Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry 29 (2005) 220–228

Computational Biology and Chemistry

www.elsevier.com/locate/compbiolchem

Analyzing functional similarity of protein sequences with discrete wavelet transform

Zhi-ning Wen[a,1], Ke-long Wang[a,1], Meng-long Li[a,b,*], Fu-sheng Nie[b], Yi Yang[c]

[a] College of Chemistry, Sichuan University, Chengdu, Sichuan 610064, PR China
[b] Software Engineering College, Sichuan University, Chengdu, Sichuan 610064, PR China
[c] College of Life Science, Sichuan University, Chengdu, Sichuan 610064, PR China

Received 9 October 2004; accepted 14 April 2005

Abstract

This paper applies discrete wavelet transform (DWT) with various protein substitution models to find functional similarity of proteins with low identity. A new metric, '*S*' function, based on the DWT is proposed to measure the pair-wise similarity. We also develop a segmentation technique, combined with DWT, to handle long protein sequences. The results are compared with those using the pair-wise alignment and PSI-BLAST.

Keywords: Discrete wavelet transform; Protein sequence analysis; Substitution model; Functional similarity evaluation

1. Introduction

As the basis of molecular biological analysis, protein sequence analysis is one of the most important tools of biologists, and has been widely used in structure and function prediction, phylogenesis studies and different conservation pattern recognition. The key to protein sequence analysis is to detect the similar segments, structure-function domains or sequences with similar function.

Nowadays, general protein sequence analysis methods are sequence alignments and database query methods. Needleman and Wunsch (1970) presented a global pair-wise alignment algorithm, and Smith and Waterman (1981) designed an algorithm for local pair-wise alignment. Multiple sequence alignment is a tool for extracting the relationship between multiple sequences, such as CLUSTAL series (Thompson et al., 1994, 1997; Chenna et al., 2003) and T-Coffee (Notredame et al., 2000). Database query methods, such as FASTA (Pearson and Lipman, 1988; Pearson, 2000) and BLAST (Altschul et al., 1990, 1997) are commonly used to recruit a set of homologous sequences.

Two major issues in traditional sequence alignment methods are accuracy and speed. An important factor that affects the accuracy is the substitution model. If the model can effectively figure out the nuance of different amino acids, sequence analysis methods can correctly find the difference of protein sequences. An important stochastic model, hidden Markov model (HMM) has been used in programs such as HMMER (Eddy, 1995) and SAM (Karplus et al., 1998), and has been employed extensively to create large databases of sequence alignments such as Pfam (Bateman et al., 2004) and ProSite (Hulo et al., 2004). HMMs are also used to refine progressive alignment to enhance the sensitivity of sequence alignments (Löytynoja and Milinkovitch, 2003). The knowledge-based refinement combined with protein's structure information can also improve the sensitivity of sequence alignments (Thompson et al., 2003; Dror et al., 2003; O'Sullivan et al., 2004). Considering that scoring function may not describe biological reality well, some improvements are accomplished in scoring strategy by designing elaborate scoring system such as position specific gap penalty (Thompson et al., 1994; Reese and Pearson, 2002). To improve the speed of sequence

[*] Corresponding author. Tel.: +86 28 85415376; fax: +86 28 85415376.
E-mail address: liml@scu.edu.cn (M.-l. Li).

[1] These authors contributed equally to this work.

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.04.007

alignment, heuristic algorithm such as FASTA and BLAST are proposed.

Although many enhancements have been achieved in protein sequence alignment, the results of it at 10–20% residue identity are still doubtful. Below the ‘twilight zone’ (Doolittle, 1981; Rost, 1999) at 10–20% residue identity, the accuracy of the best programs correctly aligning on average is lower than 47% of the residues (Thompson et al., 1999). The ‘twilight zone’ clearly constitutes a great barrier for all the programs in this study. It is badly necessary to introduce other methods to analyze protein sequences.

As an effective tool of signal processing, wavelet transform (WT) is widely used in bioinformatics and chemometrics and shows its advantage in analyzing different scale information of a signal and bioinformatical data (Liò, 2003). In chemistry, WT has been applied to data processing and analysis in spectrometry, chromatography, and nuclear magnetic resonance spectrometry (NMR) (Leung et al., 1998; Shao et al., 2003; Li et al., 2002, 2003). For protein sequence analysis, WT has been used to discriminate proteins with different tertiary structures (Mandell et al., 1997), to predict hydrophobic cores from hydropathy data (Hirakawa et al., 1999) and to locate highly conserved residues in the hormone prolactin from electron–ion interaction potential data (Hejase de Trad et al., 2000). It has also been used to predict the location and topology of helices in transmembrane proteins (Liò and Vannucci, 2000), to detect repeating motifs (Murray et al., 2002) and conserved regions (Krishnan et al., 2004), to predict protein secondary structures (Qiu et al., 2003) and allergenic proteins (Li et al., 2004). Based on DWT (Daubechies, 1992), a new concept of similarity of protein sequence, sequence-scale similarity, has been proposed (Hejase de Trad et al., 2002) to identify the functional similarity of two protein sequences.

To apply WT to functional similarity identification of protein sequences, several problems are to be solved in this paper: first, how to transform the protein sequence into a numeral signal without losing the functional or structural information; second, different wavelets and decomposition scale will have different results in the DWT method. Wavelets and decomposition scale with best performance are chosen by comparing 46 different kinds of wavelets; and at last, in cross-correlation analysis, we can only estimate the similarity of sequences with three statuses: strongly correlated, weakly correlated and no correlation. In sequence analysis, the three statuses are not accurate enough to distinguish the small differences between protein sequences. In order to figure out the nuance, an S function is designed based on the energy normalization property of WT.

2. Materials and methods

2.1. DWT

WT has been applied to signal processing in various fields since 1980s. The most attractive character of WT is the ability to elucidate simultaneously both spectral and temporal information, in contrast to the Fourier transform that only elucidates spectral information. The Fourier coefficients contain only globally averaged time-domain information, thus leading to location specific features in the signal being lost (Bentley and McDonnell, 1994). A WT is defined as the projection of a function or a signal $f(t)$ onto the wavelet function.

$$W_f(a,b)=\langle f(t),\Psi_{a,b}(t)\rangle=\left(\frac{1}{\sqrt{|a|}}\right)\int_{-\infty}^{\infty} f(t)\Psi\left(\frac{t-b}{a}\right)\mathrm{d}t, \quad (1)$$

$$\Psi_{a,b}(t)=\left(\frac{1}{\sqrt{|a|}}\right)\Psi\left(\frac{t-b}{a}\right), \quad (2)$$

where $\Psi(t)$ is the basis function and $\Psi_{a,b}(t)$ is the basis wavelet function at a particular scale a and a translation b, a, $b\in R$, $a\neq 0$. The DWT uses $a_0=2$ and $b_0=1$, so that results lead to a binary dilation of 2^{-m} and a dyadic translation of $n2^m$. Therefore,

$$\Psi_{m,n}(t)=2^{-m/2}\Psi(2^{-m}t-n). \quad (3)$$

Generally, analyzing protein sequences with DWT includes three steps: firstly, translate the sequences into numeral signals; secondly, decompose the signals by wavelet; thirdly, detect similar sequences with cross-correlation analysis.

2.2. Substitution models

There are two sorts of substitution models to transform the protein sequence: one is substitution matrix based on the mutational possibility of two amino acids, such as PAM (Dayhoff et al., 1978) and BLOUSUM (Henikoff and Henikoff, 1992), the other is based on the physicochemical properties that contribute to the function of the protein, such as the c–p–v model (Grantham, 1974), and electron–ion interaction potential (EIIP model) (Cosic, 1994). The c–p–v model takes into account the three main properties of amino acids—the composition (c), polarity (p) and molecular volume (v); the EIIP value describes the average energy states of all valence electrons in particular amino acids; the AESNN3 model is derived from orthogonal encoding scheme by artificial neural network (ANN) that employed three numbers to describe the amino acid type of one protein residue (Lin et al., 2002).

In the first kind of substitution models of amino acids, PAM or BLOUSUM, two amino acids correspond to one value and in the second one, one amino acid can be substituted by a number or a vector. The c–p–v model, the EIIP model and the AESNN3 model are compared in this work.

2.3. The Benchmark Database: BAliBASE

The BAliBASE is built as a benchmark to evaluate the accuracy of detection/prediction and alignment of these complex sequences (Bahr et al., 2001). The database contains high quality, manually constructed multiple sequence alignments together with detailed annotations. Subgroups are built

Table 1
A comparison of three substitution models (decomposed by Bior3.3 wavelets, scale 4)

Sequence	Compare pairs	EIIP		*c*, *p*, *v*		AESNN3	
		Similar pairs	%	Similar pairs	%	Similar pairs	%
Short	1237	1220	98.63	1221	98.71	1217	98.38
Medium	925	609	65.84	599	64.76	591	63.89
Long	1355	434	32.03	480	35.42	441	32.55
Total	3517	2263	64.34	2300	65.40	2249	63.95

The result of decomposing sequences of Ref_1 and Ref_3 with different substitution models shows that the *c*–*p*–*v* model is the best in the three models. In this table, we can also see the length of sequences affected the different models' performances. The long sequences have the lowest detected percentage in three models, 33.33% in average. With length of the sequence decrease, the detecting ability of DWT increase drastically, an average of 97.9% of similar sequence pairs in the short group can be found with the three substitutions.

with structurally similar sequences but with different overall sequence similarities. The multiple alignments in BAliBASE have been built according to the superposition of the structures, and these "structural alignments" are considered as being the true ones.

Reference 1 (Ref_1) of BAliBASE contains alignments of equidistant sequences. The percentage of identity between two sequences is within a specified range. All the sequences are of similar length, with no large insertions or extensions. Therefore, Ref_1 is used to check the effects of sequence length and percentage of identity on wavelet analyzing performance. Reference 3 (Ref_3) of BAliBASE contains subgroups with <25% residues identity and highly related sequences (>25% identity) between groups. This reference can be used to assess the ability of our method to correctly detect approximately equidistant divergent families (<20% identity).

2.4. Comparing substitution models of amino acids

In the *c*–*p*–*v* model, it is reasonable to consider that the *i*th amino acid of a protein sequence is assigned to a vector whose components have the composition value $c(i)$, the polarity value $p(i)$, and the volume value $v(i)$. We use the normalized forms of these values: $\hat{c}(i) = [c(i) - \overline{c}]/\sigma_c$, $\hat{v}(i) = [v(i) - \overline{v}]/\sigma_v$ and $\hat{p}(i) = [p(i) - \overline{p}]/\sigma_p$. Considering the three factors' effects, an amino acid can be substituted with $A(i)$.

$$A(i) = \hat{c}(i) + \hat{p}(i) + \hat{v}(i). \tag{4}$$

And then, a protein sequence can be transformed into a numerical signal. For further analysis, the distance between points (or say time distance) in these numerical sequences is set at an arbitrary value $d = 1$. The last element of series contains the most recent observation.

To make a comparison, the other two models, the EIIP value and the AESNN3 value, are used, respectively. For EIIP value, each amino acid is substituted by an EIIP value of normalized form. The substitution method with AESNN3 values is similar to the method of *c*, *p* and *v* values. Respectively, substituted with the three models, protein sequences of Ref_1 and Ref_3 in BAliBASE are decomposed by Bior3.3 wavelets (Cohen et al., 1992). The results are shown in Tables 1 and 2.

2.5. Choosing the wavelet and decomposition scales

Based on different basis functions, the wavelets have different families; every wavelet family has its quality fitting for different signals and has different results. In this paper, 46 kinds of wavelets are tested to select one with best performance with the benchmark protein sequences of Ref_3.

Restricted by the property of wavelet decomposition, different decomposition scales have different results in analyzing protein sequences. On one hand, decomposing a shorter sequence with too high decomposition scale will

Table 2
The performance of three substitution models with different amino acids identity (decomposed by Bior3.3 wavelets, scale 4)

Reference		Compare pairs	EIIP		*c*, *p*, *v*		AESNN3	
			Similar pairs	%	Similar pairs	%	Similar pairs	%
Ref_1								
Short	<25%	54	53	98.15	53	98.15	52	96.30
	20–40%	85	83	97.65	83	97.65	80	94.12
	>35%	92	90	97.83	90	97.83	91	98.91
Medium	<25%	64	46	71.88	41	64.06	38	59.38
	20–40%	62	39	62.90	37	59.68	37	59.68
	>35%	80	61	76.25	65	81.25	61	76.25
Long	<25%	50	14	28.00	15	30.00	17	34.00
	20–40%	88	14	15.91	21	23.83	15	17.05
	>35%	77	17	22.08	20	25.97	15	19.48

This result shows the identity of sequence has no relationship with the DWT analyzing results.

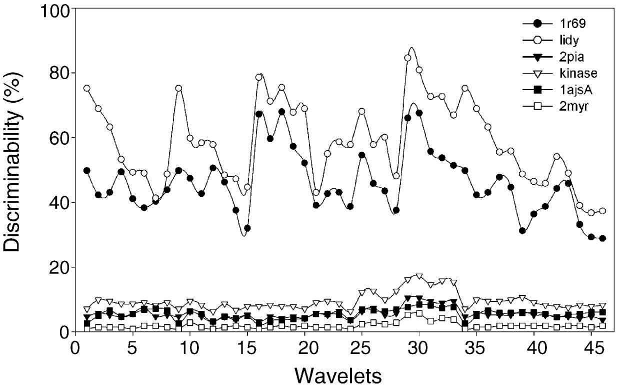

Fig. 1. Decomposing with scale 2 by different wavelets. The *x*-axis indicates the different wavelets, whereas the *y*-axis indicates the detecting ability of the wavelet to the six groups of benchmark sequences. The discriminability equals to the ratio of the detected similar pairs to the total compared protein pairs. Similar pairs in the two short sequence groups (1r69 and 1idy) can be detected better, but the similar pairs in the medium (2pia and kinase) and the long (1ajsA and 2myr) sequence groups can hardly be detected. The wavelets from 1 to 46 denote Db1, Db2, Db3, Db4, Db5, Db6, Db7, Db8, Bior1.1, Bior1.3, Bior1.5, Bior2.2, Bior2.4, Bior2.6, Bior2.8, Bior3.1, Bior3.3, Bior3.5, Bior3.7, Bior3.9, Bior4.4, Bior5.5, Rbio1.3, Rbio1.5, Rbio2.2, Rbio2.4, Rbio2.6, Rbio2.8, Rbio3.1, Rbio3.3, Rbio3.5, Rbio3.7, Rbio3.9, Rym1, Rym2, Rym3, Rym4, Rym5, Rym6, Rym7, Rym8, Coif1, Coif2, Coif3, Coif4 and Coif5.

introduce ineluctable redundancy in the decomposing process, on the other hand, decomposing a longer sequence with too low decomposition scale will omit many detail information. To choose the appropriate decomposition scale, the test sequences are decomposed with scales 2–5 separately with the above data (only the results of scales 2 and 4 are shown in Figs. 1 and 2).

2.6. *Detecting similar sequences and evaluating the similarity*

The cross-correlation coefficients are calculated at each scale to quantify the similarity between the two compared protein sequences. The cross-correlation coefficients are defined as:

$$\rho^{12}(j)=\frac{\frac{1}{N}\left[\sum_{n=0}^{N-1}A_2(n)A_1(n-j)\right]}{\frac{1}{N}\left[\sum_{n=0}^{N-1}A_1^2(n)\sum_{n=0}^{N-1}A_2^2(n)\right]^{1/2}}, \quad j=0,\pm1,\pm2,\cdots, \tag{5}$$

where N is the signal length and j is the number of lags. The maximum cross-correlation coefficient is 1 for the two same signals.

For biomedical signals, it is deemed strongly correlated if the correlation coefficient exceeds ±0.7 and weakly correlated if the correlation coefficient is between ±0.7 and ±0.5 (Oyster et al., 1987). Previous studies show that closely related proteins have a strong cross-correlation and distantly related proteins with similar biological functions have a sequence scale correlation at some scale. Distantly related proteins without common biological functions generally have no scale correlation. Strong, weak and no correlation were used to depict the similarity of protein sequences (Hejase de Trad et al., 2002).

In this paper, a new metric '*S*' is designed to quantitatively evaluate the similarity between two sequences. Similarities on different (WT) decomposition levels are calculated separately and weighted sum is taken using each level's energy as weight. With energy normalized wavelet, wavelet energy intensity $\int|\Psi_{a,b}(t)|^2\mathrm{d}t = E_{\Psi(a,b)} = E_{\Psi} = 1$, and each wavelet function $\Psi(a,b)$ gives the same contribution in energy intensity for signal analysis. Therefore, the wavelet coefficients have the same energy, and the energy of each point in different scales has relation with the width of wavelet function $\Psi_{(a,b)}(t)$ and scaling function $\Phi(t)$. $\Psi_{(a,b)}(t)$ and $\Phi(t)$ have the same width at the same scale. In a higher scale, the wavelet function is expanded in time sequence, so points of this scale have higher energy than those of the lower scale. The energy of the points in A_m, D_m, $D_{(m-1)}$, ..., D_1 has the proportion of $2^{m-1/2} : 2^{m-1/2} : 2^{m-2/2} : \cdots : 2^{m-m+1/2} : 2^0$, so an experiential formula is designed:

$$S=\frac{(2^{(m-1)/2}C_{\mathrm{cro}A_m}+2^{(m-1)/2}C_{\mathrm{cro}D_m}+2^{(m-2)/2}C_{\mathrm{cro}D_{m-1}}+\cdots+2^0C_{\mathrm{cro}D_1})\times100}{2^{(m-1)/2}+2^{(m-1)/2}+2^{(m-2)/2}+\cdots+2^0}. \tag{6}$$

The '*S*' denotes the similarity degree of two sequences, A_m denotes the approximation at scale m and D_m denotes the detail at scale m, and the C_{cro} denotes the maximum cross-correlation coefficient of each scale. The bigger the S is, the more similar the sequences are. For example, with decomposed four levels, there are five layers wavelets coefficients: A_4, D_4, D_3, D_2 and D_1.

$$S=\frac{(2^{3/2}C_{\mathrm{cro}A_4}+2^{3/2}C_{\mathrm{cro}D_4}+2^1C_{\mathrm{cro}D_3}+2^{1/2}C_{\mathrm{cro}D_2}+2^0C_{\mathrm{cro}D_1})\times100}{2^{3/2}+2^{3/2}+2^1+2^{1/2}+2^0}. \tag{7}$$

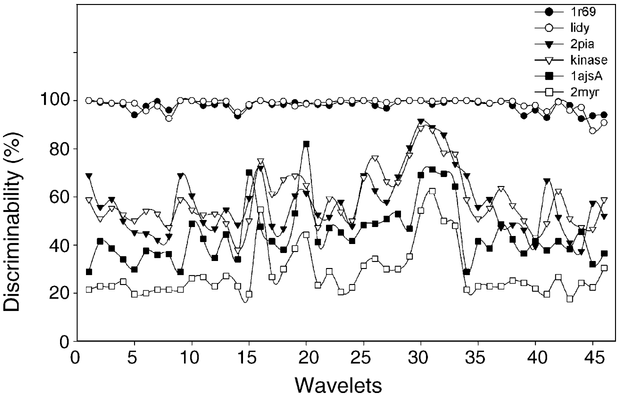

Fig. 2. Decomposing with scale 4, about 100% of the similar pairs in shorter sequence groups (1r69 and 1idy) can be detected. For the medium (2pia and kinase) and the long (1ajsA and 2myr) sequences, it still has a good performance in proper wavelets. The axis and the wavelets from 1 to 46 denote the same as in Fig. 1.

For the sequences longer than 200 amino acids, the amino acid sequences are cut into small segments in 200 amino acids. If the last segment is shorter than 100 amino acids, keep the last two segments without segmenting. It is highly possible that the segmentation may separate one motif into two parts; therefore, the later segment should overlap the end of the front one. Here, 30 amino acids are overlapped (the second segment from 171st to 370th amino acids, and the third one from 341st to 540th amino acids, etc.). The number of similar segments and *S* are recorded. The total similarity is the average of *S*. This segmentation strategy is equal to sliding the first sequence along the second one. This technique can effectively find the local similarity of two sequences. This similarity *S* is different from the identity since it can substitute the similarity of special scale information that may be caused by the conservative amino acids.

3. Results

3.1. Substitution models

Table 1 illustrates the performance of three models. As a whole, the *c–p–v* model has the best result in finding the similar protein sequences (65.40%). For the short and the long sequences, the *c–p–v* model performs better than the other two models. Especially for the long sequences, it is 3.39% higher than the EIIP model and 2.87% higher than the AESNN3 model. The EIIP model shows a better result in analyzing the medium sequences; it finds 1.08% more similar pairs than the *c–p–v* model does. The AESNN3 model is worst in analyzing sequences of any lengths.

Another important piece of information from Table 1 is that the length of sequence does affect the results whichever substitution model is used. The percentages of similar pairs decrease drastically with the increase of length of the sequence. With the *c–p–v* model, 98.71% of shorter sequence pairs can be detected while only 64.76% of medium and 35.42% of long sequence pairs can be found.

Table 2 shows the percentage of identity does not affect the results. Some sequences with higher identity cannot be detected with a higher rate than that of the sequences with low identity. For example, for the shorter sequences, the *c–p–v* model and the EIIP model can find more similar sequence pairs with identity <25% than sequences with identity >35%.

3.2. Selecting wavelets and decomposition scales

Forty six kinds of wavelets with scales 2–5 were tested, respectively. Figs. 1 and 2 show the results of scales 2 and 4. By comparing different decomposition scales, scale 4 is considered being fit for discriminating similar protein sequences. Decomposing with scale 2 is fit for the short sequences but can hardly find out the long similar sequences (Fig. 1). With high decomposing scale, the sensitivity to long sequence is improved (Fig. 2). At the same time, for the medium and long sequences, the Bior3.1 (No. 16), the Bior3.9 (No. 20), the Rbio3.3 (No. 30), the Rbio3.5 (No. 31) and the Rbio3.7 (No. 32) wavelets have better performance in the 46 wavelets (Fig. 2).

3.3. Segmentation strategy

The sequence length may affect the results of our method and the main reason is that in some long sequences the most conserved motifs that cover only a small percentage of the whole sequence are easily omitted in the cross-correlation analysis. To overcome this side effect, we introduce a segmentation comparing strategy here. The results of slicing

Table 3
The discriminability improvement after using the segmentation comparing method

Protein family	Discriminability (%)		Improvement ratio (%)
	Before segmentation	After segmentation	
1ajsA_ref3	47.62	75.93	59
1pamA_ref3	33.92	62.87	85
1ped_ref3	56.67	85.00	50
2myr_ref3	54.76	88.81	62
4enl_ref3	45.03	77.49	72
1ajsA_ref1	83.33	100.00	20
1cpt_ref1	100.00	100.00	0
1lvl_ref1	33.33	75.00	125
1pamA_ref1	70.00	80.00	14
1ped_ref1	66.67	83.33	25
2myr_ref1	33.33	58.33	75
4enl_ref1	100.00	100.00	0.00
gal4_ref1	50.00	80.00	60
1ac5_ref1	50.00	75.00	50
1adj_ref1	83.33	91.67	10
1bgl_ref1	0.00	25.00	N/A
1dlc_ref1	0.00	58.33	N/A
1eft_ref1	83.33	91.67	10
1fieA_ref1	0.00	50.00	N/A
1gowA_ref1	16.67	75.00	350
1pkm_ref1	50.00	91.67	83
1sesA_ref1	20.00	60.00	200
2ack_ref1	40.00	85.00	113
arp_ref1	50.00	70.00	40
glg_ref1	40.00	75.00	88
1ad3_ref1	100.00	100.00	0.00
1gpb_ref1	30.00	55.00	83
1gtr_ref1	20.00	60.00	200
1lcf_ref1	26.67	70.00	162
1rthA_ref1	20.00	35.00	75
1taq_ref1	20.00	40.00	100
3pmg_ref1	33.33	75.00	125
actin_ref1	70.00	95.00	36

The protein families are substituted with the *c–p–v* model and decomposed by wavelets Bior3.1 with scale 4. The segmentation technique improves the DWT's sensitivity to the similar protein. Using segmentation technique, DWT finds the similarities in three protein families (1bgl_ref1, 1dlc_ref1, 1fieA_ref1), which cannot be discriminated before segmented. After segmented, the average improvement of discriminability is 69%. The discriminating ability of DWT to 1cpt_ref1 and 1ad3_ref1 has no improvement. This is because the finding rate of the similar protein pairs had been 100% before segmenting.

Table 4
The analyzing results of Fibrillin (*Arabidopsis thaliana*) (AAC2819) in Swiss-Prot Database by pair-wise alignment, DWT (periodic-padding mode) and PSI-BLAST (without the hypothetical protein)

Sequences producing significant alignments	Global/local alignment identity (%)	*S* (%)	PSI-BLAST result	
			Score (bits)	*E* value
sp|P80471|LIPC_SOLTU	64.7/64.7	52.8	531	e−151
sp|O99019|LIPC_SOLDE	64.1/64.1	51.9	525	e−149
sp|Q8KDS2|MIAA_CHLTE	16.4/25.7	No similarity	40	0.007
sp|O34701|YOAU_BACSU	18.5/24.3	42.1	39	0.019
sp|P10248|PALY_RHORB	8.6/20.3	No similarity	36	0.13
sp|O02748|ARNT_RABIT	9.6/23.9	No similarity	35	0.2
sp|Q9UH99|U84B_HUMAN	9.9/28.8	No similarity	35	0.28
sp|Q9D309|FA3B_MOUSE	5.7/ 26.9	40.0	35	0.33
sp|Q6LR24|PEPT_PHOPR	13.4/20.4	No similarity	34	0.41
sp|O14022|ATCY_SCHPO	7.2/24.5	No similarity	34	0.57
sp|Q61165|NAH1_MOUSE	2.8/20.6	No similarity	33	1.0
sp|P14922|SSN6_YEAST	3.2/24.6	No similarity	33	1.1
sp|P27540|ARNT_HUMAN	8.2/21.0	No similarity	33	1.3
sp|Q9A2B1|MDH_CAUCR	14.9/32.1	No similarity	32	1.8

The identities are counted by Needleman–Wunsch global alignment from web server of the European Bioinformatics Institute (Harte et al., 2004). First two sequences have the highest identities. They have the highest *S* and the similar biological function too.

and without slicing in analyzing the long sequence of Ref_1 and Ref_3 are compared in Table 3. According to Table 3, the maximal improvement ratio is 350% (1gowA_ref1) after segmentation. Without segmentation, the DWT fails to detect similar pairs of proteins in three groups: 1bgl_ref1, 1dlc_ref1 and 1fieA_ref1. Using the segmentation technique, the DWT can separately find 25, 58.33 and 50% of the similar pairs in these three groups. Excluding the three groups, the average improvement of discriminating similar pairs is 69%.

3.4. *Application of S*

The major application of *S* is to quantitatively evaluate the similarity of protein sequences. Table 4 shows the analysis results of Fibrillin (*Arabidopsis thaliana*) (AAC28198, AF075598) with pair-wise alignment, PSI-BLAST and *S*. The sequence Fibrillin is queried in local Swiss-Prot database (release of Swissprot Version 46, FASTA format, download from NCBI’s FTP) (Boeckmann et al., 2003) by PSI-BLAST (2.2.8) and the query results are analyzed with DWT.

The first two sequences of Table 4 are the most similar with the query sequence in our test. The *S* value is 52.8 for the pair Fibrillin and LIPC_SOLTU and 51.9 for the pair Fibrillin and LIPC_SOLDE. These values agree with the reality. As an indicative of the functional analysis in plants, Fibrillin and these two proteins, i.e. light-induced protein C40.4 (O99019) and chloroplastic drought-induced stress protein CDSP-34 (P80471), have great similarity in their functions involved in structural stabilization of the cells and associated with chloroplasts and chromoplasts (Eymery and Rey, 1999; Gillet et al., 1998; Rey et al., 2000). For protein YOAU_BACSU and FA3B_MOUSE, their functional similarity with Fibrillin has not been reported.

Another test is taken in analyzing Amelogenin precursor (P45561). We analyzed the sequences whose PSI-BLAST score are between 49 and 100 (Table 5). The method

Table 5
The query results of P45561 in Swiss-Prot Database (by PSI-BLAST) and the analyzing result with DWT (periodic-padding mode)

Sequences producing significant alignments	Global/local alignment identity (%)	*S* (%)	PSI-BLAST result	
			Score (bits)	*E* value
sp|Q9QYX7|PCLO_MOUSE	1.3/36.8	No similarity	83	4e−016
sp|P27951|BAG_STRAG	4.0/31.9	No similarity	61	1e−009
sp|P12255|FHAB_BORPE	1.2/24.4	No similarity	55	9e−008
sp|P16053|NFM_CHICK	4.4/24.5	No similarity	55	1e−007
sp|P25384|YCB9_YEAST	2.6/33.7	No similarity	55	1e−007
sp|O97647|AMEL_TACAC	42.3/65.6	48.7	53	3e−007
sp|P16952|SSP5_STRGN	3.7/24.7	No similarity	52	8e−007
sp|Q9Y6V0|PCLO_HUMAN	1.1/22.9	No similarity	51	2e−006
sp|Q8MTI2|BSL1_TRIVA	6.9/30.3	No similarity	51	2e−006
sp|P32521|PAN1_YEAST	3.9/25.3	No similarity	50	3e−006
sp|Q9BX69|CAR6_HUMAN	5.0/27.6	No similarity	50	4e−006
sp|P17930|VP87_NPVOP	7.5/27.9	No similarity	50	4e−006
sp|Q9PU36|PCLO_CHICK	1.1/27.4	No similarity	49	6e−006
sp|O97646|AMEL_ORNAN	47.1/73.0	46.9	49	8e−006

Table 6
The similarity comparison using *S* (%)

	RGSG_BOVIN[a]	RGSB_RAT[b]	RGSA_RAT[c]	HEMA_PI3HT[d]	HEMA_PI3HU[e]	NER1_RAT[f]
RGSG_BOVIN	100	47.24	52.39	No	No	No
RGSB_RAT	47.24	100	70.26	No	No	No
RGSA_RAT	52.39	70.26	100	No	No	No
HEMA_PI3HT	No	No	No	100	98.48	39.37
HEMA_PI3HU	No	No	No	98.48	100	39.49
NER1_RAT	No	No	No	39.37	39.49	100

For the pair of same sequences, *S* is 100. The first three proteins function as GTPase-activating proteins (GAPs) that stimulate the inactivation of heterotrimeric G proteins and are responsible for the rapid turnoff of G protein-coupled receptor signaling pathways. The last three proteins share a neuraminidase function.

[a] Regulator of G-protein signaling 16.
[b] Regulator of G-protein signaling 11.
[c] Regulator of G-protein signaling 10.
[d] Hemagglutinin-neuraminidase.
[e] Hemagglutinin-neuraminidase.
[f] Sialidase 1 precursor.

detects that the Amelogenin precursor (P45561) has similarity with two sequences AMEL_TACAC (*S* = 48.7%) and AMEL_ORNAN (*S* = 46.9%). This also accords with the biological relationships of the three protein sequences. The three sequences have the same Pfam domain Amelogenin (PF02948) and have the same function in playing roles in the biomineralization of teeth and regulating the formation of crystallites during the secretory stage of tooth enamel development (Bateman et al., 2004; Wheeler et al., 2004).

The results in Table 6 generally reflect the biological relationships of those protein pairs. The *S* value of the two identical sequences is 100. Using Needleman–Wunsch algorithm, the HEMA_PI3HU and HEMA_PI3HT have the following result: length = 572; identity = 555/572 (97.0%); similarity = 562/572 (98.3%); gaps = 0/572 (0.0%); score = 2954.0. The *S* value of this pair of proteins is the highest, 98.48%, which shows the highest similarity in all the pairs of different sequences. The global alignment result of NER1_RAT and HEMA_PI3HU reveals the following result: length = 674; identity = 92/674 (13.6%); similarity = 153/674 (22.7%); gaps = 367/674 (54.5%); score = 40.0. The *S* of NER1_RAT and HEMA_PI3HU reveals the weak similarity (*S* = 39.49%). The Phylogram tree (Fig. 3) generated by Clustal W shows that the HEMA_PI3HU and HEMA_PI3HU are most closely related among these sequences; the NER1_RAT belongs to the same ancestor of HEMA_PI3HU and HEMA_PI3HU. The result of our method is accordance with the result of the Phylogram tree.

Using the Smith and Waterman algorithm for the protein pair RGSG_BOVIN and RGSA_RAT, we see the following result: length = 61; identity = 19/61 (31.1%); similarity = 35/61 (57.4%); gaps = 1/16 (1.6%); score = 88.0. The score is higher than that of the pair RGSG_BOVIN and RGSB_RAT (score = 71.5). The *S* of the two pairs shows that the RGSG_BOVIN is more similar with the RGSA_RAT (*S* = 52.39%) than with RGSB_RAT (*S* = 47.24%).

4. Discussion

In this paper, the crucial factors in analyzing functional similarity of protein sequences with DWT are discussed to build a normative DWT method. With the segmentation strategy, the efficiency of this method is improved by average rate of 69% in finding the similar sequences. It can be seen that the result of this method is greatly affected by the length of sequences.

The studies of similarity metric *S* of several groups of protein examples have been presented (Tables 4–6). Some pairs of protein sequences, which possess functional or structural similarity but have low sequence identity, are difficult to detect by sequence alignments. A protein's conservative function strongly depends on the physico-chemical properties especially the composition, polarity and molecular volume (Grantham, 1974). Based on the *c*–*p*–*v* model, the global and local similarity of protein sequences can be more effectively found.

In Tables 4 and 5, the ability of *S* to discriminate protein sequences with low similarity is compared with that of the pair-wise alignment and PSI-BLAST. The global pair-wise alignment result of Fibrillin (AAC2819) and YOAU_BACSU is: length = 379;

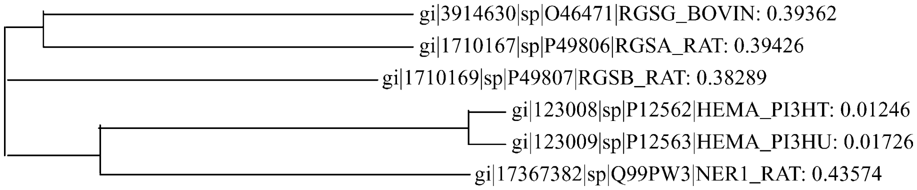
gi|3914630|sp|O46471|RGSG_BOVIN: 0.39362
gi|1710167|sp|P49806|RGSA_RAT: 0.39426
gi|1710169|sp|P49807|RGSB_RAT: 0.38289
gi|123008|sp|P12562|HEMA_PI3HT: 0.01246
gi|123009|sp|P12563|HEMA_PI3HU: 0.01726
gi|17367382|sp|Q99PW3|NER1_RAT: 0.43574

Fig. 3. The Phylogram tree generated by Clustal W on the web server of EBI (Thompson et al., 1994; Harte et al., 2004).

identity = 70/379 (18.5%); similarity = 125/379 (33.0%); gaps = 150/379 (39.6%); score = 93.5; the local pair-wise alignment result of them is: Identity = 59/243 (24.3%); score = 103.5. It is difficult to judge whether the two proteins have similar function only by the alignment result. The Pfam data show that the Fibrillin has a PAP_fibrillin domain. This family identifies a conserved region found in a number of plastid lipid-associated proteins (PAPs) and has structural molecule activity function. The YOAU_BACSU has HTH_1 (PF00126) domain and LysR_substrate (PF03466) domain. The structure of LysR substrate binding domain is known and is similar to the periplasmic binding proteins (Tyrrell et al., 1997). So, it is highly possible that the Fibrillin and the YOAU_BACSU have similar function on binding or activating proteins even though there is not report about it. The *S* shows a weak similarity on this pair of protein (*S* = 42.1%).

Based on DWT, the *S* analysis of NER1_RAT (409 amino acids) and HEMA_PI3HU (572 amino acids), which have significantly different lengths and share a common neuraminidase activity, has been performed. It is unlikely that one can link them together using sequence alignment methods such as Needleman–Wunsch global alignment (identity = 13.6%, score = 40.0) or Smith and Waterman algorithm (identity = 27.0%, score = 61.5). However, using the DWT and the *S* as defined above, we can still probe their distant connections. The *S* of this pair is 39.49%, which shows a weak similarity. This finding indicates that the *S* can reveal the weak functional similarity of protein sequences.

In the Phylogram tree, the RGSG_BOVIN and RGSA_RAT are most closely related in the three sequences (RGSA_RAT, RGSB_RAT and RGSG_BOVIN). But in our test (Table 6), the protein pair RGSA_RAT and RGSB_RAT has a highest *S* (*S* = 70.26%) in the three sequences. It means that the RGSA_RAT and RGSB_RAT are most similar. And in fact, RGSA_RAT and RGSB_RAT are more similar for they are got from the same source, *Rattus norvegicus* (Norway rat), and share the same function. They inhibit signal transduction by increasing the GTPase activity of G protein alpha subunits thereby driving them into their inactive GDP-bound form. Yet the RGSG_BOVIN is got from bovine. It can be seen that our method is reliable.

In conclusion, our new method presented in this paper can detect functional similarity of protein sequence with low identity and will nicely complement the existing sequence alignment methods. And it will be better if the more suitable substitution model and more accurate criterion should be designed. DWT and *S* can be a very promising tool for protein sequence comparison with advantage of finding the similarity of proteins with low identity. It will assist in finding the similar function of proteins more reliably.

Acknowledgements

We thank Professor Yizeng Liang, Professor Xueguang Shao, Professor Xiandeng Hou and Professor Xiao Yuan for reading the manuscript and their helpful advice. This work was partly supported by the National Natural Science Foundation of China (No. 29877016).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bahr, A., Thompson, J.D., Thierry, J.C., Poch, O., 2001. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res. 29, 323–326.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141.

Bentley, P.M., McDonnell, J.T.E., 1994. Wavelet transforms: an introduction. IEE Electron. Commun. Eng. J. 6, 175–186.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res. 31, 3497–3500.

Cohen, A., Daubechies, I., Feauveau, J.C., 1992. Bi-orthogonal bases of compactly supported wavelets. Commun. Pure Appl. Math. 45, 485–560.

Cosic, I., 1994. Macromolecular bioactivity: is it resonant interaction between macromolecules?—Theory and applications. IEEE Trans. Biomed. Eng. 41, 1101–1114.

Daubechies, I., 1992. Ten Lectures on Wavelets. 61, CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, PA.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. Atlas of Protein Sequence and Structure, vol. 5 (Suppl. 3), pp. 345–352.

Doolittle, R.F., 1981. Similar amino acid sequences: chance or common ancestry? Science 214, 149–159.

Dror, O., Benyamini, H., Nussinov, R., Wolfson, H., 2003. MASS: multiple structural alignment by secondary structures. Bioinformatics 19 (Suppl. 1), i95–i104.

Eddy, S.R., 1995. Multiple alignment using hidden Markov models. ISMB 3, 114–120.

Eymery, F., Rey, P., 1999. Immunocytolocalization of CDSP 32 and CDSP 34, two chloroplastic drought-induced stress proteins in *Solanum tuberosum* plants. Plant Physiol. Biochem. 37, 305–312.

Gillet, B., Beyly, A., Peltier, G., Rey, P., 1998. Molecular characterization of CDSP 34, a chloroplastic protein induced by water deficit in *Solanum tuberosum* L. plants, and regulation of CDSP 34 expression by ABA and high illumination. Plant J. 16, 257–262.

Grantham, R., 1974. Amino acid difference formular to help explain protein evolution. Science 185, 862–864.

Harte, N., Silventoinen, V., Quevillon, E., Robinson, S., Kallio, K., Fustero, X., Patel, P., Jokinen, P., Lopez, R., 2004. Public web-based services from the European Bioinformatics Institute. Nucleic Acids Res. 32, W3–W9.

Hejase de Trad, C., Fang, Q., Cosic, I., 2000. The resonant recognition model (RRM) predicts amino acid residues in highly conserved regions of the hormone prolactin (PRL). Biophys. Chem. 84, 149–157.

Hejase de Trad, C., Fang, Q., Cosic, I., 2002. Protein sequence comparison based on the wavelet transform approach. Protein Eng. 15, 193–203.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.

Hirakawa, H., Muta, S., Kuhara, S., 1999. The hydrophobic cores of proteins predicted by wavelet analysis. Bioinformatics 4, 141–148.

Hulo, N., Sigrist, C.J.A., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A., 2004. Recent improvements to the PROSITE database. Nucleic Acids Res. 32, D134–D137.

Karplus, K., Barrett, C., Hughey, R., 1998. Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856.

Krishnan, A., Li, K.B., Issac, P., 2004. Rapid detection of conserved regions in protein sequences using wavelets. In Silico Biol. 4, 0013.

Leung, A.K., Chau, F.T., Gao, J., 1998. A review on applications of wavelet transform techniques in chemical analysis: 1989–1997. Chemom. Intell. Lab. Syst. 43, 165–184.

Li, K.B., Issac, P., Krishnan, A., 2004. Predicting allergenic proteins using wavelet transform. Bioinformatics 20, 2572–2578.

Li, M.L., Kang, B., Qi, H.Y., Wen, Z.N., 2002. Compressibility evaluation of IR spectra wavelet compression. Chem. J. China Univ. 23, 128–1284.

Li, M.L., Qi, H.Y., Nie, F.S., Wen, Z.N., Kang, B., 2003. Application of embedded zerotree wavelet to the compression of infrared spectra. Chin. Chem. Lett. 14, 1193–1195.

Lin, K., May, A.C.W., Taylor, W.R., 2002. Amino acid encoding schemes from protein structure alignments: multi-dimensional vectors to describe residue types. J. Theor. Biol. 216, 361–365.

Liò, P., 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics 19, 2–9.

Liò, P., Vannucci, M., 2000. Wavelet change-point prediction of transmembrane proteins. Bioinformatics 16, 376–382.

Löytynoja, A., Milinkovitch, M.C., 2003. A hidden Markov model for progressive multiple alignment. Bioinformatics 19, 1505–1513.

Mandell, A.J., Selz, K.A., Shlesinger, M.F., 1997. Wavelet transformation of protein hydrophobicity sequences suggests their membership in structural families. Physica A 244, 254–262.

Murray, K.B., Gorse, D., Thornton, J.M., 2002. Wavelet transforms for the characterization and detection of repeating motifs. J. Mol. Biol. 316, 341–363.

Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G., Notredame, C., 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol. 340, 385–395.

Oyster, C.K., Hanten, W.O., Liorence, L.A., 1987. Introduction to Research: a Guide for the Health Science Professional. Lippincott, Oxford.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol. 132, 185–219.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85, 2444–2448.

Qiu, J.D., Liang, R.P., Zou, X.Y., Mo, J.Y., 2003. Prediction of protein secondary structure based on continuous wavelet transform. Talanta 61, 285–293.

Reese, J.T., Pearson, W.R., 2002. Empirical determination of effective gap penalties for sequence comparison. Bioinformatics 18, 1500–1507.

Rey, P., Gillet, B., Romer, S., Eymery, F., Massimino, J., Peltier, G., Kuntz, M., 2000. Over-expression of a pepper plastid lipid-associated protein in tobacco leads to changes in plastid ultrastructure and plant development upon stress. Plant J. 21, 483–494.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12, 58–94.

Shao, X.G., Leung, A.K., Chau, F.T., 2003. Wavelet: a new trend in chemistry. Acc. Chem. Res. 36, 276–283.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Tyrrell, R., Verschueren, K.H., Dodson, E.J., Murshudov, G.N., Addy, C., Wilkinson, A.J., 1997. The structure of the cofactor-binding fragment of the LysR family member, CysB: a familiar fold with a surprising subunit arrangement. Structure 5, 1017–1032.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thompson, J.D., Plewniak, F., Poch, O., 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27, 2682–2690.

Thompson, J.D., Thierry, J.C., Poch, O., 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19, 1155–1161.

Wheeler, D.L., et al., 2004. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res. 32, 35–40.