Prediction of Hydrophobic Cores of Globular Proteins Using Multiresolution Analysis of Wavelet Analysis

Hideki Hirakawa **Satoru Kuhara**
hirakawa@grt.kyushu-u.ac.jp kuhara@grt.kyushu-u.ac.jp
Graduate School of Genetic Resources Technology, Kyushu University,
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: wavelet analysis, multiresolution analysis, hydrophobic cores, hydropathy index, globular proteins, PDB, prediction

1 Introduction

The hydrophobic interaction is important factor for constructing a protein structure. In a globular protein structure, the hydrophobic residues tend to be assembled in interior, and the hydrophilic residues tend to be distributed on surface. But, few hydrophobic and hydrophilic residues are distributed in interior and on exterior, respectively. The hydrophobic cores are constructed by these buried residues. The information about the hydrophobic cores is considered to be efficient for structural prediction. To predict the hydrophobic cores, several methods have been devised using sequential alignment or not. With increasing the protein structures, the proteins whose amino acid sequences are not homologous to the amino acid sequences of known structures. Therefore, the method without sequential alignment is considered to be important, but their accuracy is not enough to commonly use. One of the method without sequential alignment is using smoothed hydrophobicity plot assigned the hydropathy index to amino acid residues [1].

Wavelet analysis is a mathematical method used to obtain low and high frequency from a given signal, and is used in signal/image processing, and earthquake prediction and so on. With wavelet analysis, one can use approximating functions that are contained neatly in finite domains. Wavelet analysis is well-suited for approximating data from given functions [2]. We attempted to apply the wavelet analysis to decompose the hydrophobicity plot.

In this study, we used the 67 proteins divided into 13 families selected by Wako and Blundell [3]. We investigated the relation between the continuity of the buried residues and high and low frequencies at each frequency level extracted from the hydrophobicity plot. Then we predicted the hydrophobic cores using the high and low frequencies at each frequency level.

2 Method and Results

Wavelets are local in both frequency/scale (via dilations) and in time (via translations). The most important concept in a discrete wavelet theory is multiresolution analysis. A raw function is decomposed to a low frequency and a high frequency. The wavelet function $\psi(t)$ constructed using scaling function $\phi(t)$ which satisfies the following two-scale relation

$$\phi(t) = \sum_n h_n \sqrt{2}\phi(2t-n). \tag{1}$$

Let Z be a set of integers. The coefficients h_n denote a low-pass filter (H). The wavelet function $\psi(t)$ is defined using the scaling function $\phi(t)$ as

$$\psi(t) = \sum_n g_n \sqrt{2}\phi(2t-n). \tag{2}$$

The coefficients g_n denote a high-pass filter (G). Assume that the shifted scaling functions $\{\phi(t-k),\ k \in Z\}$ and the shifted wavelet functions $\{\psi(t-k),\ k \in Z\}$ are orthonormal, respectively. Let $\{c_l^{(0)}\}$ denote a sequence of hydrophobicity values, and we define a linear combination $f(t)$ of the sequence with the scaling functions $\{\phi(t-k),\ k \in Z\}$:

$$f(t) = \sum_k c_k^{(0)} \psi(t-k). \tag{3}$$

By a wavelet theory, we have another expansion of $f(t)$:

$$f(t) = \sum_k c_k^{(-1)} \frac{1}{\sqrt{2}} \psi(2^{(-1)}t - k) + \sum_k d_k^{(-1)} \frac{1}{\sqrt{2}} \phi(2^{(-1)}t - k). \tag{4}$$

From Eqs (3) and (4) and using orthonormality of the scaling and wavelet functions, we can decompose the sequence $\{c_l^{(0)}\}$ into low frequency and high frequency components. Repeated applications of this decomposition lead us to

$$c_k^{(j-1)} = \sum_l h_{l-2k} c_l^{(j)}, d_k^{(j-1)} = \sum_l g_{l-2k} c_l^{(j)}, \quad j = 0, -1, ... \tag{5}$$

Conversely, we can derive from Eqs (3) and (4), a reconstruction formula

$$c_k^{(j)} = \sum_l [h_{k-2l} c_l^{(j-1)} + g_{k-2l} d_l^{(j-1)}], \quad j = 0, -1, ... \tag{6}$$

These formulas were found by Mallat [2]. In Eqs (5) and (6), the sequences $\{c_k^{(j-1)}\}$ and $\{d_k^{(j-1)}\}$ mean low and high frequencies.

The data in hydrophobicity plot were regarded as an input signal. Here, we used the Coiflet's wavelet with tap 18 as a mother wavelet [2].

The hydrophobic cores are defined by the degree of solvent accessibility of each amino acid side-chain. In the core of proteins, the defined buried residues tend to be continued. On the contrary, the buried residues on surface of proteins tend to be dotted.

The length of continuous buried residues corresponded until level $j = -2$. Then we attempted to use the high frequency component at level $j = -1, -2$ and low frequency at level $j = -2$ for prediction.

The residues above defined threshold were predicted as buried. The thresholds against each frequency component were defined by cross-validation. The thresholds for the high frequency at level $j = -1, -2$ and low frequency at level $j = -2$ were 1.78, 1.77 and 1.64, respectively. The prediction accuracy averaged with all proteins in Wako and Blundell dataset was 70.6%. Moreover, using the threshold for each family, the average accuracy was raised to 71.5%. Our method can offer predictions for proteins without homologous sequences or low similarity to known structures, with above 70% accuracy and without sequential alignment [3].

References

[1] Hirakawa,H., Muta,S. and Kuhara,S., The hydrophobic cores of proteins predicted by wavelet analysis, *Bioinformatics*, 15(2):141–148, 1999.

[2] Mallat, S., Multiresolution approximation and wavelets, *Trans. Am. Math. Soc.*, 315:69–88, 1989.

[3] Wako, H. and Blundell, T.L., Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes, *J. Mol. Biol.*, 238:682–692, 1994.