J. Biomedical Science and Engineering, 2008, 1, 59-63

Scientific
Research
Publishing

Hilbert Huang transform for predicting proteins subcellular location

Feng Shi * , Qiu-Jian Chen & Na-na Li

School of Science, Huazhong Agricultural University, Wuhan, Hubei, China. Correspondence should be addressed to Feng Shi (shifeng@mail.hzau.edu.cn).

ABSTRACT

Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death, and their function is related to their types. The apoptosis proteins are categorized into the following four types: (1) Cytoplasmic protein; (2) Plasma membrane-bound protein; (3) Mitochondrial inner and outer proteins; (4) Other proteins. A novel method, the Hilbert-Huang transform, is applied for predicting the type of a given apoptosis protein with support vector machine. High success rates were obtained by the re-substitute test (98/98=100%) and jackknife test (91/98 = 92.9%).

Keywords:Hilbert Huang transform; Support vector machine; Subcellular location predict

1. INTRODUCTION

Apoptosis, or programmed cell death, is a fundamental process controlling normal tissue homeostasis by regulating a balance between cell proliferation and death [1]. This process entails the autolytic degradation of cellular components, and is characterized by blebbing of cell membranes, shrinkage of cell volumes, and condensation of nuclei [2], and is currently an area of intense investigation. Cell death and renewal are responsible for maintaining the proper turnover of cells, which ensures a constant controlled flux of fresh cells. Programmed cell death and cell proliferation are tightly coupled. When apoptosis malfunctions, a variety of formidable diseases can ensue: blocking apoptosis is associated with cancer and autoimmune disease, whereas unwanted apoptosis can possibly lead to ischemic damage or neurodegenerative disease [3]. Apoptosis is considered to have a key role in these several devastating diseases and, in principle, provides many targets for therapeutic intervention [4]. To understand the apoptosis mechanism and functions of various apoptosis proteins, it will be helpful to obtain information about their subcellular location. This is because the subcellular location of apoptosis proteins is closely related to their function [5,6]. It has been known that there are 732 archetypical proteins with “apoptosis” domains [7], and only 98 of these proteins are known to be the apoptosis protein (for more details, one can visit: http://www.apoptosis-db.org). Scientists usually deal with a number of protein sequences already known belonging to apoptosis proteins. However, it is both time-consuming and costly to determine which specific subcellular location a given apoptosis protein belongs to. Confronted with such a situation, can we develop a fast and effective way to predict the subcellular location for a given apoptosis protein based on its amino acid sequence? Recently, Guo-ping Zhou [7] attempted to identify the subcellular location of apoptosis proteins according to their sequences by means of the covariant discriminant function, which was established on the basis of the Mahalanobis distance and Chou's invariance theorem [7,8,9].The results were quite promising, indicating that the subcellular location of apoptosis proteins are predictable to a considerably accurate extent if a good vector representation of protein can be established. It is expected that, with a continuous improvement of vector representation methods by incorporating amino acid properties, and by using more powerful mathematics methods, some theory predicting method might eventually become a useful tool in this area because the function of an apoptosis protein is closely related to its subcellular location. The present study was initiated in an attempt to address this problem.

Chou and Elrod made an extensive research in predicting subcellular location mainly based on the amino acid composition. Subsequently, in order to take into account the sequence-order effects and improved the prediction quality, Chou has further incorporated the quasi-sequence order effect [5] and introduced the concept of “pseudo-amino-acid composition” [9]. For example, Chou [10] classified membrane proteins into five different types and proposed a covariant discriminant algorithm to predict the types of membrane proteins. Recently, Cai *et al.* [11] applied neural network to this problem. To improve the prediction quality, Chou [5] proposed a new method in which the covariant discriminate algorithm was augmented to incorporate the quasi-sequence-order effect. This method uses the amino acid composition and the sequence-order-coupling numbers (reflecting the sequence order effect) in order to improve the prediction quality. Feng [12] proposed a new representation of unified attribute vector, that each protein can be represented by a vector, which is 20-D vector in Hilbert space with unified length. Hence, all of proteins have their representative points on the surface of the 20-D globe. The representative points of the proteins in the same family or with the higher sequence identity are closer on the surface. The overall predictive accuracy could be improved from 3% to 5% for different databases [12] with this simply modification of the usage of the amino acid composition. Recently, a series of new powerful approaches have been developed by Chou and his co-workers [13]. Encouraged by the great successes of the previous invertigators in the area, here we would like to use a different strategy, the support vector machines, to approach this very important but also very difficult problem in the hope that our approach can play a complementary role to the existing methods.

2. HILBERT HUANG TRANSFORM

The HHT consists of two parts: empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). This method is potentially viable for nonlinear and nonstationary data analysis, especially for time-frequency-energy representations. It has been tested and validated exhaustively, but only empirically. In all the cases studied, the HHT gave results much sharper than those from any of the traditional analysis methods in time-frequency-energy representations. Additionally, the HHT revealed true physical meanings in many of the data examined. Powerful as it is, the method is entirely empirical. In order to make the method more robust and rigorous, many outstanding mathematical problems related to the HHT method need to be resolved. In this section, a brief introduction to the methodology of the HHT will be given. Readers interested in the complete details should consult [14].

2.1. The empirical mode decomposition method (the sifting process)

In this method any time series, including non-linear and non-stationary series, can be decomposed into a finite number of intrinsic mode functions (IMFs) through empirical mode decomposition (EMD) process. An IMF is a function which must follow two conditions: (1) the difference between the numbers of extrema and zero-crossings is of $\leqslant 1$; and (2) the mean of the upper envelop (linked by local maxima) and the lower envelop (linked by local minima) are zero at every point.

The EMD process is as follows. According to Hilbert-Huang transform(HHT)[14], once the extrema of a time series $x(t)$ are identified, all the local maxima and minima are connected by two special lines as the upper and lower envelopes respectively. Their mean is designated as m_1, and the difference between $x(t)$ and m_1 is $x(t)$-m_1=h_1 . If h_1 is not an IMF, h_1 is treated as the data and undergoes the procedure above, then h_1-m_{11}=h_{11} . Repeat this sifting procedure k times until h_{1k} is an IMF, that is $h_{1(k-1)}$-m_{1k}=h_{1k}, thus the first IMF component is obtained, i.e. . Then separate IMF_1 from the original time series by $x(t)$- IMF_1=r_1. Treat r_1 as the new data and subject it to the same sifting process above. Repeat this procedure on all the subsequent r_j , i.e. r_1-IMF_2=r_2 , r_2-IMF_3=r_3,···, r_{n-1}-IMF_n=r_n .

So the result is:

$$x(t) = \sum_{j=1}^{n} IMF_j(t) + r_n(t)$$

2.2. Hilbert transform

Having obtained the intrinsic mode function components IMF_i (denoted as c_i), one will have no difficulty in applying the Hilbert transform to each IMF component,

$$H(c_i(t)) = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{c_i(t)}{t-\tau} d\tau$$

in which the PV indicates the principal value of the singular integral. With the hilbert transform, the analytic signal is defined as

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

Here, $a_i(t)$ is the instantaneous amplitude, and $\theta_i(t)$ is the phase function,

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

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

and the instantaneous frequency is simply

$$\varpi_i(t) = \frac{d\theta_i(t)}{dt}$$

With the Hilbert Spectrum defined, we can also define the marginal spectrum $h(\varpi)$ as

$$h(\varpi) = \int_0^T H(\varpi, t)dt$$

The marginal spectrum offers a measure of the total amplitude (or energy) contribution from each

Table 1. Comparative summary of Fourier, Wavelet and HHT analyses.

	Fourier	Wavelet	Hilbert
Basis	A priori	a priori	adaptive
Frequency	Convolution: global Uncertainty	convolution: regional uncertainty	differentiation local, certainty
Presentation	energy -frequency	energy-time-frequency	energy-time-frequency
Nonlinear	no	no	yes
Nonstationary	no	yes	yes
Feature Extraction	no	discrete: no continuous: yes	yes
Theoretical base	theory complete	theory complete	empirical

frequency value. This spectrum represents the accumulated amplitude over the entire data span in a probabilistic sense.

The combination of the empirical mode decomposition and the Hilbert spectral analysis is also known as the "Hilbert-Huang transform" (HHT) for short. Empirically, all tests indicate that HHT is a superior tool for time-frequency analysis of nonlinear and nonstationary data. It is based on an adaptive basis, and the frequency is defined through the Hilbert transform. Consequently, there is no need for the spurious harmonics to represent nonlinear waveform deformations as in any of the priori basis methods, and there is no uncertainty principle limitation on time or frequency resolution from the convolution pairs based also on a priori basis.

A comparative summary of Fourier, wavelet and HHT analyses is given in the **Table1**:

This table shows that the HHT is indeed a powerful method for analyzing data from nonlinear and nonstationary processes: it is based on an adaptive basis; the frequency is derived by differentiation rather than convolution; therefore, it is not limited by the uncertainty principle; it is applicable to nonlinear and nonstationary data and presents the results in time-frequency-energy space for feature extraction.

Support Vector Machine (SVM) is one type of learning machines based on statistical learning theory. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book.[15]. SVMs have been used in a range of bioinformatics problems including protein fold recognition [16]; proteinprote in interactions prediction [17]; prediction of protein subcellular location [17, 18], protein secondary structure prediction,T-cell epitopes prediction, Classification of protein quaternary structure [19].

In this paper, we apply Vapnik's support vector machine for predicting the types of apoptosis proteins. We have used the OSU_SVM, a Matlab SVM toolbox (http://www.ece.osu.edu/~maj/osu_svm), which is an implementation of SVM for the problem of pattern recognition.

3. TRAINING AND PREDICTION

According to their subcellular location [12], apoptosis proteins are classified into the following four types: (1) type I: Cytoplasmic protein; (2) type II: Plasma membrane-bound protein; (3) type Ⅲ: Mitochondrial inner and outer proteins; (4) type Ⅳ: Other proteins (see **Table 2**).

In this research, we first translate every aminoacid sequence *s* into a numerical sequence *f* by hydrophobicity index, then, decompose it into a finite number of intrinsic mode functions (IMFs) through empirical mode decomposition (EMD) process, we just select the 2nd to 4th components (IMF2, IMF3, IMF4), because first IMF just reflects the rand composition and the last is just the trendences composition of the numerical sequence *f*. Then applying the Hilbert

Table 2. List of the acession numbers for the 98 apoptosis proteins classified into four categories according to their subcellular locations. (Type I: 43 Cytoplasmic proteins; Type II: 30 Plasma membrane-bound proteins; Type III: Mitochondrial inner and outer proteins ; Type IV: 12 Other proteins).

Type	Type I	Type II	Type III	Type IV
proteins	NP_033941, NP_033940, NP_033939, NP_031637, NP_031570, NP_031563, NP_031490, NP_033447, ,NP_036246, NP_001218, NP_004041, NP_065209, NP_001151, NP_071610,NP_071567, NP_066961, NP_037054, NP_036894, NP_005649, NP_004392, NP_004315, NP_001187, NP_001159, NP_001157, NP_001156, P55212,P42574, P39429, P55867, P22366, P55866,P55214, P55269, P29466, P55865,P29452, Q02357, O54786, Q60989,Q62210, Q60431, O70201, XP_013050,	NP_037223, NP_037275, NP_032013, NP_032612, NP_037315, NP_005916, NP_005579, NP_000034, NP_001056, NP_003781, NP_002498, NP_036742, NP_031553, NP_031549, P50555, P25118,P18519, P51867, O19131, Q63199, O77736, , O02703,Q13014, Q63690, Q07820, Q91828, Q91827, Q07812, P28825, NP_001179	P10417, P53563, Q07816, P49950, Q07817, O95831, Q9OX1, Q9JM53, Q9VQ79, O77737, Q00709, XP_008738, NP_033873,	Q63369, Q90660, Q00653, Q04861, P19838, NP_032715, P98150, Q15121, Q62048, NP_033872, NP_004040, NP_005736

a.Derived from SWISS-PROT data bank.

b.Of the 12 other apoptosis proteins, five are located in nucleus, two in endoplasmic reticulum, one in microtubule, and one in lysosome [7].

transform to each IMF component, we get the instantaneous amplitude $a_i(t)$, then get the energy value $e_i = \sum_t a_i^2(t)$, (t=2, 3, 4). Next, get its energy ratio $g_i = \frac{e_i}{e_2 + e_3 + e_4}$ .Last every protein was represented as a point or a vector in a 23-D space. The first 20 components of its vector were supposed to be the occurrence frequencies of the 20 amino acids in the protein concerned, the last three components were its energy ratio times a weight, there, we set the weight is 0.2.

The computations were carried out on a PC. Also for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension. After being trained, the hyper-plane output by the SVM was obtained. The SVM method is applied to two-class problems. In this paper, for the four-class problems, we have used a simple and effective method: "one-against-others" method [16] to transfer it into two-class problems. We first test the selfconsistency and leave-one-out cross-validation (jackknife test) of the method, followed by testing the method by prediction of an independent dataset. As a result, the rates of self-consistency, cross-validation of prediction were quite high.

In addition to the prediction algorithm, we also need to construct a training data set to complete the establishment of a statistical prediction method. To realize this, based on the SWISS-PROT data bank, 98 apoptosis proteins (the date were taken from Zhou [7]) were classified into the following four subcellular locations: (1) cytoplasmic, (2) plasma membrane-bound, (3) mitochondrial, and (4) other (**Table 1**).

4. RESULTS AND DISCUSSION

By means of the SVM algorithm described in the last section, a statistical prediction was performed for the 98 apoptosis proteins listed in **Table 2**. The prediction was conducted by two different approaches, the re-substitution test and the jackknife test. The results are given in **Table 3**.

4.1. Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of a prediction method[7]. When the re-substitution test was performed for the current study, the type of each apoptosis protein in a data set was in turn identified using the rule parameters derived from the same data set, the so-called training data set. As shown in **Table 3**, the overall success rate thus obtained for the 98 apoptosis proteins in **Table 1** was 100%, indicating an excellent self-consistency.

However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of a prediction method in practical application. This is important especially for checking the validity of a training data set-whether it contains sufficient information to reflect all the important features concerned so as to field a high success rate in application.

4.2. Jackknife test

As is well known, the independent data set test, sub-sampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most effective and objective one for a comprehensive discussion about this). During jackknifing, each protein in the data set is in turn singled out as a tested protein and all the rule parameters are calculated based on the remaining proteins. In other words, the subcellular location of each apoptosis protein is identified by the rule parameters derived using all the other apoptosis proteins except the one that is being identified. During the process of

Table 3. Tested results for the 98 apoptosis prtoeins in Table 2 by both Re-substitution test and Jackknife test. All use Gauss RBF kernel function, while the value C =15, and the gama= 80.

Test method		Success Rate				
		Type Ⅰ	**Type Ⅱ**	**Type Ⅲ**	**Type Ⅳ**	**Overall**
Re-substitute	covariant	43/43=100%	30/30=100%	9/13=60.2%	7/12=58.3%	89/98=90.8%
	SVM	42/43=97.70%	30/30=100%	13/13=100%	12/12=100%	97/98=99.0%
	HHT	43/43=100%	30/30=100%	13/13=100%	12/12=100%	98/98=100%
Jack-knife	covariant	42/43=97.7%	22/30=73.3%	4/13=30.8%	3/12=25.0%	71/98=72.5%
	SVM	39/43=91.4%	28/30=93.3%	12/13=92.5%	9/12=75.0%	88/98=89.8%
	HHT	41/43=95.3%	29/30=96.7%	12/13=96.7%	9/12=75.7	91/98=92.9%

jackknifing, both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. As expected, the success prediction rates by jackknife test were decreased in comparison with those by the re-substitution test. Such a decrement is particularly more remarkable for small subsets. This is because the cluster-tolerant capacity for small subsets is usually low. And hence the information loss resulting from jackknifing will have a greater impact on the small subsets than the large ones. Nevertheless, as shown in **Table 2**, the overall jackknife rate for the data set of the 98 apoptosis proteins could still reach 93%. It is expected that the success rate for identifying the subcellular location of apoptosis proteins can be further enhanced by improving the training data of small subsets by adding into them more new proteins that have been found belonging to the subcellular location defined by these subsets.

5. CONCLUSIONS

The above results, together with those obtained by the covariant discriminant prediction algorithm [7], have indicated that the types of apoptosis proteins are predictable with a considerable accuracy. It is anticipated that the HHT, and the SVM, if effectively complemented with each other, will become a powerful tool for predicting the types of apoptosis proteins. The current study has further demonstrated that the datasets originally constructed by Zhou[7] will be very useful for the area of apoptosis study. It is expected that the prediction quality can be further improved if the current HHT can be properly combined with pseudoamino acid composition[9] and function domine composition and with other amino acid properties.

ACKNOWLEDGMENTS

The authors thank Dr. Guo-Ping Zhou for providing the amino acid sequences of the apoptosis proteins and some helpful discussions. The work was partly supported by Huazhong Agricultural University, P. R. China

REFERENCE

[1]Zhou, P., Chou, J. J., Olea RS, Yuan, J. & G. Wagner. Solution structure of Apaf-1 CARD and its interaction with caspase-9 CARD: a structural basis for specific adaptor/caspase interaction. Proc Natl Acad Sci USA 1999, 96:11265-11270.

[2]Kerr J.F., Wyllie A. H. & A. R. Currie. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer 1972, 26:239-257.

[3]Schulz J. B., Weller M. & M. A. Moskowitz. Caspases as treatment targets in stroke and neurodegenerative diseases. Ann Neurol 1999, 45:421-429.

[4]Barinaga M. Stroke-damaged neurons may commit cellular suicide. Science 1998, 281:1302-1303.

[5]Chou, K. C. A new branch of proteomics: prediction of protein cellular attributes. Gene Cloning and Expression Technologies 2002, 4:57-70,

[6]Huang, J. & Shi, F. Support vector machines for predicting apoptosis proteins types. Acta bioinformatics 2005, 53:39-47.

[7]Zhou, G. P. & Doctor. K. Subcelluar location of Apoptosis proteins. Proteins:Structure, Function, and Genetic 2003, 50:44-48.

[8]Chou, K C. A. novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins:Structure, Function and Genetics 1995, 21:319-344.

[9]Chou, K. C. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins :Structure, Function, and Genetics 2001, 43:246-255 (Erratum:ibid., 2001, vol. 44, 60).

[10]Chou, J. J., Li, H., Salvesen G.S., Yuan, J. & G. Wagner. Solution structure of BID, an intracellular amplifier of apoptotic signaling. Cell 1999, 96:615-624.

[11]Cai, Y. D., Liu, X. J. & Chou, K. C. Artificial neural network model for predicting membrane protein types. J. Biomol. Struct. Dyn. 2001, 18:607-610.

[12]Feng, Z. P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 2001, 58:491-499.

[13]Cai, Y. D. & Chou, C. Nearest neighbour algorithm for predicting protein subcellular location by combing functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm. 2003, 305:407-411.

[14]Huang, N. E., Shen, Z., Long, S. R., Wu, M. L., H.H. Shih, Zheng, Q., N.C. Yen, C.C. Tung & Liu, H. H. The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. Proc. Roy. Soc. London A 1998, 454:903-995.

[15]Vapnik V. Statistical Learning Theory. WileyInterscience 1998.

[16]Ding, C. H. & I. Dubchak. Multiclass protein fold recognition using support vector machines and neural networks. Bioinformatics 2001, 17:349-358.

[17]Cai, Y. D., Liu, X. J., Xu, X. B.& Chou, K. C. Support vector machines for prediction of protein subcellular location by incorporating quasisequenceorder effect. J. Cell. Biochem. 2002, 84:343-348.

[18]Hua, S. J. & Sun, Z. R. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001, 17:721-728.

[19]Hua, S. J. & Sun, Z. R. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 2001, 308:397-407.