J Math Chem (2009) 46:562–568
DOI 10.1007/s10910-008-9480-6

ORIGINAL PAPER

Feature analysis of protein structure by using discrete Fourier transform and continuous wavelet transform

Shengli Zhang · Tianming Wang

Received: 10 July 2008 / Accepted: 6 October 2008 / Published online: 25 October 2008

Abstract In this paper, discrete Fourier transform (DFT) and continuous wavelet transform (CWT) are used to predict the protein structure. Hydrophobicity plays a key role in the form of protein structure. The amino acid sequence is first mapped into hydrophobicity sequence, and then process it by DFT and CWT so that power spectral density is gained. The results show that continuous wavelet transform can extract the features of protein structure effectively and availably and has a tremendous development foreground.

Keywords Discrete Fourier transform · Continuous wavelet transform · α-Helices · Hydrophobicity · Power spectral density

1 Introduction

Protein is composed of amino acids, and it is the amino acid sequence that determines the chemical structure of protein. The biological function of a protein depends mostly on its spatial structure, so the identification of the spatial structure of protein is the basis of the study in its biological function. The three-dimensional structure of a protein is uniquely dictated by its primary sequence (the amino acid sequence), the so-called primary structure. The interaction among the components of amino acid composition is an important driving force in determining the structural class of a protein during the sequence folding process [1]. Therefore, we are able to predict

S. Zhang (✉) · T. Wang
Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China
e-mail: shengli0201@163.com

T. Wang
e-mail: wangtm@dlut.edu.cn

protein structure from its amino acid sequence. The chain can form secondary structure with the hydrophobicity of amino acid. Hydrogen bonds between amino acids in the sequence also play important roles in forming secondary structures. These secondary structures then fold into a three dimensional structure [2].

Different side chains make different amino acid, and different amino acids have different properties, among various properties hydrophobicity of amino acid sequence is the most crucial factor in influencing the stability of protein structure. It is well known to determine protein secondary structure such as α-helix, β-sheet, turn, loop and so on. Kauzmann [3] had established the work that hydrophobic interactions were a driving force of protein folding, hydrophobic residues did occur on the molecular surface and hydrophilic residues could be found in the interiors. Hirakawa [4] predicted hydrophobic cores of proteins by utilizing discrete wavelet transform (DWT). Mandell [5–7] used Morlet wavelet transform of protein hydrophobicity sequences to suggest their memberships in protein structural families. In this paper, we adopt discrete Fourier transform (DFT) and continuous wavelet transform (CWT) to realize the analysis and prediction of protein structure by utilizing the hydrophobicity of amino acid sequence.

2 Discrete Fourier transform and continuous wavelet transform

The Fourier transform provides the means of transforming a signal defined in the time domain into one defined in the frequency domain. The discrete Fourier transform is used in the case where both the time and the frequency variables are discrete (which they are if digital computers are being used to perform the analysis).

Let $x(j)$ represent the discrete time signal, and let $F(k)$ represent the discrete frequency transform. The discrete Fourier transform is given by

$$F(k) = \sum_{j=1}^{N} x(j)\omega_N^{(j-1)(k-1)} \tag{2.1}$$

where

$$\omega_N = \exp(-2\pi i/N)$$

where N is the length of signal sequence.

Wavelet transform, the signal processing tool, has been applied widely in many fields since it was established. It has the property of time–frequency domain localization, in which time and frequency domain can be changed arbitrary. That is, in low frequency region it has high resolution to the frequency and low resolution to the time, and in high frequency region it has low resolution to the frequency and high resolution to the time [8,9]. Continuous wavelet transform has continuous scales which make it more abundant and comprehensive in extracting signals than that of discrete wavelet transform. The formula is given below

$$W_f(a,b) = |a|^{-1/2} \int f(t)\overline{\Psi\left(\frac{t-b}{a}\right)}\,dt \quad a,b \in R, a \neq 0. \tag{2.2}$$

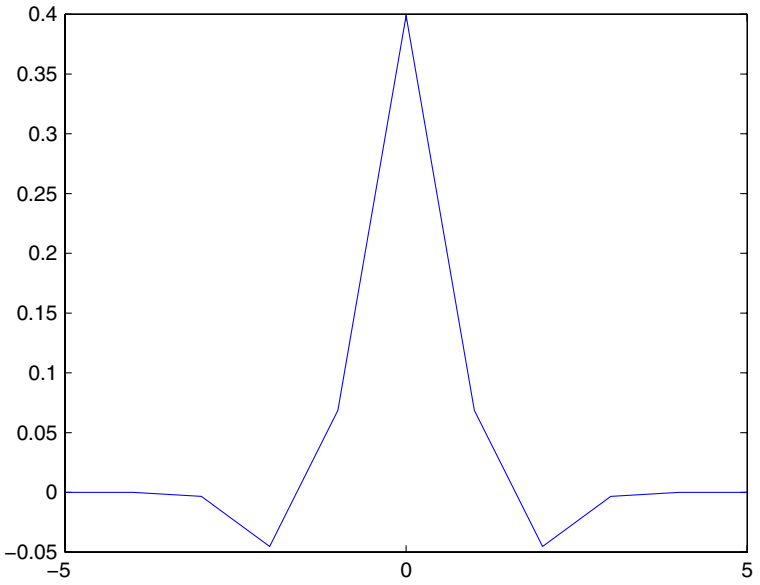

Fig. 1 Morlet wavelet

where a is the scale factor, b is the shift factor, $f(t)$ is the original signal, $\Psi(t)$ is wavelet core function. Considering the local performance in time and frequency domain, the Morlet wavelet function is adopted in this article, the definition is given below:

$$\Psi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \cos 5t \tag{2.3}$$

The Morlet wavelet is shown in Fig. 1.

3 Methods and steps

According to the contents and rankings of the α-helices and β-sheets in the protein secondary structure, we can sort the protein structure into five classes: all α-class, all β-class, $\alpha + \beta$-class, α/β-class and irregular class. Different classes have different typical features, we adopt DFT and CWT to process the hydrophobic value sequences to pick up the high-frequency parts which responds to the features of protein secondary structure, then through the frequency–spectrum graph we can analyze the periodic features of α, β, α, $\alpha + \beta$, α/β-class and the influences of the hydrophobicity in protein structure.

Firstly, we map protein amino acid sequences into hydrophobicity sequences. Hydrophobic value reflects the interaction between residues or the force between water molecule and amino acid. There are some methods to calculate the hydrophobic value

Table 1 The hydrophobic values of 20 amino acids

Amino acid	Hydrophobicity	Amino acid	Hydrophobicity
A	0.62	M	0.64
C	0.29	N	−0.85
D	−1.05	P	0.12
E	−0.87	Q	−0.78
F	1.19	R	−1.37
G	0.48	S	−0.18
H	−0.4	T	−0.05
I	1.38	V	1.08
K	−1.35	W	0.81
L	1.06	Y	0.26

[5,10], we choose the Eisenberg hydrophobic values of the 20 amino acids showed in Table 1.

Secondly, we apply the fast Fourier transform (FFT) to hydrophobicity sequences to obtain the frequency-spectrum graph according to the formula (2.1). FFT is simply a class of special algorithms which implement the DFT with considerable savings in computational time.

Lastly, we process the hydrophobicity sequences at a small scale with CWT according to the formula (2.2) and, subsequently, apply FFT algorithm to the new signal sequences. From the figures we can analyze and abstract the features of protein secondary structure.

4 Results and discussion

From the Brookhaven Protein Data Bank (PDB) database, we randomly choose 150 protein sequences of four classes respectively as the examples to analyze the features by using DFT and CWT. The similarity between sequences is less than 40%. First we can find the power spectral density of each protein sequence, then average statistically all protein sequences of four classes. It is difficult to gain the relationships when we analyze each protein sequences of each class, so we research protein sequences by using the average statistically. This can enhance the relationship between protein sequences if the sequences which belong to the same class have strong relationship at the same period. Figures 3 and 5 show the power spectral density of α-class and β-sheet protein sequences, respectively. Figures 2 and 4 show the power spectral density of α-class and β-sheet protein sequences through wavelet transform, respectively.

Compared with the results in Figs. 3 and 5, we can find the results in Figs. 2 and 4 are better:

1. As Figs. 2 and 3 show, there is an aiguille around 3.6, this corresponds to the periodic feature of α-class amino acid sequence. Because the α-helices of protein

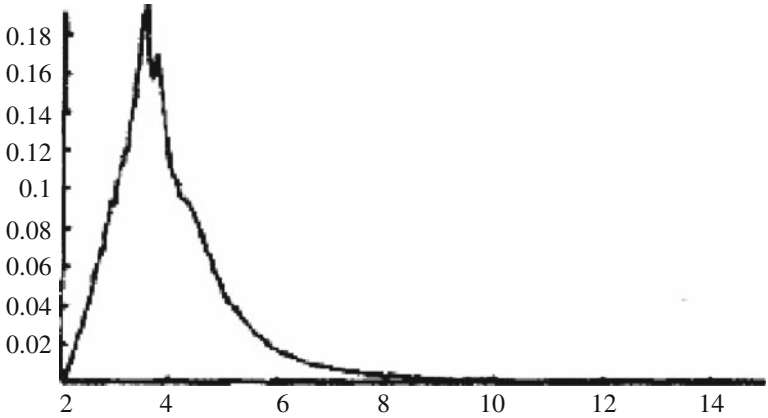

Fig. 2 Power spectral density of α-class protein sequence through continuous wavelet transform

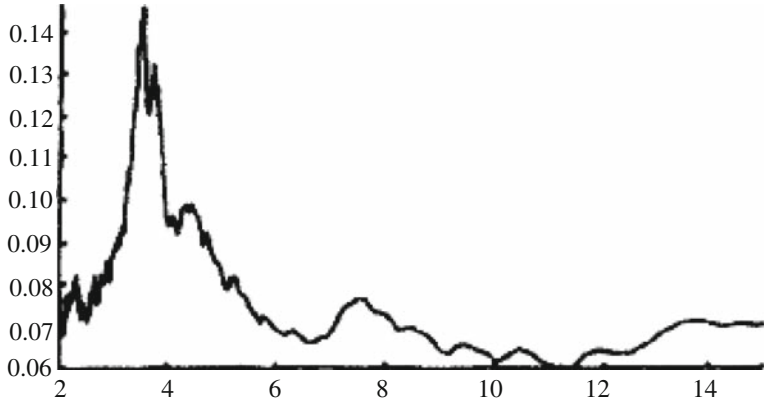

Fig. 3 Power spectral density of α-class protein sequence

is composed entirely of turns which are the hydrophobic modes of $\omega_{-1} \approx 3.6$ amino acids per turn [11].

2. The aiguille is more obvious in Fig. 2 about the hydrophobic sequences of α-class through continuous wavelet transform. This implies the continuous wavelet transform can pick up the features of protein structure effectively and availably.
3. Because the bigger the hydrophobic value is, the stronger the hydrophobic features of amino acid sequences are, we can see that residues in α-helix are most hydrophobic while in β-sheet involves both hydrophobic and hydrophilic residues.
4. This similar features about protein sequences of all β-class is shown in the Figs. 4 and 5, its aiguille appears around 2.0, because β-sheet is an another conformation unit in protein structure, it only has two amino acids per turn.
5. Figure 5 shows there are some other maximum points besides the point 2.0, this does not correspond to the structural feature of β-sheet.

We can process the amino acid sequences of $\alpha + \beta$-class and α/β-class according to the same method, they have the features of both α-helix and β-sheet.

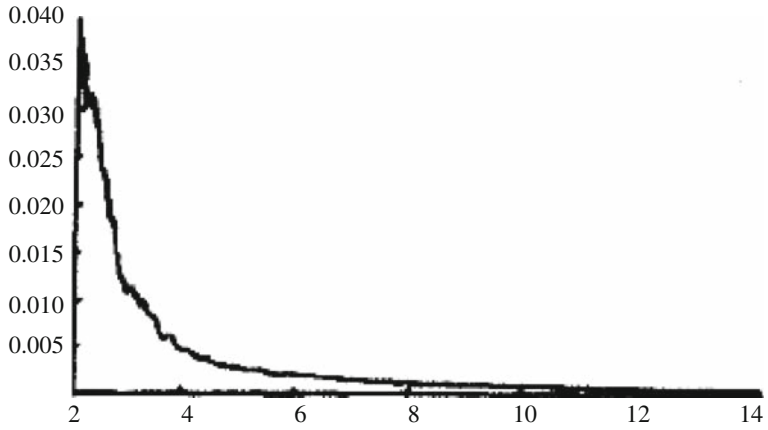

Fig. 4 Power spectral density of β-class protein sequence through continuous wavelet transform

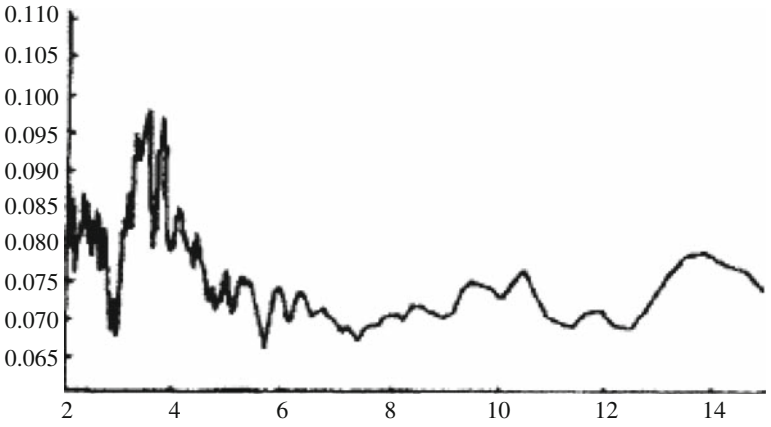

Fig. 5 Power spectral density of β-class protein sequence

Although the prediction of our method is fine, there are still some problems. On the one hand, in order to investigate the period and the relationships about the α-helix and β-sheet, we can process them based on the statistic methods. If the amino acid sequences belong to the same protein class, the periodic feature will strength so that we can see the aiguilles clearly in the graphs. On the other hand, although hydrophobicity of amino acid is the most important factor to determine the stability of protein structure, it is not the only factor, besides hydrophobicity, there are hydrogen bonds, salt linkage, disulfide bonds at the interior of peptide chains. Also, we can improve the hydrophobic value, the hydrophobicity of amino acids in different types of protein secondary structure must be statistically calculated, respectively. If we can integrate the all factors which influence the protein structure, our results will more credible and convictive.

Acknowledgement This work was supported by the National Natural Science Foundation of China (Grant No.10571019).

References

1. K.C. Chou, Biochem. Biophys. Res. Commun. **264**, 216 (1999)
2. D.W. Mount, *Bioinformatics sequence and genome analysis* (Cold Spring Harbor Laboratory Press, 2002), pp. 440–460
3. W. Kauzamnn, Adv. Protein Chem. **14**, 1 (1959)
4. H. Hirakawa, S. Muta, S. Kuhara, J. Bioinform. **15**, 141 (1999)
5. A.J. Mandell, K.A. Selz, M.F. Shlesinger, Phys. A **244**, 254–262 (1997)
6. A.J. Mandell, K.A. Selz, M.F. Shlesinger, J. Phys. Chem. B **104**, 3953 (2000)
7. K.A. Sell, A.J. Mandell, M.F. Shelesinger, Biophys. J. **75**, 2332 (1998)
8. J. Qiu, R. Liang, X. Zou, J. Mo, Talanta **61**, 285–293 (2003)
9. Y. Lou, Bull. Sci. Technol. **22**, 2 (2006)
10. D. Eisenberg, R.M. Weiss, T.C. Terwilliger, W. Wilcox, Faraday Symp. Chem. Soc. **17**, 109–120 (1982)
11. D. Eisenberg, R.M. Weiss, T.C. Terwilliger, Proc. Natl. Acad. Sci. USA **82**, 140 (1984)