Journal of Molecular Graphics and Modelling 30 (2011) 129–134

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

OligoPred: A web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition

Jian-Ding Qiu[a,b,*], Sheng-Bao Suo[a], Xing-Yu Sun[a], Shao-Ping Shi[a], Ru-Ping Liang[a]

[a] Department of Chemistry, Nanchang University, Nanchang 330031, China
[b] Department of Chemical Engineering, Pingxiang College, Pingxiang 337055, China

ARTICLE INFO

Article history:
Received 6 April 2011
Received in revised form 18 June 2011
Accepted 30 June 2011
Available online 7 July 2011

Keywords:
Homo-oligomers
Discrete wavelet transform
Support vector machine
Jackknife test
Classification

ABSTRACT

In vivo, some proteins exist as monomers (single polypeptide chains) and others as oligomers. Not like monomers, oligomers are composed of two or more chains (subunits) that are associated with each other through non-covalent interactions and, occasionally, through disulfide bonds. These proteins are the structural components of various biological functions, including cooperative effects, allosteric mechanisms and ion-channel gating. However, with the dramatic increase in the number of protein sequences submitted to the public data bank, it is important for both basic research and drug discovery research to acquire the possible knowledge about homo-oligomeric attributes of their interested proteins in a timely manner. In this paper, a high-throughput method, combined support vector machines with discrete wavelet transform, has been developed to predict the protein homo-oligomers. The total accuracy obtained by the re-substitution test, jackknife test and independent dataset test are 99.94%, 96.17% and 96.18%, respectively, showing that the proposed method of extracting feature from the protein sequences is effective and feasible for predicting homo-oligomers. The online service is available at http://bioinfo.ncu.edu.cn/Services.aspx.

1. Introduction

The actions of proteins are at the core of biological processes, and their function can only be understood based on the structure of their constituent polypeptide chains. The structure hierarchy of proteins is defined in terms of four levels: primary, secondary, tertiary, and quaternary. The quaternary structure was first introduced by Bernal in 1958 [1,2], which is the interaction of non-covalently bound monomeric protein subunits to form oligomers. In vivo, many proteins exist as oligomers with various different quaternary structural attributes rather than as single individual chains. For example, hemoglobin is a heterotetramer of two α chains and two β chains, and the four chains must be aggregated into one construct to perform their cooperative function during the oxygen-transporting process [3]. Phospholamban is formed by a homopentamer [4,5], while the γ-aminobutyric acid type A ($GABA_A$) receptor [6,7] and the α7 nicotinic acetylcholine receptor [8] are formed by heteropentamers. According to the number of subunits aggregated together in an oligomeric complex, protein quaternary structures can be further classified into: monomer, dimer, trimer, tetramer, and so forth. In addition, from an evolutional point of view, the oligomeric proteins have more advantages than the monomers [9,10]. For example, they are easier for multi-subunit proteins to repair their defects by simply replacing the flawed subunit [11]. Besides, they contribute significantly to evolutionary stability in that changes in the quaternary structure can occur through each individual chain or through their reorientation relative to each other [9,10,12]. Given a polypeptide chain, it is not clear whether it will form a dimer, trimer or any other oligomer. To the best of our knowledge, no report in literature has systematically addressed this question.

Although the protein spatial structure can be determined by various experiments, it is both time consuming and costly to acquire this kind of knowledge through experimentation [13]. What's more, the number of newly found protein sequences has increased explosively in the post-genomic era. For instance, the Swiss-Prot databank contained only 3939 protein sequence entries in 1986, but the number has jumped to 525,997 in March 2011, according to the UniProtKB/Swiss-Prot at http://www.expasy.org/sprot/relnotes/relstat.html, meaning that the current number of protein sequence entries is more than 133 times the number of entries approximately two decades ago. Therefore, facing the tremendous challenge, many statistical and sequence-based tools

* Corresponding author at: Department of Chemistry, Nanchang University, Nanchang 330031, China. Tel.: +86 791 3969518.
E-mail address: jdqiu@ncu.edu.cn (J.-D. Qiu).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.06.014

have been developed for detecting homo-oligomers. These tools have made certain progress, but most methods were mainly based on different amino acid compositions to extract features. For example, in 2001, Garian [14] predicted primarily homodimer and non-homodimer using decision-tree models and a feature extraction method (simple binning function), and found that protein sequences contain the quaternary structure information. Zhang et al. [15] used the support vector machine (SVM) and the covariant discriminant algorithms to predict homo-oligomeric proteins from the protein primary sequences, the results showed that the SVM method was superior to both the covariant discriminant algorithm and the method based on the decision tree. Moreover, several methods based on different amino acid composition (AAC) have been developed and applied successfully for predicting protein quaternary structures as well [16–19]. Although, the results could be improved by using different amino acid composition to extract feature, all of the order information of sequences were lost by using the AAC model.

To avoid losing important information hidden in protein sequences, the pseudo amino acid composition (PseAAC) was proposed [20,21] to replace the simple AAC for representing the sample of a protein. For a concise introduction about Chou's PseAAC, click the link at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition to see a Wikipedia article about PseAAC. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems [22–31]. For a summary about its development and applications, see a recent comprehensive review [32]. Similarly, Zhang et al. [33–35] and Xiao and Wei [36] also used Chou's PseAAC to predict protein quaternary structures. In these methods, the most total accuracy was 85.7% for predicting homo-oligomers [18], but the greatest prediction accuracy was about 68% for homo-hexamers, only 34.68%, 59.71% and 51.11% for homo-dimers, homo-trimers and homo-tetramers, respectively [33]. Therefore, developing an effective method to predict the homo-oligomers attributes of proteins based on their sequence information can not only save much time, but can also be very helpful for the design of drugs in treating certain diseases related with quaternary structure attributes defects. Hence it has become a crucial issue to develop some reliable computational methods for identifying protein quaternary structures. In this study, a simple and high-throughput information extraction method that couples discrete wavelet transform (DWT) with classifier algorithms based on the amino acid (AA) hydrophobic index was developed for the multi-class prediction of oligomeric proteins. This method consisted of three main steps. First, the protein sequences were transformed into numerical signals using the physicochemical properties of amino acids. Then these numerical sequences were further processed by DWT to extract salient frequency-band features from signals. Following this step, using the statistical method, a series of statistical feature vectors were constructed to represent the protein sequences. Finally, different classifier algorithms were applied to deal with the classified problem of protein quaternary structures using these statistics feature vectors as input. The influences of wavelet functions, classifier algorithms, sequence identity and the size of dataset on the results were discussed. The predictive results of the jackknife cross-validation test show significant improvements compared to results obtained with earlier methods.

According to a recent comprehensive review [37], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

2. Materials and methods

2.1. Datasets

To investigate the feasibility of our method, three datasets of proteins as benchmarks have been used. The first dataset, as the training dataset, was originally constructed by Garian [14]. It contains of 1568 highly sequence identity protein sequences, 914 of which are homo-dimers (2EM), 139 homo-trimers (3EM), 407 homo-tetramers (4EM) and 108 homo-hexamers (6EM), it was denoted as dataset R_{1568}. To avoid homology bias and remove the redundant sequences from the benchmark dataset, a cutoff threshold of 25% was recommended [38–41] to exclude those proteins from the benchmark datasets that have equal to or greater than 25% sequence identity to any other. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance. Meanwhile, in order to investigate the influence of the dataset size, we randomly extracted four small datasets from each class (2EM, 3EM, 4EM and 6EM) of the dataset R_{1568} to construct dataset R_{400}, R_{320}, R_{280}, and R_{160}, which consisted of 400, 320, 280 and 160 homo-oligomeric protein sequences, respectively. The second dataset, R_{1283}, as an independent testing dataset, was constructed by Bairoch and Apweiler [42]. It contains 1283 protein sequences, 759 of which are homo-dimers (2EM), 105 homo-trimers (3EM), 327 homo-tetramers (4EM) and 92 homo-hexamers (6EM), it was denoted as dataset R_{1283} [42]. The R_{1568} and R_{1283} databases limited to the prokaryotic, cytosolic subset of homo-oligomers in order to eliminate membrane proteins and other specialized proteins. To investigate the impact of sequence identity on estimation of the classification accuracy, we selected another subset R_{2581} dataset as our third dataset, which was established by Chou and coworkers [43] and the pairwise average sequence identity was less than 15%.

2.2. Methods

A protein sequence can be represented as a series of amino acids (AAs) 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:

$$R_1R_2R_3R_4R_5R_6R_7R_8 \ldots R_L \quad (1)$$

Suppose $H(R_1)$ is the physicochemical feature value of the first residue R_1, $H(R_2)$ is the physicochemical feature values 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. DWT analysis can decompose digit signal into coefficients at different dilations and then remove the noise component from the profiles, so it can give us local structures of sequences that more effectively reflect the sequence-order effects. DWT operates two sets of function, which can be viewed as low- frequency and high-frequency filters. The high-frequency components are more noisy and hence only the low-frequency components are more important [44]. This is just like the case of protein internal motions where the low-frequency components are functionally more important [45]. The original modulated signal $f(x)$ [46] is passed through low-pass and high-pass filters, and approximation $A^j(n)$, and detail $D^j(n)$ coefficients signals are obtained. The approximation coefficient represents the high-scale and low-frequency components of the signal, and the detail coefficient represents the low-scale and high-frequency components of the signal. The approximation coef-

ficients and detail coefficients of the DWT for the digital signal at level j can be expressed as:

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

$$D^j(n) = \sum_{k \in z} h_{k-2n} D^{j-1}(k) \quad (3)$$

In this study, to further decrease the dimensionality of the extracted feature vectors, statistics were used over the set of the wavelet coefficients. The following statistical features calculated from the approximation coefficients and the detail coefficients were used for the classification of protein quaternary structures: (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 paper, the decomposition level $j = 4$ [46] was chosen to classify the protein quaternary structures, and the obtained 20 dimension feature vectors with Kyte–Doolittle hydrophobicity [47] scales were then fed to classifiers for classification.

In general, the problem of classifying protein types can be formulated as follows. Consider a dataset containing X proteins (P_1, $P_2, \ldots, P_X$) that have been classified into M subsets,

$$S = S_1 \cup S_2 \cup S_3 \cup \cdots \cup S_M \quad (4)$$

where each subset Sm ($m = 1, 2, \ldots, M$) is composed of proteins with the same protein quaternary structure and its size (the number of proteins therein) is Nm. We then have $X = N_1 + N_2 + \cdots + N_M$. For a query protein P, how can we identify the subset to which it belongs? Many different prediction algorithms have been developed to address this problem. Here we focused on the K nearest neighbor (KNN) [48] with Euclidean distance, Bayes [49], decision trees [50,51] and support vector machines (SVM) [52]. In this paper we discussed mainly the pattern classification principle of SVM. Others classifier's detailed classification principle please refer to the literature sections [48–52].

The SVM introduced by Vapnik and coworkers [52] has proven to be a useful learning machine, especially for classification. The basic idea of applying SVM to pattern classification can be stated briefly as follows [16]. To classify the two classes of samples, SVM maps the input vectors into a higher dimension feature space. Then, within this feature space, construct a hyperplane which maximizes the distance of the closest vectors belonging to the two classes to the hyperplane. The mapping function will involve only the relatively low-dimensional vectors in the input space and dot products in the feature space. These dot products are represented by kernel functions. Thus the 'curse of dimensionality' can be avoided in this condition. SVM training always seeks a globally optimized solution and avoids over-fitting, so it is of the ability to deal with a large number of features. For actual implementation, we used the LIBSVM package (version 2.81) [53]. To obtain an SVM classifier with optimal performance, radial basis function (RBF) was 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.

3. Results and discussion

3.1. Effect of wavelet functions

Based on different basis functions, the wavelets have different families; every family has its unique quality fitting for certain signals. As the characteristics of the analyzing wavelet influence the performance of DWT, the better the analyzing wavelet matches the underlying structure in the signal, the better feature values can be extracted from the sequences [54]. Therefore, selection of a suitable wavelet basis which possesses desirable properties such as compactly support, orthogonality, symmetry, smoothness and high order of vanishing moments is necessary for the signal processing [55]. It is well known that for wavelet function selections, there are many conflict conditions that restrict the selection of a wavelet function. It is hoped that the wavelet functions would hold properties such as orthogonality, symmetry, high order of vanishing moments and smoothness. However, none of wavelet basis has simultaneously these desirable properties. Therefore, we investigated the effect of the different wavelets on classification accuracy, nine wavelet functions: Meyer, Daubechies of number 1 (Db1) and number 10 (Db10), Biorthogonal of number 2.4 (Bior2.4) and number 6.8 (Bior6.8), Symlets of number 1 (Sym1) and number 8 (Sym8), Coifman of number 1 (Coif1) and number 5 (Coif5) were chosen for test in the research. The predictive results of database R_{1568} performed by different wavelet functions with Kyte–Doolittle hydrophobicity scales were listed in Fig. 1. As can be seen from Fig. 1, in jackknife test by using SVM, the training accuracy can reach 96.17% when using Bior2.4 wavelet to extract feature. However, when using other wavelet functions, the training accuracy only ranged from 84.82% to 92.66%. This is because wavelet of Bior2.4 is a class of symmetric and biorthogonal wavelets that have good ability to reduce dimension by Mallat decomposition [56] and can remove high-frequency noises by selecting several or even one approximation sub-band. Therefore, Bior2.4 wavelet was selected as the appropriate wavelet function to input SVM to predict homo-oligomeric proteins in this study.

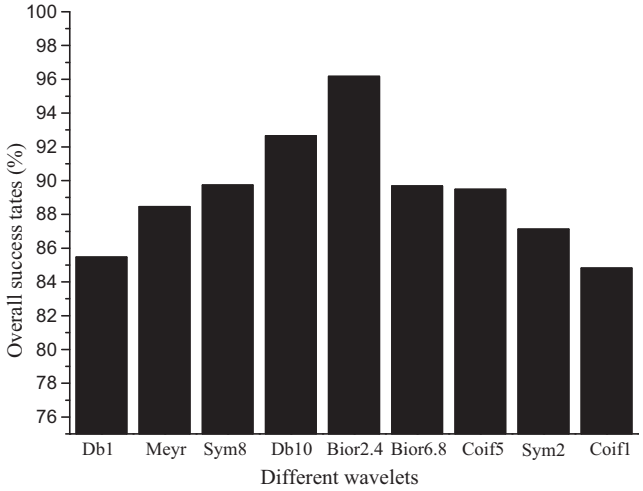

Fig. 1. The performance of different wavelet functions by using SVM and Kyte–Doolittle hydrophobicity scales on the R_{1568} database.

3.2. Comparison of different classifier algorithms

In order to further optimize classifier algorithms, the performance of the following four classifier algorithms were compared: KNN, Bayes, Decision Trees and SVM. These four classifier algorithms are state-of-the-art methods within the field of protein prediction. The detailed setup procedures of these methods are described in the literature section [48–52]. Table 1 summarized the performance of the various methods. It can be seen from Table 1 that the accuracy by SVM was 96.17%, which was 7.71, 26.21 and 11.16 percentile higher than those of KNN, Bayes and Decision Tree, respectively. It is obvious that SVM outperforms the other methods for predicting protein. Similarly, the predictive success

Table 1
The performance of different classifiers by using Bior2.4 wavelet and Kyte–Doolittle hydrophobicity scales on the R_{1568} database.

Classifiers	Overall success rate for each class (%)				
	2EM	3EM	4EM	6EM	Overall
SVM	99.12	92.09	92.87	88.89	96.17
KNN	98.14	67.63	83.29	52.78	88.46
Bayes	62.14	84.89	83.04	67.59	69.96
Decision tree	99.89	58.27	70.02	50.00	85.01

rates of 3EM, 4EM and 6EM classes were remarkably enhanced and reached to 92.09%, 92.87% and 88.89% (shown in Table 1), respectively, about 10–30% higher than other algorithms. As is well-known, SVM has been widely applied in the study about the domain of biological information [43] due to its well-founded statistical learning theory and attractive features including effective avoidance of over-fitting, ability of handle large feature space and absence of local minima. Therefore, SVM was selected as the appropriate classifier algorithm and combined with Bior2.4 wavelet function and Kyte–Doolittle hydrophobicity scales to construct the DWT_SVM model for the prediction of homo-oligomeric proteins in this study.

3.3. Effect of sequence identity in datasets

Sequence identity is one of the main factors that significantly impact prediction accuracy. Sequence identity is defined as the percentage of AAs in the protein sequence that is identical after aligning the sequence with other sequence from a given dataset (gaps between consecutive AA may be introduced during alignment, if necessary). Just as an example about sequence identity, one of the most often used dataset R_{1568}, is highly sequence identity (<80%) and thus the corresponding accuracy of the DWT_SVM model showed over 96% (shown in Table 1). However, for low sequence identity (<15%) of dataset R_{2581}, the overall predictive accuracy achieved by DWT_SVM model for the jackknife test still reached to 90.93%. Although sequence identity is known to impact the prediction accuracy, no standards are imposed when it comes to performing tests [57]. However, when we compared with other published results based on the same R_{2581} dataset, our result was 10.9, 4.9 and 3.2 percentile higher than those obtained by covariant discriminant (CD) [11], SVM [35] and *K*-nearest neighbor (KNN) classifiers [36] (shown in Table 2). According to the above discussion, we can draw that our method not only enhances significantly the accuracy of prediction for the datasets with high sequence identity, but also possesses obvious and effective character in the aspect of resistant sequences identity.

Table 2
Comparison of different methods by the jackknife test with using Bior2.4 wavelet and Kyte–Doolittle hydrophobicity scales on the R_{2581} dataset.

Method	Input	Success rate (%)				
		2EM	3EM	4EM	6EM	Overall
CD [11]	PseAA	85.7	72.5	85.4	62.7	80.0
SVM [35]	PseAA	93.4	79.4	90.8	63.8	86.0
KNN [36]	PseAA	85.3	86.5	93.7	78.4	87.7
DWT_SVM (ours)	KDH	96.1	92.2	84.1	94.8	90.9

3.4. Effect of the size of datasets

In order to investigate the influence of the dataset size for the DWT_SVM model, we randomly extracted four small datasets from each class (2EM, 3EM, 4EM and 6EM) of the dataset R_{1568} to form datasets R_{400}, R_{320}, R_{240} and R_{160}, which consisted of 400, 320, 280 and 160 homo-oligomeric protein sequences, respectively. With the datasets decrease, the variation of accuracy was only about 3% in the self-consistency test, while the performance of jackknife test decreased obviously from 85.25%to 66.88% (given in Table 3). Obviously, jackknife test is more objective to reflect the power of a prediction model. For above surprisingly results, Chou and Maggiora [58] considered this was because the algorithm needs more training data to make its prediction mechanism work properly; with the training data decreasing, the small datasets not only lose a host of sequence information but also make its prediction mechanism work improperly. So it would have a greater impact on the predicted results in jackknife test. In conclusion, jackknife test is more effective and objective than re-substitution test for reflecting the power of a prediction model and the algorithms need more training data to make the prediction mechanism work properly.

3.5. Comparison with other methods

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [59]. However, as elucidated in [60] and demonstrated by Eqs. (28)–(32) of [37], among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors [61–70]. The success rates by re-substitution test and jackknife test for the dataset R_{1568} were 99.94% and 96.17%, respectively. It can be seen from Table 4 that the overall success rate by jackknife test on the dataset R_{1568} was 96.17%, which was 10.46, 11.03, 14.73, 18.81, 32.78, and 52.68 percentile higher than those of nearest neighbors algorithm (NNA) with Subsequence length (SL) [18], SVM with subsequence length [17], SVM with PLIV (pseudo-amino acid composition was extracted based on amino acid residue index of Pliska) and COMP (this set was composed of amino acid compositions) [33], covariant discriminant (CD) [71] with PLIV and COMP, respectively. Similarly, the predictive success rates of 3EM, 4EM and 6EM classes were remarkably enhanced and reach up to 92.09%, 92.87% and 88.89%, respectively, about 20–40% higher than others. The results demonstrate that DWT_SVM possesses a stronger predictive capacity than other existing models.

Table 3
Comparison of prediction results for different size datasets by using DWT_SVM model.

Protein type	Re-substitution test (%)				Jackknife test (%)			
	R_{400}	R_{320}	R_{240}	R_{160}	R_{400}	R_{320}	R_{240}	R_{160}
2EM	100	96.25	95.00	100	75.00	70.00	56.67	52.50
3EM	95.00	100	100	100	88.00	87.50	80.00	75.00
4EM	94.00	97.50	100	100	85.00	81.25	53.33	77.50
6EM	100	98.75	100	100	93.00	77.50	95.00	62.50
Overall (%)	97.25	98.13	98.75	100	85.25	79.06	72.92	66.88

Table 4
Comparison of different methods by the jackknife test on the dataset R_{1568}.

Classifiers	Input	Success rate for each class (%)				Overall (%)
		2EM	3EM	4EM	6EM	
SVM [33]	COMP	89.93	57.55	64.13	46.30	77.36
	PLIV	93.22	62.59	68.55	54.63	81.44
CD [71]	COMP	34.68	59.71	51.11	68.25	43.49
	PLIV	59.24	59.71	77.15	47.22	63.39
SVM [17]	SL (l = 4)	97.05	63.31	72.24	61.11	85.14
NNA [18]	SL (l = 4)	97.16	64.75	73.46	62.04	85.71
DWT_SVM (ours)	KDH	99.12	92.09	92.87	88.89	96.17

3.6. Predictive performance of independent test

Moreover, as a demonstration for practical applications, the DWT_SVM model was also used for predicting independent proteins (dataset R_{1283}), based on the rule parameters derived from the training dataset. The overall accuracy of the independent dataset test by this new algorithm was 96.18%, which was higher about 15% than that of Naive Bayes feature fusion algorithm [34]. The predictive success rates of 3EM, 4EM and 6EM classes were increased to 87.62%, 93.88% and 82.96%, respectively, about 20–30% higher than the Naive Bayes feature fusion algorithm of Zhang et al. [34]. The experiment results show that DWT_SVM approach is convenient to extract valuable information from protein sequences, which may be a useful tool in other assignment problems in proteomics and genome research. Moreover, this new approach provides a superior prediction performance with a relatively simple formalism that is more easily generalized to large databases.

Why could the success rate be improved so much by introducing the DWT? SVM is a kind of learning machine based on statistical leaning theory and has many attractive features, including effective avoidance of over-fitting, the ability to handle large feature space and absence of local minima. However, as a machine learning technique, SVM requires a fixed length of pattern, and it is not possible to use this technique in case of proteins with depths that are too small or too large length [72]. These problems can be overcome by DWT. DWT is a useful tool for analyzing the protein sequences from both time and frequency localization, which 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 it can give us local structures of sequences [73]. With these properties, the current method can more effectively reflect the sequence-order effects. Thus, using DWT as a novel feature extraction tool, the success rate in prediction protein structure has been significantly enhanced.

3.7. Web server

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors [74], here we have provided the web-server for the prediction method presented in this paper. Through the above optimization, DWT_SVM model fused SVM and Bior2.4 wavelet based on Kyte–Doolittle hydrophobicity scale was constructed and was represented via web server of system OligoPred, the parameter of which is completely the same as DWT_SVM model. The web server available at http://bioinfo.ncu.edu.cn/Services.aspx also provides user-friendly input and output interfaces. Several proteins in FASTA format could be inputted to the system. And the system OligoPred presents in a diagram that includes the information of feature extraction and the classification error rate. Users can download the predicted results in tab-delimited format for further analysis, and the programs can also be downloaded from the proposed web site.

4. Conclusions

In this work, a novel predictive method (DWT_SVM) which is far more easily extracted sequence features and more effectively to deal with the problems with many long and complicated sequences has been proposed for the prediction of homo-oligomeric proteins. The predictive results demonstrate that DWT can reduce dimension of input vector, improve calculating efficiency, and effectively extract important classified information. In comparison with previous literatures, the predictive performance has been significantly enhanced. The results indicate that the DWT_SVM approach is an effective tool for the prediction of homo-oligomeric proteins and might also become a useful high-throughput tool in characterizing other attributes of proteins, such as enzyme class, membrane protein type, and nuclear receptor subfamily according to their sequences.

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] I.M. Klotz, D.W. Darnall, N.R. Langerman, Quaternary structure of proteins , in: H. Neurath, R.L. Hill (Eds.), The Proteins, vol. 1, Academic Press, New York, 1975, pp. 293–411.
[2] H. Sund, K. Weber, The quaternary structure of proteins , Angew. Chem. Int. Ed. 5 (1966) 231–245.
[3] M.F. Perutz, The hemoglobin molecule , Sci. Am. 211 (1964) 65–76.
[4] K. Oxenoid, J.J. Chou, The structure of phospholamban pentamer reveals a channel-like architecture in membranes , Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 10870–10875.
[5] K. Oxenoid, A.J. Rice, J.J. Chou, Comparing the structure and dynamics of phospholamban pentamer in its unphosphorylated and pseudo-phosphorylated states , Protein Sci. 16 (2007) 1977–1983.
[6] K.C. Chou, Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5 , Biochem. Biophys. Res. Commun. 316 (2004) 636–642.
[7] V. Tretter, N. Ehya, K. Fuchs, W. Sieghart, Stoichiometry and assembly of a recombinant GABAA receptor subtype , J. Neurosci. 17 (1997) 2728–2737.
[8] K.C. Chou, Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor , Biochem. Biophys. Res. Commun. 319 (2004) 433–438.
[9] N.C. Price, Assembly of multi-subunit structure , in: R.H. Pain (Ed.), Mechanisms of Protein Folding, Oxford University Press, New York, 1994, pp. 160–193.
[10] I.M. Klotz, N.R. Langerman, D.W. Darnall, Quaternary structure of proteins , Annu. Rev. Biochem. 39 (1970) 25–62.
[11] K.C. Chou, Y.D. Cai, Predicting protein quaternary structure by pseudo amino acid composition , Proteins 53 (2003) 282–289.
[12] E. Einstein, H.K. Schachman, Determining the roles of subunits in protein function , in: T.E. Creighton (Ed.), Protein Function: A Practical Approach, IRL, London, 1989, pp. 135–176.
[13] X. Xiao, P. Wang, K.C. Chou, Predicting the quaternary structure attribute of a protein by hybridizing functional domain composition and pseudo amino acid composition , J. Appl. Crystallogr. 42 (2009) 169–173.
[14] R. Garian, Prediction of quaternary structure from primary structure , Bioinformatics 17 (2001) 551–556.

[15] S.W. Zhang, Q. Pan, H.C. Zhang, Y.L. Zhang, H.Y. Wang, Classification of protein quaternary structure with support vector machine , Bioinformatics 19 (2003) 2390–2396.
[16] J. Song, H.W. Tang, Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy , J. Chem. Inf. Model 44 (2004) 1324–1327.
[17] J. Song, H.W. Tang, Support vector machines for classification of homo-oligomeric proteins by incorporating subsequence distributions , J. Mol. Struct-Theochem. 722 (2005) 97–101.
[18] H. Song, Prediction of homo-oligomeric proteins based on nearest neighbour algorithm , Comput. Biol. Med. 37 (2007) 1759–1764.
[19] A. Carugo, structural proteomics filter: prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences , J. Appl. Crystallogr. 40 (2007) 986–989.
[20] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition , Proteins 43 (2001) 246–255.
[21] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes , Bioinformatics 21 (2005) 10–19.
[22] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses , J. Theor. Biol. 263 (2010) 203–209.
[23] Q. Gu, Y.S. Ding, T.L. Zhang, Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns , Protein Peptide Lett. 17 (2010) 559–567.
[24] X. Xiao, P. Wang, K.C. Chou, GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions , Mol. Biosyst. 7 (2011) 911–919.
[25] F.M. Li, Q.Z. Li, Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach , Protein Peptide Lett. 15 (2008) 612–616.
[26] H. Lin, The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition , J. Theor. Biol. 252 (2008) 350–356.
[27] J.D. Qiu, J.H. Huang, S.P. Shi, R.P. Liang, Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform , Protein Peptide Lett. 17 (2010) 715–722.
[28] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction , Comput. Biol. Chem. 34 (2010) 320–327.
[29] L. Yu, Y. Guo, Y. Li, G. Li, M. Li, et al., SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition , J. Theor. Biol. 267 (2010) 1–6.
[30] D. Zou, Z. He, J. He, Y. Xia, Supersecondary structure prediction using Chou's pseudo amino acid composition , J. Comput. Chem. 32 (2011) 271–278.
[31] X. Xiao, P. Wang, K.C. Chou, Quat-2L: a web-server for predicting protein quaternary structural attributes , Mol. Divers. 15 (2011) 149–155.
[32] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology , Curr. Proteomics 6 (2009) 262–274.
[33] S.W. Zhang, Q. Pan, H.C. Zhang, Y.H. Wu, J.Y. Shi, Support vector machines for predicting protein homo-oligomers by incorporating pseudo-amino acid composition , Int. Electron. J. Mol. Des. 2 (2003) 392–402.
[34] S.W. Zhang, Q. Pan, H.C. Zhang, Z.C. Shao, J.Y. Shi, Prediction of protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and Naive Bayes Feature Fusion , Amino Acids 30 (2006) 461–468.
[35] S.W. Zhang, W. Chen, F. Yang, Q. Pan, Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach , Amino Acids 35 (2008) 591–598.
[36] X. Xiao, Z.L. Wei, Application of protein grey incidence degree measure to predict protein quaternary structural types , Amino Acids 37 (2009) 741–749.
[37] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review) , J. Theor. Biol. 273 (2011) 236–247.
[38] K.C. Chou, H.B. Shen, Review: recent progresses in protein subcellular location prediction , Anal. Biochem. 370 (2007) 1–16.
[39] K.C. Chou, H.B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0 , PLoS ONE 5 (2010) e9931.
[40] K.C. Chou, H.B. Shen, Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization , PLoS ONE 5 (2010) e11335.
[41] K.C. Chou, Z.C. Wu, X. Xiao, iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins , PLoS ONE 6 (2011) e18258.
[42] Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its new supplement TrEMBL , Nucleic Acids Res. 24 (1996) 21–25.
[43] Y.D. Cai, G.P. Zhou, K.C. Chou, Support vector machines for predicting membrane protein types by using functional domain composition , Biophys. J. 84 (2003) 3257–3263.
[44] K.C. Chou, The biological functions of low-frequency phonons: III. Helical structures and microenvironment , Biophys. J. 45 (1984) 881–890.
[45] K.C. Chou, Low-frequency motions in protein molecules: beta-sheet and beta-barrel , Biophys. J. 48 (1985) 289–297.
[46] J.D. Qiu, S.H. Luo, J.H. Huang, R.P. Liang, Using support vector machines for prediction of protein structural classes based on discrete wavelet transform , J. Comput. Chem. 30 (2009) 1344–1350.
[47] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic charactor of a protein , J. Mol. Biol. 157 (1982) 105–132.
[48] B.V. Dasarathy, Nearest neighbor (NN) Norms: NN Pattern Classification Techniques , McGraw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, CA, 1991.
[49] M. James, Classification Algorithms , Collins, London, UK, 1985.
[50] J.R. Quinlan, Induction of decision trees , Mach. Learn. 1 (1986) 81–106.
[51] J.R. Quinlan, C4.5 Programs for Machine Learning , Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[52] C. Cortes, V. Vapnik, Support-vector networks , Mach. Learn. 20 (1995) 273–297.
[53] C.C. Chang, C.J. Lin, LIBSVM: A Library For Support Machines [Software], 2001. www.csie.ntu.edu.tw/∼cjlin/libsvm.
[54] S.M. Ahmeda, M. Abo-Zahhad, A new hybrid algorithm for ECG signal compression based on the wavelet transformation of the linearly predicted error , Med. Eng. Phys. 23 (2001) 117–126.
[55] D.F. Li, G.C. Wu, Construction of a class of Daubechies type wavelet bases , Chaos Soliton Fract. 42 (2009) 620–625.
[56] L. Holm, C. Sander, Mapping the protein universe , Science 273 (1996) 595–602.
[57] L.A. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains – impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy , Pattern Recogn. 39 (2006) 2323–2343.
[58] K.C. Chou, G.M. Maggiora, Domain structural class prediction , Protein Eng. 11 (1998) 523–538.
[59] K.C. Chou, C.T. Zhang, Prediction of protein structural classes , Crit. Rev. Biochem. Mol. 30 (1995) 275–349.
[60] 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.
[61] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition , J. Theor. Biol. 271 (2011) 10–17.
[62] Y.H. Zeng, Y.Z. Guo, R.Q. Xiao, L. Yang, L.Z. Yu, et al., Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach , J. Theor. Biol. 259 (2009) 366–372.
[63] M. Masso, I.I. Vaisman, Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms , J. Theor. Biol. 266 (2010) 560–568.
[64] K.K. Kandaswamy, K.C. Chou, T. Martinetz, S. Moller, P.N. Suganthan, et al., AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties , J. Theor. Biol. 270 (2011) 56–62.
[65] H. Mohabatkar, Prediction of cyclin proteins using Chou's pseudo amino acid composition , Protein Peptide Lett. 17 (2010) 1207–1214.
[66] C. Chen, L. Chen, X. Zou, P. Cai, Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine , Protein Peptide Lett. 16 (2009) 27–31.
[67] H. Ding, L. Luo, H. Lin, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition , Protein Peptide Lett. 16 (2009) 351–355.
[68] X. Jiang, R. Wei, T.L. Zhang, Q. Gu, Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy , Protein Peptide Lett. 15 (2008) 392–396.
[69] L. Hu, T. Huang, X. Shi, W.C. Lu, Y.D. Cai, et al., Predicting functions of proteins in mouse based on weighted protein–protein interaction network and protein hybrid properties , PLoS ONE 6 (2011) e14556.
[70] Z. He, J. Zhang, X.H. Shi, L.L. Hu, X. Kong, et al., Predicting drug–target interaction networks based on functional groups and biological features , PLoS ONE 5 (2010) e9603.
[71] K.C. Chou, D.W. Elrod, Protein subcellular location prediction , Protein Eng. 12 (1999) 107–118.
[72] D.Z. Zhu, B.P. Ji, C.Y. Meng, B.L. Shi, Z.H. Tu, Z.S. Qing, Study of wavelet denoising in apple's charge-coupled device near-infrared spectroscopy , J. Agric. Food Chem. 55 (2007) 5423–5428.
[73] J.D. Qiu, R.P. Liang, X.Y. Zou, J.Y. Mo, Prediction of transmembrane proteins based on the continuous wavelet transform , J. Chem. Inf. Model 44 (2004) 741–747.
[74] K.C. Chou, H.B. Shen, Review: recent advances in developing web-servers for predicting protein attributes , Nat. Sci. 2 (2009) 63–92.