as the PSI-BLAST profile. The frequency of AA pairs can be computed from the binary coding matrix. For a given protein sequence $A_1A_2, \ldots, A_N$

A_iA_{i+1} is the $\mathrm{AA}_m\mathrm{AA}_n$ dipeptide
$\Leftrightarrow A_i = \mathrm{AA}_m$ and $A_{i+1} = \mathrm{AA}_n$
$\Leftrightarrow b_{i,m} = 1,\ b_{i+1,n} = 1,\ b_{i,p} = 0,\ b_{i+1,q} = 0$, where $p \neq m$ and $q \neq n$
Given that $c_{s,t} = \min(b_{i,s}, b_{i+1,t})$, then

$$c_{s,t} = \begin{cases} 1 \text{ (iff } s = m, t = n) \\ 0 \text{ (else)} \end{cases}$$

which means that $\mathrm{AA}_m\mathrm{AA}_n$ was counted once while all other dipeptides were counted 0 times. Matrix $[c_{s,t}\ ]$ stores the frequencies of all dipeptides. The count of the AA pairs along the entire sequence can be computed as

$$c_{s,t} = \sum_{i=1}^{N-1} \min(b_{i,s}, b_{i+1,t})$$

The PSI-BLAST profile-based collocation of AA pairs is calculated in similar way. The only difference is that the binary coding matrix $[b_{i,j}]$ is replaced by the PSI-BLAST profile $[a_{i,j}]$. The frequency of $\mathrm{AA}_s\mathrm{AA}_t$ dipeptide is computed as $c_{s,t} = \sum_{i=1}^{N-1} \min(a_{i,s}, a_{i+1,t})$ and matrix $[c_{s,t}]$ stores the frequencies of all dipeptides.

Since the PSI-BLAST profile values can be negative, whereas using negative values to represent the frequencies of AA pairs could lead to misleading results, we compute of $c_{s,t}$ as follows

$$c_{s,t} = \sum_{i=1}^{N-1} \max(0, \min(a_{i,s}, a_{i+1,t}))$$

in which the negative value of $\min(a_{i,s}, a_{i+1,t})$ is replaced by 0. On the other hand, longer protein sequence will result larger frequencies of AA pairs. As a result, we also normalize $c_{s,t}$ as follows:

$$c_{s,t} = \frac{1}{N-1} \sum_{i=1}^{N-1} \max(0, \min(a_{i,s}, a_{i+1,t}))$$

Finally, the frequencies of p-collocated AA pairs are defined as

$$d_{s,t,p} = \frac{1}{N-p-1} \sum_{i=1}^{N-p-1} \max(0, \min(a_{i,s}, a_{i+p+1,t}))$$

The AXXXC and CXXXA pairs are symmetrical, i.e., they concern AAs A and C, which are separated by three AAs. Therefore, the corresponding symmetrical pairs could be combined together for the purpose of classifying the five types of IMP. By calculating the information gain (defined in the "Feature Selection" section) that quantifies the strength of relation between a given feature and the class label for each pair, we found that given AA pair and its symmetrical form usually provide similar information gain values, i.e., the information gain of pair IXXXF and FXXXI pairs are equal 0.552 and 0.514, respectively; the information gain of MXXXF equals 0.492, whereas for FXXXM it equals 0.489. This may imply that the symmetrical pairs have comparable ability to classify the five types of IMP. To reduce the dimensionality of the feature space, each pair and its symmetrical form are merged into one feature as follows:

$$c'_{s,t} = \begin{cases} c_{s,t} + c_{t,s} & (\text{if } s < t) \\ c_{s,t} & (\text{if } s = t) \\ 0 & (\text{if } s > t) \end{cases} \qquad d'_{s,t,p} = \begin{cases} d_{s,t,p} + d_{t,s,p} & (\text{if } s < t) \\ d_{s,t,p} & (\text{if } s = t) \\ 0 & (\text{if } s > t) \end{cases}$$

As a result, the matrixes $[c'_{s,t}]$ and $[d'_{s,t,p}]$, are upper triangular and the values below the main diagonal are set to 0. The dimensionality of $[c'_{s,t}]$ and $[d'_{s,t,p}]$ is 210. We generate PSI-BLAST profile-based collocation of AA pairs for $p = 0, 1, 2, 3$, and 4, which results in 1050 features per each sequence.

Autocorrelation Based on Hydrophobicity Indices

Prior research shows that hydrophobicity-based autocorrelation is related to protein structural class and the secondary structure content.[33,34] This study evaluates the effectiveness of autocorrelation, which is based on hydrophobicity index, in classification of the IMP types. Three sets of hydrophobicity indices, which include the Fauchere-Pliska's,[35] the Eisenberg's,[36] and the hydropathy index[37] were tested. The autocorrelation function is defined as

$$r_n = \frac{1}{N-n}\sum_{j=1}^{N-n} \text{index}(AA_i)\text{index}(AA_{i+n})$$

where N is the length of the sequence, n is a parameter that defines the autocorrelation step, index(AA_i) is the hydrophobicity index value for the i^{th} AA in the sequence. In this article, r_n was generated for $n = 1, 2, \ldots, 20$ for the three aforementioned indices, which resulted in $20 \times 3 = 60$ features.

To study the properties of different parts of IMP sequences, i.e., the N-terminus, the internal segment, and the C-terminus, the membrane protein sequences were divided into three equal-size subsequences, and the autocorrelation was calculated for each of these segments separately. This resulted in 180 features, i.e., 60 features for each terminus and another 60 for the internal segment.

Feature Selection

Since the proposed representation includes relatively large number of features, three feature selection methods, Information Gain based method (IG),[28,38] Chi-Squared method (CHI),[39] and the Relief algorithm (REL)[40] were used to reduce the dimensionality and potentially improve the prediction. We used three different methods to reduce bias introduced by each of the methods. In all three algorithms, each feature was ranked based on its merit (information gain in IG, the value of the χ^2 statistic in CHI, and the weights in REL), and next they were sorted by their average rank across the three algorithms. The measurement of the merit for the three algorithms is defined later.

IG: Information Gain

Information gain measures the decrease in entropy when a given feature is used to group values of another (class) feature. The entropy of a feature X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

where $\{x_i\}$ is a set of values of X and $P(x_i)$ is the prior probability of x_i. The conditional entropy of X, given another feature Y (in our case the IMP type) is defined as

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j))$$

where $P(x_i|y_j)$ is the posterior probability of X given the value y_i of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain

$$IG(X|Y) = H(X) - H(X|Y)$$

According to this measure, Y has stronger correlation with X than with Z if $IG(X|Y) > IG(Z|Y)$.

CHI: Chi-Squared Statistic

CHI: Chi-Squared statistic is the common statistical test that measures divergence from the expected distribution assuming that the occurrence of a given feature is independent of the class value. Let X be a discrete random variable (which corresponds to a feature in this article) with m possible outcomes $x_1, x_2, \ldots, x_m$ (which correspond to five IMP types) with probability of each outcome $P(X = x_i) = p_i$. Pearson-χ^2 statistic is defined as

$$\chi^2 = \sum_{i=1}^{m}\frac{(n_i - np_i)^2}{np_i}$$

where n_i is the number of instances, which will result the outcome x_i. A feature that gives higher value of χ receives lower rank.

REL: Relief Algorithm

REL: Relief algorithm is based on the feature weighting approach, which estimates the attributes according to their performance in distinguishing similar instances. REL searches the two nearest neighbors for each instance: one from the same class (nearest hit) and another from any other class (nearest miss). The algorithm to calculate the weights is as follows:

```
1. Initialization: given $D = \{(x_n, y_n)\}$ ($n = 1, 2, \ldots, N$) where
   $x_n$ is the feature space, $y_n$ is class label, and $N$ is the number
   of instances, set $w_i = 0$, $1 \le i \le I$, where $I$ is the number of
   features and $T$ is the number of iterations.
2. For $t = 1{:}T$
     Randomly select an instance $x$ from $D$;
       Find the nearest hit NH($x$) and miss NM($x$) of $x$;
       For $i = 1{:}I$
         Calculate: $w_i = w_i + |x^{(i)} - \mathrm{NM}^{(i)}(x)| - |x^{(i)} - \mathrm{NH}^{(i)}(x)|$
       End
     End
```

The feature selection was performed using 10-fold cross validation on training set 1 for all three algorithms. The features were sorted by their average rank obtained with the three algorithms. Next, we successively added one feature at the time and performed prediction of IMP type using 10-fold cross validation test on the training dataset using classifier described in "Classification Algorithm" section. Figure 1 shows that the prediction accuracy increases when adding up to 120 features, and later it saturates. Therefore, the top 120 features which have the lowest average ranks were selected; see Table 1. Among the selected features, 109 which are based on the PSI-BLAST profile-based collocation of AA pairs are shown in Table 2. We observe that the selected pairs are consistent for different number of gaps expressed by the p value, as well as across individual p values, i.e., the same Ala Ile (AI) pair is selected for $p = 0, 1, \ldots, 4$ and Ala (Ile) is also included in 3 (6) other selected pairs for $p = 0$. A detailed discussion of the selected features is provided in the "Results and Discussion" section. Among the remaining 11 features, 10 are hydrophobicity-based autocorrelations for entire sequence and only one feature corresponds to the hydro-

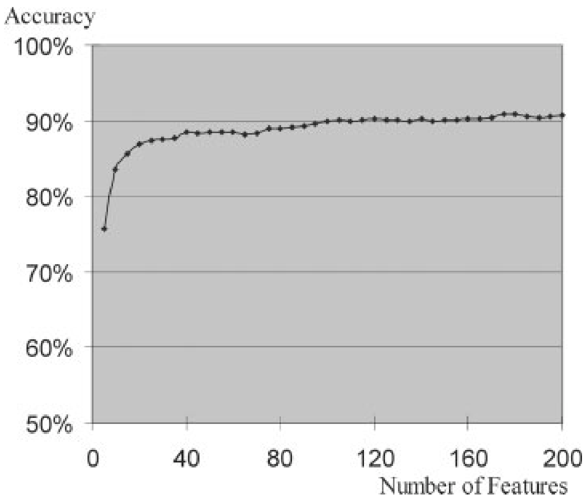


Figure 1. Prediction accuracy (*y*-axis) when selecting top features (*x*-axis) according to the ranked feature list generated by feature selection methods.

phobicity-based autocorrelation for internal segment. The latter result suggests that the hydrophobicity of termini and internal segments of IMP proteins could not be used to discriminate the five types of IMP. One of the main reasons for this finding is that the transmembrane segments of IMP are embedded in various locations in the sequence,[41,42] and thus they could not be captured using the division of the sequence into segments which was applied in our article. We plan to develop an improved scheme to divide the sequence into the transmembrane segments, which could be used to compute hydrophobic autocorrelations, as our future work.

Previous works show that GXXXG motif is frequently associated with β-branched residues at neighboring positions in TM proteins,[43,44] whereas this pair was not included among the selected features. We investigate the effect of GXXXG motif in distinguishing the five types of IMPs by comparing this pair with the top five selected features, which include LXXF, IXXXF, LXXXF, MXXXF, and FXXXV. The average feature values and the corresponding standard deviations of these five top pairs and the GXXXG motif for each IMP type are given in Table 3. The average feature values of LXXF pair varies

Table 1. Summary of the Feature-Based Sequence Representation and Results of the Feature Selection.

Feature set	Total number of features	Selected features
PSI-BLAST profile based collocation of AA pairs	1050	109
Hydrophobicity autocorrelations for entire sequence	60	10
Hydrophobicity autocorrelations for termini and internal segments	180	1
Total	1290	120

Table 2. Features Selected from the Set of 1050 PSI-BLAST Profile-Based Collocation of AA Pairs.

	PSI-BLAST profile based collocation of AA pairs																								
$p = 0$	AI	AM	AF	II	IM	IF	IW	LW	MF	MV	FF	FW	FV	LF	LL	LM	IL	LV	IV	CL	AL				
$p = 1$	AI	AM	AF	II	IM	IF	IW	LW	MF	MV	FF	FW	FV	LF	WV	CL	IL								
$p = 2$	AI	AM	AF	II	IM	IF	IW	LW	MF	MV	FF	FW	FV	LF	LL	LM	IL	LV	IV	VV	AL				
$p = 3$	AI	AM	AF	II	IM	IF	IW	LW	MF	MV	FF	FW	FV	LF	LL	LM	VV	LV	IV	WV	AL	LY	MY	FY	IY
$p = 4$	AI	AM	AF	II	IM	IF	IW	LW	MF	MV	FF	FW	FV	LF	LL	LM	IL	LV	IV	WV	MW	LY	MY	FY	IY

Each row corresponds to a different number of gaps between the pair of amino acids.

Table 3. The Average Feature Values and the Standard Deviations of the Top 5 Selected Features and the GXXXG Motif for 5 Types of IMPs.

	Avg. (Std.)					
IMP type	LXXF	IXXXF	LXXXF	MXXXF	FXXXV	GXXXG
Type-I	2.73 (1.35)	2.79 (1.39)	2.66 (1.33)	2.03 (1.10)	2.90 (1.35)	1.14 (0.67)
Type-II	4.46 (2.77)	4.31 (3.03)	4.64 (3.17)	3.63 (2.20)	3.75 (2.36)	0.95 (0.65)
Type-III	8.50 (2.59)	9.27 (2.12)	9.09 (2.61)	7.56 (0.60)	8.17 (2.22)	1.73 (1.28)
Type-IV	1.83 (1.23)	1.86 (1.35)	1.89 (1.32)	1.47 (1.26)	2.00 (1.36)	1.92 (2.13)
Type-V	2.57 (1.20)	2.26 (1.02)	2.53 (1.12)	1.81 (1.08)	2.34 (1.02)	1.54 (1.72)

between 1.83 and 8.50 for the five types of IMPs and similar spread is observed for the other four selected pairs. At the same time, the average feature values of GXXXG motif varies between 0.95 and 1.92, which shows that this motif occurs at similar rates for all IMP types and thus it does not allow differentiating between the different types. Based on a study by Engelman and coworkers,[44] we observe that Gly occurs less frequently than Leu, Ile, and Val in TM proteins and the GXXXG motif is less frequent than some other pairs, e.g., LXL pair occurs 7509 times and GXXXG occurs 1641 times. Additionally, we note that another study by Eisenberg and coworkers shows that GXXXG motif also frequently occurs in the α-helix of soluble proteins and it stabilizes helix–helix interactions.[45] This suggests that this pair may not be specific for TM proteins, but it rather constitutes a strong helical pattern.

Classification Algorithm

We use SVM classifier[46] that was previously applied to predict IMP types.[15,16] Given a training set of data point pairs (x_i, c_i), $i = 1, 2, \ldots n$, where x_i denotes the feature vector, $c_i = \{-1, 1\}$ denotes binary class label, n is the number of training data points, finding the optimal SVM is achieved by solving:

$$\min\|w\|^2 + C\sum_i \xi_i$$

such that

$$c_i(wz_i - b) \geq 1 - \xi_i \quad \text{and} \quad 1 \leq i \leq n$$

where w is a vector perpendicular to $wx - b = 0$ hyperplane that separates the two classes, C is a user-defined complexity constant, ξ_i are slack variables that measure the degree of misclassification of x_i for a given hyperplane, b is an offset that defines the size of a margin that separates the two classes, and $z = \phi(x)$, where $k(x,x') = \phi(x) \cdot \phi(x')$ is a user-defined kernel function.

The SVM classifier was trained using Platt's sequential minimal optimization algorithm,[47] which was further optimized by Keerthi et al.[48] The IMP type prediction that includes multiple classes is solved using pairwise binary classification, namely, a separate classifier is build for each pair of classes. Two popular families of kernel functions including polynomials and radial basis functions (RBF) were used. The kernel function selection and parameterization as well as selection of the complexity constant value were performed based on 10-fold cross validation on the training dataset using 120 features. The final classifier uses $C = 6$ and the RBF kernel

$$k(x_i, x_i') = e^{-\gamma\|x-x'\|^2} \quad \text{where } \gamma = 2.0$$

The classification algorithm and feature selection algorithms used to develop and compare the proposed method were implemented in Weka.[49]

Results and Discussion

Prediction Results

The proposed method was evaluated using three test types, i.e., by applying resubstitution and jackknife tests on the training dataset and by testing on an independent test dataset. The choice of the out-of-sample test was motivated by recent results that suggest that the jackknife test is more rigorous and objective than subsampling tests such as 5- or 10-fold cross-validation.[50,51] The evaluation setup closely follows previous studies[8,9,20] to allow for a consistent and objective comparison. Summary of classification results, which includes the corresponding classification accuracy values of IMP-TYPE and the competing methods, is shown in Table 4.

For training set 1 and test set 1, IMP-TYPE obtained 99.5% accuracy for the resubstitution test (on training set 1), 90.6% accuracy for the jackknife test (on training set 1), and 97.4% for the test on the independent test set 1. For the two out-of-sample tests (the jackknife test and the test on the independent set), IMP-TYPE provides substantial improvement over the best competing method, i.e., 4.8 and 0.6% higher accuracy, which corresponds to 4.8/14.2 = 33.8% and 0.6/3.2 = 18.8% reduction of the corresponding error rates.

Using the training set 2 and test set 2, which concern sequences with low, 40%, identity, we compared the proposed method with two representative, competing methods, see Table 4. They include best-performing ensemble model[20] and another SVM-based prediction method that uses composition vector to represent the sequences.[15] The comparison is limited to the above

Table 4. Summary of Experimental Comparison Between IMPTYPE and Competing Methods for Prediction of Membrane Protein Type.

			Accuracy (%)		
Dataset	Algorithm	Sequence representation	Resubstitution (training set)	Jackknife (training set)	Independent set (test set)
Training set 1 and test set 1	Covariant-discriminant[8]	AA composition	81.1	76.4	79.4
	Augmented covariant discriminant[9]	PseAA composition	90.9	80.9	87.5
	Weighted SVM[10]	PseAA composition	99.9	82.4	90.3
	SVM[15]	AA composition	Not reported	80.4	85.4
	Supervised locally linear embedding and KNN[11]	PseAA composition	Not reported	82.3	95.7
	OET-KNN[17]	PseAA composition	99.5	84.7	94.2
	SVM[16]	PseAA composition	99.0	78.3	86.6
	Stacking[19]	PseAA composition	98.7	85.4	94.3
	Fuzzy KNN[18]	PseAA composition	Not reported	85.6	95.7
	Ensemble of 40 NN[20]	PseAA composition	Not reported	85.8	96.8
	IMPTYPE (SVM) (this paper)	Custom (120 features)	99.5	90.6	97.4
Training set 2 and test set 2	SVM[15]	AA composition	99.9	77.8	87.6
	Ensemble of 40 NN[20]	PseAA composition	100	76.2	89.7
	IMPTYPE (SVM) (this paper)	Custom (120 features)	99.6	84.2	94.3

two methods since prior predictors were not tested on datasets with controlled, low identity, and thus we had to implement the empirical evaluation on these two sets. IMP-TYPE obtained 99.6% accuracy for the resubstitution test (on training set 2), 84.2% accuracy for the jackknife test (on training set 2), and 94.3% for the test on the independent test set 2. For both out-of-sample tests IMP-TYPE provides substantial improvement over the best competing ensemble based method, i.e., 6.4 and 4.6% higher accuracy, which corresponds to 6.4/22.2 = 28.8% and 4.6/10.3 = 44.7% reduction of the corresponding error rates. The accuracy of membrane protein type prediction for the datasets of low sequence identity (training and test sets 2) is lower than when considering sequences with high similarity (training and test sets 1). This is expected as the former datasets are more challenging. At the same time, our method is still capable of producing highly accuracy predictions for the low identity sequence, which are consistently better than the corresponding predictions of the competing methods. Most importantly, after sequences with high similarity were removed, the improvement provided by IMP-TYPE over the competing methods became bigger, i.e., 4.8% (accuracy improved by IMP-TYPE for training set 1) versus 6.4% (accuracy improved by IMP-TYPE for training set 2), 0.6% (accuracy improved by IMP-TYPE for test set 1) versus 4.6% (accuracy improved by IMP-TYPE for test set 2). This indicates that IMP-TYPE is characterized by an improved ability to provide accurate predictions in case when the query sequence shares low similarity with sequence used to derive the prediction model.

For test set 3 that includes recently resolved membrane proteins, IMP-TYPE obtains 90.3% accuracy for the independent test, whereas the ensemble model and the composition vector-based SVM obtain 84.8 and 84.2%, respectively. Detailed analysis shows that among 165 proteins in this set, the three methods provided consistent predictions for 148 chains, which include 133 correct predictions and 15 incorrect predictions. The remaining 17 samples include five type-I IMP, two type II IMP, and 10 type III IMP proteins. The proposed IMP-TYPE method made correct predictions for all five samples of type-I, one sample of type-II, and all 10 samples of type-III. The best competing ensemble-based method correctly predicted two samples of type-I, one of type-II, and four of type-III, whereas the composition vector-based SVM correctly classified one sample of type-I, two of type-II, and three of type-III. We note that except the two samples of type-II IMP, all correct predictions made by the competing ensemble and SVM models were also correctly predicted by IMP-TYPE, whereas IMP-TYPE provided additional nine correct predictions. This shows that when applied to new IMPs, our method duplicates virtually all correct predictions of the competing methods and adds a number of additional correct predictions for which the competing methods fail to provide correct outcomes.

We also provide a detailed comparison of performance, which includes sensitivity, specificity, and Matthews correlation coefficient (MCC) for each type of IMP, between IMP-TYPE and the best-performing ensemble model[20]; see Table 5. For test set 1, IMP-TYPE obtained comparable sensitivity and MCC for all five types of IMP, and consistently higher specificity than the ensemble model. For test set 2, which includes sequences with low similarity, except for the sensitivity for type-IV IMP, all other sensitivity, specificity, and MCC values obtained with IMP-TYPE are higher than the corresponding values provided by the competing ensemble model. This again confirms that IMP-TYPE constitutes and improvement, rather than being a complementary method, when compared with the best existing method.

IMP-TYPE obtained relatively high sensitivities and specificities, i.e., between 94 and 100%, for type-I and type-III IMP for both out-of-sample tests. This indicates that the proposed feature representation can accurately characterize these two types of IMPs.

Table 5. Comparison Between IMP-TYPE and the Ensemble Model[20] for Individual IMP Types on Test Set 1 and Test Set 2.

	Independent test set 1						Independent test set 2					
	Ensemble model[20]			IMPTYPE			Ensemble model[20]			IMPTYPE		
IMP type	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
Type-I	0.96	0.96	0.95	0.98	0.97	0.97	0.90	0.86	0.85	0.96	0.94	0.94
Type-II	0.79	0.94	0.86	0.81	0.97	0.88	0.56	0.82	0.66	0.71	0.91	0.78
Type-III	0.99	0.97	0.93	1.00	0.98	0.95	0.95	0.93	0.83	0.98	0.95	0.90
Type-IV	0.57	0.80	0.68	0.43	0.86	0.61	0.50	0.40	0.44	0.25	1.00	0.50
Type-V	0.91	0.91	0.92	0.89	0.98	0.93	0.75	0.69	0.71	0.92	0.85	0.88
Average	0.84	0.92	0.87	0.82	0.95	0.87	0.73	0.74	0.70	0.76	0.93	0.80

It includes a subset of sequence from test set 1 that is characterized by low sequence identity.

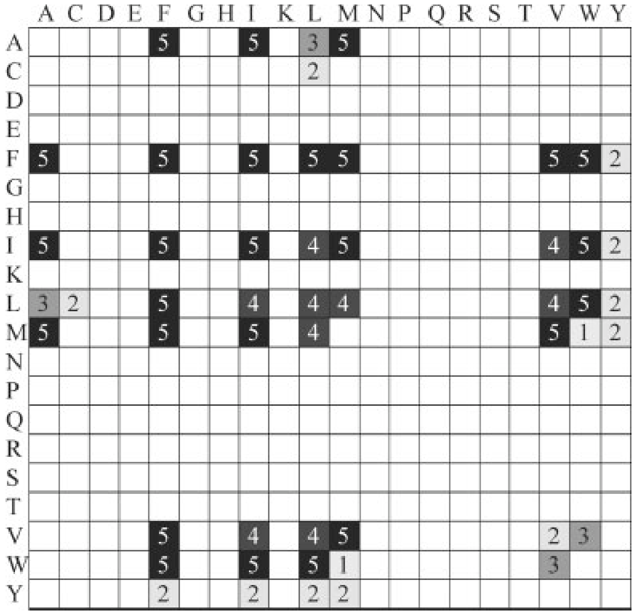
Analysis of Biological Patterns in the Proposed Features Representation

The proposed method not only provides accurate predictions for the five types of IMPs, but is also characterized by a transparent feature set that can be used to analyze the underlying biological patterns. The selected AA pairs include mainly hydrophobic residues, which is consistent with a recent study concerning TM protein that indicates that formation of TM helices is associated with hydrophobic residues.[52]

The selected 109 AA pairs are summarized in Figure 2, which shows that over 90% of the selected AA pairs are based on only seven AAs. Using the occurrences of AAs pairs in the proposed representation, the 20 AAs can be divided into two groups: (1) the first one includes AAs that occur in the pairs at least 18 times (see Table 6); (2) the second group includes the remaining residues which occur in the pairs less than 10 times. The first group includes Phe (F), Ile (I), Leu (L), Met (M), Val (V), Ala (A), and Trp (W). Among the seven AAs, Ala, Ile, Leu, and Val are traditionally classified as hydrophobic residues; Phe and Trp are among aromatic residues; and Met is a polar residue. However, according to the three commonly used hydrophobicity index tables,[35–37] Phe and Met are consistently assigned positive index values which indicate hydrophobic propensity, whereas Trp is assigned positive values in two of these hydrophobicity index tables. Overall, the values in the aforementioned hydrophobicity index tables suggest that all seven residues that occur frequently in the selected pairs share hydrophobic propensity. At the same time, our predictions show that the same seven hydrophobic residues are crucial for obtaining accurate classification of the five types of IMP.

The specific roles of hydrophobic residues in the formation of TM segments have been discussed in several studies:

- The distribution of the 20 AAs in the TM segments was systematically examined in study by Landolt-Marticorena et al.[53] Their dataset contained 115 human type-I membrane proteins. Based on contrasting the AA composition of the TM segments and the AA composition of the entire sequences, the authors found that Ile, Leu, Val, Ala, and Phe occur mostly in TM segments. These five residues are among the seven residues used in our proposed representation. Another study by Ulmschneider and Sansom indicates that Leu, Ala, Gly, Val, Ile, and Phe occur more frequently than other residues in TM helices, whereas Gly, Thr, Ala, Tyr, Leu, and Val occur more frequently than other residues in TM strands.[52] Both of these two groups of AAs include some hydrophobic residues, i.e., Ala, Val, and Leu. We note that the latter study is limited in its scope since it was based on only 29 structures.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A					5			5		3	5									
C										2										
D																				
E																				
F	5				5			5		5	5							5	5	2
G																				
H																				
I	5				5			5		4	5							4	5	2
K																				
L	3	2			5			4		4	4							4	5	2
M	5				5			5		4								5	1	2
N																				
P																				
Q																				
R																				
S																				
T																				
V					5			4		4	5							2	3	
W					5			5		5	1							3		
Y					2			2		2	2									

Figure 2. A summary of the selected features that correspond to AA pairs. Each cell in the figure corresponds to a given AA pairs defined by AAs in the corresponding row and column. The value shown for each shaded cell corresponds to the number of times a given pair occurs in the proposed feature representation (across all value of p). For example, pair AI occurs five times (for p = 0, 1, 2, 3, 4) in Table 2. Since a given pair and its symmetrical form were combined into one feature, the Figure is also symmetrical. Darker shading corresponds to larger value for the corresponding cell.

Table 6. Occurrences of the 20 AAs in the Selected AA Pairs.

Amino acid	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Occurrence in the pairs	18	2	0	0	42	0	0	40	0	37	27	0	0	0	0	0	0	25	19	8

- A study by Ulmschneider et al.[54] calculated the potentials of mean force for membrane protein based on 46 all-α-helical membrane protein with structures resolved with resolutions greater than 4A. They reported that hydrophobic residues (Ala, Ile, Val, and Leu) display a potential energy near the center of the membrane region and extending into the interfacial regions. They classified Met into polar residue group and Phe into aromatic residue group, although we note that the graphs that show the potentials of mean force of these two residues are characterized by a different shape when compared with the graphs of other polar and aromatic residues, and that this shape is similar to the graphs of hydrophobic residues (see Fig. 6 in ref. [54]). This supports our claim of hydrophobic tendency of Met and Phe in IMP. We again emphasize that the hydrophobic propensity of Met and Phe is also confirmed by all three of the aforementioned hydrophobicity index tables and the fact that they constitute the sequence representation in the proposed method that is characterized by high prediction accuracy for TM proteins. We hypothesize that Met and Phe display a certain degree of hydrophobicity, which may be more crucial than their polarity or aromaticity in TM segments.
- In another study, Punta and Maritan[55] generated two scales for AA membrane propensity. The scales were based on different datasets, one dataset included three-dimensional structures determined by X-ray diffraction or NMR method, and the other contained helix-bundle membrane proteins in which TM helices have been identified by experimental techniques other than X-ray and NMR. Coincidently, the seven residues identified in this work based on the proposed sequence representation (Phe, Ile, Leu, Met, Val, Ala, and Trp) correspond to the seven highest values of membrane propensity in both of the above scales, i.e., less than −0.17 in scale MPS(1D_r) and less than −0.15 in scale MPS(3D), in which negative value indicates a high membrane propensity (see Table 2 in ref. 55).

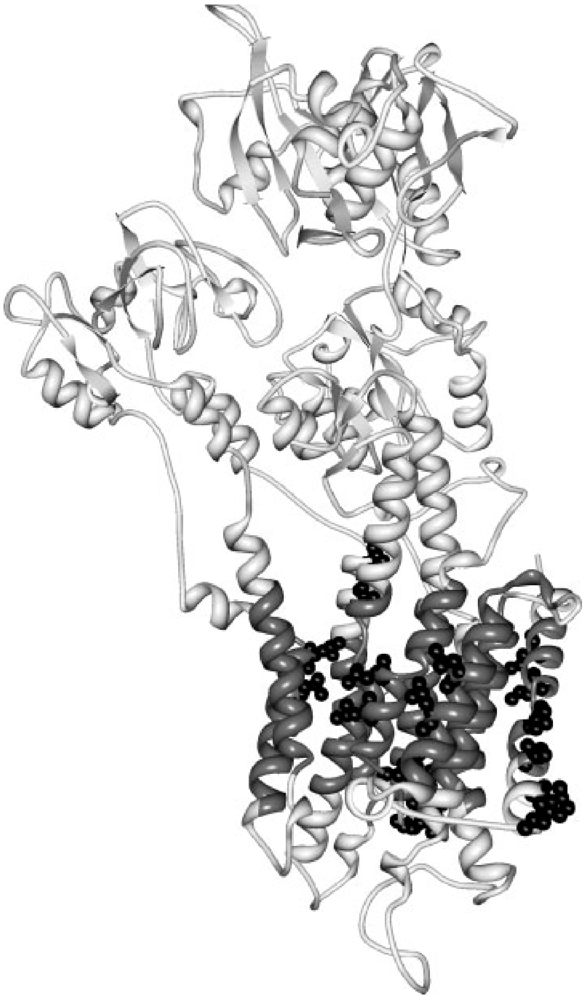


Figure 3. Ribbon structure of chain A of 2AGV. The transmembrane segments (shown in gray) were identified using TMDET.[56] The LXXXL pairs (shown in black) is displayed using ball-and-stick representation. Among the total of nine LXXXL pairs, seven are located in the transmembrane segments and the other two pairs are located in the helical regions close to the transmembrane region.

This discussion shows that the proposed sequence representation is tightly correlated with structural patterns that occur in TM helices, which provides further validation for the high quality of the obtained classifications. Finally, an example that illustrates how one of the selected AA pairs can be used to find transmembrane segments is given in Figure 3. The LXXXL pair occurs nine times in the chain A of 2AGV, which is a multipass TM protein. Seven out the nine LXXXL pairs are buried in the membrane and the other two pairs are located in the helical regions near the membrane.

Conclusions

An accurate computational model, IMP-TYPE, is proposed for the classification of the five types of IMP from protein sequences. When contrasted with the best-performing competing method, IMP-TYPE obtains 4.8 and 0.6% higher accuracy, which translates into 34 and 19% error reduction rate for the out-of-sample tests on the training set 1 and test set 1, respectively. Our empirical tests conducted with a large sets of membrane proteins that is characterized by low sequence similarity (training set 2 and test set 2) shows that IMP-TYPE is characterized by even sore substantial improvements in accuracy when the predicted sequences share low identity with the sequences used for prediction, i.e., IMP-TYPE obtains 6.4 and 4.6% higher accuracy, which translates into 29 and 45% error reduction rate, when compared with the best performing competing method. Our evaluation performed with a large set of recently resolved IMP proteins also shows that the proposed method duplicates correct predictions of competing methods while providing additional correct predictions, i.e., IMP-TYPE can be used standalone, i.e., without the need to use results of other membrane protein type prediction methods.

The proposed method not only provides accurate classification, but most importantly is characterized by a transparent model that shows interesting biological patterns. The AA pairs selected to develop sequence representation used by IMP-TYPE are consistent with the results of several studies concerning membrane proteins,[52–55] which in turn were based on a variety of different datasets. Our results show that hydrophobic AA pairs, which are used by IMP-TYPE, can be successfully used as makers of IMP and that they, at the same time, can be applied to distinguish between different types of IMP. We also hypothesize that hydrophobic propensity of Met and Phe may be more crucial than their polarity or aromaticity in IMP.

References

1. von Heijne, G. Nat Rev Mol Cell Biol 2006, 7, 909.
2. Wallin, E.; von Heijne, G. Protein Sci 1998, 7, 1029.
3. Grisshammer, R.; Tate, C. G. Q Rev Biophys 1995, 28, 315.
4. Schnell, J. R.; Chou, J. J. Nature 2008, 451, 591.
5. Oxenoid, K.; Chou, J. J. Proc Natl Acad Sci USA 2005, 102, 10870.
6. Call, M. E.; Schnell, J. R.; Xu, C.; Lutz, R. A.; Chou, J. J.; Wucherpfennig, K. W. Cell 2006, 127, 355.
7. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2007, 360, 339.
8. Chou, K. C.; Elrod, D. W. Proteins 1999, 34, 137.
9. Chou, K. C. Proteins 2001, 43, 246.
10. Wang, M.; Yang, J.; Liu, G. P.; Xu, Z. J.; Chou, K. C. Protein Eng Des Sel 2004, 17, 509.
11. Wang, M.; Yang, J.; Xu, Z. J.; Chou, K. C. J Theor Biol 2005, 232, 7.
12. Cai, Y. D.; Zhou, G. P.; Chou, K. C. Biophys J 2003, 84, 3257.
13. Chou, K. C.; Cai, Y. D. J Chem Inf Model 2005, 45, 407.
14. Chou, K. C.; Cai, Y. D. Biochem Biophys Res Commun 2005, 327, 845.
15. Cai, Y. D.; Pong-Wong, R.; Feng, K.; Jen, J. C. H.; Chou, K. C. J Theor Biol 2004, 226, 373.
16. Liu, H.; Yang, J.; Wang, M.; Xue, L.; Chou, K. C. Protein J 2005, 24, 385.
17. Shen, S.; Chou, K. C. Biochem Biophys Res Commun 2005, 334, 288.
18. Shen, H. B.; Yang, J.; Chou, K. C. J Theor Biol 2006, 240, 9.
19. Wang, S. Q.; Chou, K. C. J Theor Biol 2006, 242, 941.
20. Shen, S.; Chou, K. C. Amino Acids 2007, 32, 483.
21. Chou, K. C.; Shen, H. B. Biochem Biophys Res Commun 2007, 360, 339.
22. Hirokawa, T.; Boon-Chieng, S.; Mitaku, S. Bioinformatics 1998, 14, 378.
23. Cserzo, M.; Eisenhaber, F.; Eisenhaber, B.; Simon, I. Bioinformatics 2004, 20, 136.
24. Yuan, Z.; Mattick, J. S.; Teasdale, R. D. J Comput Chem 2004, 25, 632.
25. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. Nucleic Acids Res 1997, 25, 3389.
26. Chen, K.; Kurgan, L.; Rahbari, M. Biochem Biophys Res Commun 2007, 355, 764.
27. Li, W.; Godzik, A. Bioinformatics 2006, 22, 1658.
28. Chen, K.; Kurgan, L. A.; Ruan, J. BMC Struct Biol 2007, 7, 25.
29. Jones, D. T. J Mol Biol 1999, 292, 195.
30. Kim, H.; Park, H. Proteins 2004, 54, 557.
31. Jones, D. T. Bioinformatics 2007, 23, 538.
32. Chen, K.; Kurgan, L. A. Bioinformatics 2007, 23, 2843.
33. Homaeian, L.; Kurgan, L. A.; Ruan, J.; Cios, K. J.; Chen, K. Proteins 2007, 69, 486.
34. Kurgan, L. A.; Homaeian, L. Pattern Recognit 2006, 39, 2323.
35. Fauchere, J. L.; Pliska, V. Eur J Med Chem 1983, 18, 369.
36. Eisenberg, D.; Weiss, R. M.; Trewilliger, T. C. Proc Natl Acad Sci USA 1984, 81, 140.
37. Kyte, J.; Doolitle, R. F. J Mol Biol 1982, 157, 105.
38. Chen, K.; Kurgan, L. A.; Ruan, J. J Comput Chem 2008, 29, 1596.
39. Forman, G. J Mach Learn Res 2003, 3, 1289.
40. Kira, K.; Rendell, L.A. Proceedings of the Ninth International Workshop on Machine Learning, Aberdeen, Scotland, UK, 1992; p. 249.
41. Cuthbertson, J. M.; Doyle, D. A.; Sansom, M. S. Protein Eng Des Sel 2005, 18, 295.
42. Gromiha, M. M.; Ahmad, S.; Suwa, M. J Comput Chem 2004, 25, 762.
43. Russ, W. P.; Engelman, D. M. J Mol Biol 2000, 296, 911.
44. Senes, A.; Gerstein, M.; Engelman, D. M. J Mol Biol 2000, 296, 921.
45. Kleiger, G.; Grothe, R.; Mallick, P.; Eisenberg, D. Biochemistry 2002, 41, 5990.
46. Vapnik, V. The Nature of Statistical Learning Theory; New York: Springer-Verlag, 1999.
47. Platt, J. Fast training of support vector machines using sequential minimal optimization. In Advances in Kernel Methods—Support Vector Learning; Schoelkopf, B.; Burges, C.; Smola, A., Eds.; MIT: Cambridge, MA, 1998.
48. Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murphy, K. R. K. Neural Comput 2001, 13, 637.
49. Witten, I.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques; Morgan Kaufmann: San Francisco, 2005.
50. Chou, K. C.; Shen, H. B. Nature Protoc 2008, 3, 153.
51. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1.
52. Ulmschneider, M. B.; Sansom, M. S. Biochim Biophys Acta 2001, 1512, 1.
53. Landolt-Marticorena, C.; Williams, K. A.; Deber, C. M.; Reithmeier, R. A. J Mol Biol 1993, 229, 602.
54. Ulmschneider, M. B.; Sansom, M. S.; Di-Nola, A. Proteins 2005, 59, 252.
55. Punta, M.; Maritan, A. Proteins 2003, 50, 114.
56. Tusnady, G. E.; Dosztanyi, Z.; Simon, I. Bioinformatics 2005, 21, 1276.