Protein Engineering vol.15 no.3 pp.193–203, 2002

Protein sequence comparison based on the wavelet transform approach

Chafia Hejase de Trad, Qiang Fang and Irena Cosic[1]

BioElectronics Group, Department of Electrical and Computer Systems Engineering, PO Box 35, Monash University, VIC 3800, Australia

[1]To whom correspondence should be addressed. Present address: School of Electrical and Computer Engineering, RMIT University, GPO Box 2476V, Melbourne 3001, Australia. E-mail: irena.cosic@rmit.edu.au

A protein's chemical properties, the chain conformation, the function of the protein and its species specificity are determined by the information contained in the amino acid sequence. Proteins of similar functions have at some level sequential identical amino acid sequences. The closer the phylogenetic relationship, the more similar are the sequences. To find the similarities between two or more protein sequences is of great importance for protein sequence analysis. The differences in the amino acid sequences permit the construction of a family tree of evolution. In this work, a comparison method was devised that is capable of analysing a protein sequence 'hierarchically', i.e. it can examine a protein sequence at different spatial resolutions. Based on a wavelet decomposition of protein sequences and a cross-correlation study, a sequence–scale similarity concept is proposed for generating a similarity vector, which renders the comparison of two sequences feasible at different spatial resolutions (scales). This new similarity concept is an expansion of the conventional sequence similarity, which only takes into account the local pairwise amino acid match and ignores the information contained in coarser spatial resolutions.

Keywords: heme proteins/resonance recognition model/ wavelet transform

Introduction

Protein comparison and alignment still represent one of the most important and widely used methods of protein sequence analysis (Bishop and Rawlings, 1996). The aim of protein comparison and alignment is to find the similarities or differences between two or more protein sequences. These comparative studies have provided new insights into the structure–function relationships of the active site of a protein.

Often within a protein class, only a few amino acid residues could be designated as 'invariable'. A substitution of such amino acids would destroy both biological activity and function. A consequence of the different amino acid sequences of functionally similar proteins is their immunological diversity, their species specificity. The similarities can help to identify individual amino acids crucial for the biological function, protein–target interaction and structure maintenance. The structurally essential similarities can be most effectively deduced from amino acid exchange frequencies in proteins of different species. It is not only the local similarity but also the global similarity between sequences needs to be found (Bishop and Rawlings, 1996). The similarities can be expressed as a template or a motif, the determinant of a specific structure and function.

Derivation of the three-dimensional structure from the amino acid sequence would be worthwhile, since it cannot be expected that all proteins would produce suitable crystals for X-ray analysis (Goffin *et al.*, 1996). Thus, if a protein sequence of unknown function and unknown structure is compared with other known sequences, its functional and structural information may be revealed by their similarity pattern.

Previous approaches such as FASTA, BLAST and PROSRCH (Pearson and Lipman, 1988; Bishop and Rawlings, 1996) are mainly based on sequence comparison and alignment. The concept of the similarity (a sequence similarity) for those approaches only means how many identical pairs of amino acids exist for the query sequence and the subject sequence.

However, two protein sequences with low sequential identity may show similarities in their physicochemical properties, tertiary structure, resonance recognition model (RRM) spectra and biological functions (Lesk, 1988; Cosic, 1994, 1997). This similarity concept can be enriched by incorporating the notion of similarity in other contexts.

The RRM multiple-cross spectral function can be regarded as a measurement of the similarity among different protein sequences in the frequency domain when each protein sequence is treated as a numerical series (Cosic, 1994, 1997). The most prominent peak frequencies show the spectral similarity of the protein sequences. Furthermore, the similarity can be either a local similarity or long-range similarity, the overall sequence similarity. For those traditional sequence comparison approaches, to find the local similarity is relatively easy but to find the global similarity is a difficult task (Bishop and Rawlings, 1987). The significance of the similarity is also hard to assess by those approaches. The spectrum similarity determined by the RRM is a global similarity because the spectrum is a contribution of all individual amino acids in the sequence.

Another analytical approach is the wavelet transform (WT) representation. It is a signal processing method efficient for multi-resolution analysis and local feature extraction (Daubechies, 1988, 1992). If the WT is introduced to a protein sequence, the similarity can be measured at different resolution scales based on a space-scale analysis. This sequence–scale similarity may reveal more information than other conventional methods.

Materials and methods

The sequence–scale similarity measurement introduced here is based on the discrete wavelet transform (DWT) (Daubechies, 1988) and a cross-correlation analysis of numerical representations of protein sequences as explained later (see the last section). The comparing sequences are initially 'converted' into numerical series using the RRM (Cosic and Nesic, 1988;

Table I. Electron–ion interaction potential (EIIP) values for amino acids (from Cosic, 1994)

Amino acid	EIIP
Leu	0.0000
Ile	0.0000
Asn	0.0036
Gly	0.0050
Val	0.0057
Glu	0.0058
Pro	0.0198
His	0.0242
Lys	0.0371
Ala	0.0373
Tyr	0.0516
Trp	0.0548
Gln	0.0761
Met	0.0823
Ser	0.0829
Cys	0.0829
Thr	0.0941
Phe	0.0946
Arg	0.0959
Asp	0.1263

Cosic *et al.*, 1989, 1991; Cosic, 1990, 1994, 1995, 1996, 1997; Cosic and Hearn, 1991).

The RRM model

The RRM (Cosic and Nesic, 1988; Cosic *et al.*, 1989, 1991; Cosic, 1990, 1994, 1995, 1996, 1997; Cosic and Hearn, 1991) is a physical and mathematical model which interprets protein sequence linear information using signal analysis methods. It comprises two stages. The first involves the transformation of the amino acid sequence into a numerical sequence. Each amino acid is represented by the value of the electron–ion interaction potential (EIIP) (Veljkovic and Slavic, 1972; Pirogova and Cosic, 1999), which describes the average energy states of all valence electrons in particular amino acids (Table I). The EIIP values for each amino acid were calculated using the following general model pseudopotential (Veljkovic and Slavic, 1972; Pirogova and Cosic, 1999):

$$<\mathbf{k} + \mathbf{q}|w|\mathbf{k}> = 0.25Z\sin(\pi\times 1.04Z)/(2\pi) \quad (1)$$

where $\mathbf{q}$ is the change of momentum $\mathbf{k}$ of the delocalized electron in the interaction with potential w and

$$Z = (\Sigma Z_i)/N \quad (2)$$

where Z_i is the number of valence electrons of the ith component of each amino acid and N is the total number of atoms in the amino acid. Each amino acid or nucleotide, irrespective of its position in a sequence, can thus be represented by a unique number (Table I). Numerical series obtained in this way are then analysed by digital signal analysis methods in order to extract information pertinent to the biological function. The original numerical sequence is transformed to the frequency domain using the discrete Fourier transform (DFT). As the distance between consecutive CA atoms in a protein sequence is 3.8 Å, it can be assumed that the points in the numerical sequence derived from the amino acid sequence are equidistant. For further numerical analysis the distance between points in these numerical sequences is set at an arbitrary value $d = 1$. Then the maximum frequency in the spectrum is $F = 1/2d = 0.5$. The total number of points in the sequence influences the resolution of the spectrum only. Thus, for an N-point sequence the resolution in the spectrum is equal to $1/N$. The nth point in the spectral function corresponds to the frequency $f = n/N$.

In order to extract common spectral characteristics of sequences having the same or similar biological function, the following cross-spectral function was used:

$$S_n = X_n Y_n^* \; n = 1, 2, ..., N/2 \quad (3)$$

where X_n are the DFT coefficients of the series $x(m)$ and Y_n^* are complex conjugate DFT coefficients of the series $y(m)$. Peak frequencies in the amplitude cross-spectral function define common frequency components of the two sequences analysed. The whole procedure, protein sequence → numerical series → amplitude spectra → cross spectra, is represented in Figure 1 using the example of human α- and β-hemoglobins.

To determine the common frequency components for a group of protein sequences, we calculated the absolute values of multiple cross-spectral function coefficients M, which are defined as follows:

$$|M_n| = |X1_n||X2_n| \ldots |XM_n| \; n = 1, 2, ..., N/2 \quad (4)$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analysed. The signal-to-noise ratio (S/N) for each peak is defined as a measure of similarity between sequences analysed. S/N is calculated as the ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. The extensive experience gained from previous research (Cosic, 1994, 1995, 1996, 1997) suggests that an S/N of at least 20 can be considered significant. The multiple cross-spectral function for a large group of sequences with the same biological function has been named 'consensus spectrum'. The presence of a peak frequency with significant S/N in a consensus spectrum implies that all of the analysed sequences within the group have one frequency component in common. This frequency is related to the biological function provided that the following criteria are met:

(1) one peak only exists for a group of protein sequences sharing the same biological function;
(2) no significant peak exists for biologically unrelated protein sequences;
(3) peak frequencies are different for different biological functions.

In our previous studies (Table II), the above criteria were tested with over 1000 proteins from 28 functional groups (Cosic, 1994, 1995, 1996, 1997; Trad *et al.*, 2000). Multiple cross-spectral functions of four different functional groups of proteins are represented in Figure 2. The following fundamental conclusion was drawn from our studies: each specific biological function of protein or regulatory DNA sequence(s) is characterized by a single frequency. Once the RRM characteristic frequency for a particular biological function has been determined, it is possible to identify the individual amino acid so-called 'hot spots' [using Fourier transformation (FT)] or domains [using the continuous wavelet transform (CWT) (Fang and Cosic, 1998, 1999; Trad *et al.*, 2000, 2001)] that contribute mostly to the characteristic frequency and thus also to the protein's biological function.

The physical meaning of the characteristic frequency

The correlation between the amplitude spectrum of numerical representation of genetic sequences and the corresponding

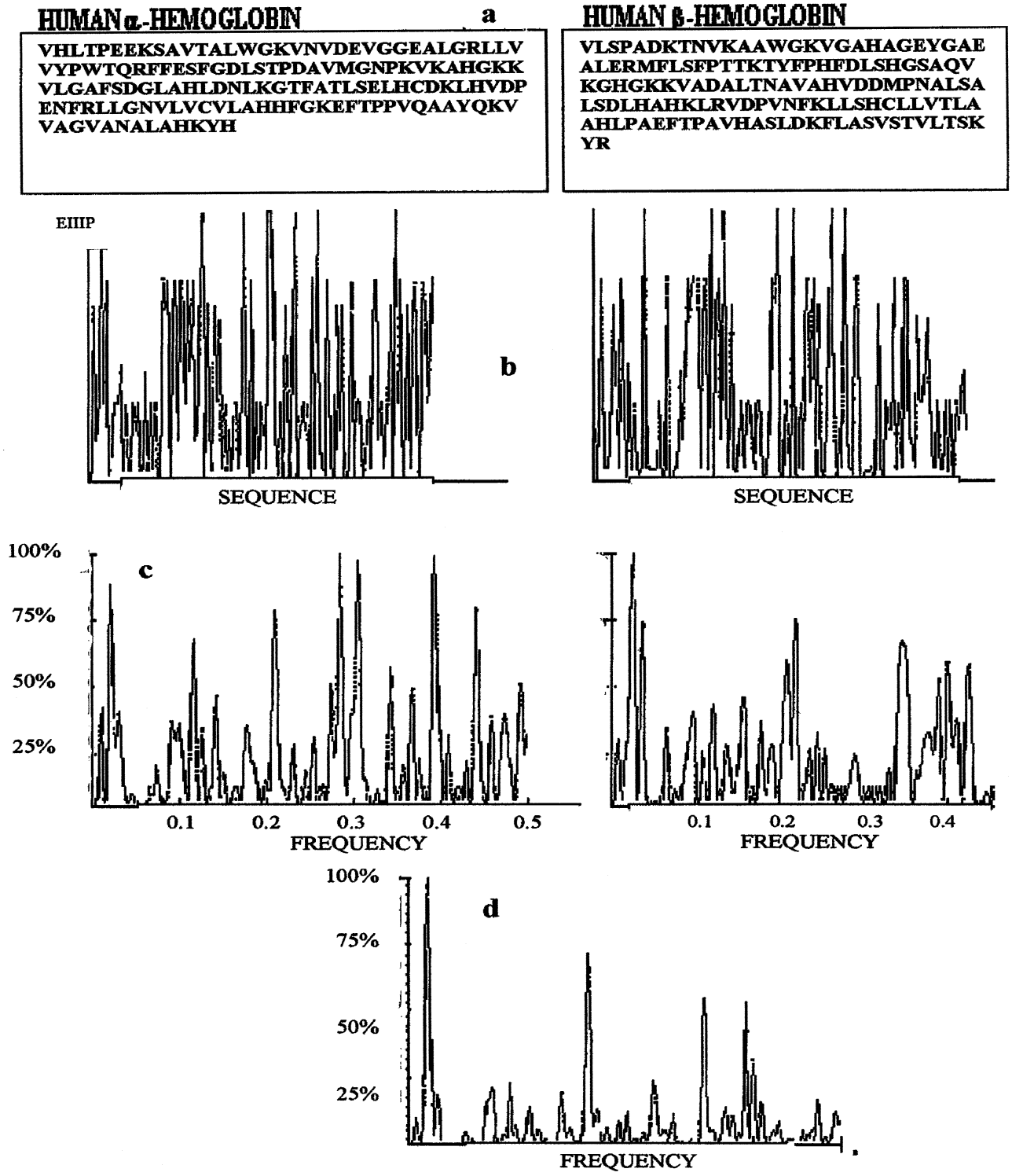

Fig. 1. The RRM procedure: (**a**) sequences of α- and β-hemoglobins; (**b**) graphical representation of the corresponding numerical sequences obtained by replacing each amino acid with its EIIP value; (**c**) spectra of both α- and β-hemoglobins; (**d**) cross-spectral function of the spectra presented in (c). The prominent peaks denote common frequency components. The abscissa represents RRM frequencies and the ordinate is normalized intensity (from Cosic, 1994).

biological function presented previously can lead to a completely new approach to protein dynamics. Each frequency in the RRM characterizes one biological function (Figure 2). Each biological process involves a number of interactions between proteins and their targets (other protein, DNA regulatory segment or small molecule). Each of these processes involves energy transfer between interacting molecules. These interactions are highly selective and this selectivity is defined within the protein structure. The selectivity of these interactions is proposed to be the resonant energy transfer between interacting molecules (Cosic, 1994). Consequently, the characteristic resonant frequencies for a number of different

Table II. Characteristic RRM frequencies for different functional groups of proteins and DNA regulatory sequences (from Cosic, 1994; Trad *et al.*, 2000).

Molecule type	Frequency	No. of sequences	S/N
Oncogenes	0.03130	46	468
Kinases	0.42969	8	71
Fibrinogens	0.44230	5	99
ACH receptors	0.49219	21	137
Phage repressors	0.10547	4	51
Bacterial repressors	0.08398	4	56
Heat-shock proteins	0.09473	10	326
Interferons	0.08203	18	117
Hemoglobins	0.02340	187	119
Signal proteins	0.14063	5	31
Protease inhibitors	0.35550	27	203
Proteases	0.37700	80	511
Restriction enzymes	0.29102	3	36
Amylases	0.41211	12	170
Neurotoxins	0.07031	16	60
Growth factors	0.29297	105	200
Ins.-like (IGF I, II)	0.49220	12	72
Fgfs	0.45120	7	121
Glucagons	0.32030	13	71
Prolactin	0.00780	10	161
	0.28520	10	30
Homeo box proteins	0.04590	9	100
Cytochromes *b*	0.05900	16	201
Cytochromes *c*	0.47656	38	252
Myoglobin	0.08200	49	128
Lysozymes	0.32810	15	124
Phospholipases	0.04300	29	115
Actins	0.48000	12	163
Myosins	0.34000	11	201
RNA polymerases	0.35693	10	256

interactions, i.e. biological functions, were theoretically calculated (Table II). These calculations were based on the following key finding: proteins with the same biological functions have common periodicities in the distribution of energies of delocalized electrons along the protein. With this in mind and taking into account the conductive properties of the protein backbone, the theoretical model of biologically relevant protein resonances was established (Cosic, 1990, 1994, 1997).

The discrete wavelet transform

The wavelet transform (WT) is a relatively new signal processing tool efficient for multi-resolution analysis and local feature extraction of non-stationary signals (Daubechies, 1988, 1992). The wavelet transform can be viewed as an inner product operation that measures the similarity or cross-correlation between the signal and the wavelets.

The sequence–scale similarity measurement introduced here is based on the discrete wavelet transform (DWT) and a cross-correlation analysis. The comparing sequences are initially 'converted' into numerical series using the RRM (Cosic *et al.*, 1989; Cosic, 1997). These numerical series are normalized to zero mean and unit standard deviation and zero-padded to have an identical sequence length. Then they are decomposed to M levels with details from level 1 to level M and an approximation at level M by the DWT. Because a correlation function quantifies the degree of interdependence of one process upon another or establishes the similarity between one set of data and another (Oppenheim and Schafer, 1997), the cross-correlation coefficients are calculated at each level to establish and quantify the similarity between the two compared protein sequences. There are a total of $M + 1$ correlation coefficients. The value of a correlation coefficient lies between –1 and +1; +1 means 100% correlation in the same sense and –1 means 100% correlation in the opposing sense (Oppenheim and Schafer, 1997). The cross-correlation coefficient is defined as

$$\rho^{12}(j) = \frac{r^{12}(j)}{\frac{1}{N}\left[\sum_{n=0}^{N-1} s_1^2(n) \sum_{n=0}^{N-1} s_2^2 s(n)\right]^{\frac{1}{2}}} \quad j = 0, \pm 1, \pm 2, \pm 3, \ldots \quad (5)$$

where N is the signal length, j is the number of lag and $r_{xy}(n)$ is an estimate of the cross-covariance and defined as:

$$r^{12}(j) = \frac{1}{N}\sum_{n=0}^{N-1} s_2(n) s_1(n-j) \quad (6)$$

The maximum absolute value of the correlation coefficient at each decomposition level is regarded as the similarity score for these two proteins at that level. Therefore, a total of $M + 1$ maximum values are taken out to form a sequence–scale similarity vector. The sequence–scale similarity vector depicts the similarity of two protein sequences at different scales or different frequency bands. More specifically, this vector describes the correlation with a multiresolution point of view.

The underlying property of wavelets is that they are localized in both time and frequency (Strang and Nguyen, 1996). The product of the uncertainties of both time and frequency is bound by the Heisenberg's uncertainty principle; no filter can have a width product smaller than $1/\pi$. The Gaussian filters attain this theoretical limit.

In this work we used the Bior3.3 biorthogonal wavelets (Cohen *et al.*, 1992) for the protein signal decomposition for all cases. Biorthogonal discrete wavelet transform uses two wavelets, one for decomposition and the other for reconstruction. Hence the analysis and synthesis tasks can be separated (Cohen *et al.*, 1992). Biorthogonal wavelets are symmetrical wavelets and have linear phase.

Results

Figure 3 illustrates Bior3.3 wavelet functions and scaling functions. It is found that the Bior3.3 decomposition wavelet function and scaling function are very rugged and have many abrupt changes. Therefore, this particular wavelet is chosen here to analyse a protein signal, which is also very rugged and contains many sharp variations. These wavelets are applied initially here in the example of hemoglobin. Hemoglobin is a tetrameric molecule whose quaternary structure is composed of two α- and two β-peptide chains. Each of the four subunits of the hemoglobin molecule take up one oxygen atom. The relative positions of the subunits alter according to their state of oxidation. The subunits are capable of cooperation and the uptake and evolution of oxygen causes an allosteric conformational change in each of the subunits. Although the sequences of hemoglobin β-chain and α-chain are not completely identical, they have exactly the same biological function (Lehninger *et al.*, 1993).

The discrete wavelet transform up to level 4 of a protein signal, hemoglobin human α-chain, an oxygen-carrying heme protein, is shown in Figure 4. The original protein signal is represented by S, An denotes the approximation at level n and Dn denotes the detail at level n. It can be seen that D1 and D2 contain more energy than any other level, i.e. the

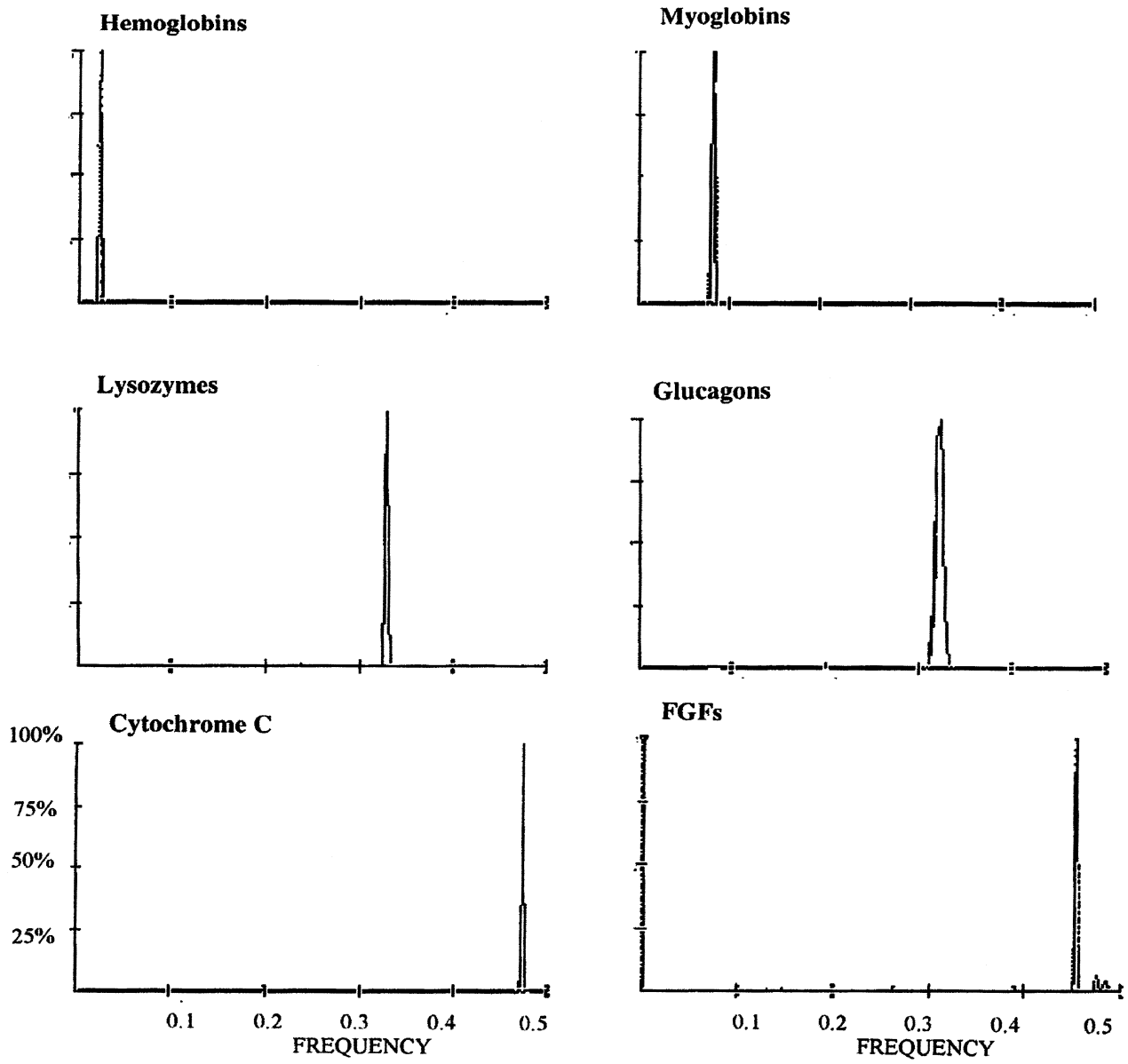

Fig. 2. Multiple cross-spectral function of six different functional groups of proteins. The multiple cross-spectral function of each group of proteins has a prominent peak representing the frequency characteristic for the biological function (from Cosic, 1994).

energy of this protein signal is most concentrated in these two frequency bands. Once the multi-level decompositions of two protein sequences have been obtained in such a way, their cross-correlation coefficients at different details and approximation can be calculated using Equation 5.

Figure 5 shows the cross-correlation coefficients between the DWTs of hemoglobin human and horse α-chains (hahu and haho). 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). As expected, the computed sequence–scale similarity vector, which is (0.92 0.83 0.90 0.89 0.89) following the order from A4, D4, D3, D2 to D1, reveals a strong correlation at all resolution levels between these two homologous protein sequences.

Figure 6 shows the sequence–scale similarity analysis of human α- and β-hemoglobins. The similarity vector is (0.97 0.62 0.44 0.39 0.23), revealing one strongly correlated frequency band A4 (= 0.97) and one weakly correlated frequency band D4 (= 0.62). Because these two polypeptides share the oxygen-carrying function, it is reasonable to consider that these two frequency bands are essential to this biological function. This result is also consistent with that from RRM: according to the RRM, a resonant frequency at 0.0234 characterizes the common biological function of hemoglobins (Table II). The frequency range of A4 is from 0 to 0.03125. Thus the RRM characteristic frequency of hemoglobin (f = 0.02340) is inside the frequency band A4, which shows the strongest correlation at 0.97.

Similarity measurement of functional related sequences

The similarity of closely related sequences is obvious. However, it is difficult to find the sequence similarity for proteins which are distantly related but have similar biological function or tertiary structure. For example, sperm whale myoglobin and lupine leghemoglobin have only 15% identical

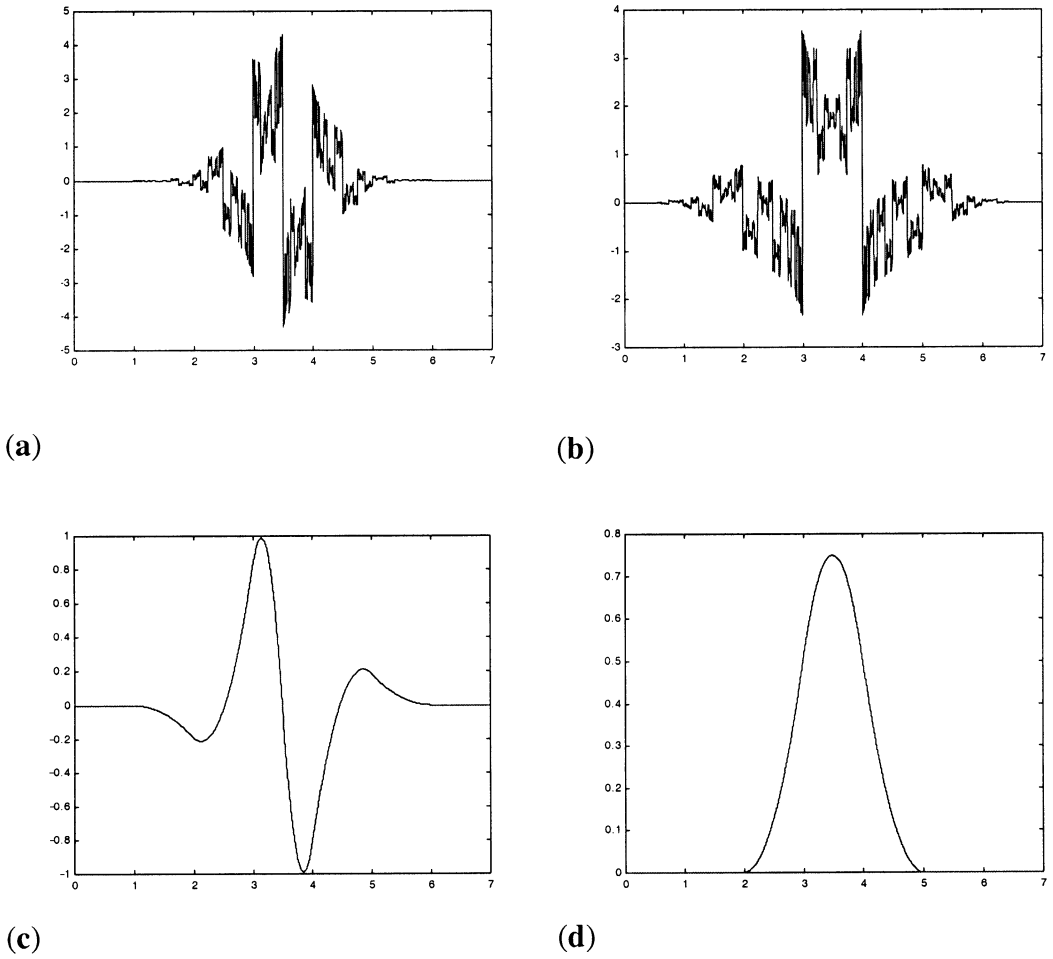

Fig. 3. (**a**) Bior3.3 decomposition wavelet function; (**b**) Bior3.3 decomposition scaling function; (**c**) Bior3.3 reconstruction wavelet function; (**d**) Bior3.3 reconstruction scaling function.

residues, which is far below the twilight zone of sequence identity, although they both contain a heme group, have similar secondary and tertiary structures and bind oxygen (Doolittle, 1981). The cross-correlation analysis of lupine leghemoglobin and sperm whale myoglobin revealed the sequence-similarity vector (0.40, 0.53, 0.44, 0.36, 0.25), showing a weak correlation in D4 (= 0.53). It is reasonable to deduce that this correlation is related to their sharing biological function, the oxygen binding capability.

Another example is chymotrypsin and subtilisin. These two proteins have a very low sequence identity, only 12% even using an optimal alignment method. However, they share a common proteolytic function and a common catalytic mechanism as an example of convergent evolution (Lesk, 1988). Because of the low sequence identity of these two pairs of proteins, it is unlikely that they can be linked together using the sequence alignment methods. However, using the sequence–scale similarity as defined above, we still can probe their distant connections. The sequence–scale similarity analysis of chymotrypsin and subtilisin revealed the sequence-similarity vector (0.35, 0.60, 0.42, 0.25, 0.18). At D4 (= 0.60), there is also a weak correlation for these two distantly related proteins.

Myoglobin is an oxygen-carrying globular heme protein like hemoglobin involved in oxygen storage and transport in vertebrate muscle. The myoglobin molecule is built up of eight helices, which compose a box-like structure with a hydrophobic pocket. The heme group responsible for oxygen binding (Fe^{2+}-porphyrin) is fixed in this pocket only by weak bonding. Myoglobin and hemoglobin are composed of an association of smaller subunits (α- and β-chains) and are thought to be evolutionarily related (Lehninger *et al.*, 1993). The sequence similarity of myoglobin and hemoglobin is very poor. However, Figure 7 indicates that hemoglobin and myoglobin are not dissimilar in the sense of the sequence–scale similarity. There are two weakly correlated frequency bands A4 and D4 which have correlation coefficients 0.63 and 0.60, respectively. Moreover, hemoglobin α-chain and β-chain are also correlated in these two frequency bands (see Figure 6). These two proteins have a strong correlation (correlation coefficient 0.97) and a weak correlation (correlation coefficient 0.62). This gives more evidence that frequency bands A4 and/or D4 contain the information related to the oxygen-carrying function of those proteins (hemoglobin, sigmoid oxygen saturation curve; myoglobin, hyperbolic saturation curve).

Cytochrome *c* is another heme-containing protein. Cytochrome *c* transfers electrons from the QH2–cytochrome *c* reductase complex to the cytochrome *c* oxidase complex. Figure 8 shows the cross-correlation of human hemoglobin α-chain and pig cytochrome *c*. There is no strong cross-correlation but a weak correlation at D3. Although cytochromes and

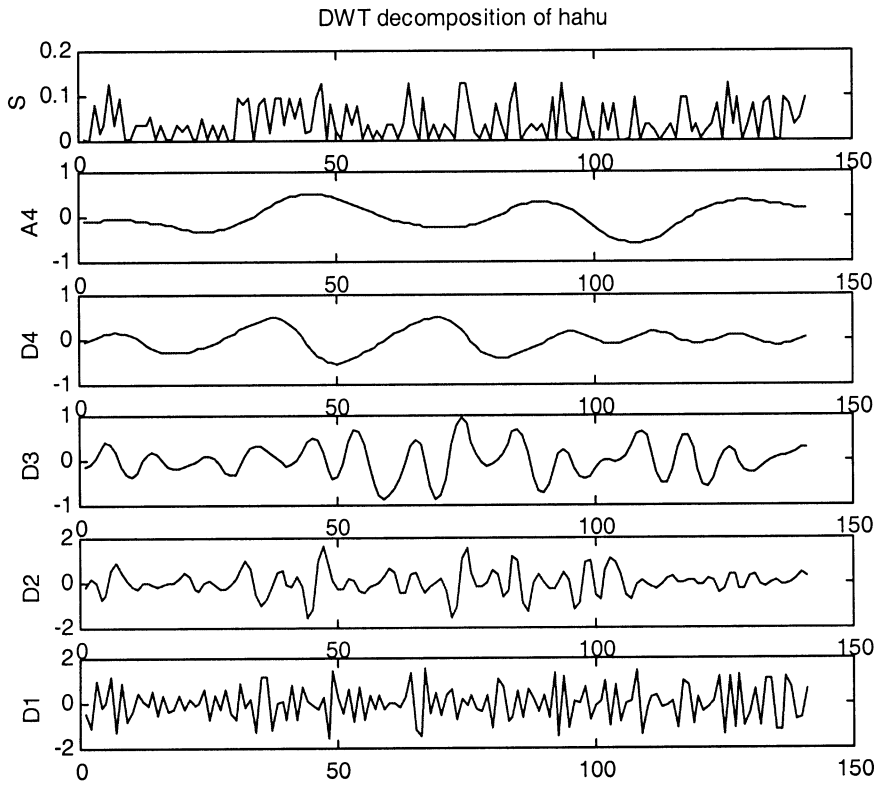

Fig. 4. The discrete wavelet transform up to level 4 of a protein signal for human hemoglobin α-chain. S is the original protein signal. The abscissa is the amino acid position along the protein backbone and the ordinate is the magnitude of the DWT coefficient. The Bior3.3 wavelets was used.

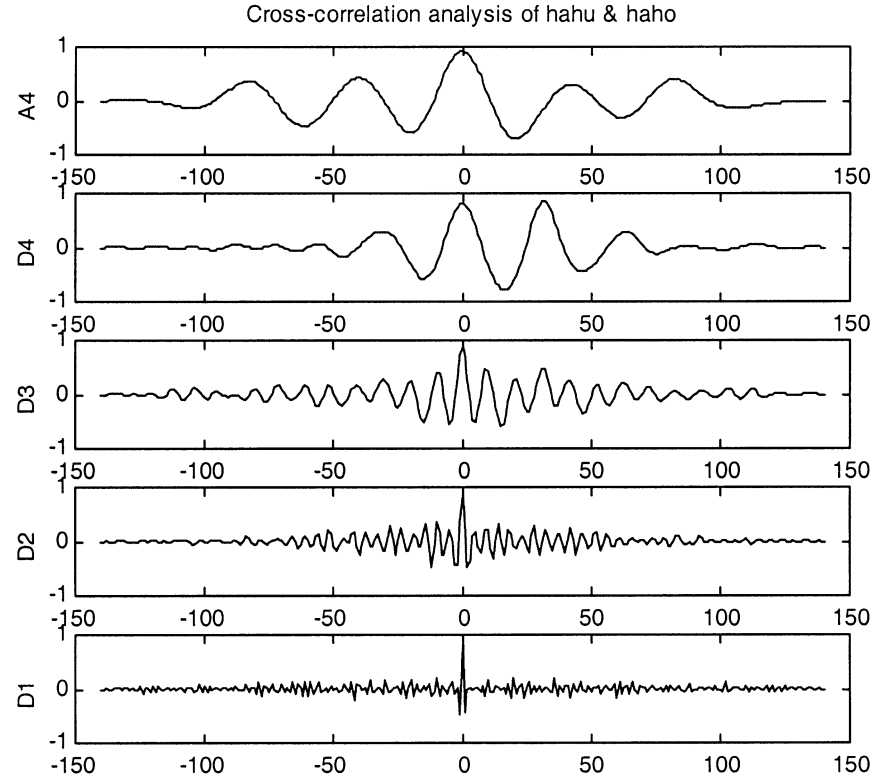

Fig. 5. The cross-correlation coefficients between the DWTS of hemoglobin human and horse α-chains from D1 to D4 and A4. The abscissa is the amino acid position along the protein backbone and the ordinate is the magnitude of the cross-coefficient. The similarity vector is (0.92 0.83 0.90 0.89 0.89) following an order (A4 D4 D3 D2 D1), which shows a strong sequence–scale similarity between these two homologous protein sequences. The Bior3.3 wavelet was used.

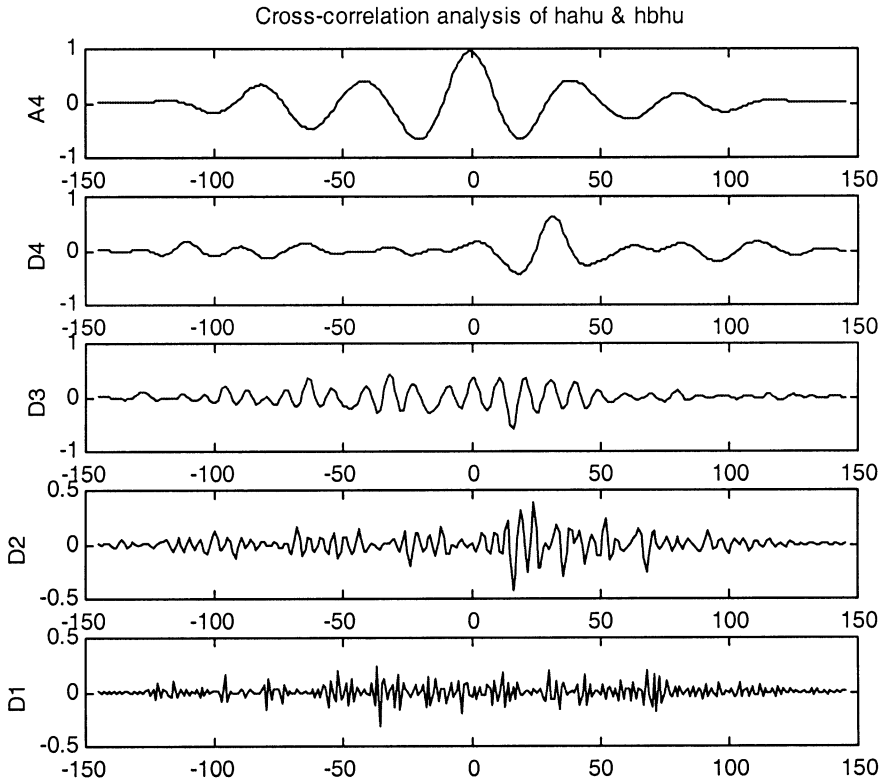

Fig. 6. The cross-correlation coefficients between the DWTs of human hemoglobin α-chain and β-chain from D1 to D4 and A4. The sequence–scale similarity vector is (0.97 0.62 0.44 0.39 0.23), which shows a strong correlation at A4 and a weak correlation at D4.

hemoglobins have very low sequence similarity, they do both possess a heme prosthetic group. Whether or not the weak correlation in D3 is due to the common heme group still needs further exploration.

Similarity measurement of non-functional related sequences

The sequence–scale similarity vector shows a strong cross-correlation between two closely related proteins and a certain correlation for two functionally related proteins. One requirement for choosing an appropriate analysis tool for protein sequence is to have a direct relationship with the underlying processing. This requires that a self-contained similarity measurement scheme shall give no-correlation results for functionally and/or structurally unrelated proteins.

Lysozyme is a widespread enzyme found especially in animal secretions, in egg white and in some microorganisms. It splits the glycosidic bond between certain residues in mucopolysaccharides and mucopeptides of bacterial cell walls. Lysozyme and hemoglobin do not share any biological function. This is also shown (Figure 9) by the cross-correlation study of their DWTs. In Figure 9, there is no peak that exceeds the weak correlation boundary 0.5.

A further study was carried out to calculate the sequence–scale similarity vectors among eight arbitrarily chosen different protein sequences. The comparison result is given in Table III. S denotes a strong correlation (>0.7), W denotes a weak correlation (0.5–0.7) and N denotes no correlation (<0.5). A bold entry indicates a complete correlation (correlated in each scale), an italic entry indicates a clear correlation (strong correlation in at least one scale or weak correlation in at least two scales), an underlined entry indicates a marginal correlation (only one weak correlation) and an ordinary entry indicates no correlation. The number before S, W and N is the number of the scales.

Discussion and conclusion

Sequence–scale similarity studies of several pairs of protein examples (Figures 5–9) have been presented. A certain number of these protein pairs have low sequence identity but possess functional or structural similarities which are difficult to detect by sequence alignments. However, certain correlations are still found in one or two frequency bands for each of these four pairs. The sequence–scale similarity analysis of chymotrypsin (83 amino acids) and subtilisin (56 amino acids), which have significantly different lengths and share a common proteolytic function and a common catalytic mechanism, has been performed. Because of the low sequence identity of these two pairs of protein, it is unlikely that one can link them together using sequence alignment methods such as BLAST (result = no significant similarity was found). However, using the sequence–scale similarity as defined above, we can still probe their distant connections. The sequence–scale similarity vector is (0.35 0.60 0.42 0.25 0.18), which shows a weak correlation at D4 for these two distantly related proteins.

This finding indicates that the functional or structural similarity of two protein sequences could be revealed by the sequence–scale study. One important judgement to compare different computational approaches is how well they perform in finding low degrees of similarity (Bishop and Rawlings, 1996). Hence the sequence–scale similarity can be a very promising tool for sequence comparison with the important advantage of not requiring indels.

These comparative studies have provided new insights into the structure–function relationships of certain groups of

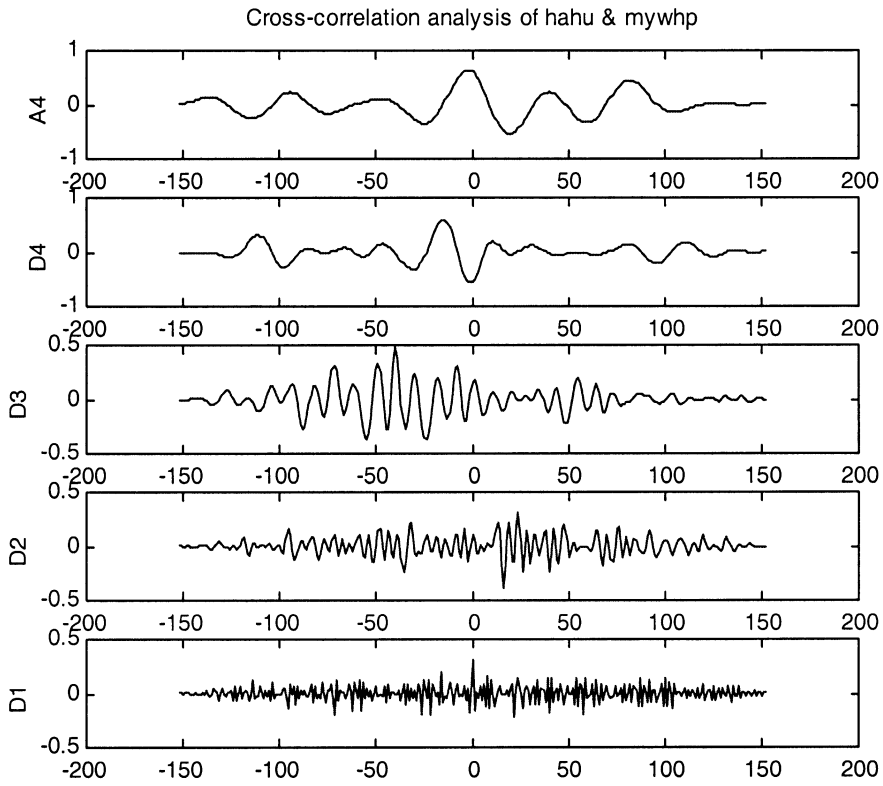

Fig. 7. The cross-correlation coefficients between DWTs of human hemoglobin α-chain and sperm whale myoglobin from D1 to D4 and A4. The sequence–scale similarity vector is (0.63 0.60 0.48 0.31 0.30).

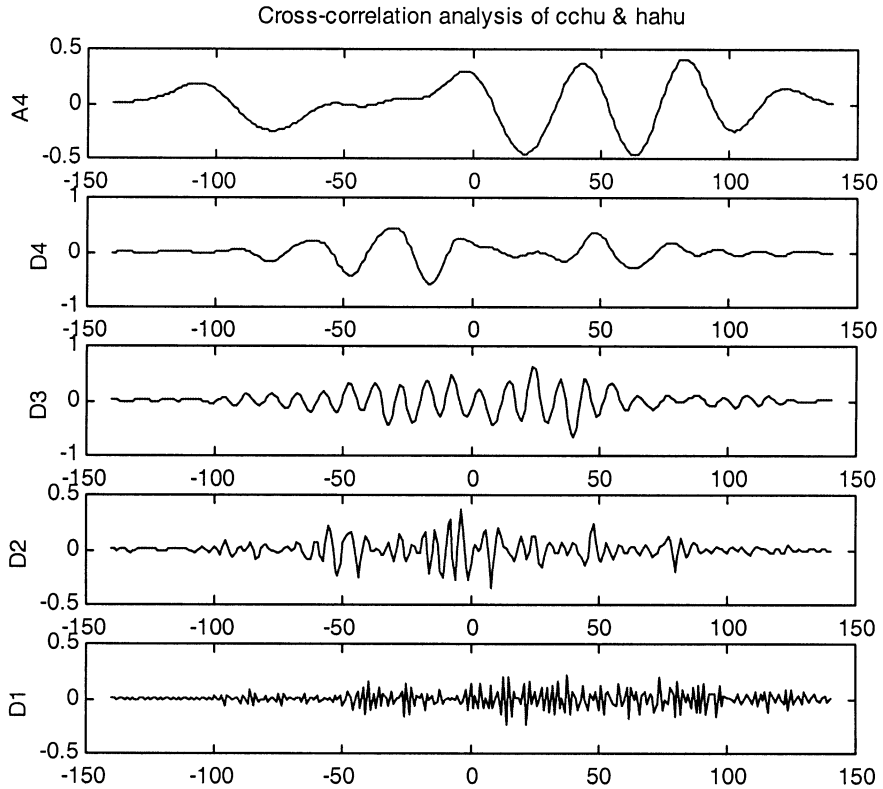

Fig. 8. The cross-correlation coefficients of human hemoglobin α-chain and pig cytochrome *c* from D1 to D4 and A4. The weak sequence–scale similarity vector is (0.41 0.45 0.63 0.37 0.21).

proteins. The results in Table III generally match the biological relationships of each protein pair. Using BLAST for the protein pair hemoglobin α-chain (hahu: 142 amino acids) and itself revealed the following results: score = 286 bits; identities = 142/142 (100%); positives = 142/142 (100%). The sequence–scale similarity vector shows complete correlation in all five

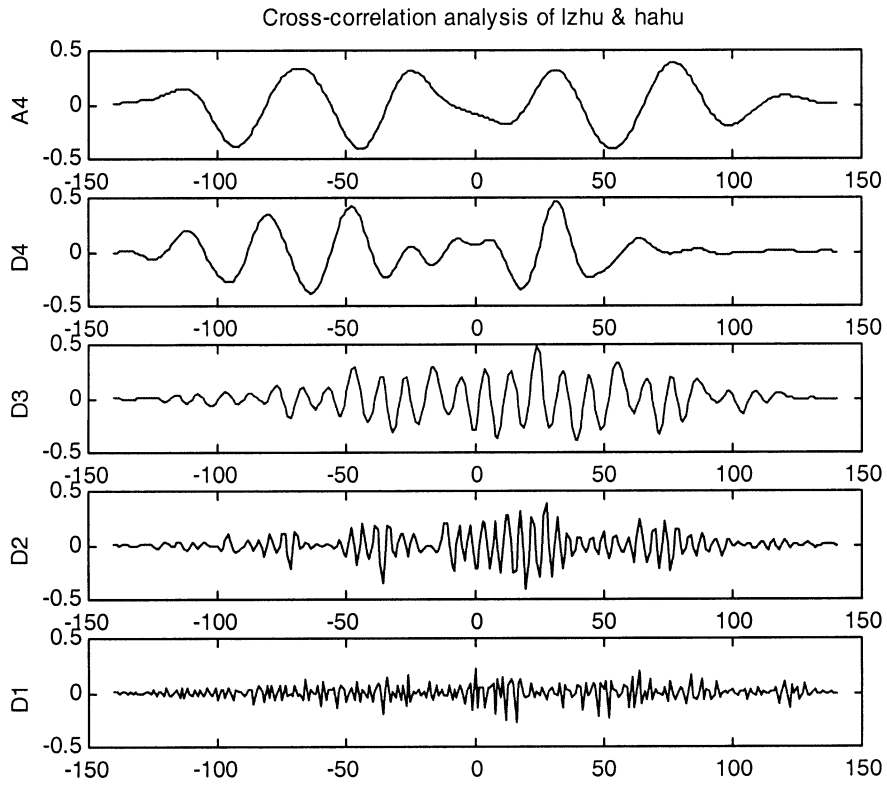

Fig. 9. The cross-correlation coefficients between DWTs of human hemoglobin α-chain and lysozyme rat from D1 to D4 and A4. The similarity vector is (0.38 0.46 0.48 0.38 0.22), which does not show an evident sequence–scale similarity between these two protein sequences.

Table III. Similarity comparison using sequence–scale cross-correlation method

	Hahu[a]	Haho[b]	Hbhu[c]	Ccpg[d]	Legh[e]	Lzrt[f]	Mwhp[g]	Fgfbh[h]
Hahu	**5S**[i]	**5S**	*1S1W3N*[j]	1W4N	*2W3N*	5N[k]	*2W3N*	5N
Haho	**5S**	**5S**	*1S1W3N*	1W4N	*2W3N*	5N	*2W3N*	5N
Hbhu	*1S1W3N*	*1S1W3N*	**5S**	1W4N	*2W3N*	5N	*2W3N*	5N
Ccpg	1W4N	1W4N	1W4N	**5S**	1W4N	1W4N	5N	5N
Legh	*2W3N*	*2W3N*	*2W3N*	1W4N	**5S**	1W4N	1W4N	*2W3N*
Lzrt	5N	5N	5N	1W4N	1W4N	**5S**	5N	1W4N
Mwhp	*2W3N*	*2W3N*	*2W3N*	5N	1W4N	5N	**5S**	*1W4N*
Fgfb	5N	5N	5N	5N	*2W3N*	1W4N	1W4N	**5S**

[a]Human hemoglobin α-chain.
[b]Horse hemoglobin α-chain.
[c]Human hemoglobin β-chain growth factor.
[d]Pig cytochrome *c*.
[e]Lupine leghemoglobin correlation.
[f]Rat lysozome.
[g]Sperm whale myoglobin.
[h]Basic human FGF.
[i]S denotes strong.
[j]W denotes weak.
[k]N denotes no correlation.
A bold entry indicates complete correlation, an italic entry indicates a strong correlation in at least one scale or weak correlation in at least two scales, an underlined entry indicates one weak correlation and an ordinary entry indicates no correlation. The number before S, W and N is the number of scales.

scales (**5S**). Using BLAST for the protein pair hemoglobin α-chain (hahu: 142 amino acids) and sperm whale myoglobin (mwhp: 153 amino acids) revealed the following results: score = 46.2 bits; identities = 37/147 (25%); positives = 59/147 (39%); Gaps = 6/147 (4%). The sequence–scale similarity vector (0.63 0.60 0.48 0.31 0.30) shows weak correlations at A4 and D4 expressed as 2W3N. It is reasonable to deduce that this correlation is related to their sharing biological function, the oxygen binding capability. Only the fgfbh (basic human growth factor) and legh (lupine leghemoglobin) have a clear correlationalthough no reported common biological properties of them have been found. The reason that causes this exception is still not clear.

Thus a fundamental and empirical conclusion for sequence–scale similarity measurement is reached:

(1) For closely related proteins, e.g. homologous proteins, there is a strong sequence–scale cross-correlation.
(2) For proteins that are distantly related but with similar biological functions, there is a clear sequence–scale correla-

tion. The correlation need not necessarily appear in each scale. The correlated scales are deemed to contain the information crucial to the common biological functions.

(3) For proteins that are distantly related and have no common biological functions, there is generally no sequence–scale correlation.

There are two additional advantages of the sequence–scale similarity measurement. First, the significance of the similarity is given directly by the correlation value rather than an alignment score as shown in the discussion above. The results derived from a sequence comparison scheme measure the quality of the alignment. Thus with the sequence–scale similarity vectors, the similarity significance can be compared, assessed and interpreted easily. For the conventional comparison methods, the comparison score needs to be processed using various empirical and statistical methods before it can be evaluated (Bishop and Rawlings, 1987; Lesk, 1988). Second, with the introduction of a cross-correlation function, the deletion and insertion which are often used in other conventional sequence comparison and alignment schemes are no longer needed. All the drawbacks derived from the gap insertion and deletion are not inherent to this method at all. Therefore, proteins with different sequence lengths can be compared easily.

Having in mind that the majority of theoretically predicted biological properties of proteins in this paper are functionally important, we can conclude that this study confirms our earlier hypothesis that the WT method could be established as a novel approach to examine protein sequences at different spatial resolutions.

References

Bishop,M. and Rawlings,C. (1987) *Nucleic Acid and Protein Sequence Analysis – A Practical Approach*. IRL Press, Oxford.

Bishop,M. and Rawlings,C. (1996) *DNA and Protein Sequence Analysis – A Practical Approach*. IRL Press, Oxford.

Cohen,A., Daubechies,I. and Feauveau,J.C. (1992) *Commun. Pure Appl. Math.*, **45**, 485–560.

Cosic,I. (1990) In Wise,D. (ed.), *Bioinstrumentation and Biosensors*. Marcel Dekker, New York, pp. 475–510.

Cosic,I. (1994) *IEEE Trans. Biomed. Eng.*, **41**, 1101–1114.

Cosic,I. (1995) *Bio/Technology*, **13**, 236–238.

Cosic,I. (1996) *Med. Biol. Eng. Comput*., **34**, 139–140.

Cosic,I. (1997) *The Resonant Recognition Model of Macromolecular Activity.* Birkhauser, Basel.

Cosic,I. and Hearn,M.T.W. (1991) *J. Mol. Recognit*., **4**, 57–62.

Cosic,I. and Nesic,D. (1988) *Eur. J. Biochem.*, **170**, 247–252.

Cosic,I., Pavlovic V. and Vojisavljevic,V. (1989) *Biochimie*, **71**, 333–342.

Cosic,I., Hodder,A., Aguilar,M. and Hearn,M.T.W. (1991) *Eur. J. Biochem.*, **198**, 113–119.

Daubechies,I. (1988) *Commun. Pure Appl. Math.*, **41**, 909–996.

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

Doolittle,R.F. (1981) *Science*, **214**, 149–159.

Fang,Q. and Cosic,I. (1998) *Aus. Phy. Eng. Sci. Med*., **21**, 179–185.

Fang,Q. and Cosic,I. (1999) In *Proceedings of the Inaugural Conference of the Victorian Chapter of the IEEE EMBS*. pp. 211–214.

Goffin,V., Martial,J.A. and Summers,N.L. (1996) *Protein Eng*., **8**, 1215–1231.

Lehninger,A.L., Nelson,D.L. and Cox,M.M. (1993) In *Principles of Biochemistry.* Worth, New York.

Lesk,A.M. (1988) In *Computational Molecular Biology.* Oxford University Press, Oxford.

Oppenheim,A.V. and Schafer,R.W. (1997) In *Discrete-time Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

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

Pearson,W.R. and Lipman,D. J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444.

Pirogova,E. and Cosic,I. (1999) In *Proceedings of the Inaugural Conference of the Victorian Chapter of the IEEE EMBS*. pp. 203–206.

Strang,G. and Nguyen,T. (1996) In *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley.

Trad,C.H., Fang,Q. and Cosic,I. (2000) *Biophys. Chem.*, **84**, 149–157.

Trad,C.H., Fang,Q. and Cosic, I (2001) In *Proceedings of the 2nd Conference of the Victorian Chapter of the IEEE EMBS*. pp. 115–119.

Veljkovic,V. and Slavic,I. (1972) *Phys. Rev. Lett.*, **29**, 105–108.

Received October 24, 2001; revised December 18, 2001; accepted January 4, 2002