Predicting the subcellular location of apoptosis proteins based on recurrence quantification analysis and the Hilbert–Huang transform*

Han Guo-Sheng(韩国胜)[a], Yu Zu-Guo(喻祖国)[a)b)†], and Anh Vo[b)]

[a)] *School of Mathematics and Computational Science, Xiangtan University, Xiangtan* 411105, *China*

[b)] *Discipline of Mathematical Science, Faculty of Science and Technology, Queensland University of Technology, GPO Box* 2434, *Brisbane, Queensland* 4001, *Australia*

(Received 2 June 2011; revised manuscript received 15 June 2011)

Apoptosis proteins play an important role in the development and homeostasis of an organism. The elucidation of the subcellular locations and functions of these proteins is helpful for understanding the mechanism of programmed cell death. In this paper, the recurrent quantification analysis, Hilbert–Huang transform methods, the maximum relevance and minimum redundancy method and support vector machine are used to predict the subcellular location of apoptosis proteins. The validation of the jackknife test suggests that the proposed method can improve the prediction accuracy of the subcellular location of apoptosis proteins and its application may be promising in other fields.

Keywords: apoptosis proteins, subcellular location, recurrent quantification analysis, Hilbert–Huang transform

PACS: 05.45.Df, 64.60.al, 87.15.Qt **DOI:** 10.1088/1674-1056/20/10/100504

1. Introduction

In the post-genomic era, with an increasing number of prokaryotic and eukaryotic genomes being sequenced, more and more protein sequences are available for analysis. It is not viable for the relevant information about a protein to be determined by conducting time-consuming and costly chemical experiments alone. Obviously, it is imperative to develop fast computational methods to accurately and reliably determine the structures and functions of uncharacterized proteins,[1−3] which can be utilized for both basic research and drug discovery.[4] The function of a protein is closely related to its subcellular location. In the past few years, different methods have been used to determine subcellular locations of proteins, including neural networks,[5] support vector machines,[6] complexity measures,[4,7] the fuzzy k-NN method,[8] naive Bayes classification,[9] pseudo amino acid composition,[10,11] Markov chain models,[12] and semi-supervised learning.[13] These methods integrate various biological data sources.[14−16]

Apoptosis, also called programmed cell death, is a type of cell death that plays an important role in normal tissue homeostasis by regulating a balance between cell proliferation and death.[17] When cells undergo apoptosis, they usually exhibit a characteristic morphology including fragmentation of the cell into membrane-bound apoptotic bodies, nuclear and cytoplasm condensation and hemolytic cleavage of the DNA into small oligonucleosomal fragments.[18] When apoptosis goes wrong, many diseases can ensue: blocking apoptosis is associated with cancer and autoimmune diseases,[19,20] while unwanted apoptosis can possibly lead to ischemic damage[21,22] or neurodegenerative diseases.[23] Therefore, determining subcellular locations of apoptosis proteins contributes to the understanding of the apoptosis mechanism and functions of proteins.[23,24]

In 2003, Zhou and Doctor[25] first explored a dataset with 98 apoptosis proteins (DB98), which contains four types of subcellular locations and used the

*Project supported by the National Natural Science Foundation of China (Grant No. 11071282), the Chinese Program for New Century Excellent Talents in University (Grant No. NCET-08-06867), the Natural Science Foundation of Hunan Province of China (Grant No. 10JJ7001), the Lotus Scholars Program of Hunan Province of China, the Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province of China, the Australian Research Council (Grant No. DP0559807), and the Postgraduate Research and Innovation Project of Hunan Province of China (Grant No. CX2010B243).

†Author to whom any correspondence should be addressed. E-mail: yuzg@hotmail.com

© 2011 Chinese Physical Society and IOP Publishing Ltd

covariant discrimination function to predict their subcellular locations. The resulting overall prediction accuracies in the re-substitution test and jackknife test are 90.8% and 72.5%, respectively. Bulashevska and Eils[26] further improved the jackknife test accuracy by using a hierarchical ensemble of a Bayesian classifier. In 2006, Zhang *et al.*[27] built a new dataset with 225 apoptosis proteins (DB225) and developed a new method by combining a grouped weight of protein sequences and a support vector machine (EBGW_SVM). Their jackknife test result achieved 83.1% for this dataset. Ding and Zhang[28] applied the fuzzy k-nearest neighbour algorithm to dataset DB225 and improved the overall prediction accuracy. In 2007, Chen and Li[29] constructed another new apoptosis protein dataset with 317 proteins (DB317) and proposed a hybrid approach based on the increment of diversity algorithm and support vector machine (ID_SVM). Since most of the previous methods extracted n-peptide composition information from a protein sequence, the sequence-order information is lost. Although Chen and Li[30] considered local sequence-order information based on Chou's concept of pseudo amino acid composition, the overall prediction accuracy was not significantly improved. Zhang *et al.*[31] developed a new representation for protein sequences and the jackknife test result reached 88% on dataset DB317 by combining with the support vector machine. Gu *et al.*[32] extracted ensemble features and then used the binary particle swarm optimization method to extract effective features, the jackknife results achieved 91.5% and 88.0% for datasets DB317 and DB225, respectively.

In this paper, we intend to predict the subcellular locations of apoptosis proteins on three benchmark datasets (DB98, DB225, DB317).[25,27,29] For each protein sequence, the amino acid composition (AAC) is first calculated and 20 characteristic parameters are generated. Then, the protein sequences are converted into two different time series by using the hydrophobic free energies and solvent accessibility scales of 20 amino acids, which have been used to predict structural classes of large proteins.[33] Measured in kcal/mol, the hydrophobic free energies of 20 amino acids are A=0.87, R=0.85, N=0.09, D=0.66, C=1.52, Q=0.0, E=0.67, G=0.0, H=0.87, I=3.15, L=2.17, K=1.65, M=1.67, F=2.87, P=2.77, S=0.07, T=0.07, W=3.77, Y=2.76 and V=1.87.[34] The solvent accessibilities for a solvent exposed area larger than 30 AA^2 are S=0.70, T=0.71, A=0.48, G=0.51, P=0.78, C=0.32, D=0.81, E=0.93, Q=0.81, N=0.82, L=0.41, I=0.39, V= 0.40, M=0.44, F=0.42, Y=0.67, W=0.49, K=0.93, R= 0.84 and H=0.66.[35] Next, the recurrence quantification analysis (RQA)[36] and Hilbert–Huang transformation (HHT)[37] are used to analyse these time series. For each time series, RQA and HHT output 14 and 7 parameters respectively. Thus, for a given protein sequence, 42 ($2 \times 14 + 2 \times 7$) characteristic parameters are obtained, yielding 62 ($20 + 42$) characteristic parameters in total. Then, the maximum relevance and minimum redundancy (mRMR) method, which was originally developed by Peng *et al.*[38] to perform microarray data processing, is used to rank the parameters. In our method, each parameter is ranked according to its relevance to the class vector and redundancy with other parameters. Finally, based on the ranked parameters obtained from the mRMR method, we can construct 62 feature components by adding one parameter at a time in an ascending order. Optimal feature subsets are selected by the best jackknife test accuracy. The jackknife test is used to evaluate and compare our method with other existing methods. The jackknife test accuracies of our method reach 93.9%, 92.9%, and 92.1% for datasets DB98, DB225, DB317, respectively. The method proposed here has good performance, indicating that it may be used in conjunction with the existing methods.

2. Method

2.1. Amino acid composition

The amino acid composition is the simplest discrete representation of a protein sequence.[39] A protein sequence can be expressed as the occurrence frequencies f_i ($i = 1, 2, \ldots, 20$) of 20 amino acids. In much pioneering work the amino acid composition model was used to predict structural classes of proteins,[40] protein subcellular locations,[25,41,42] and other attributes.

2.2. Recurrence plot and recurrence quantification analysis

Recurrence plot (RP) is a purely graphical tool originally proposed by Eckmann *et al.*[43] in 1987 to visualize the patterns of recurrence in the data. For a time series $\{x_1, x_2, \ldots, x_N\}$ with length N, we can embed it into the space R^m with embedding dimension m and a time delay τ according to nonlinear dynamic theory.[44] Supposing that $\{\boldsymbol{y}_i\}_1^{N_m}$ represents a

trajectory in the corresponding phase space,

$$\boldsymbol{y}_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}),$$
$$i = 1, 2, \ldots, N_m, \quad (1)$$

where $N_m = N - (m-1)\tau$. Both embedding dimension m and time delay τ must be fixed according to nonlinear dynamic theory.[44]

Once a norm function is selected (e.g., the commonly chosen Euclidean norm[45]), we can calculate the distance matrix (DM) from the above N_m points. DM is an $N_m \times N_m$ square matrix whose elements are the distances between any pair of points. DM can be transformed into a rescaled distance matrix (RDM) through dividing each element in the DM by the maximum value of DM.[44] After obtaining RDM, it can be further transformed into a recurrence matrix (RM) whose elements are 0 or 1 by choosing a threshold ε. The elements of RM are calculated by the following equation:

$$R_{ij}(\varepsilon) = H(\varepsilon - D_{ij}), \quad i, j = 1, 2, \ldots, N_m, \quad (2)$$

where H is the Heaviside function

$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases} \quad (3)$$

RP is obtained by visualizing RM using different colours for its binary elements (e.g., plotting a black dot at the coordinates (i, j), if $R_{ij} = 1$ and a white dot, if $R_{ij} = 0$). For any ε, since $R_{ii} = 1$ $(i = 1, 2, \ldots, N_m)$ by definition, the RP always has a black main diagonal line. Furthermore, the RP is symmetric with respect to the main diagonal as $R_{ij} = R_{ji}$ $(i, j = 1, 2, \ldots, N_m)$. The RPs of hydrophobic free energy and solvent accessibility time series of the protein with Swiss–Prot access number Q9H1Y0 are shown in figure 1. An important step is to choose the parameter ε of RP. If ε is not chosen appropriately, we will not extract sufficient information about the underlying system.[46,47] The selection of ε will be discussed numerically below.

In order to overcome the physical limitation of RP, Webber and Zbilut[46] proposed a new nonlinear technique, namely recurrence quantification analysis (RQA), to quantify the information in the RP based on diagonal structures and vertical structures. In recent years, the RQA has been successfully applied in many different fields.[45] Many efforts have been made to use RQA to analyse protein sequences.[47−52] But, as far as we know, the RQA has not been used to predict the subcellular locations of proteins. There are 14 recurrence variables used to quantify RP.[44−46] The detailed description about these recurrence variables can be found in Refs. [44] and [46] and the references therein. They can also be calculated by the CRP toolbox for Matlab, which is provided by TOCSY (http://tocsy.agnld.uni-potsdam.de).

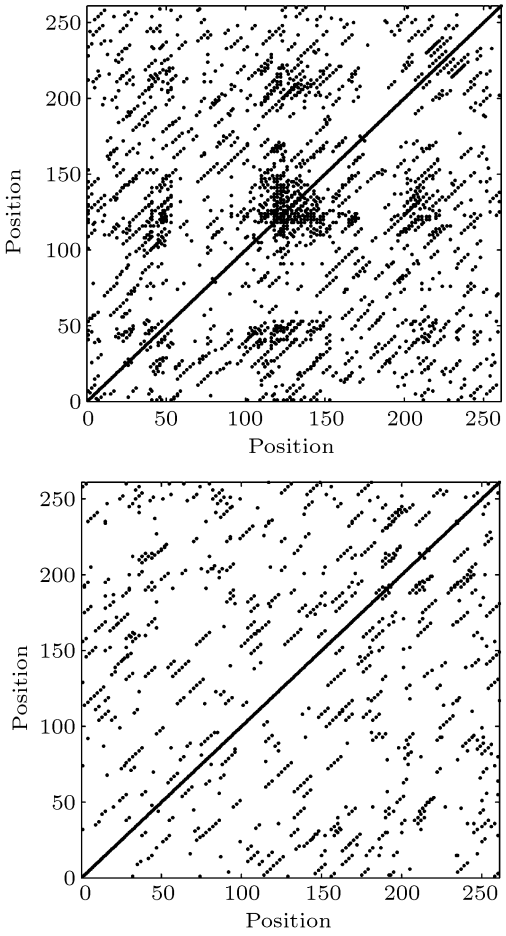


Fig. 1. The RPs for the two time series of the protein Q9H1Y0. The parameters used are $m = 8$, $\tau = 2$ and $\varepsilon = 30\%$.

We now discuss the determination of parameters. In order to obtain the recurrence plot, the three parameters of embedding dimension m, time delay τ and radius ε need to be determined in advance. Different approaches to the estimation of the smallest sufficient embedding dimension m[44] as well as an appropriate time delay τ[44,53] have been proposed. In this paper, we use an approach similar to that of Riley *et al.*[44] and Yang *et al.*[47,52] The embedding dimension m is set to be between 6 and 9, and we produce 4 surface plots for time delay τ from 1 to 4 and radius ε from 10 to 39. A general guideline is that ε should be selected such that $\%REC$ remains low (often smaller

than 5%).[44,47,52] The variable $\%REC$ means the density of recurrence points in the RP.[36]

We consider the apoptosis protein with Swiss–Prot access number Q9H1Y0 for example. We transform it into two time series based on hydrophobic free energies and the solvent accessibility of 20 amino acids. For the hydrophobic free energy time series, we show 4 surface plots in Fig. 2. From Fig. 2, we set $m = 8$ since the changes of $\%REC$ are relatively small for $m = 7, 8, 9$ and the changes are smooth for $m = 8$ with the increases of radius ε and τ; we set $\tau = 2$ since the values of $\%REC$ remain almost constant (or change smoothly) with τ when $\tau \geq 2$. For ε, it is set at 30% to ensure that $\%REC$ is smaller than 5%.[44,47] Similarly, for the solvent time series, m, τ and ε are set at 8, 2, and 30%, respectively.

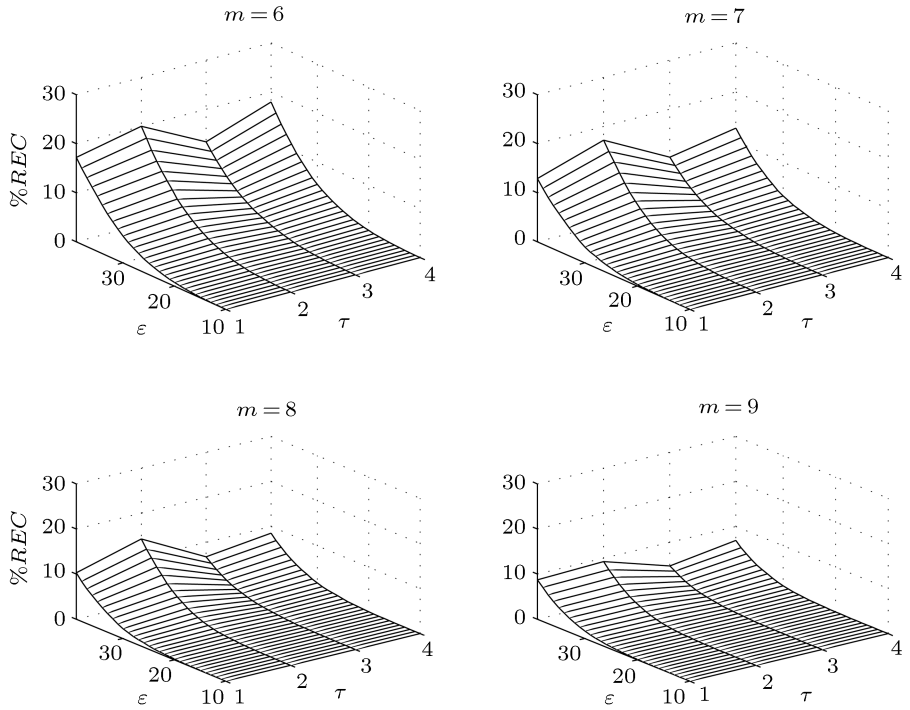


Fig. 2. Surface plots for embedding dimensions m from 6 to 9 and delays τ from 1 to 4. The values of $\%REC$ for different values of parameters in RQA are calculated for the upper panel in figure 1. From the above panels, we set $m = 8\varepsilon$ ranges between 10% and 39%.

2.3. Hilbert–Huang transformation

The Hilbert–Huang transformation consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). They are briefly described below.

2.3.1. Empirical mode decomposition

The empirical mode decomposition is a time-frequency analysis method and was originally proposed by Huang *et al.*[37] for the study of ocean waves. Similar to other time-frequency methods, such as Fourier analysis and wavelet analysis, EMD decomposes a time series into a finite set of oscillatory bases.[37] However, the base functions of Fourier and wavelet analyses are pre-determined and are not suitable for nonlinear systems.[37] In EMD, the base functions, which are called intrinsic mode functions (IMFs), are obtained adaptively from the original signal. Since the characteristic scale for EMD is defined as the distance between two successive maxima (or minima), EMD has a local ability.[37] The principle of EMD is to decompose data into finite IMFs to approximate the mono-component data. An IMF must satisfy the following two conditions: (i) the difference between the number of local extrema and the number of zero-crossings must be zero or one; (ii) the mean value of the upper envelope defined by the local maxima and the lower envelope defined by the local minima is nearly zero. The first condition makes IMFs narrow-banded; the second condition renders IMFs roughly symmetric with respect to the time axis. These two conditions fulfill the physically necessary conditions to define a meaningful instantaneous frequency. Those

IMF components are obtained from the original signal by means of an algorithm called sifting process. The detailed information about sifting process can be found in Ref. [37]. After sifting, the original signal can be reconstructed as follows:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t), \tag{4}$$

where n is the number of IMFs, $r_n(t)$ denotes the final residual and $c_i(t)$ $(i = 1, 2, \ldots, n)$ is an IMF that is nearly orthogonal to each other. The EMD for a hydrophobic free energy time series of the protein Q9H1Y0 is shown in Fig. 3. In this figure, the IMFs of the highest frequency (i.e., at the smallest time scale) are extracted first. Sometimes an IMF may contain more than one mono-function; it is then called a mixing mode, which can be separated by applying a more stringent stopping criterion or using the intermittence test.[54] Moreover, since the cubic spline can have large swings near the end points,[37] it induces boundary effects. The boundary processing procedures then need to be used to improve EMD. The EMD method has been used by our group to simulate geomagnetic field data recently.[55]

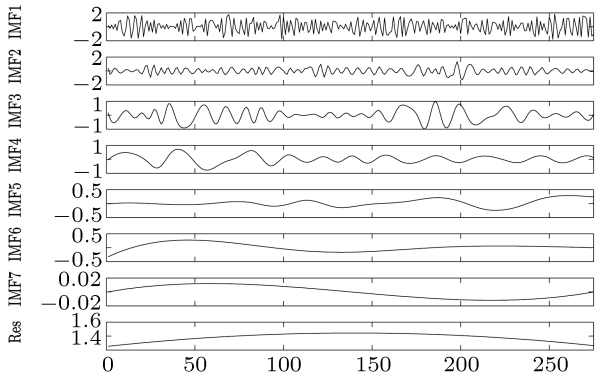


Fig. 3. Seven IMF components and the residual obtained by EMD of hydrophobic free energy time series of the protein Q9H1Y0.

2.3.2. Hilbert spectral analysis

After EMD, we apply the Hilbert transform to each IMF component and obtain an analytic signal as

$$H(c_i(t)) = \frac{1}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{c_i(t)}{t - \tau} \mathrm{d}\tau, \tag{5}$$

$$A(c_i(t)) = c_i(t) + jH(c_i(t)) = a_i(t)\, \mathrm{e}^{\,\mathrm{j}\,\theta_i(t)}, \tag{6}$$

where P indicates the principle value of the singular integral, $a_i(t)$ is the instantaneous amplitude and $\theta_i(t)$ is the phase function:

$$a_i(t) = \sqrt{c_i^2(t) + H^2(c_i(t))}, \tag{7}$$

$$\theta_i(t) = \arctan \frac{H(c_i(t))}{c_i(t)}. \tag{8}$$

The instantaneous frequency can thus be calculated as

$$\omega_i(t) = \frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t}. \tag{9}$$

The marginal spectrum $h(\omega)$ (or the Hilbert–Huang spectrum) can be obtained by integrating with respect to the time variable as

$$h(\omega) = \int H(\omega, t)\, \mathrm{d}t. \tag{10}$$

The marginal spectrum offers a measure of total amplitude contribution from each frequency.

If the Hilbert–Huang spectrum is denoted as a function of frequency f instead of angle frequency ω, the marginal spectrum can be calculated for each IMF and then normalized by

$$\hat{h}(f) = \frac{h(f)}{\sum\limits_{f} h(f)}, \tag{11}$$

(c.f. Ref. [56]). Then, by applying the Shannon entropy theory to the normalized marginal spectrum, the Hilbert–Huang spectral entropy (HHSE) is obtained as

$$H = -\sum_{f} \hat{h}(f) \log(\hat{h}(f)), \tag{12}$$

$$\mathrm{HHSE} = H/\log(N), \tag{13}$$

where N is the number of frequency components and the value of HHSE varies between 0 (complete regularity) and 1 (maximum irregularity).

2.4. Maximum relevance and minimum redundancy method and feature selection

The maximum relevance and minimum redundancy (mRMR) method was originally proposed by Peng *et al.*[38] to analyse microarray data. The relevance R of two vectors is measured by the following mutual information (MI):

$$\begin{aligned} R(x, y) &= I(x, y) \\ &= \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, \mathrm{d}x\, \mathrm{d}y, \end{aligned} \tag{14}$$

where x, y are two vectors, $p(x, y)$ is the joint probability density, $p(x)$ and $p(y)$ are the marginal probability densities.

Let Ω denote the feature space which contains all of the aforementioned 62 parameters; each feature is represented by a vector consisting of the values of this feature (parameter) for all proteins in the training set. A and B are two disjoint subsets of Ω and $\Omega = A \cup B$. The redundancy $D(f, B)$ of the feature f in A with all the features in B is calculated from

$$D(f, B) = \frac{1}{\sharp B} \sum_{f_i \in B} I(f, f_i), \tag{15}$$

where $\sharp B$ denotes the cardinality of subset B and the same below.

In order to select a feature f_j in A with maximum relevance and minimum redundancy with B, the following function is used by combining Eq. (13) and Eq. (14):

$$\max_{f_j \in A} [I(f_j, \theta) - D(f_j, B)] \quad (j = 1, 2, \ldots, \sharp A), \tag{16}$$

where θ is a vector indicating the class of all proteins in the training set.

If we select one feature as A and set the remaining features as B each time, the entire feature space Ω can be ranked by the mRMR method through $\sharp\Omega$ rounds. Suppose that S denotes the ranked feature set

$$S = \left\{ f_{i_1}, f_{i_2}, \ldots, f_{i_{\sharp\Omega}} \right\} \quad (1 \le i_k \le \sharp\Omega). \tag{17}$$

After all features are ranked, we can determine the optimal feature components by an incremental feature selection (IFS) method.[57] For the ranked feature set S, we can construct $\sharp\Omega$ feature component sets by adding one component at a time in an ascending order as follows:

$$S_k = \{ f_{i_1}, f_{i_2}, \ldots, f_{i_k} \} \quad (1 \le k \le \sharp\Omega). \tag{18}$$

For each feature component set, a predictor is constructed and its jackknife success rate is obtained. Finally, we choose the feature component set for the best jackknife test accuracy as the optimal feature set.

2.5. Support vector machine

In 1995, Vapnik[58] introduced the support vector machine (SVM) method to solve pattern recognition problems. In recent years, it has been used to predict subcellular locations of proteins (e.g. Ref. [6]). SVM was originally designed for binary classification. But predicting protein subcellular locations is a multiclass classification problem. In order to solve the latter problem, three methods have been developed based on binary classification: one-against-all, one-against-one and directed acyclic graph SVM (DAGSVM). In this paper, we use the one-against-one strategy to predict apoptosis protein subcellular locations. For a k classification problem, SVM designed by one-against-one will construct $k \times (k-1)/2$ classifiers, each of which is trained on data from two different classes. The class of a test protein sequence is determined by a voting strategy from $k \times (k-1)/2$ classifiers. The selection of parameters for SVM is also crucial. In this paper, we use RBF kernel function and select parameters by the jackknife test and use the software toolbox LIBSVM[59] to implement SVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/cjlin/libsvm.

2.6. Evaluation of the prediction performance

In statistical prediction, three cross-validation tests are often used to evaluate the prediction performance: independent dataset test, sub-sampling test and jackknife test. Among these three tests, the jackknife test is thought to be the most rigorous and objective.[40] Hence, we use the jackknife test in this paper. That is, each protein sequence in the samples is singled out in turn as a test sample and the remaining protein sequences are used as training samples. Therefore, the jackknife test is also called the leave-one-out test.

Based on SVM, we calculate the prediction accuracies for the datasets DB317, DB225 and DB98, which are listed in Tables 2, 3, and 4, respectively. The overall prediction accuracy A_c, individual sensitivity $S_{\rm in}$, individual specificity $S_{\rm ip}$ and Matthew's correlation coefficient MCC_i are used to evaluate the prediction performance of our work. Their definitions are

$$S_{\rm in} = TP_i/(TP_i + FN_i), \tag{19}$$

$$S_{\rm ip} = TP_i/(TP_i + FP_i), \tag{20}$$

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i) \times (TP_i + FN_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}}, \tag{21}$$

$$A_c = \frac{\sum_i TP_i}{N}, \quad i = 1, 2, 3, \ldots, n, \tag{22}$$

where true positives TP = number of positive events that are correctly predicted; true negatives TN = number of negative events that are correctly predicted; false positives FP = number of negative events that are incorrectly predicted to be positive; and false negative FN = number of subjects that are incorrectly predicted to be negative.

3. Results and discussion

In order to test the performance of the methods, three benchmark data sets have been used.[25,27,29] The DB98 dataset[25] consists of 98 apoptosis protein sequences, which include 43 cytoplasmic proteins, 30 plasma membrane-bound proteins, 13 mitochondrial proteins and 12 other proteins. The DB225 dataset[27] contains 41 nuclear proteins, 70 cytoplasmic proteins, 25 mitochondrial proteins and 89 membrane proteins. The third dataset DB317[29] consists of 317 apoptosis protein sequences divided into six subcellular locations with 112 cytoplasmic proteins, 55 membrane proteins, 34 mitochondrial proteins, 17 secreted proteins, 52 nuclear proteins and 47 endoplasmic reticulum proteins. All the protein sequences in three datasets are extracted from SWISS-PROT and the access numbers can be found in Refs. [25], [27], and [29].

Table 1. Prediction results for dataset DB317 using our method for re-substitution and jackknife tests (SVM: $\gamma = 0.5$, $c = 8$).

Location	Re-substitution test			Jackknife test		
	S_n/%	S_p/%	MCC	S_n/%	S_p/%	MCC
Cyto	99.1	100.0	0.99	95.5	89.9	0.88
Memb	100.0	98.2	0.98	94.6	91.2	0.91
Nucl	100.0	100.0	1.0	90.4	90.4	0.88
Endo	100.0	100.0	1.0	91.5	93.5	0.91
Mito	100.0	100.0	1.0	97.1	100.0	0.98
Secr	100.0	100.0	1.0	58.8	100.0	0.76
Acc(%)	99.7			92.1		

For each apoptosis protein sequence, we first calculate the AAC to obtain 20 parameters. Then, two time series are derived by transforming amino acids into their hydrophobic free energies and solvent accessibilities.[33] We use RQA and HHT to analyse these time series and obtain a total of 42 parameters. Finally, each apoptosis protein sequence is represented by 62 features which can be used to predict its subcellular locations. Further, mRMR is used to rank 62 features and then select optimal features by the best jackknife accuracy. The best parameters are also determined by SVM. Detailed predictive results for the datasets DB317, DB225, and DB98 are listed in Tables 1, 2, and 3, respectively. The information about the optimal feature subsets of these three datasets is shown in Table 4.

Table 2. Prediction results for dataset DB225 using our method for re-substitution and jackknife tests (SVM: $\gamma = 0.03$, $c = 16$).

Location	Re-substitution test			Jackknife test		
	S_n/%	S_p/%	MCC	S_n/%	S_p/%	MCC
Cyto	98.6	93.2	0.94	97.1	88.3	0.89
Memb	97.8	98.9	0.97	92.1	94.3	0.89
Nucl	100.0	100.0	1.0	100.0	100.0	1.0
Mito	88.0	100.0	0.93	72.0	90.0	0.78
Acc(%)	97.3			92.9		

Table 3. Prediction results for dataset DB98 using our method for re-substitution and jackknife tests (SVM: $\gamma = 0.02$, $c = 64$).

Location	Re-substitution test			Jackknife test		
	S_n/%	S_p/%	MCC	S_n/%	S_p/%	MCC
Cyto	100.0	95.6	0.96	97.7	89.4	0.88
Memb	100.0	100.0	1.0	96.7	100.0	0.98
Mito	92.3	100.0	0.96	76.9	91.0	0.81
Other	91.7	100.0	0.95	91.7	100.0	0.95
Acc(%)	98.0			93.9		

Table 4. Optimal feature subsets for the datasets DB317, DB225 and DB98.

Dataset	The number of feature			Overall
	AAC	RQA	HHT	
DB317	16	14	8	38
DB225	20	23	12	55
DB98	13	10	2	25

Table 1 shows that the overall accuracies for dataset DB317 by our method achieve 99.7% and 92.1% in the re-substitution and jackknife tests, respectively. It shows RQA and HHT complement sequence-order information to some extent. Since RQA and HHT can capture recurrence patterns and periodic changes in different frequency domains, respectively, it is feasible to integrate RQA and HHT with AAC. For dataset DB317, the results show that when AAC, RQA, and HHT are used in conjunction with mRMR, the overall accuracies are 88.0%, 78.6%, and 60.9%, respectively, which are higher than those without using mRMR. In other words, AAC, RQA,

and HHT all make positive contributions to this prediction problem. In order to evaluate the performance of our method, we also use datasets DB225 and DB98. The detailed prediction results for these datasets are shown in Tables 2 and 3, respectively.

Table 5 shows a detailed comparison of our method with the other methods on the same dataset DB317. We find that the overall accuracy of our method is higher than the other methods in the jackknife test. The prediction accuracies for the mitochondrial proteins and membrane proteins are much improved. But, the prediction accuracies for secreted proteins are low, which may be due to their small samples. The values of specificity and MCC further indicate that our method is significantly effective for location prediction except for secreted proteins. Tables 6 and 7 show a detailed comparison of our method with the other methods on datasets DB225 and DB98. Especially, from Table 6, we can see that the total prediction accuracy by our method is nearly 5% higher than the second highest one. Even for the nuclear proteins, our prediction accuracy is 100%. In Table 7, we find that the overall accuracy on dataset DB98 using our method is the highest, which reaches 93.9% in the jackknife test. Except for mitochondrial proteins, the accuracies for the other proteins are all over 90%. Comparison with the method proposed by Zhang *et al.*[31] shows that two methods have similar results for cytoplasmic proteins and membrane proteins, but for mitochondrial proteins and the other proteins, the results are significantly different. Their jackknife test results achieve 92.3% and 75.7% for mitochondrial proteins and the other proteins respectively. On the other hand, our jackknife test results are 76.9% and 91.7%. From Tables 5, 6, and 7, we can see that our method achieves the highest prediction accuracy for the jackknife test in all three datasets, but some other methods can achieve the highest jackknife test success rate in one or two datasets. Therefore, our method is effective and useful for predicting apoptosis protein subcellular locations.

Table 5. Comparisons of prediction performance for different methods on dataset DB317 with a jackknife test.

Methods	Sensitivity/%						Acc	Number of features
	Cy	Me	Nu	En	Mi	Se		
Ref. [29]	81.3	81.8	82.7	83.0	85.3	88.2	82.7	18
Ref. [30]	91.1	89.1	73.1	87.2	79.4	58.8	84.2	18
Ref. [60]	90.2	90.9	86.5	91.5	82.4	88.2	89.0	40
Ref. [32]	98.2	83.6	90.4	97.9	79.4	82.4	91.5	8420
Ref. [31]	92.9	85.5	86.5	93.6	76.5	76.5	88.0	396
Our method	**95.5**	**94.6**	**90.4**	**91.5**	**97.1**	**58.8**	**92.1**	**38**

Table 6. Comparisons of prediction performance for different methods on dataset DB225 with a jackknife test.

Methods	Sensitivity/%				Acc	Number of features
	Me	Cy	Nu	Mi		
Ref. [27]	93.3	90.0	63.4	60.0	83.1	120
Ref. [30]	91.0	92.9	73.2	68.0	85.8	18
Ref. [28]	93.3	84.3	85.5	72.0	85.8	40
Ref. [32]	94.4	94.3	80.5	60.0	88.0	8420
Ref. [31]	92.1	87.1	73.2	64.0	84.0	225
Our method	**92.1**	**97.1**	**100.0**	**72.0**	**92.9**	**55**

Remark The performance results of other methods cited above and in Tables 5–7 are from the original papers. We use exactly the same way to perform the validation for comparison. The number of parameters in our methods and other methods are given in Tables 6 and 7 for comparison. One can find that the number of parameters in our method is relatively small, but the prediction performance is better. So our improvement is not due to the increase in the number of parameters.

Table 7. Comparisons of prediction performance for different methods on dataset DB98 with a jackknife test.

Methods	Sensitivity/%				Acc	Number of features
	Cy	Mi	Me	Other		
Ref. [25]	97.7	30.8	73.3	25.0	72.5	19
Ref. [61]	76.8	92.3	83.3	50.0	77.6	21
Ref. [27]	97.7	92.3	90.0	83.3	92.9	120
Ref. [30]	95.3	84.6	93.3	58.3	88.8	18
Ref. [26]	95.3	92.3	90.0	66.7	89.8	–
Ref. [56]	95.3	92.3	96.7	75.7	92.9	23
Ref. [31]	97.7	92.3	96.7	75.7	93.9	225
Our method	**97.7**	**76.9**	**96.7**	**91.7**	**93.9**	**25**

4. Conclusions

The subcellular location of a protein is closely related to its function. The amino acid composition is important for predicting subcellular locations of proteins.[39] For each protein sequence, we first calculate its AAC and generate 20 parameters. By using hydrophobic free energies and solvent accessibilities of amino acids, we transform every protein sequence into two time series and apply HHT and RQA methods to these two time series, respectively. A total of 42 parameters can be obtained from these two methods. Finally, each protein sequence is represented by a 62-dimensional feature vector. By using mRMR, all 62 features are ranked. We use SVM as the prediction engine and the jackknife test to select optimal feature subsets and evaluate the method. For dataset DB317, the jackknife test accuracy reaches 92.1%, which is slightly higher than that in Ref. [32]. For dataset DB225, the prediction accuracy for the jackknife test achieves 92.4% and it is nearly 5% higher than the second highest score.[32] For dataset DB98, our jackknife test result is similar to that in Ref. [31]. The above results suggest that our method may complement the existing methods. If more information, such as structure information, sequence-order sequence and so on, is added to the current method, the results may be further improved. However, we only consider that every protein has a subcellular location; on the other hand, many proteins have multiple subcellular locations. Hence more work is needed in this direction.

References

[1] Jensen L J, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt H H, Rapacki K, Workman C, Andersen C A F, Knudsen S, Krogh A, Valencia A and Brunak S 2002 *J. Mol. Biol.* **319** 1257

[2] Yu Z G, Xiao Q J, Shi L, Yu J W and Anh V 2010 *Chin. Phys.* B **19** 068701

[3] Zhu S M, Yu Z G and Anh V 2011 *Chin. Phys.* B **20** 010505

[4] Chou K C and Cai Y D 2004 *Biochem. Biophys. Res. Commun.* **320** 1236

[5] Reinhardt A and Hubbard T 1998 *Nucleic Acids Res.* **26** 2230

[6] Hua S J and Sun Z R 2001 *Bioinformatics* **17** 721

[7] Liu T G, Zheng X Q and Wang J 2010 *Amino Acids* **38** 721

[8] Huang Y and Li Y D 2001 *Bioinformatics* **20** 21

[9] Briesemeister S, Rahnenführer J and Kohlbacher O 2010 *Bioinformatics* **26** 1232

[10] Xiao X, Shao S H, Ding Y S, Huang Z D and Chou K C 2006 *Amino Acids* **30** 49

[11] Chou K C and Shen H B 2006 *J. Cell Biochem.* **99** 517

[12] Yuan Z 1999 *FEBS Lett.* **451** 23

[13] Xu Q, Hu D H, Xue H, Yu W C and Yang Q 2009 *BMC Bioinformatics* (Suppl. 1) **10** S47

[14] Tung T Q and Lee D 2009 *BMC Bioinformatics* (Suppl. 1) **10** S43

[15] Höglund A, Dönnes P, Blum T, Adolph H W and Kohlbacher O 2006 *Bioinformatics* **22** 1158

[16] Chou K C and Shen H B 2010 *PLoS ONE* **5** e9931

[17] Jacobson M D, Weil M and Raff M C 1997 *Cell* **88** 347

[18] Kerr J F R, Wyllie A H and Currie A R 1972 *Brit. J. Cancer* **26** 239

[19] Adams J M and Cory S 1998 *Science* **281** 1322

[20] Evan G and Littlewood T 1998 *Science* **281** 1317

[21] Reed J C and Paternostro G 1999 *Proc. Natl. Acad. Sci. USA* **96** 7614

[22] Raff M 1998 *Nature* **396** 119

[23] Schulz J B, Weller M and Moskowitz M A 1999 *Ann. Neurol.* **45** 421

[24] Suzuki M, Youle R J and Tjandra N 2000 *Cell* **103** 645

[25] Zhou G P and Doctor K 2003 *Proteins Struct. Funct. Genet.* **50** 44

[26] Bulashevska A and Eils R 2006 *BMC Bioinformatics* **7** 298

[27] Zhang Z H, Wang Z H, Zhang Z R and Wang Y X 2006 *FEBS Lett.* **580** 6169

[28] Ding Y S and Zhang T L 2008 *Pattern Recogn. Lett.* **29** 1887

[29] Chen Y L and Li Q Z 2007 *J. Theor. Biol.* **245** 775

[30] Chen Y L and Li Q Z 2007 *J. Theor. Biol.* **248** 377

[31] Zhang L, Liao B, Li D C and Zhu W 2009 *J. Theor. Biol.* **259** 361

[32] Gu Q, Ding Y S, Jiang X Y and Zhang T L 2010 *Amino Acids* **38** 975

[33] Yu Z G, Anh V, Lau K S and Zhou L Q 2006 *Phys. Rev.* E **73** 031920

[34] Selz K A, Mandell A J and Shlesinger M F 1998 *Biophys. J.* **75** 2332

[35] Bordo D and Argos P 1991 *J. Mol. Biol.* **217** 721

[36] Webber J C L and Zbilut J P 1994 *J. Appl. Physiol.* **76** 965

[37] Huang N E, Shen Z, Long S R, Wu M C, Shih H H, Zheng Q A, Yen N C, Tung C C and Liu H H 1998 *Proc. R. Soc.* **454** 903

[38] Peng H, Long F and Ding C 2005 *IEEE Tran. Pattern Anal.* **27** 1226

[39] Chou K C and Shen H B 2007 *Anal. Biochem.* **370** 1

[40] Chou K C 1995 *Proteins Struct. Funct. Genet.* **21** 319

[41] Nakashima H and Nishikawa K 1994 *J. Mol. Biol.* **238** 54

[42] Cedano J, Aloy P, Perez-Pons J A and Querol E 1997 *J. Mol. Biol.* **266** 594

[43] Eckmann J P, Kamphorst S O and Ruelle D 1987 *Europhys. Lett.* **4** 973

[44] Riley M A and van Orden G C 2005, Retrieved March 1, 2005, online available from website http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

[45] Giuliani A, Benigni R, Zbilut J P, Webber J C L, Sirabella P and Colosimo A 2002 *Chem. Rev.* **102** 1471

[46] Marwan N, Romano M C, Thiel M and Kurths J 2007 *Phys. Rep.* **438** 237

[47] Yang J Y, Peng Z L, Yu Z G, Zhang R J, Anh V and Wang D S 2009 *J. Theor. Biol.* **257** 618

[48] Giuliani A, Sirabella P, Benigni R and Colosimo A 2000 *Protein Eng.* **13** 671

[49] Manetti C, Ceruso M A, Giuliani A, Webber J C L and Zbilut J P 1999 *Phys. Rev.* E **59** 992

[50] Webber J C L, Giuliani A, Zbilut J P and Colosimo A 2001 *Proteins* **44** 292

[51] Zhou Y, Yu Z G and Anh V 2007 *Phys. Lett.* A **368** 314

[52] Yang J Y, Peng Z L and Chen X 2010 *BMC Bioinformatics* (Suppl. 1) **11** S9

[53] Cao L 1997 *Physica* D **110** 43

[54] Huang N E, Wu M L C, Long S R, Shen S S P, Qu W D and Gloersen P and Fan K L 2003 *Proc. R. Soc.* **459** 2317

[55] Yu Z G, Anh V, Wang Y, Mao D and Wanliss J 2010 *J. Geophys. Res.* **115** A10219

[56] Shi F, Chen Q J and Li N N 2008 *J. Biomed. Sci. Eng.* **1** 59

[57] Huang T, Shi X H, Wang P, He Z S, Feng K Y, Hu L L, Kong X Y, Li Y X, Cai Y D and Chou K C 2010 *PLoS ONE* **5** e10972

[58] Vapnik V N 1995 *The Nature of Statistical Learning Theory* (Berline: Springer)

[59] Chang C C and Lin C J *LIBSVM: a Library for Support Vector Machines* Software available online at http://www.csie.ntu.edu.tw/cjlin/libsvm

[60] Shi J Y, Zhang S W, Pan Q, Cheng Y M and Xie J 2007 *Amino Acids* **33** 69

[61] Huang J, Shi F and Zhou H B 2005 *China J. Bioinformatics* **3** 121