Chinese Chemical Letters Vol. 16, No. 8, pp 1133-1136, 2005 1133
http://www.imm.ac.cn/journal/ccl.html

A New Hybrid Model of Amino Acid Substitution for Protein Functional Classification

Ke Long WANG[1,2], Zhi Ning WEN[1,2], Fu Sheng NIE[1,2], Meng Long LI[1,2]*

[1] College of Chemistry, Sichuan University, Chengdu 610064
[2] State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha 410082

Abstract: In this paper, a new hybrid model of amino acid substitution is developed and compared with the others in previous works. The results show that the new hybrid model can characterize the protein sequences very well by calculating Fisher weights, which can denote how much the variants contribute to the classification.

Keywords: Hybrid model of amino acid substitution, protein functional classification, Fisher weights.

In recent years, methods of protein sequences analysis have been gradually evolved into two directions: One is based on the models of probability and statistics[1-4]; the other is based on the digital signal processing technologies [5-8]. The latter mainly converts the protein character sequences into digital signals and uses some signal processing methods to analyze them, *i.e.*, fast Fourier transform (FFT). However, it is still unsolved how to characterize the protein sequences accurately with digital signals.

To tackle this problem, researchers have designed some substitution models. But all of those works only consider one single feature from many of what can contribute to the function of the protein, such as physico-chemical properties[6], mutability[9], and electron-ion interaction potential[8] (EIIP). A new hybrid model of amino acid substitution developed in this paper combines three physico-chemical properties with three structural parameters of amino acid, including polarity (*p*), volume (*v*), component (*c*)[10], electron-ion interaction potential (*E*), hydrophobicity (*h*)[11] and relative stability of α helix conformation (*s*)[12]. We use the range scaling forms of these values calculating as

$$X_i^{'} = \frac{X_i - X_{i(\min)}}{X_{i(\max)} - X_{i(\min)}} \qquad (1)$$

where X_i is the vector of the i^{th} character of 20 amino acids, $X_{i(\min)}$ is the minimum value in X_i and $X_{i(\max)}$ is the maximum value in X_i. Then an amino acid is

* E-mail: liml@scu.edu.cn

represented as

$$AA(a) = \sum_{i=1}^{n} X_i^{'}(a) \qquad (2)$$

where $X_i^{'}(a)$ is the range scaling form of the i^{th} property value of amino acid *a*, *i.e.*, $v^{'}(a)$. So a protein sequence can be converted to a digital signal and transformed by fast Fourier transform (FFT). Then, we can calculate the Fisher weights with the power spectral density (PSD) of two protein sequences for functional classification. The Fisher weights defined as

$$F = \frac{\left(\bar{x}_{PSD,1} - \bar{x}_{PSD,2}\right)^2}{S_{PSD,1} + S_{PSD,2}} \qquad (3)$$

where $\bar{x}_{PSD,1}$, $\bar{x}_{PSD,2}$ denote the power spectra's average values of class 1 and class 2 respectively, and $S_{PSD,1}$, $S_{PSD,2}$ denote the power spectra's standard deviation of class 1 and class 2 respectively.

In Fisher weights calculation, the less the differences of the character in two classes are, the smaller the Fisher weight value is. Therefore, in the functional similarity comparison of the protein sequence pairs, the Fisher weight will be small when two protein sequences are functionally similar and the Fisher weight will be large when they are dissimilar. **Table 1** shows the comparing results of the Fisher weights of three substitution models. The hybrid model performs very well when the similarity in the protein sequence pairs decreases. The model formed with physico-chemical properties lacks stability and the model formed with structural parameters does not well represent the protein sequences.

Table 1 Results of applying Fisher weights method to five protein sequences pairs with three substitution models compared with T-COFFEE score

Protein sequence pairs	Origin of the proteins	T-COFFEE Score	Fisher weights		
			Hybrid model	Physico-chemical properties	Structural parameters
Nif-specific regulatory protein/ Nif-specific regulatory protein	Rhodobacter capsulatus/ Azo-rhizobium caulinodans	99	2.29e-005	6.14e-005	2.05e-004
Lysozyme/ Lysozyme	Bacteriophage N15/ Bacterio-phage PA-2	97	4.29e-004	9.37e-004	4.93e-004
Pol polyprotein/ Pol polyprotein	Human immunodeficiency *virus* type 1 (ELI isolate) (HIV-1)/ Human immuno-deficiency *virus* type 1 (MN isolate) (HIV-1)	97	1.99e-004	5.37e-006	4.49e-006
NGFI-A binding pro-tein 1/ Phosphoribosyl-amine-glycine ligase	Rattus norvegicus/ Lactococcus lactis	50	0.0032	0.0024	5.97e-004
Phenylalanyl-tRNA synthetase beta chain/ Nonstructural protein NS-S	Rhizobium loti/ Maguari virus	49	0.0261	0.0769	0.0332

Figure 1 Fisher weight values in HTH-type transcriptional repressor (a) and Lysozyme (b). All the values are calculated by comparing the sequences one with each other in the two test sets

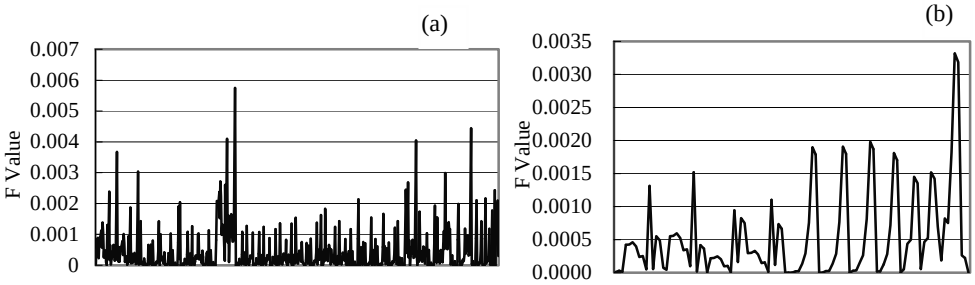
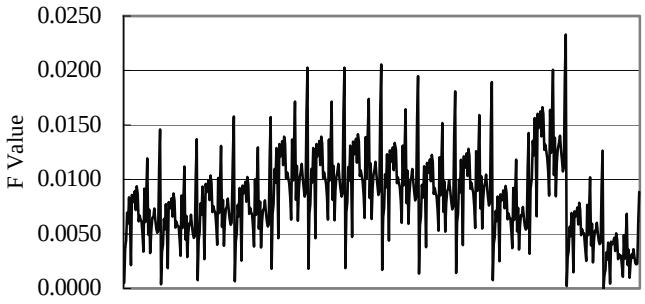


Figure 2 Fisher weight values between the two test sets.



The F values are calculated by comparing one of the sequences in HTH-type transcriptional repressor with those in Lysozyme.

In order to test the new hybrid model's ability of classifying the different protein sequences, two protein sequence sets are chosen in our paper. One is HTH-type transcriptional repressor including 37 protein sequences and the other is Lysozyme including 15 protein sequences. The former is a kind of transcriptional repressor that binds to some operator, *e.g.* the *purF* operator, or co-regulates other genes for de novo purine nucleotide synthesis and the latter is essential for lysis of bacterial cell wall by showing cell wall hydrolyzing activity. All of them come from SWISS-PROT release 44.0 and are selected by BLASTp release 2.0. **Figure 1a** and **Figure 1b** show the calculation results of Fisher weight values in HTH-type transcriptional repressor and Lysozyme respectively. The Fisher weight values between the two test sets are showed in **Figure 2**. Given a lower limit value, *e.g.* 0.003, we can attain recognition accuracy of 99% in HTH-type transcriptional repressor and that of 99% in Lysozyme. As for the case between the two test sets, the recognition accuracy is 92%.

In conclusion, the new hybrid model is superior to the model only consisting of physico-chemical properties of amino acids and the model only combining the structure information. Moreover, in protein functional classification, the new hybrid model can characterize the protein sequences well in frequency domain and achieve high recognition accuracy.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No. 29877016).

References

1. H. O. Smith, T. M. Annau, S. Chandrasegaran, *Proc. Natl. Acad. Sci.*, **1990**, *87*, 826.
2. P. C. Ng, S. Henicoff, *Genome Res.*, **2001**, *11*, 863.
3. S. Sunyaev, V. Ramensky, I. Koch, *et al.*, *Hum. Mol. Genet.*, **2001**, *10*, 591.
4. A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, *Nucleic Acids Res.*, **2002**, *30*, 2478.
5. A. J. Mandell, K. A. Selz, M. F. Shlesinger, *Physica A*, **1997**, *244*, 254.
6. K. Katoh, K. Misawa, K. Kuma, T. Miyata, *Nucleic Acids Res.*, **2002**, *30*, 3059.
7. J. D. Qiu, L. P. Liang, X. Y. Zou, J. Y. Mo, *Talanta*, **2003**, *61*, 285.
8. C. Hejase de Trad, Q. Fang, I. Cosic, *Protein Eng.*, **2002**, *15*, 193.
9. J. Majewski, J. Ott, *Gene*, **2003**, *305*, 167.
10. R. Grantham, *Science*, **1974**, *185*, 862.
11. J. L. Fauchere, *Eur. J. Med. Chem. Chim. Ther.*, **1983**, *18*, 369.
12. K. T. O'Neil, *Science*, **1990**, *250*, 646.

Received 6 September, 2004