



Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: An approach from discrete wavelet transform

Jian-Ding Qiu^{a,*}, Jian-Hua Huang^a, Ru-Ping Liang^a, Xiao-Quan Lu^{b,*}

^a Department of Chemistry, Nanchang University, Nanchang 330031, People's Republic of China

^b College of Chemistry and Chemical Engineering, Northwest Normal University, Lanzhou 730070, People's Republic of China

ARTICLE INFO

Article history:

Received 15 January 2009

Available online xxxx

Keywords:

Discrete wavelet transform

Support vector machine

Hydrophobicity scale

G-protein-coupled receptors

Pseudo amino acid composition

ABSTRACT

Being the largest family of cell surface receptors, G-protein-coupled receptors (GPCRs) are among the most frequent targets. The functions of many GPCRs are unknown, and it is both time-consuming and expensive to determine their ligands and signaling pathways by experimental methods. It is of great practical significance to develop an automated and reliable method for classification of GPCRs. In this study, a novel method based on the concept of Chou's pseudo amino acid composition has been developed for predicting and recognizing GPCRs. The discrete wavelet transform was used to extract feature vectors from the hydrophobicity scales of amino acid to construct pseudo amino acid (PseAA) composition for training support vector machine. The prediction accuracies by the current method among the major families of GPCRs, subfamilies of class A, and types of amine receptors were 99.72%, 97.64%, and 99.20%, respectively, showing 9.4% to 18.0% improvement over other existing methods and indicating that the proposed method is a useful automated tool in identifying GPCRs.

© 2009 Elsevier Inc. All rights reserved.

G-protein-coupled receptors (GPCRs)¹ play a key role in cellular signaling pathways that regulate many basic physiological processes such as neurotransmission, secretion, growth, cellular differentiation, and inflammatory and immune responses [1]. GPCRs consist of a single polypeptide that crosses the membrane seven times [2]. The N terminal of these proteins is located extracellularly, and the C terminal is extended in the cytoplasm. This arrangement makes these proteins capable of transducing an extracellular signal into the cell via a guanine binding protein (G protein) [3]. Therefore, structural and functional annotation of these proteins is useful in understanding the processes of signal transduction. Unfortunately, it is both time-consuming and expensive to determine their structure by experimental methods because GPCRs are difficult to crystallize and most of them do not dissolve in normal solvents [4]. With the rapid accumulation of biological data produced by many large-scale genome sequencing projects, it is of great practical significance

to develop automatic and reliable methods to predict the structure and function of GPCRs [5].

Many methods have been proposed for the prediction of GPCRs during the past few years. One commonly used method is sequence similarity searching in protein databases by sequence alignment tools based on pairwise similarity such as BLAST and FASTA [6,7]. Several pattern databases have been constructed by some investigators for the prediction of GPCRs [8,9]. However, these methods are not always successful when the query protein sequences have no significant sequence similarity to the database sequences [10]. Therefore, some statistical and machine learning methods have been proposed, including the statistical analysis method [11], covariant discriminant algorithm (CDA) [12,13], support vector machine (SVM) [14,15], hidden Markov models [16,17], and bagging classification tree [18]. Among them, SVM has been widely used to solve various biological problems such as initiation sites [19], membrane protein types [20], protein–protein interactions [21], protein subcellular localization [22], protein fold [23], and gene expression pattern discovery [24] due to its attractive features, including effective avoidance of overfitting, ability to handle large feature space, and absence of local minima. However, as a machine learning technique, SVM requires a fixed length of feature vectors. One commonly used feature vector is conventional amino acid (AA) composition, where the sample of a protein is represented by 20 discrete numbers, with each number representing the occurrence frequency of one of the 20 constituent native amino

* Corresponding authors. Fax: +86 791 3969963 (J.-D. Qiu).

E-mail addresses: jdqiu@ncu.edu.cn, qiujianding@163.com (J.-D. Qiu), luxq@nwnu.edu.cn (X.-Q. Lu).

¹ Abbreviations used: GPCR, G-protein-coupled receptor; CDA, covariant discrimination algorithm; SVM, support vector machine; AA, amino acid; PseAA, pseudo amino acid; WT, wavelet transform; DWT, discrete wavelet transform; SV, support vector; Db10, Daubechies of number 10; Db6, Daubechies of number 6; Bior2.4, Biorthogonal of number 2.4; Bior3.3, Biorthogonal of number 3.3; Coif4, Coiflet of number 4; Sym10, Symlets of number 10; RBF, radial basis function.

acids. Obviously, if one uses the conventional AA composition to represent the sample of a protein, all of its sequence order will be lost. To include these effects, the concept of pseudo amino acid (PseAA) composition was proposed [25,26] and has been widely used in many research studies [27–31]. Stimulated by the concept of PseAA composition, the current study was initiated in an attempt to introduce a novel approach—wavelet transform (WT) analysis to formulate the PseAA composition. WT is a local time–frequency analysis method with both a changeable time window and a frequency window. Because of its character of multiresolution, WT has been applied in bioinformatics recently to analyze and process biological data [32–41].

In this article, a novel method based on discrete wavelet transform (DWT) and SVM is proposed to recognize and classify the GPCRs. DWT analysis can decompose the AA sequences into coefficients at different dilations and then remove the noise component from the profiles, so that it can provide local structures of sequences that can more effectively reflect the sequence order effects. This method consists of three main steps. First, the AA residues were translated into numerical signal representation by hydrophobicity scales. Second, the hydrophobic profile was decomposed into wavelet coefficients by using DWT. Finally, SVM was applied to deal with the problem of multiclassification using the feature coefficients produced by DWT as input. The influence of AA hydrophobic values, decomposition levels, wavelets, and kernel functions are optimized for prediction.

Materials and methods

Data sets

Three data sets were used in this work. The first data set contains 1238 GPCR sequences that can be classified into three major families: 1103 class A–rhodopsin like, 84 class B–secretin like, and 51 class C–metabotropic/glutamate/pheromone [13]. The average sequence identity percentages for classes A, B, and C are 18.05%, 22.67%, and 26.94%, respectively [13]. The second data set used to recognize the subfamilies of class A–rhodopsin like was generated by Strope and Moriyama [42]. It contains 574 sequences that can be classified into three subfamilies: 126 amine/rhodopsin subfamily, 139 peptide subfamily, and 309 olfactory subfamily. The third data set, with an average sequence identity percentage of 26.25%, was used to classify the different types of amine receptors [12,15,18]. It contains 167 sequences; of these, 31 are of acetylcholine-type receptors, 44 are of adrenoceptor-type receptors, 38 are of dopamine-type receptors, and 54 are of serotonin-type receptors.

PseAA composition and WT

A protein sequence can be represented as a series of amino acids by their single-character codes A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y, formulated as

$$R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_8 \dots R_L \quad (1)$$

Suppose that $H(R_1)$ is the hydrophobic value of the first residue R_1 , $H(R_2)$ is the hydrophobic value of the second residue R_2 , and so forth. In terms of these hydrophobic values, the protein sequence of Eq. (1) can be converted to a digit signal from which we can generate several groups of wavelet coefficients. The WT of a signal $f(x)$ is defined as the sum over all of the time of the signal $f(x)$ multiplied by a scaled and shifted version of the wavelet function $\psi(x)$. The coefficients $T(a,b)$ of the WT of the signal $f(x)$ can be expressed as

$$T(a,b) = \frac{1}{\sqrt{a}} \int_0^x f(x) \psi\left(\frac{x-b}{a}\right) dx \quad (2)$$

where a and b are the scaling and shifting parameters (they belong to a set of real numbers $[R]$, and $a > 0$), x is the AA sequence length of the protein, and $\psi[(x-b)/a]$ is the analyzing wavelet function. The DWT uses $a_0 = 2$ and $b_0 = 1$, so that the results can lead to a binary dilation of 2^{-m} and a dyadic translation of $n2^m$. Therefore,

$$\Psi_{m,n}(x) = 2^{-m/2} \psi(2^{-m}x - n) \quad (3)$$

where $m = 1, 2, \dots$, and $n = 0, 1, 2, \dots$. The wavelet coefficients of the signal $f(x)$ are obtained by following formula:

$$T(a,b) = \langle f(x), \Psi_{a,b}(x) \rangle = 2^{-m/2} \int_0^x f(x) \psi(2^{-m} \cdot x - n) dx \quad (4)$$

The coefficients $T(a,b)$ of the DWT can be divided into two parts; one is the approximation coefficient $A^j(n)$, which represents the high-scale and low-frequency components of the signal $f(x)$, and the other is the detail coefficient $D^j(n)$, which represents the low-scale and high-frequency components of the signal $f(x)$ [43]. The approximation coefficients and detail coefficients of the DWT for the digital signal $f(x)$ at level j can be expressed as

$$A^j(n) = \sum_{k \in \mathbb{Z}} h_{k-2n} A^{j-1}(k) \quad (5)$$

$$D^j(n) = \sum_{k \in \mathbb{Z}} h_{k-2n} D^{j-1}(k) \quad (6)$$

According to both experimental and theoretical progress in protein dynamics, it is clear that low-frequency internal motions do exist in protein and DNA molecules and indeed play a significant role in biological functions [44–46]. Using the low-frequency wavelet coefficients to formulate the sample of a protein can better reflect its overall sequence order effect. Here we chose a GPCR sequence from class A–rhodopsin like (accession number Q15760) as an example to describe the process of extracting the information hidden in AA sequence by using DWT. In Fig. 1, S denotes the hydrophobicity sequence of Q15760; D_1 , D_2 , and D_3 denote three detail scales at levels from $j = 1$ to $j = 3$; and A_3 denotes the approximation scale at level $j = 3$ after using DWT. It can be clearly seen that two types of coefficients—approximation coefficient A_3 and detail coefficients D_1 , D_2 , and D_3 —can be obtained from DWT decomposition for each GPCR sequence in Fig. 1.

To further decrease the dimensionality of the extracted feature vectors, statistics over the set of the wavelet coefficients were used [47]. The following statistical features calculated from the approximation coefficients and detail coefficients were used for the classification of GPCRs: (i) maximum of the wavelet coefficients in each sub-band, (ii) mean of the wavelet coefficients in each sub-band, (iii) minimum of the wavelet coefficients in each sub-band, and (iv) standard deviation of the wavelet coefficients in each sub-band. So, a protein x can be characterized as a $4(j+1)$ dimension feature vector. In this study, the decomposition level 3 was chosen to classify the GPCRs, and the obtained 16 dimension feature vectors were then inputted to SVM for classification.

Support vector machine

SVM is a kind of learning machine based on statistical learning theory. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, large datasets, and large input spaces [48]. Details about the theory of SVM can be found in the literature [49,50]. Basically, for a given dataset $x_i \in R^n$ ($i = 1, \dots, N$) with corresponding labels y_i ($y_i = +1$ or -1 , representing the two classes to be classified), SVM gives a decision function:

$$f(x) = \sum_{i=1}^N y_i a_i K(x_i, x_j) + b. \quad (7)$$

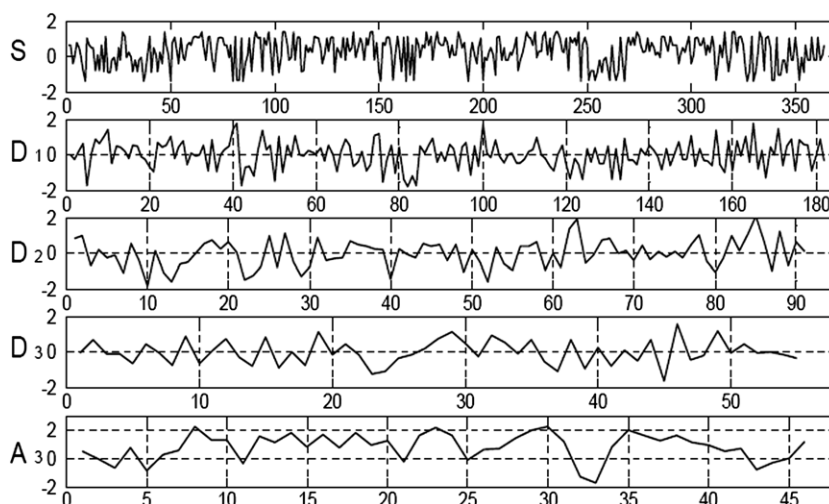


Fig. 1. Hydrophobicity scales plot of the GPCR Q15760 and the wavelet decomposition process from levels $j = 1$ to $j = 3$.

In Eq. (7), a_i is the coefficient to be learned and K is a kernel function. Parameter a_i is trained through maximizing the Lagrangian expression given below:

$$\max_{a_i} = \sum_{i=1}^N a_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j K(x_i, x_j) \quad (8)$$

Subject to $\sum_{i=1}^N a_i y_i = 0$, $0 \leq a_i \leq C$, $i = 1, \dots, N$. In addition, C is the penalty parameter, which is now the upper bound on a_i . Only if $0 < a_i \leq C$ are the corresponding data points called support vectors (SVs).

For actual implementation, we used the LIBSVM package (version 2.81) [51]. To obtain an SVM classifier with optimal performance, three kernel functions (linear, polynomial, and radial basis function) were tested in our research, and the penalty parameter C and kernel parameter were tuned based on the training set using the grid search strategy in LIBSVM.

Design and implementation of the prediction system

SVM was originally designed for binary classification [52], whereas prediction of GPCRs is a multiclass classification problem. In this study, the one versus one (o-v-o) SVM training strategy was adopted to decompose multiclass into a series of binary SVMs to solve this problem [23]. For an N class classification, $N \times (N - 1) / 2$ classifier needs to be trained, covering all possible different pairwise combinations (i, j) , $i < j$, such that when training the (i, j) classifier patterns belonging to class i are used as positive samples and those from class j are taken as negative samples. Finally, one unknown sample is classified into the class obtained by accumulating the binary decisions and selecting as the winning class the one with more votes [23].

Prediction assessment

Among the independent dataset test, subsampling (e.g., 5- or 10-fold cross-validation) test, and jackknife test, which are most often used for examining the accuracy of a statistical prediction method [53], the jackknife test was deemed the most objective test that can always yield a unique result for a given benchmark dataset [54,55]. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods [25–31]. However, because jackknife cross-validation would take a much longer time to test an SVM-based predictor, here we chose to use the 10-fold cross-validation test

to test the current method, in which the data set was divided into 10 subsets of approximately equal size [18,23]. This means that the data were partitioned into training and testing data in 10 different ways. After training the SVMs with a collection of 9 subsets, the performance of the SVMs was tested against the 10th subset. This process was repeated 10 times, so that every subset was once used as the test data. The average of the 10 estimates of the accuracy rate was reported as the cross-validation accuracy. The prediction quality was then evaluated by the Q percentage accuracy, which was defined as the number of instances correctly predicted over the total number of instances in the test set: $Q_i = (\sum Z_i / N) \times 100\%$ [23,56], where N is the total number of GPCRs in the test set and Z_i represents the true positives. Individual Q_i relates to the overall Q in a very simple way. An individual class contributes to the overall accuracy in proportion to the number of GPCRs in its class and, thus, has a weighted $W_i = n_i / N$. Therefore, the overall accuracy equals the weighted average over individual classes:

$$Q = \sum_{i=1}^K W_i Q_i \quad (9)$$

Results and discussion

Selecting wavelet functions

Based on different basis functions, the wavelets have different families; every family has its quality fit for different signal and has different results. Because the characteristics of the analyzing wavelet control the performance of the WT, the better the analyzing wavelet matches the underlying structure in the signal, the more concise and sparse the WT representation. It has been clearly stated that the amount of signal compression and the reconstruction quality are highly dependent on the selection of the mother wavelet. To investigate the effect of the wavelets on classification accuracy, eight wavelet functions—Haar, Daubechies of number 10 (Db10), Daubechies of number 6 (Db6), Biorthogonal of number 2.4 (Bior2.4), Biorthogonal of number 3.3 (Bior3.3), Coiflet of number 4 (Coif4), Symlets of number 10 (Sym10), and Meyer—were chosen for testing in the research. The performances for GPCRs with different types of wavelet functions are summarized in Table 1. As can be seen from the table, by using the Coif4 wavelet, the prediction accuracies for major families of GPCRs, subfamilies of class A, and different types of amine receptors reached 99.72%, 97.64%, and 99.20%, respectively. In general, the Coif4 wavelet per-

Table 1

Performance of our method based on different wavelet functions.

Wavelet function	Predicted accuracy (%)		
	Families	Subfamilies	Types
Db1	99.96	95.77	97.80
Db6	99.43	94.16	99.00
Db10	99.43	90.33	96.21
Coif4	99.72	97.64	99.20
Bior2.4	99.23	90.76	98.80
Bior3.3	99.34	91.99	97.80
Sym10	99.55	95.65	96.21
Meyer	99.50	97.12	97.81

formed better than other wavelet functions. Therefore, the Coif4 wavelet was selected as the appropriate wavelet function for the prediction of GPCRs in this study.

Selecting kernel functions

The kernel function selection is one of the key problems for SVM; different kernel function can produce different SVMs. Three kernel functions—linear kernel [57], polynomial kernel [48], and radial basis function (RBF) kernel [58]—were tested in the study. The prediction accuracies for GPCRs with different types of kernel functions are summarized in Table 2. In general, polynomial kernel function performed better than linear and RBF kernel functions, as indicated by the overall accuracies in Table 2. Therefore, the polynomial kernel function was selected as the appropriate kernel function in the study.

Selecting amino acid hydrophobicity scales

In an aqueous environment, hydrophobic molecules, including the hydrophobic AA side chains, are forced together to minimize the disruptive effect on the hydrogen-bonded water molecule networks. Thus, distribution of hydrophobic AA side chains has a significant impact on the protein structure. In this work, the hydrophobicity scales were used to map the protein AA sequences into protein hydrophobicity sequences for classifying and recognizing the GPCRs. Quantitative estimates of the hydrophobicity can be derived from their relative concentrations in organic versus water bulk phase of a binary solution. Different experimental conditions, solvents, and computational schemes have led to different sets of hydrophobicity scales [59–69]. In fact, only several AA hydrophobicity scales are available to transform AA sequences into real numbers. Three sets of AA hydrophobicity scales that have been commonly used in previous studies [28,55]—Kyte–Doolittle hydrophobicity scales (KDHΦ) [70], Mandell hydrophobicity scales (MHΦ) [71], and Fauchereand hydrophobicity scales (FHΦ) [65]—were investigated in the current study. The performances based on the three hydrophobicity scales are listed in Table 3. It can be seen from the table that the performance based on FHΦ achieved the highest overall accuracies of 99.72%, 97.64%, and 99.20% for major families of GPCRs, subfamilies of class A, and different types of amine receptors, respectively.

Table 2

Performance of our method based on three different kernel functions.

Kernel function	Predicted accuracy (%)		
	Families	Subfamilies	Types
RBF	99.46	95.99	99.59
Linear	96.14	90.06	90.01
Polynomial	99.72	97.64	99.20

Table 3

Performance of our method based on three different hydrophobicity scales.

Hydrophobic scale	Predicted accuracy (%)		
	Families	Subfamilies	Types
FHΦ	99.72	97.64	99.20
KDHΦ	99.46	92.25	98.01
MHΦ	98.93	92.35	98.60

Comparison with other methods

The data set that contains 167 GPCRs of amine receptor for four types [12,15,18] was used to ascertain the quality of the proposed method. It can be seen from Table 4 that the overall accuracy obtained by the current approach was 99.2%, which is 11.8% higher than the result of the bagging tree method, and 16% higher than that of the CDA method. Furthermore, the success rates of dopamine and serotonin receptors were remarkably enhanced. The success rate by the current approach for dopamine receptor was 100%, approximately 18% and 16% higher than the success rates by the CDA and bagging tree algorithms, respectively. The success rate by the current approach for serotonin receptor was 98.90%, approximately 10% and 18% higher than the success rates by the CDA and bagging tree algorithms, respectively. Bhasin and Raghava [15] developed the GPCRclass method, where the SVM was performed with the conventional AA composition as input for predicting four types of amine receptors and the GPCRclass predictive accuracy was improved to 89.2%, which was still 9.4% lower than ours. As for dopamine and serotonin receptors, the predictive accuracies were 92.1% and 85.3%, respectively, approximately 8% and 14% lower than our respective results, demonstrating that DWT was effective and helpful for the prediction of GPCRs.

So far as we know, most of the existing algorithms for predicting the GPCRs were based on the AA composition, where the sample of a protein was represented by 20 discrete numbers, with each number representing the occurrence frequency of one of the 20 constituent native amino acids. Obviously, if one uses the conventional AA composition to represent the sample of a protein, all of its sequence order will be lost. This problem can be overcome by DWT. DWT is a useful tool for analyzing the protein sequences from both time and frequency localization that is similar to mathematic microscopy and has the ability of amplification and translation. In other words, DWT analysis can decompose the hydrophobic value sequences into coefficients at different dilations and then remove the noise component from the hydrophobicity profiles, so that it can provide local structures of sequences. With these properties, the current method can more effectively reflect the sequence order effects. Thus, using DWT as a novel feature extraction tool followed by pairwise SVM, the success rate in prediction GPCRs was significantly enhanced.

Table 4

Comparison of our method with other methods for the types of amine receptors.

Amine receptor	Predicted accuracy (%)			
	CDA ^a	Bagging ^b	GPCRclass ^c	Our method
Acetylcholine	67.7	96.8	87.1	100
Adrenoceptor	88.6	90.9	95.5	99.1
Dopamine	81.6	84.2	92.1	100
Serotonin	88.9	81.5	85.3	98.9
Overall accuracy	83.2	87.4	89.8	99.2

^a Results are from Ref. [12].^b Results are from Ref. [18].^c Results are from Ref. [15].

Conclusion

In this work, a novel predictive method has been proposed for the prediction of GPCRs by coupling SVM with DWT. The predictive results demonstrate that WT can reduce dimension of input vector, improve calculating efficiency, and effectively extract important classified information. In comparison with previous literature methods, the predictive performance was significantly enhanced, indicating that the current method is an effective tool for the prediction of GPCRs. The establishment of such a fast and accurate prediction method will speed up the pace of identifying proper GPCRs to facilitate drug discovery.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (20605010, 20865003, and 20805023), the Jiangxi Province Natural Science Foundation (2007JZH2644), and the Opening Foundation of State Key Laboratory of Chem/Biosensing and Chemometrics of Hunan University (2006022).

References

- [1] K.T. Attwood, M.D. Croning, A. Gaulton, Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors, *Protein Eng.* 15 (2002) 7–12.
- [2] F. Horm, J. Weare, M.W. Beukers, S. Horsch, A. Bairoch, W. Chen, Q. Edvardsen, F. Campagne, G. Vriend, GPCRDB: an information system for G protein-coupled receptors, *Nucleic Acids Res.* 26 (1998) 277–281.
- [3] J.M. Baldwin, Structure and function of receptors coupled to G proteins, *Curr. Opin. Cell Biol.* 6 (1994) 180–190.
- [4] D.C. Teller, T. Okada, C.A. Behnke, K. Palczewski, R.E. Stenkamp, Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs), *Biochemistry* 40 (2001) 7761–7772.
- [5] N. Vaidehi, W.B. Floriano, R. Trabaino, S.E. Hall, P. Freddolino, E.J. Choi, G. Zamanakos, W.A. Goddard, Prediction of structure and function of G protein-coupled receptors, *Proc. Natl. Acad. Sci. USA* 99 (2002) 12622–12627.
- [6] S.F. Altschul, T.L. Adden, A.A. Schaffer, Z. Zhang, W. Miller, D.D. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [7] L.P. Miller, P.M. Nadkarni, N.M. Carriero, Parallel computation and FASTA: confronting the problem of parallel database search for a fast sequence comparison algorithm, *Bioinformatics* 7 (1991) 71–78.
- [8] M. Lapinsh, A. Gutcaits, P. Prusis, C. Post, T. Lundstedt, J.E.S. Wikberg, Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences, *Protein Sci.* 11 (2002) 795–805.
- [9] M.I. Sadowski, J.H. Parish, Automated generation and refinement of protein signatures: case study with G-protein coupled receptors, *Bioinformatics* 19 (2003) 727–734.
- [10] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai, M. Suwa, GRIFFIN: a system for predicting GPCR–G-protein coupling selectivity using a support vector machine and a hidden Markov model, *Nucleic Acids Res.* 33 (2005) W148–W151.
- [11] K.C. Chou, D.W. Elrod, Bioinformatical analysis of G-protein-coupled receptors, *J. Proteome Res.* 1 (2002) 429–433.
- [12] D.W. Elrod, K.C. Chou, A study on the correlation of G-protein coupled receptor types with amino acid composition, *Protein Eng.* 15 (2002) 713–715.
- [13] K.C. Chou, Prediction of G-protein-coupled receptor classes, *J. Proteome Res.* 4 (2005) 1413–1418.
- [14] R. Karchin, K. Karplus, D. Haussler, Classifying G-protein coupled receptors with support vector machines, *Bioinformatics* 18 (2002) 147–159.
- [15] M. Bhasin, G.P.S. Raghava, GPCRclass: a web tool for the classification of amino type of G protein-coupled receptors, *Nucleic Acids Res.* 33 (2005) W143–W147.
- [16] B. Qian, O.S. Soyler, R.R. Neubig, R.A. Goldstein, Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs, *FEBS Lett.* 554 (2003) 95–99.
- [17] P.K. Papasaikas, P.G. Bagos, Z.I. Litou, V.J. Promponas, S.J. Hamodrakas, PRED-GPCR: GPCR recognition and family classification server, *Nucleic Acids Res.* 32 (2004) W380–W382.
- [18] Y. Huang, J. Cai, L. Ji, Y. Li, Classifying G-protein coupled receptors with bagging classification tree, *Comput. Biol. Chem.* 28 (2004) 275–280.
- [19] A. Zien, G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, K.R. Muller, Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics* 16 (2000) 799–807.

- [20] Y.D. Cai, X.J. Liu, X. Xu, K.C. Chou, Support vector machine for predicting membrane protein types by incorporating quasi-sequence-order effect, *Internet Electron. J. Mol. Des.* 1 (2002) 219–226.
- [21] J.R. Bock, D.A. Gough, Predicting protein–protein interactions from primary structure, *Bioinformatics* 17 (2001) 455–460.
- [22] S.J. Hua, Z.R. Sun, Support vector machine for protein subcellular localization prediction, *Bioinformatics* 17 (2001) 721–728.
- [23] C.H.Q. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics* 17 (2001) 349–358.
- [24] P. Paclidis, J. Weston, J. Cai, W.S. Noble, Learning gene functional classifications from multiple data types, *J. Comput. Biol.* 9 (2002) 401–411.
- [25] L.P. Miller, P.M. Nadkarni, N.M. Carriero, ~~Parallel computation and FASTA: confronting the problem of parallel database search for a fast sequence comparison algorithm~~, *Bioinformatics* 7 (1991) 71–78.
- [26] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics* 21 (2005) 10–19.
- [27] C. Chen, X.B. Zhou, Y.X. Tian, X.Y. Zou, P.X. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, *Anal. Biol.* 357 (2006) 116–121.
- [28] C. Chen, X.B. Zhou, Y.X. Tian, X.Y. Zou, P.X. Cai, J.Y. Mo, Using pseudo-amino acid composition and support vector machine to predict protein structural class, *J. Theor. Biol.* 243 (2006) 444–448.
- [29] K.C. Chou, Y.D. Cai, Predicting protein quaternary structure by pseudo amino acid composition, *Proteins Struct. Funct. Genet.* 53 (2003) 282–289.
- [30] K.C. Chou, Y.D. Cai, Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition, *J. Cell. Biochem.* 90 (2003) 1250–1260.
- [31] K.C. Chou, Y.D. Cai, Predicting protein structural class by functional domain composition, *Biochem. Biophys. Res. Commun.* 321 (2004) 1007–1009.
- [32] X.Q. Lu, H.D. Liu, Z.H. Xie, Q. Zhang, Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal, *J. Chem. Inform. Comput. Sci.* 44 (2004) 1228–1237.
- [33] A.J. Mandell, K.A. Selz, M.F. Shlesinger, Mode homologies and their locations in the hydrophobic free energy sequences of peptide ligands and their receptor eigenfunctions, *Proc. Natl. Acad. Sci. USA* 94 (1997) 13576–13581.
- [34] A.J. Mandell, K.A. Selz, M.F. Shlesinger, Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families, *Phys. A* 244 (1997) 254–262.
- [35] A.J. Mandell, M.J. Owens, K.A. Selz, W.N. Morgan, M.F. Shlesinger, Mode matches in hydrophobic free energy eigenfunctions predict (neurotensin, cholecystokinin) peptide–protein interactions, *Biopolymers* 46 (1998) 89–101.
- [36] K.A. Selz, A.J. Mandell, M.F. Shlesinger, Hydrophobic free energy eigenfunctions of pore, channel, and transporter proteins contains β -burst patterns, *Biophys. J.* 75 (1998) 2332–2342.
- [37] K.A. Selz, A.J. Mandell, M.F. Shlesinger, V. Arguragi, M.J. Owens, Designing human m1 muscarinic receptor-targeted hydrophobic eigenmode matched peptides as functional modulators, *Biophys. J.* 86 (2004) 1308–1332.
- [38] P. Lio, Wavelets in bioinformatics and computational biology: state of art and perspectives, *Bioinformatics* 19 (2003) 2–9.
- [39] K.B. Li, P. Issac, A. Krishnan, Predicting allergenic proteins using wavelet transform, *Bioinformatics* 20 (2004) 2572–2578.
- [40] R. Gupta, A. Mittal, K. Singh, A novel and efficient technique for identification and classification of GPCRs, *IEEE Trans. Inform. Technol. Biomed.* 12 (2008) 541–548.
- [41] K.C. Chou, D.W. Elrod, ~~Bioinformatical analysis of G-protein-coupled receptors~~, *J. Proteome Res.* 1 (2002) 429–433.
- [42] P.K. Strope, E.N. Moriyama, Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors, *Genomics* 89 (2007) 602–612.
- [43] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 11 (1989) 674–693.
- [44] K.C. Chou, Low-frequency collective motion in biomacromolecules and its biological functions, *Biophys. Chem.* 30 (1988) 3–48.
- [45] K.C. Chou, Low-frequency vibration of DNA molecules, *Biochem. J.* 221 (1984) 27–31.
- [46] K.C. Chou, Low-frequency motions in protein molecules: β -Sheet and β -barrel, *Biophys. J.* 48 (1985) 289–297.
- [47] A. Kandaswamy, C.S. Kumar, R.P. Ramanathan, S. Jayaraman, N. Malmurugan, Neural classification of lung sounds using wavelet coefficients, *Comput. Biol. Med.* 34 (2004) 523–537.
- [48] N. Zavaljevski, F.J. Stevens, J. Reifman, Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions, *Bioinformatics* 18 (2002) 689–696.
- [49] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Data Min. Knowl. Discov.* 2 (1998) 121–167.
- [50] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.
- [51] C.C. Chang, C.J. Lin, LIBSVM: a library for support machines [software], 2001, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [52] V. Vapnik, *Statistical Learning Theory*, John Wiley, New York, 1998.
- [53] K.C. Chou, C.T. Zhang, Prediction of protein structural classes, *Crit. Rev. Biochem. Mol. Biol.* 30 (1995) 275–349.
- [54] K.C. Chou, H.B. Shen, Cell-PLOC: a package of web-servers for predicting subcellular localization of proteins in various organisms, *Nat. Protoc.* 3 (2008) 153–162.
- [55] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, *Anal. Biochem.* 370 (2007) 1–16.
- [56] B. Rost, C. Sandwe, Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.* 232 (1993) 584–599.

- [57] K.J. Park, M. Kanehisa, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics* 19 (2003) 1656–1663.
- [58] D. Sarda, G.H. Chua, K.B. Li, A. Krishnan, PSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties, *BMC Bioinform.* 6 (2005) 152.
- [59] G.D. Rose, Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature* 272 (1978) 586–590.
- [60] V.I. Lim, Structural principles of the globular organization of protein chains: a stereochemical theory of globular protein secondary structure, *J. Mol. Biol.* 88 (1974) 857–862.
- [61] V.I. Lim, Algorithms for prediction of α -helical, β -structural regions in globular proteins, *J. Mol. Biol.* 88 (1974) 873–894.
- [62] P.Y. Chou, G.D. Fasman, Conformational parameters for amino acids in helical, β -sheet, and random coil regions calculated from proteins, *Biochemistry* 13 (1974) 211–222.
- [63] P.Y. Chou, G.D. Fasman, Prediction of protein conformation, *Biochemistry* 13 (1974) 222–245.
- [64] B.M. Bull, K. Breese, Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues, *Arch. Biochem. Biophys.* 161 (1974) 665–670.
- [65] J.L. Fauchereand, V. Pliska, Hydrophobic parameters of amino-acid side chains from the partitioning of *N*-acetyl-amino-acid amides, *Eur. J. Med. Chem.* 18 (1983) 369–375.
- [66] J.M.R. Parker, D. Guo, R.S. Hodges, New hydrophobicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites, *Biochemistry* 27 (1986) 5425–5432.
- [67] P.K. Ponnuswamy, M. Prabhakaran, P. Manavalan, Hydrophobic packing, spatial arrangement of amino acid residues in globular proteins, *Biochim. Biophys. Acta* 623 (1980) 301–316.
- [68] R.M. Sweet, D. Eisenberg, Correlation of sequence hydrophobicities measure similarity in three dimensional protein structure, *J. Mol. Biol.* 171 (1983) 479–488.
- [69] D.H. Wertz, H. Scheraga, Influence of water on protein structure: an analysis of the preferences of amino acids residues for the inside or outside and for specific conformations in a protein molecule, *Macromolecules* 11 (1978) 9–15.
- [70] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* 157 (1982) 105–132.
- [71] T.P. Hopp, K.R. Woods, Predicting of protein antigenic determinants from amino acid sequences, *Proc. Natl Acad. Sci. USA* 78 (1981) 3824–3828.