Front. Biol. China 2007, 2(3): 268–271
DOI 10.1007/s11515-007-0038-y

RESEARCH ARTICLE

A modified resonant recognition model to predict protein-protein interaction

LIU Xiang, WANG Yifei (✉)

College of Sciences, Shanghai University, Shanghai 200444, China

Abstract Proteins are fundamental components of all living cells and the protein-protein interaction plays an important role in vital movement. This paper briefly introduced the original Resonant Recognition Model (RRM), and then modified it by using the wavelet transform to acquire the Modified Resonant Recognition Model (MRRM). The key characteristic of the new model is that it can predict directly the protein-protein interaction from the primary sequence, and the MRRM is more suitable than the RRM for this prediction. The results of numerical experiments show that the MRRM is effective for predicting the protein-protein interaction.

Keywords protein-protein interaction, resonant recognition model, modified resonant recognition model, discrete wavelet transform, characteristics frequency

1 Introduction

The protein-protein interaction (P-P interaction) is the foundation of biological function diversities. Thousands of proteins participate in the regulation of the metabolic pathway, immunological recognition and DNA replication, and one of the foundations to realize these biological functions is a P-P interaction. However, the final goal of discovering the P-P interaction is to reveal the biological functions of all the proteins. To achieve this goal, we must first predict the P-P interaction accurately. Based on these interactions, we can then construct the P-P interaction network. And then the metabolic pathway can be concluded from the network; finally, the protein function can be predicted by the metabolic pathway. Therefore predicting P-P interaction is the most basic and the most important part of this final goal.

There are some experimental approaches trying to predict the P-P interaction, such as the Yeast-Two-Hybrid System (Chafia, Qing and Cosic, 2002), which can be applied to predict the dynamic interaction between proteins (Joel and David, 2001; Uetz et al., 2000). However, all these approaches are tedious, labor intensive and inaccurate (Enright et al., 1999), forcing researchers to find computational approaches to predict the P-P interaction. Many research groups have also exploited many computational approaches based on different principles in order to predict the P-P interaction, such as the approach based on the interaction protein sequences (Pazos et al., 1997); the approach based on the genomic context to predict the protein interaction (Huynen et al., 2000); or the approach predicting the interaction by searching the homological sequences in the genome (Marcotte et al., 1999) and also using the Support Vector Machine (SVM) based on the primary structure of proteins (Bartel and Fields, 1997).

Since all of the protein information is incorporated in the primary structure, we proposed prediction of the P-P interaction from the protein sequence itself. Based on E.W. Prohofsky's solid-state physics experiment (Prohofsky, 1987), Cosic designed the Resonant Recognition Model (RRM) (Cosic, 1999). This paper introduces the Wavelet Transform (WT) into the RRM to create the Modified Resonant Recognition Model (MRRM). The character of MRRM is to predict the P-P interaction directly from the protein sequence.

2 Materials and methods

To check the effectiveness of the MRRM approach, 50 pairs of interaction proteins from DIP (Database of Interaction Protein, http://dip.doe-mbi.ucla.edu) and 12 pairs of interaction proteins from PDB (http://www.rcsb.org/pdb/) are downloaded. These 62 pairs of proteins are used as the test set of the MRRM.

2.1 Resonant recognition model

The resonant recognition model is a physical-mathematical model, which assumes that the recognition (communication)

Translated from *Journal of Shanghai University* (*Natural Science*), 2006, 12(1): 69–73 [译自: 上海大学学报 (自然科学版)]

E-mail: yifei_wang@staff.shu.edu.cn

between biological molecules (including protein and DNA) is implemented by the resonant energy transfer. The RRM introduces some physical parameters of the amino acids and Digital Signal Analysis (DSP) approaches such that the analysis of protein and gene becomes global.

The RRM comprises three major steps as follows.

1) The protein sequence is converted into a numerical sequence by assigning each amino acid residue a value of Electron-Ion Interaction Pseudopotential (EIIP) (Ladik, 1974).

2) The numerical sequence is analyzed by the Discrete Fourier Transform (DFT) to extract the information about the biological function.

3) To find the common character of a group of proteins, the multiple cross spectrum function is introduced to calculate the strength of energy at each frequency in the RRM.

The interaction between proteinase and proteinase-inhibitor (Cosic, 1999) has been predicted by the RRM quite well. However, this method requires that the predicted proteins have an amount of homological sequences, bringing an obstacle to the prediction. Thus the RRM is modified to overcome this problem.

2.2 Modified resonant recognition model

In the RRM, the main purpose of searching the homological sequences is to extract the characteristic frequency of the proteins, i.e., to magnify common details of the proteins. If the homological proteins are lacking, then the RRM is ineffective. To exceed this limitation, the Discrete Wavelet Transform (DWT) (Daubechies, 1988; Daubechies, 1992) is adopted in the MRRM. By using the DWT to pre-process the protein sequences, the information of the protein at different levels is extracted, and then the information is subjected to spectrum analysis. The EIIP value is acquired by an ideal model, so the EIIP is substituted with the Ionization Constant (IC), which is acquired by biological experiment (Pirogova, 1999) to achieve a higher reliability.

The computational procedure of the MRRM comprises the following six major steps.

1) The two protein sequences are both assigned the IC value in order to acquire the numerical sequences, and the IC value is shown in Table 1.

2) These two numerical sequences are processed by the DWT respectively. The Biorthogonal Wavelet 3.3 (Fig. 1) at level 3 (Fig. 2) is adopted in the MRRM. Then the original numerical sequence was divided into A3 (Approximation at level 3), D3 (Detail at level 3), D2 (Detail at level 2) and D1 (Detail at level 1), where D3, D2 and D1 contain the detailed signals at different levels, and A3 just contains a little low-frequency signal.

The characteristic frequency is searched within the four different levels above.

3) Assume that the four level numerical signals of two proteins are $S^{K,p}$, $K = A3,D3,D2,D1$ was obtained. By using the DFT, these numerical signals are transformed into the time

Table 1 Ionization constant of amino acids

Amino acid	IC value
Leu	2.40
Ala	2.30
Asn	2.20
Cys	1.96
Ile	2.40
Pro	2.00
Trp	2.37
Thr	2.09
Tyr	2.20
His	2.30
Gln	2.06
Phe	1.98
Gly	2.46
Lys	2.20
Met	2.17
Arg	1.82
Val	2.35
Glu	2.30
Ser	2.10
Asp	1.88

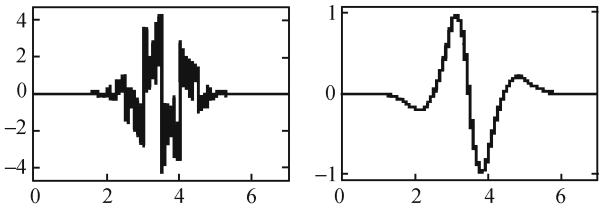

Fig. 1 Biorthogonal Wavelet 3.3

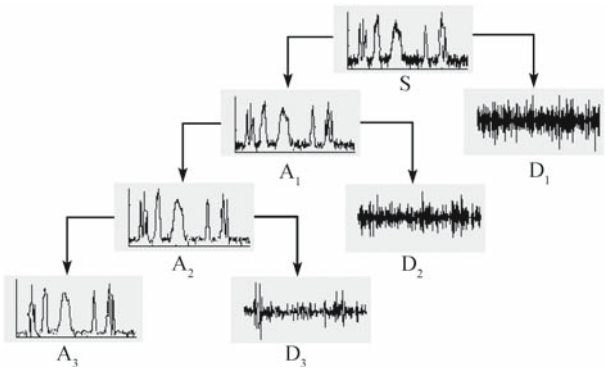

Fig. 2 Using discrete wavelet to transform the signal at level 3

phase to acquire the spectrum information of these signals. If the lengths of the corresponding signals are not equal, we append enough '0's to the terminal of shorter ones to make them the same length and maintain the resolution of the spectrum sequences. Thus, with each element of the DFT coefficient vector as $C_n^{K,p}$, $K = A3,D3,D2,D1$, $p = 1,2$ and the corresponding phase $\varphi_n^{K,p}$, $K = A3,D3,D2,D1$, $p = 1,2$ can be calculated by formulas (2.2.1) and (2.2.2)

$$C_n^{K,p} = \sum_{m-1}^{M} S_m^{K,p} \exp(-\frac{2\pi mn}{N} i) \qquad (2.2.1)$$

$$C_n^{K,p} = \left|C_n^{K,p}\right| \exp(-i\varphi_n^{K,p}) \quad (2.2.2)$$

$$n = 1,2,...N/2;\ p = 1,2;\ K = A3,D3,D2,D1$$

where $S_m^{K,p}$ is the mth element of the pth numerical sequence at level K, N is the length of the sequence, i is the image unit. The DFT coefficients represent the original signal by its amplitude, frequency and phase. The modulus of the DFT coefficients is the amplitude of the signal, namely $\left|C_n^{K,p}\right|$.

4) The DFT coefficients vector is normalized by formula (2.2.3)

$$NOM_n^{K,p} = \frac{C_n^{K,p}}{\max\limits_n(C_n^{K,p})} \quad (2.2.3)$$

$$n = 1,2,...N/2;\ p = 1,2;\ K = A3,D3,D2,D1$$

where $NOM_n^{K,p}$ is the normalized sequence, and p is the sequence number of the proteins.

5) The cross-spectrum function is used to calculate the cross-spectrum coefficients CV_n^K

$$CV_n^K = \prod_{p=1}^{2} NOM_n^{K,p} \quad (2.2.4)$$

$$n = 1,2,...N/2;\ p = 1,2;\ K = A3,D3,D2,D1$$

6) The characteristic frequency is searched within

$$CV_n^K,\ K = A3,D3,D2,D1,\quad n = 1,2,...N/2$$

The characteristic frequency must satisfy the following condition.

If $f = F$ is characteristic frequency in level K, the cross function $CV_F^K > 0.36$, and the phase $2 < \left|\varphi_F^{K,1} - \varphi_F^{K,2}\right| < 4$.

The threshold value is set to 0.36, which means that the strength of the energy at this frequency is greater than 0.36 for both proteins and the average energy strength of the two proteins is greater than 0.6.

To find the corresponding characteristic frequency, this progress is repeated to the numerical sequences of A3, D2 and D1. According to previous research and numerical experiments (Prohofsky, 1987), the information relevant to the P-P interaction is mostly contained in D3, D2 and D1. However, information in A3 is mainly about the frame information of protein, so the characteristic frequency can be found in D3, D2 and D1. Consequently, the common characteristic frequency of the two proteins can be discovered in levels D3, D2 and D1, enabling the interaction between the two proteins. The criteria are as follows.

1) Satisfactory conditional characteristic frequencies can be found in two levels in D3, D2 and D1.

2) There is no characteristic frequency in A3 because the signal in A3 lacks interaction information.

The MRRM approach is implemented with the Wavelet Toolbox of MATLAB software—a product of the MathWorks company.

3 Results and discussion

The MRRM is employed to predict the previous 62 pairs of interaction proteins. The analysis results of dip:5446e are shown in Fig. 3. As the tertiary structure of PDB sequences has been determined, the predicting results of 12 PDB interaction pairs are detailed in Table 2.

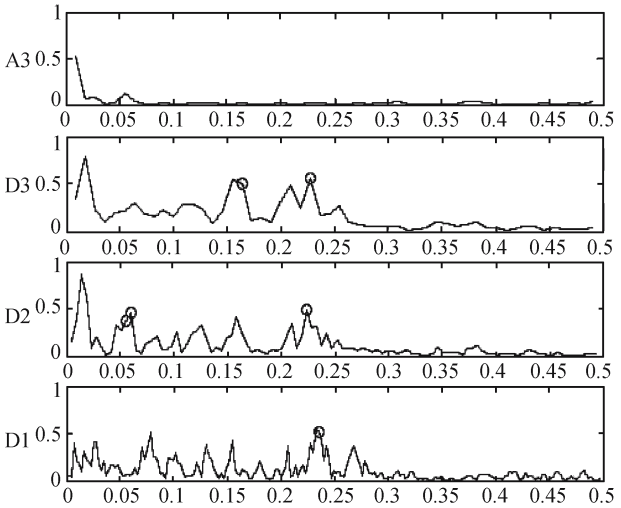

Fig. 3 Analysis results of dip:5446e (the characteristic frequency are marked by 'o')

Table 2 Predicting results of twelve PDB entries (P-Positive, N-Negative)

PDB code	A3	D3	D2	D1
1ATN	N	P	P	P
1CHO	N	P	P	P
1CSE	N	P	P	P
1HRP	N	P	P	N
1LPA	N	P	P	N
1MCT	N	P	P	P
1STF	N	P	P	N
1TAB	N	P	P	P
1TGS	N	P	P	P
2BTF	N	N	N	N
2PTC	N	P	P	P
2SIC	N	P	N	N

From the above table, it is clear that there is no characteristic frequency in level A3. There are only two entries where the characteristic frequency occurs only in one of the three levels, setting prediction accuracy at 83%.

The statistical predicting results of the above 62 pairs of interaction proteins are reported as follows. In all the 62 pairs, there are 34 pairs where the characteristic frequency occurs in the three levels D3, D2 and D1, which is 54%. There are 16 pairs where the characteristic frequency occurs in two levels of D3, D2 and D1, which is 25%. However, there are 10 pairs where the characteristic frequency only occurs in one of D3, D2 and D1, which is 16%. Only two pairs have no characteristic frequency, which is only 3%. Thus, prediction accuracy is about 79% (Fig. 4).

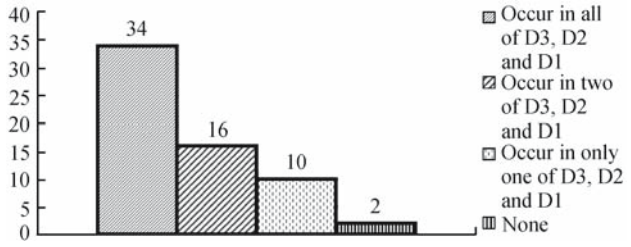

Fig. 4 Statistics of 62 interactions

Recently, computational approaches to predict the P-P interaction have emerged. Most approaches are based on the protein domain theory, and some other approaches require the tertiary structure of proteins. However, the structure of most proteins remains unresolved. Since information about the protein function and its structure is contained in the primary structure, why not predict the P-P interaction from their primary structures? This character is just the advantage of the MRRM—the approach predicts the P-P interaction from the primary structure directly. Since the RRM can predict the protein domain itself (Cosic, 1999), the MRRM does not need the additional information about the protein domain. The algorithm of the MRRM is simple, effective, and feasible. Because of these new characters, the MRRM is much more suitable for the prediction of the P-P interaction.

Although there are mistakes in prediction, it is believed that with the development of the model and the adoption of the biological wavelet, it is able to enhance prediction precision and prevent the occurrence of false positive prediction. Thus, we believe that the method will serve as a useful tool for studying the P-P interaction.

Acknowledgements This work was supported by National High-Tech Research and Development Program of China (No. 2002AA234021).

References

Bartel P L, Fields S (1997). The yeast two-hybrid system. In: Advances in Molecular Biology. NewYork: Oxford University Press

Chafia H T, Qing F, Cosic I (2002). Protein sequence comparison based on the wavelet transform approach. Protein Engineering, 15(3): 193–203

Cosic I (1999). The resonant recognition model of macromolecular bioactivity. BioMethod, 8:

Daubechies I (1988). Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math, 41(7): 909–996

Daubechies I (1992). Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics. Philadephia.

Enright A J, Ililopoulos I, Kyrpides N C, Ouzounis C A (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature, 402(6757): 86–90

Huynen M, Snel B, Lathe W, Bork P (2000). Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Res, 10(8): 1204–1210

Joel R B, David A G (2001). Predicting protein-protein interactions from primary structure. Bioinformatics, 17(5): 455–460

Ladik J (1974). All valence electron band structures of simple periodic protein models. Int J Quantum Chemistry Quantum Biol Symp, 1: 65–69

Marcotte E, Pellegrini M, Ng H L, Rice D W, Yeates T O, Eisenberg D (1999). Detecting protein function and protein-protein interactions from genome sequences. Science, 285(5428): 751–753

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997). Correlated mutations contain information about protein-protein interaction. J Mol Biol, 271(4): 511–523

Pirogova E, Cosic I (1999). Proc.of IEEE EMBS VIC Australia, 203–206

Prohofsky E W (1987). Vibrational modes of a DNA polymer at low temperature. Physical Review B, 36(6): 3449

Uetz P, Giot L, Cagney,G, Mansfield T A, Judson R S, Knight J R, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin,B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J M (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403(6770): 623–627