doi:10.1006/jmbi.2001.5332 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 316, 341–363

JMB

Wavelet Transforms for the Characterization and Detection of Repeating Motifs

Kevin B. Murray[1]*, Denise Gorse[2] and Janet M. Thornton[3]

[1]*Department of Biochemistry and Molecular Biology University College London and Department of Computer Science, University College London, UK*

[2]*Department of Computer Science, University College London, UK*

[3]*Department of Biochemistry and Molecular Biology University College London and Department of Crystallography Birkbeck College, London, UK*

**Corresponding author*

The role of repeating motifs in protein structures is thought to be as modular building blocks which allow an economic way of constructing complex proteins. In this work novel wavelet transform analysis techniques are used to detect and characterize repeating motifs in protein sequence and structure data, where the Kyte-Doolittle hydrophobicity scale (HΦ) and relative accessible surface area (rASA) data provide residue information about the protein sequence and structure, respectively. We analyze a variety of repeating protein motifs, TIM barrels, propellor blades, coiled coils and leucine-rich repeat structures. Detection and characterization of these motifs is performed using techniques based on the continuous wavelet transform (CWT). Results indicate that the wavelet transform techniques developed herein are a promising approach for the detection and characterization of repeating motifs for both structural and in some instances sequence data.

Keywords: protein repeats; hydropathy; accessibility; wavelets; Fourier transforms

Introduction

Repeating sequences and structures are common in nucleic acids and proteins. A recent survey indicates that 14% of proteins contain sequence repeats, half the number which are contained in nucleic acid sequences.[1] Protein repeats come in considerable variety, ranging from repeats of a single residue, through heptad repeats in coiled coils, motif repeats (e.g. propellor blades) and finally to the repetition of homologous domains of 100 or more residues. Here, we are primarily concerned with motif repeats which are by definition secondary or supersecondary structural units: α-helices, β-strands, β-sheets, Rosmann folds, etc., connected together by short, sometimes variable, lengths of peptide in a repeating pattern. In terms of tertiary structure, protein motif repeats can be viewed as modular building blocks which allow an economic way of constructing complex protein topologies.[2] Motif repeats are commonly observed in proteins as a single motif repeated in tandem fashion along the protein sequence. A compendium of repeats is displayed in Figure 1(a) to (f).

Current protein repeat detection methods from sequence utilize standard sequence comparison algorithms adapted to find repeats. Andrade *et al.*[10] use optimal and sub-optimal score distributions from profile analysis to find homologous families for 11 kinds of tandem repeats, which once detected, may be used to identify additional repeats in any other sequence. Repeats are identified based on the probabilities of finding matches of different sub-optimal alignments when compared to random sequences. Pellegrini *et al.*[11] utilize multiple alignment techniques, based on a modified version of the Smith-Waterman dynamic programming algorithm,[12] where a sequence is aligned against itself enabling internal repeats to be found. Heger & Holm[13] use a similar but more refined technique which can validate distant repeats by profile alignment and optimizes repeat borders to yield a maximal integer number of repeats. For these methods detection of repeats is straightforward when the repeat in question is perfect. However, detection is complicated when evolution erodes any sequence similarity and when insertions, deletions and substitutions corrupt the regularity of the repeating pattern. Furthermore, repeats may be incomplete, widely spaced and be of multiple types interspersed throughout a sequence. As a consequence of these complications

Abbreviations used: rASA, relative accessible surface area; CWT, continuous wavelet transform; *MM*, modulus maxima; TIM, triosphosphate isomerase; LLR, leucine-rich repeat.

E-mail address of the corresponding author: murray@biochem.ucl.ac.uk

0022-2836/02/020341–23 $35.00/0

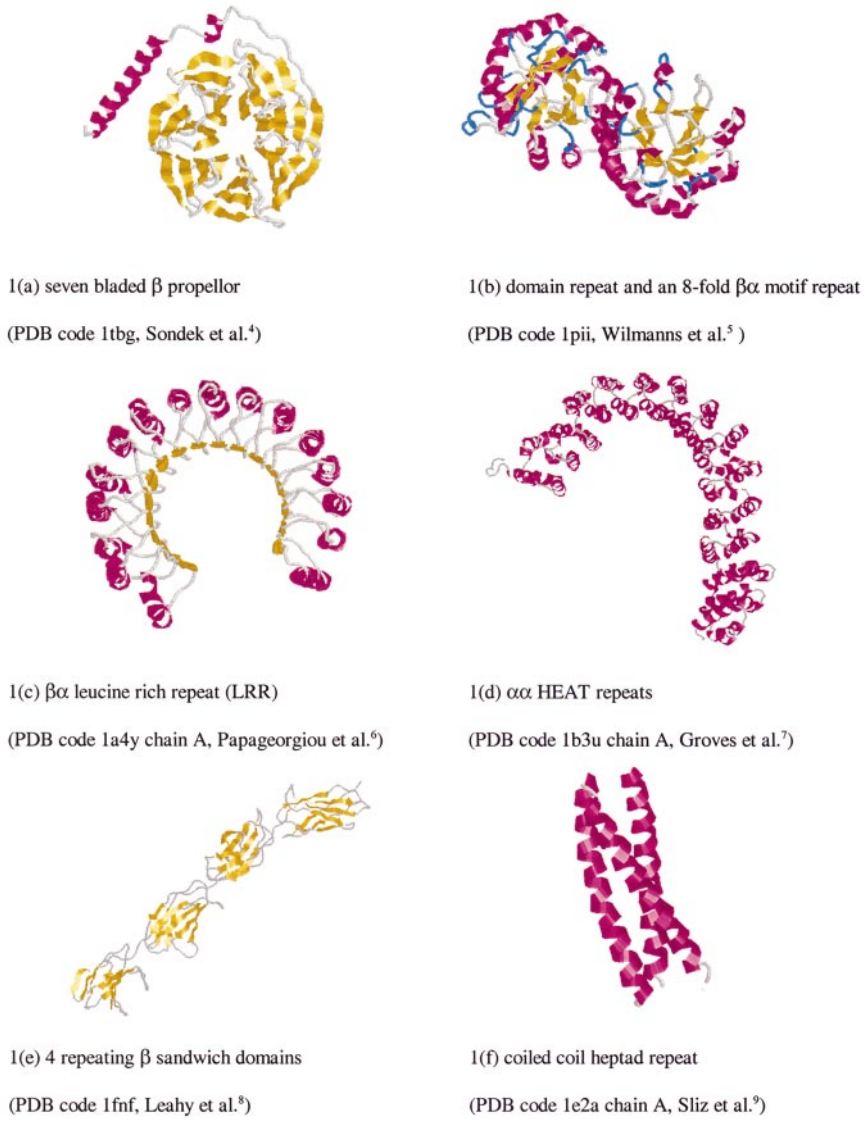


Figure 1. Rasmol[3] cartoons of repeating motifs and domains. (a) Seven bladed β propellor (PDB code 1tbg, Sondek *et al.*[4]). (b) Domain repeat and an 8-fold βα motif repeat (PDB code 1pii, Wilmanns *et al.*[5]). (c) βα Leucine-rich repeat (LRR) (PDB code 1a4y chain A, Papageorgiou *et al.*[6]). (d) αα HEAT repeats (PDB code 1b3u chain A, Groves *et al.*[7]). (e) Four repeating β sandwich domains (PDB code 1fnf, Leahy *et al.*[8]). (f) Coiled coil heptad repeat (PDB code 1e2a chain A, Sliz *et al.*[9]).

some repeats are not detected by current methods. To our knowledge, there is not yet even an automated method to assign repeats from 3D structure, which would provide valuable comparative data for assessing the performance of sequence-based structural repeat predictors.

In this work an alternative approach to repeat detection is adopted. A suite of continuous wavelet transform analysis techniques will be used to detect and characterize a selection of repeating protein motifs from both sequence and structural data. For sequence data wavelet transform analysis can be considered as an *ab initio* approach to motif detection and characterization. Introduced in the early 1980s, wavelets have become a popular signal analysis tool due to their ability to elucidate simultaneously both spectral and temporal information within the signal. This overcomes the basic shortcoming of Fourier analysis, which is that the Fourier coefficients contain only globally averaged information, thus leading to location specific features in the signal being lost.[14] Applications of wavelet analysis are now widespread and cover many fields of scientific research, including medical science, geophysics, engineering testing, image analysis, financial signal analysis and the topic of interest herein, proteins, where the dimension of ''time'' becomes that of sequence distance.

The body of literature concerning wavelet transform analysis and proteins (and DNA) is relatively small and comparatively recent. For proteins, wavelets have been used to predict hydrophobic

cores from hydropathy data,[15] the location of highly conserved residues in the hormone prolactin from electron ion interaction potential data,[16] the structural families of protein hydrophobicity sequences[17] and the location and topology of helices in transmembrane proteins.[18] Other wavelet-based research has focused on DNA, where wavelets have identified regular features in nucleotides,[19] the genome of Chinese hamster cells,[20] transcriptive yeast cell cycle microchip data,[21] and non-coding sequences.[22] We now look at three of the above references in more depth. Hirakawa *et al*.[15] define a wavelet-based method to predict the hydrophobic cores of globular proteins from hydropathy sequence data with 70% accuracy. This method predicts hydrophobic cores by thresholding the smallest wavelet scale to eliminate hydrophilic/neutral regions. It is worth noting that for this data sequential alignment techniques can predict hydrophobic cores with 76% accuracy,[23] but tend to perform poorly when there are no homologues or low sequence similarity, for these cases the wavelet-based method can predict cores at nearly 70% accuracy. Mandell *et al*.[17] apply the continuous wavelet transform to protein hydrophobicity sequences. This technique suggests which structural family the sequence belongs to, for one example each of α, β and αβ proteins. The protein structure and it's fractal dimension are used as reference criteria for the analysis. Lio & Vannucci[18] have developed a discrete wavelet threshold technique to predict the location and topology of helices in transmembrane proteins. Predictions are made by discrete wavelet thresholding, a new propensity scale generated from 1087 transmembrane domain sequences. This method works by wavelet transforming the data to generate wavelet coefficients, then coefficients below a certain size being shrunk or set to zero. A denoised signal is recovered by inverse transforming these thresholded coefficients. When compared to empirical methods based on hydrophobicity and/or helical propensity data for a test set of 83 proteins it was found that this method permits an improvement in the automatic location of transmembrane helices.

Wavelet Transforms

Wavelet theory

In this work the continuous wavelet is the preferred wavelet representation; justification for its use can be found in the Appendix. A brief summary of continuous wavelet transform theory and a description of some wavelet tools which assist interpretation of wavelet coefficients is presented; more details are given in the Appendix. Both wavelet theory and tools are illustrated by simple examples which outline some of the key concepts of this technique.

The continuous wavelet transform (CWT)

The wavelet transform of a continuous distance signal $x(t)$ is defined as:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} g\left(\frac{t-b}{a}\right) x(t)dt \qquad (1)$$

where $g((t-b)/a)$ is the analyzing wavelet function. The transform coefficients $T(a,b)$ are found for both specific locations on the signal, $t = b$, and for specific wavelet periods (which are a function of a). It is usual to plot $T(a,b)$ against a and b in either a surface or contour plot known as a scalogram. One of the most popular continuous wavelets is the Mexican hat,[24,25] plotted in Figure 2. This wavelet is particularly good at highlighting periodic structures and it will be used here to analyze a number of protein repeat motifs. The Mexican hat wavelet is defined as:

$$g\left(\frac{t-b}{a}\right) = \left(1 - \left(\frac{t-b}{a}\right)^2\right) \exp^{-\frac{1}{2}\left(\frac{t-b}{a}\right)^2} \qquad (2)$$

where a is the dilation parameter and b is the location parameter. A discussion is given in the Appendix concerning the functional properties of analyzing wavelets and the relationship between wavelet scale and frequency.

The ability of the wavelet transform to locate specific features temporally and spatially is now illustrated with the aid of a simple test signal. Consider the signal $Y(t)$ of, length N (512 distance units), containing two waveforms: a sine wave of period 64 distance units and a low amplitude cosine wave of period 16 distance units superimposed upon the first waveform (Figure 3(a)):

$$Y(t) = \begin{cases} \sin\left(\frac{64t}{2\pi}\right) t = 0 \ldots 129, \;\; t = 301 \ldots 512 \\ \sin\left(\frac{64t}{2\pi}\right) + 0.1\cos\left(\frac{16t}{2\pi}\right) t = 130 \ldots 300 \end{cases} \qquad (3)$$

Note that the higher frequency wave form only occurs in the region 130-300 units, and is difficult to discriminate from the low frequency high amplitude waveform. Fourier spectra analysis of this waveform reveals the two dominant frequency components in the test signal (Figure 3(b)), but no information on their location. Figure 3(c) displays

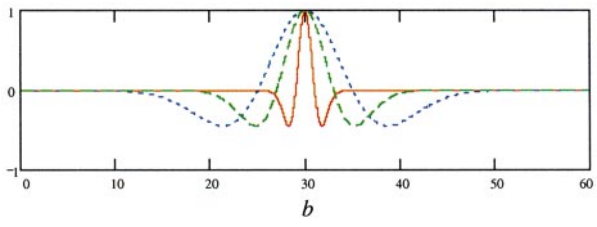
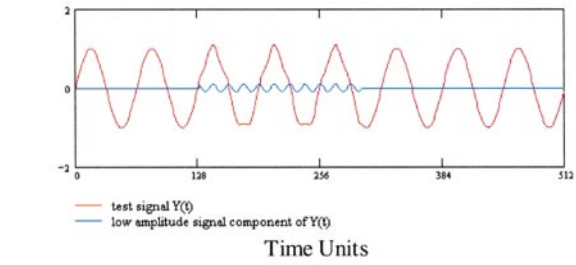
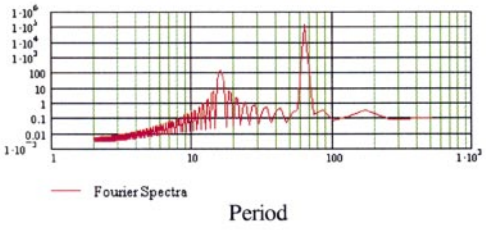
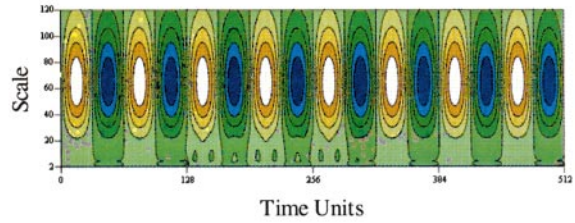


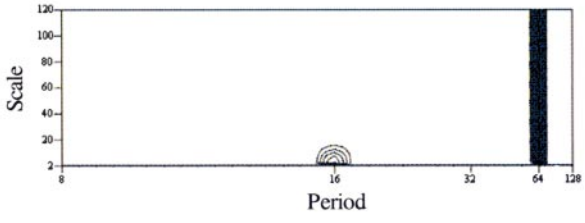
Figure 2. The Mexican hat wavelet plotted for three a values (1, 3 and 5) at $b = 30$.



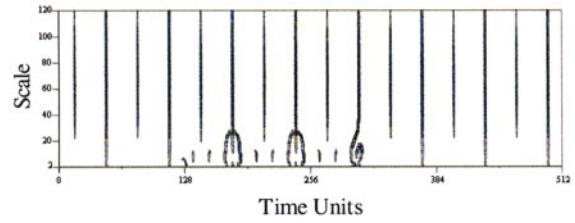


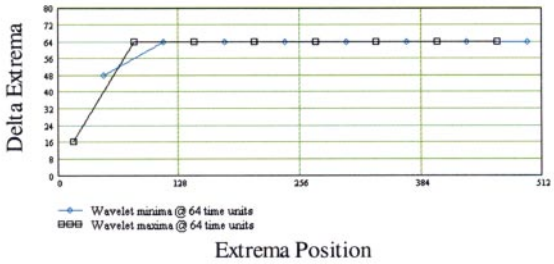
3(a, b) Test Signal and it's respective Fourier Spectra





3(c, d) Scalogram (Mexican Hat Analyzing Wavelet) and Fourier scale transform





3(e, f) Modulus Maxima and Maxima and Minima slices

Figure 3. Wavelet transform analysis of test signal. (a) and (b) Test signal and it's respective Fourier spectra. (c) and (d) Scalogram (Mexican hat analyzing wavelet) and Fourier scale transform. (e) and (f) Modulus maxima and maxima and minima slices.

the wavelet scalogram of the test signal with scale (or period) on the Y axis and translation (or distance location) on the X axis. At a scale centered at 64 distance units the wavelet transform detects the low frequency high amplitude waveform as a regular continuous series of features. At smaller scales evidence of the high frequency, low amplitude waveform is depicted as a series of regular features over the region 130 to 300.

Wavelet tools

It is possible to characterize the relationship between wavelet scale and frequency by taking the Fourier transform of each wavelet scale. This measure enables frequency detection at the expense of locality, but allows association of the dominant frequencies with wavelet scale, when plotted as a contour plot. Such a measure may prove useful in resolving protein repeat motifs. The Fourier coefficients at each wavelet scale a are given by:

$$\hat{P}(\omega, a) = \int_{-\infty}^{\infty} T(a, b) e^{-2i\omega b} db \qquad (4)$$

As some scales contain much more energy than others, we can normalize the $\hat{P}(\omega, a)$ term so that each scale contains the same Fourier transformed energy, giving rise to the new set of coefficients:

$$C(\omega, a) = \frac{|\hat{P}(\omega, a)|^2}{N^{-1}\sum_{\omega}|\hat{P}(\omega, a)|^2} \quad (5)$$

where $C(\omega,a)$ is the normalized Fourier power spectrum of the wavelet coefficients indexed by scale. An example of the Fourier scale transform is exhibited in Figure 3(d) for equation (3). The Figure shows that all investigated scales contain evidence of the low frequency waveform at a period of 64 units. A second smaller feature, associated with the high frequency waveform is evident on the scale frequency plot over the scale range 20 to 0 distance units at a period of 16 units.

As an aid to interpreting wavelet scalograms the positions of extrema across scales can be used to ''denoise'' the wavelet coefficients and thereby obtain dominant features of interest from background information.[26] This method uses continuous wavelet transform modulus maxima (*MM*); these can be defined as any point in wavelet space which corresponds to a local maximum of the modulus $|T(a,b)|$ i.e.:

$$WTMM(a,b) = \begin{cases} |T(a_0,b)| \text{ if } |T(a_0,(b-1))| \\ \quad < |T(a_0,b)| < T(a_0,b(b+1))| \\ 0 \text{ Otherwise} \end{cases} \quad (6)$$

Modulus maxima are local extrema at any point (a_0,b_0) such that $(\partial T(a_0,b))/(\partial b)$ has a zero crossing at $b = b_0$. Maxima lines can be drawn linking common extrema across scales. As an extension to the scalogram method, the *MM* technique is also useful for detecting singularities in the signal which can often characterize irregular structures and transient phenomena. An example of the *MM* technique is exhibited in Figure 3(e) for equation (3), where the elucidation of both waveforms in the distance-scale space becomes obvious with a series of features representative of the waveform modulus maxima.

If a scale of the *MM* is of interest, then the positive or negative maxima can be represented on a plot as a modulus slice (ΔMM *versus MM*), where ΔMM, the separation between extrema, is given by:

$$\Delta MM_{a,j} = MM_{a,j+1} - MM_{a,j} \quad (7)$$

and $j = 0,1,2 \ldots M-1$, where M is the total number of positive or negative maxima at a specific scale a. This method developed herein allows identification of globally periodic, locally periodic and transient behavior at the wavelet scale of interest. Figure 3(f) shows an example of wavelet maxima and minima slices at a scale of 64 distance units. Both extrema display ΔMM values of 64 distance units, indicative of the dominant waveform for most of the signal apart from the first values at 16 and 48 distance units. This discrepancy results from these extrema having no smaller neighbors, consequently they are paired with 0. The high frequency waveform is not displayed here as the scale of investigation (64 distance units) is too coarse to detect this feature. Below we apply these tools to analyze protein sequence and structure data.

Data Types

The data utilized in this study are relative accessible surface area (rASA) and simple hydrophobicity (HΦ) which, for each residue of a protein, provide information derived from the protein structure and sequence, respectively. rASA is generated using Hubbard's NACCESS program,[27] which implements Lee & Richards accessibility calculation.[28] This measures the relative accessibility of every residue to solvent in the 3D protein structure. More specifically rASA calculates the accessible surface of the residue which is then divided by the maximum accessibility of that amino acid in a extended tripeptide. A value of rASA between 0 and 100 % is then assigned to each residue where 0 % represents buried residues and 100 % freely accessible to solvent. The Kyte-Doolittle hydropathy scale[29] is based on the free energy transfer of each amino acid between organic solvent and water. This measure is often used to detect regions of sequences that can be inserted into cell membranes, commonly known as transmembrane helix predictions. In this work the hydropathy scale is inverted to allow comparison with rASA data; where hydrophilic residues should display a positive correlation with large exposed values of rASA. Observed secondary structure (SS) derived from the structure is also utilized here as an aid to assess the quality of wavelet structure and sequence detections. SS is represented numerically as a simple code: helix = 1, strand = −1 and coil = 0.

The choice of rASA and HΦ as data in this work is not arbitrary. It is hoped that rASA data will give a strong indication of the protein's overall geometry, and more specifically repeating motifs. Similarly for HΦ data important structural information may be contained in these 1D sequence-derived data sets. Other workers have used accessibility and hydrophobicity data to evaluate model structures.[30]

The aims of this work are: (1) to identify and classify motif repeats from structural data automatically, forming a database of all repeats; and (2) to develop automatic detection of these repeats from sequence data alone. As a preliminary attack to these problems wavelet transforms for rASA and HΦ are applied to a variety of proteins known to contain repeating motifs. By wavelet transforming these data and employing appropriate thresholding techniques we aim to identify motif repeats in structural and sequence data automatically.

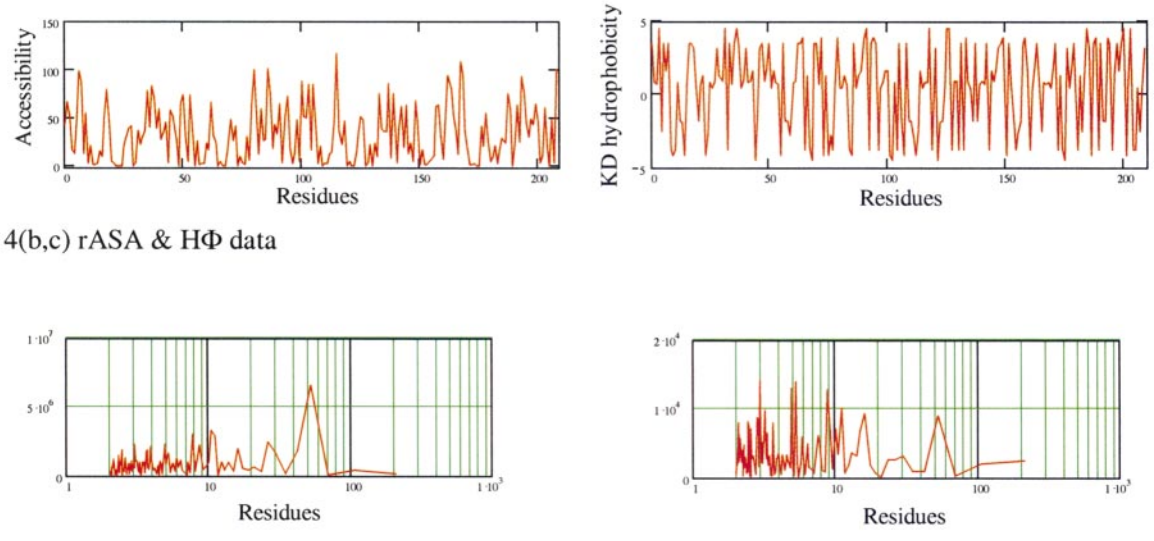
Wavelet analysis of four proteins displaying different types of repeats is presented below.

Results

Propellor

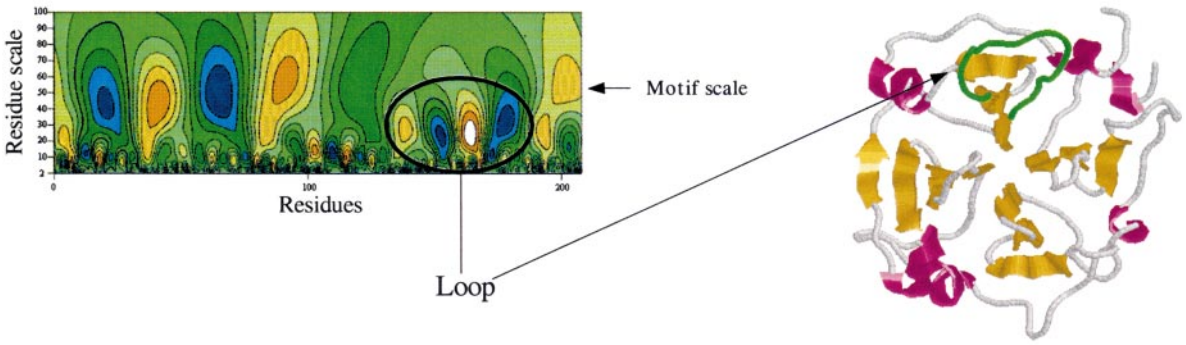
In this section wavelet-transformed rASA and HΦ data for the C-terminal four-bladed propellor domain of rabbit serum haemopexin (RCSB PDB code 1hxn, Faber *et al.*[31]), are used to detect the location of each propellor blade in the protein and provide information about the topology within the blade. This protein ''is a serum glycoprotein that binds heam reversibly and delivers it to the liver where it is taken up by receptor mediated endocytosis''[31] (see Figure 4(a)). This domain consists of a 4-fold repeat of four/five stranded antiparallel β-strands linked to each other by β-turns. The sheets pack side-to-side and are twisted and radially positioned around their central tunnel, which serves as an ion-binding site. A structural summary with residues numbered from 0 is displayed for 1hxn in Table 1. In this work the repeating motif is assumed to terminate after the last β-strand in the propellor blade. The average length of each motif is 50 residues, with the longest and shortest repeats being 44 and 57 residues, respectively. Generally the first element in the repeating motif is an α-helix situated on the circumference of the structure. This is followed for most motifs by four β-strands which compose the propellor blades with the fourth β-strand residing on the circumference of the protein. Motifs 3 and 4 have α-helices inserted between β-strands 3 and 4. There is a long loop insertion (residues 155-170) between strands 1 and 2 of motif 3. From the sequence data of haemopexin two repeats are identified by RADAR,[13] an algorithm based on multiple alignments of the protein sequence against itself. The repeats are 51 residues long and occur between residues 10-66 and 108-161. These limits correspond well to the length of repeating motif unit but straddle the neighboring motifs β-strands, also

4(a) Rasmol cartoon of haemopexin

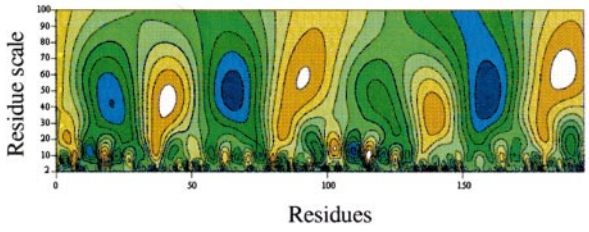


4(b,c) rASA & HΦ data

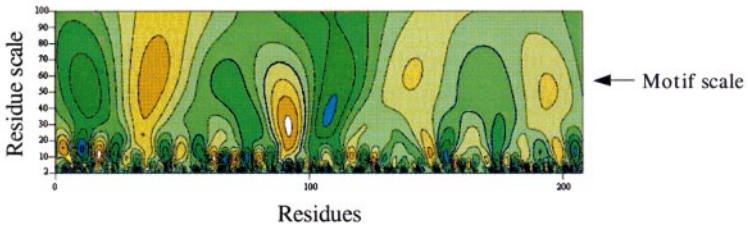
Figure 4 (*legend shown on page 348*)



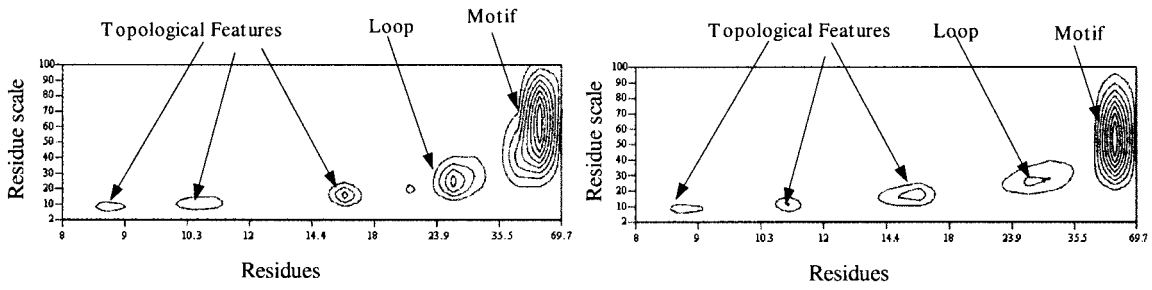
4(f,g) rASA scalogram & Rasmol Loop Insertion (colored green) residues 155-170



4(h) rASA scalogram with residues 158-170 deleted.

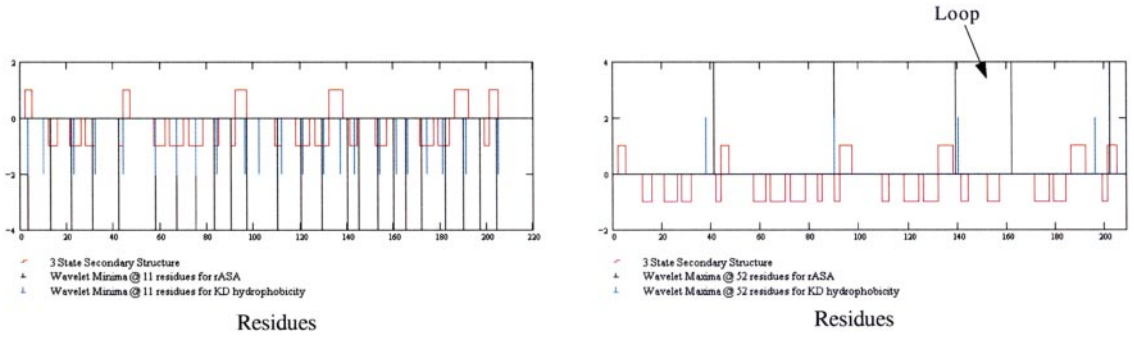


4(i). HΦ Scalogram

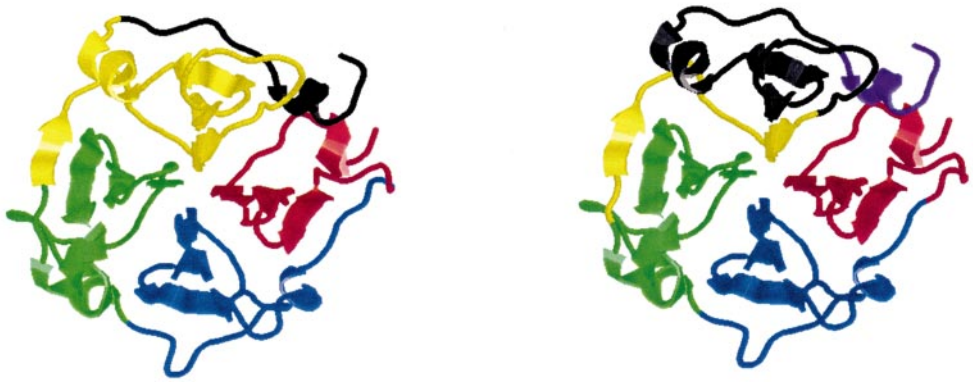


4(j,k) Scale frequency measure for rASA and HΦ

Figure 4 (*legend shown on page 348*)



4(l,m) Extrema for HΦ and rASA at 11 residues (l) and 52 residues (m)



4(n,o) Rasmol representations of HΦ and rASA maxima

Figure 4. Fourier and wavelet analysis of haemopexin. (a) Rasmol cartoon of haemopexin. (b) and (c) rASA and HΦ data. (d) and (e) rASA and HΦ Fourier spectra. (f) and (g) rASA scalogram and Rasmol loop insertion (colored green) residues 155-170. (h) rASA scalogram with residues 158-170 deleted. (i) HΦ Scalogram. (j) and (k) Scale frequency measure for rASA and HΦ. (l) and (m) Extrema for HΦ and rASA at 11 residues (l) and 52 residues (m). (n) and (o) Rasmol representations of HΦ and rASA maxima.

blades 2 and 4 are not recognized. An alternative method is required to identify the repeating motif consistently.

Figure 4(b) and (c) depicts the input data (rASA and HΦ); no obvious evidence of repeating motifs is apparent from visual inspection of this data. However, Fourier analysis displays a dominant frequency at approximately 50 residues for both data sets (Figure 4(d) and (e)); this is indicative of the 50-residue propellor motif. Additionally, some evidence of motif topology is given by large amplitude Fourier spikes in the range 10-30 residues for both data sets. These spectral peaks may characterize the repeats within each blade. It should be noted that the Fourier transform gives no information on the location of the repeat motifs, as it is a global measure. In contrast, as wavelet transforms are localized, repeating motifs should be identified with more confidence.

The rASA wavelet scalogram (Figure 4(f)) unfolds the distance-scale organization of the data for scales of 2 to 100 residues. Yellow and orange regions on the scalogram indicate exposed/hydrophilic regions, blue and dark green regions display buried/hydrophobic regions. A range of features is displayed prominently at a scale corresponding to 50 residues in length over the entire domain: this reflects the repeating propellor sheet. Between 124-185 residues there is irregularity with a series of features centered at a scale of 26 residues. These result from a large hydrophilic loop insertion of approximately 15 residues in length. The loop insertion (highlighted in green, Figure 4(g)) is displayed in Fourier space as a large amplitude spike at 26 residues (see Figure 4(e)). Removing residues 158-170 (i.e. 12 residues from the inserted loop region) for rASA results in a scalogram where the insertion features resident at a scale of 26 residues

Table 1. Structural summary for 1hxn

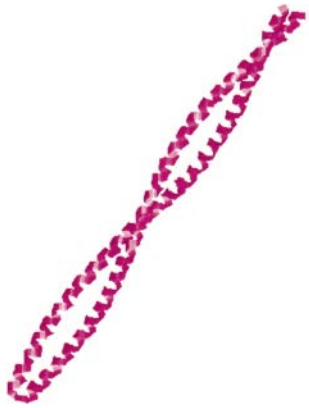
Motif	Motif range & length[a]	α 1	β 1	β 2	β 3	α 2	α 3	β 4	β 5
1	225-268 44	2-4 TRC	12-15 AMVS	21-25 TYVFS	28-31 HYWR			42-43 WP	
2	269-316 48	44-46 IAH	57-61 AAFSW	64-69 KLYLIQ	72-77 KVYVFL			83-84 TL	90-91 KR
3	317-368 52	92-96 LEKEL	109-111 AAF	118-123 RLHIMA	126-131 RLWWLD	132-134 LKS	135-137 GAQ	141-143 TEL	
4	369-425 57		152-156 GALCM	171-176 NLYLIH	179-183 NLYCY	186-191 VDKLNA		199-200 QR	
Avg. length	50.25	3.67	4.25	5.75	5.2	4.5	3	2.25	3

[a]In this column residues are numbered as in the RCSB PDB, for the rest of this work residues start from 0, i.e. residue 227 in the PDB corresponds to residue 2 in this instance.

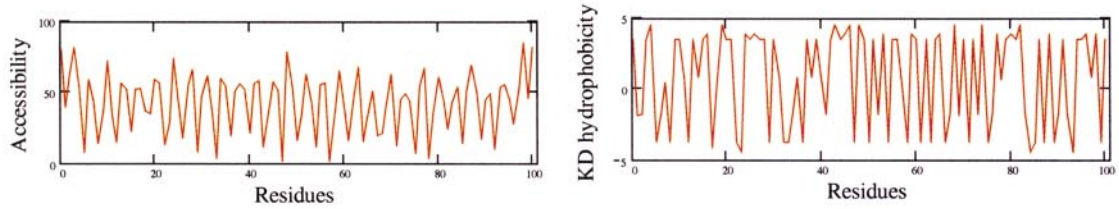
are absent (see Figure 4(h)). The dominant features over the region 124-185 residues now correspond to the propellor sheet period centered at a scale of approximately 50 residues. Also evident on the scalogram over residue scales 5 to 30 are a range of features which correlate strongly with the topology of each motif when compared to the 3D structures of the polypeptide. Detailed analysis of these features is undertaken below with the aid of wavelet extrema, which can allow a more legible representation of the wavelet data by denoising the scalogram to reveal the ''skeleton'' structure of the data (see Figure 3(e)). The scalogram for HΦ data (Figure 4(i)) displays some correlation with rASA, apart from the insertion region, which is much less apparent in this case. The frequency content of both scalograms is given by the Fourier scale transform (Figure 4(j) and (k)), where at a period of 50 residues the motif repeat is displayed from scales 20 to 100 residues for both sets of data. The correlation between Figure 4(j) and (k) is striking. At smaller scales (≈30-7 residues) a number of features are common to both data sets, these are assumed to provide information on the topological components of each motif (i.e. periodicities related to the location of β-hairpins). More precise information on these elements is given by comparing wavelet extrema at the scale of interest to three state secondary structure, this allows association of a wavelet/signal frequency with structural and topological features. For example the wavelet minima at a scale of 11 residues (Figure 4(l)), set to equal −2 and −4 for HΦ and rASA, indicate that the center of the β-sheet is the most hydrophilic region at this scale. Wavelet maxima at this scale map to β-hairpins, which connect the β-strand structural subunits. This indicates that these regions are locally the most exposed. It is a useful property of the wavelet transform to unfold topological information contained in both sequence and structural data. More specifically rASA identifies 16 of the 17 strands. HΦ is not as effective, detecting 14 strands, both data having four minima outwith the strand regions. Using the same approach wavelet maxima at the dominant scale of 52 residues (as indicated by Fourier and wavelet scale frequency analysis) are now analyzed to quantify the motif detection for both data types (see Figure 4(m)). Although, we note that strictly speaking it is impossible to define the beginning and end of repeats due to their circular nature. Positive maxima are chosen in this case, as the end of the propellor motif is likely to be the most exposed or hydrophilic region of the motif. All maxima are set to 2 and 4 for HΦ and rASA accordingly. Most rASA *MM* spikes delimit the β-sheets contained in each propellor blade, apart from the spike at 163 residues which results from the insertion discussed above. The HΦ maxima detect all four propellor blades. Mapping the maxima limits into a Rasmol protein cartoon for each data type shows that the detected repeats correlate well with those observed (Figure 4(n) and (o)).

Coiled coils

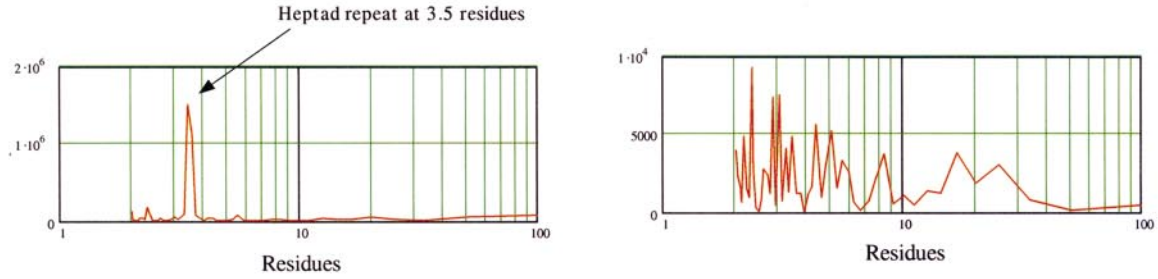
In this section we attempt to predict the location of heptad repeats in chain A of the coiled coil dimerisation domain from cortexillin I (Figure 5(a); RCSB PDB code 1d7m; Burkhard *et al.*[32]). Cortexil-



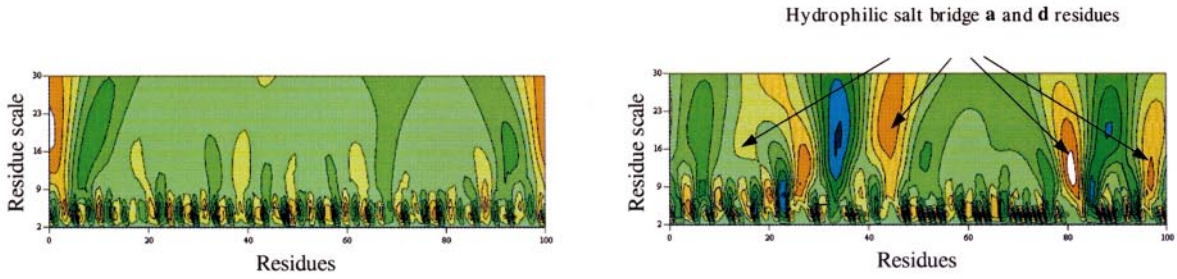
5(a) Rasmol cartoon of cortexillin I



5(b,c) rASA & HΦ Data

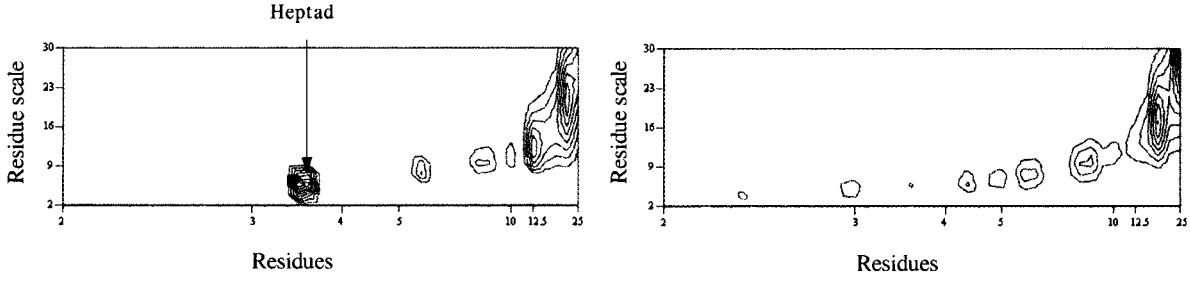


5(d,e) rASA & HΦ Fourier Spectra

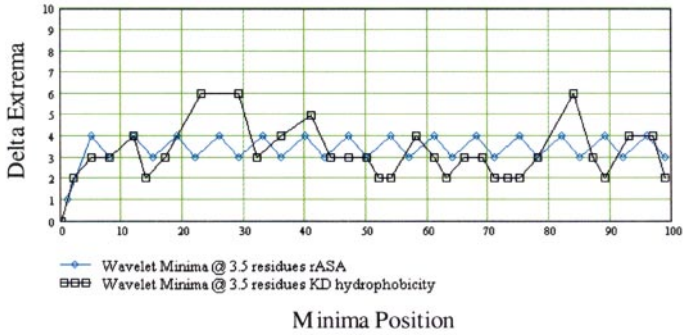


5(f,g) rASA & HΦ scalograms

Figure 5 (*legend shown opposite*)



5(h,i) rASA & HΦ Fourier scale transforms



5(j) Wavelet minima slice for rASA & HΦ at 3.5 residues.

Figure 5. Fourier and wavelet analysis of the coiled coil 1d7m. (a) Rasmol cartoon of cortexillin I. (b) and (c) rASA and HΦ data. (d) and (e) rASA and HΦ Fourier spectra. (f) and (g) rASA and HΦ scalograms. (h) and (i) rASA and HΦ Fourier scale transforms. (j) Wavelet minima slice for rASA and HΦ at 3.5 residues.

lin I and II are actin bundling proteins which play a major role in dictyostellion cytokinesis.[33] Coiled coils are very common in the eukaryotic genomes and represent one of the simplest tertiary structures. Most coiled coils display seven-residue patterns called ''heptad repeats'' containing amino acid residues denoted ***abcdefg***, in which the ***a*** and ***d*** residues are generally hydrophobic. The coiled coil structure is then formed by the component helices twisting together to bury their hydrophobic seams, which results in the helices themselves coiling around each other. It is this ''knobs-into-holes'' packing that defines a domain as a coiled coil.[34] RADAR[13] detects four 19 residue sequence repeats, although after extensive investigation of the data, the RADAR program, and the literature it is unclear why these repeats occur.

Figure 5(b) and (c) displays rASA and HΦ data. For rASA data there is a very strong periodic pattern at approximately three to four residues, the HΦ data is much more complex with no obvious indication of a global periodicity throughout the signal. The Fourier spectra of both the rASA and HΦ data are displayed in Figure 5(d) and (e) where rASA data exhibit a dominant frequency of 3.5 residues, which is often indicative of the heptad repeat observed in coiled coils.[35] Note that this differs from α-helices which generally exhibit a repeat frequency of 3.6 residues. However, the HΦ spectra depict noisy multi-period spikes with no evidence of periodicity at 3.5 residues. As expected the scalogram for the rASA data (Figure 5(f)) shows a series of features centered at three to four residues for the entire length of the data; also shown are large high amplitude features at large residue scales at the beginning and ends of the scalogram, which result from the ends of the polypeptide being more exposed than the center. Visual inspection of the rASA data supports this with residues generally tending to becoming more exposed at the chain termini. The HΦ scalogram (Figure 5(g)) depicts a more complicated organization with large features over the range 10 to 25 residues scale scattered without obvious organization along the signal. Inspection of the literature (Burkhard *et al.*[32]) reveals that the location of these features corresponds to hydrophilic residues situated in the **a** and **d** positions of the heptad. These residues tend to form interhelical salt bridges with chain B, corrupting the possibility of hydrophobic residues at positions **a** and **d** of the heptad. At smaller scales there are a number of features scat-

tered intermittently at varying periods (two to eight residues) these are consistent with the multi-period spikes displayed on Figure 5(e). These features give little indication of the heptad associated with coiled coils and their role in the protein structure is unclear.

The Fourier scale transform of the rASA wavelet coefficients (Figure 5(h)) reveals low frequency features associated with the exposed ends of the protein, and a strong indication of the heptad repeat at 3.5 residues over scales two to ten residues. Results for HΦ (Figure 5(i)) exhibit little indication of a heptad repeat. A number of low frequency features are also displayed; as discussed above these result from interhelical saltbridges.

For the rASA data more detailed evidence of heptad repeats can be seen in Figure 5(j), where a modulus slice at a wavelet scale corresponding to 3.5 residues reveals heptad repeats for the full length of the structure. These form an alternate repeating pattern of three and four residues. Also illustrated here are results for HΦ data; as expected the quality of this detection is poor, with little indication of heptad repeats displayed. Further investigation of other coiled coils not displayed herein, reveals similar results with rASA being an accurate detector of heptad repeats, and HΦ being a poor detector. The methods presented here appear to be poor predictors of heptad repeats from sequence; existing homology-based coiled coil predictors are much more powerful.[36–38] Although, scalogram analysis of the sequence data reveals the location of hydrophilic residues associated with interhelical salt bridges.

TIM Barrels

Chain A of the enzyme triosephosphate isomerase (Figure 6(a), RCSB PDB code 1tim; Banner *et al*.[39]) is involved in glycolysis where it catalyzes the conversion of dihydroxyacetone phosphate into glyceraldehyde-3-phosphate. The core of this protein consists of eight tightly packed twisted β-strands arranged in a barrel-like fashion. The α-helices that join the β-strands are on the circumference of the barrel. A structural summary is shown for 1tim chain A in Table 2. The average length of the eight βα motifs is 30.5 residues, with the largest and shortest repeat lengths being 19 and 50 residues, respectively, resulting from the high degree of variability in the length of secondary structure repeat elements and numerous α-helix insertions. The strands are all connected by at least one helix, which may be split into two or three short helices, resulting in 42.9% of the structure being helix, only 17% of the structure being β-strand. No sequence repeats were identified by RADAR (Heger & Holm[13]).

Little evidence of the βα repeat motif is displayed in the rASA and HΦ data sets (Figure 6(b) and (c)). Similarly no obvious information on the repeat motif is depicted in the rASA spectra (Figure 6(d)), which displays a dominant period at approximately 17 residues; this might be indicative of long α-helices in the structure. The HΦ spectra reveal a large spike at 40 residues (Figure 6(e)) which may represent elongated βα motifs or motifs with insertions. These results indicate that, due to the high variability in the motif length, finding a scale of investigation where most of the repeating motif information is contained might be problematic. For example investigation of the HΦ data at a scale of 40 residues may fail to elucidate the shorter repeat motifs displayed at smaller scales and *vice versa* for the rASA data. This is reflected in the scalograms of both data having a propensity to display motifs of varying lengths (Figure 6(f) and (g)). Also as might be expected the similarity between both sets of wavelet scalogram data is reflected in a correlation between their Fourier scale transforms (Figure 6(h) and (i)). Both rASA and HΦ data display a series of peaks at 40, 28 and 20 residues. These features tend to smear together for the structure data, indicating that the motif is multi-scaled. Also shown are strong features at 17 and 13 residues. Investigation of wavelet extrema with the aid of secondary structure at these scales (not shown herein) revealed there is periodicity between the locally exposed hydrophilic ends of the protein and also the hyrdophobic center residues of the β-strands and α-helices. Results for the sequence data are more defined with a strong feature at 40 residues indicative of the long motifs; this is also displayed in the scalogram and Fourier transform of this data. The task now is to select an appropriate scale which contains information on all repeating motifs. Visual inspection of both scalograms confirms that an appropriate scale of investigation is approximately 30 residues, as most motifs are evident, but not necessarily centered, at this scale. This is supported by both Fourier scale transforms which display peaks at 28 residues. Figure 6(k) depicts wavelet minima at this scale and three state secondary structure for validation of the wavelet data. The rASA data successfully delimits the βα motif, with the exception of an outlying minima spike at 185 residues. Interestingly, the HΦ data derived solely from sequence offers good agreement with the rASA results, although HΦ detections are generally less precise at discerning the location of the motif, and the number of maxima unassociated with the protein structure is greater than the rASA case. Rasmol cartoons of the limits derived from the minima data are displayed in Figure 6(l) and (m). It should be noted that this approach fails to detect the large α-helix insertions in this structure. To locate these a scale of 20 residues corresponding to the next peak on the Fourier scale transforms is investigated. The rASA maxima (Figure 6(n)) at this scale locates the βα repeat and the associated insertions with a high degree of accuracy apart from three errors, namely erroneous structure detections at 17 and 70 residues and a failure to delimit the end of the helix insertion at 203 residues. In contrast HΦ performs poorly with the

Table 2. Structural summary for 1tim chain A

Motif	Motif range & length[a]	β 1	α 1	α 2	α 3
1	7-37 31	5-9 FVGGN	16-28 RKSLGELIHTLDG		
2	38-58 21	36-40 EVVCG	45-52 YLDFARQK		
3	59-89 31	57-62 IGVAAQ	78-84 PAMIKDI		
4	90-122 33	88-91 WVIL	94-100 SERRHVF	104-116 DELIGQKVAHALA	
5	123-159 37	121-127 VIACIGE	129-134 LDEREA	137-151 TEKVVFQETKAIAD N	
6	160-204 45	158-164 VVLAYEP	165-167 VWA	176-194 PQQAQEVHEKLRG WLKTHV	196-201 DAVAVQ
7	205-227 23	203-206 RIIY	214-219 NCKELA		
8	228-248 21	226-229 GFLV	231-234 GASL	237-242 EFVDII	
Avg. length	30.25	5.25	6.75	13.25	6

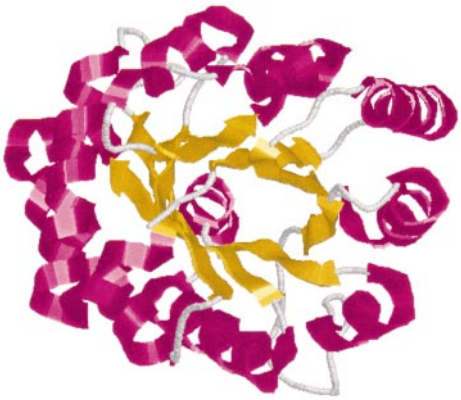
[a]Numbered according to RCSB PDB.

positive maxima only able to detect one repeat motif clearly. This results from the extra difficulty of having numerous insertions in the protein structure. Further investigation of other TIM barrels (not displayed herein) reveals similar results, with rASA being an accurate detector of both the βα motif and insertions. HΦ is generally much poorer as it is unable to distinguish between insertions and deletions consistently.

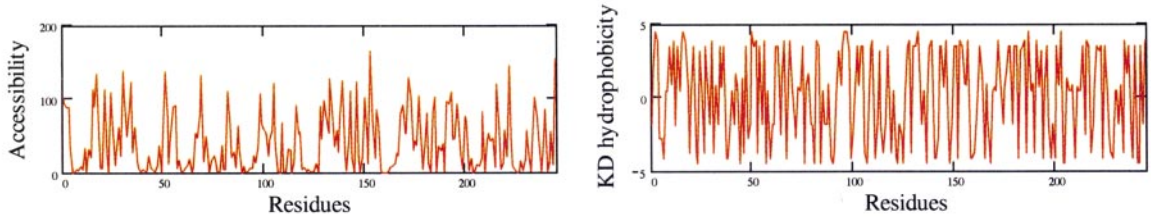
LRR (leucine rich repeats)

In this section we attempt to predict the length and location of a repeating βα motif in chain I of the LRR ribonuclease inhibitor (Figure 7(a); RCSB PDB code 1dfj; Kobe & Desienhofer[40]), which inhibits the activity of RNase A-type enzymes. LRR repeats are tandem homologous amino acid sequences of about 20-30 residues. In this protein there are 15 repeating motifs of two variants which alternate along the protein chain: type A, with 29 residues, and type B with 28 residues. Furthermore, at the chain termini there are two short regions with non-homologous sequences. This repeat produces a βα motif, with tandem repeats arranged consecutively and parallel to a common axis, causing the structures to adopt a curved shape akin to a horseshoe, where parallel β-sheets line the inner circumference and α-helices line the circumference. A structural summary is illustrated for 1dfj (chain I) in Table 3. The β-sheet component of this motif has a constant length of three residues, whereas the α-helix region displays more variability with an average length of 12.53 residues and maximum and minimum lengths of 11 and 13 residues, respectively. Unlike the previous examples this protein has no obvious insertions or deletions.

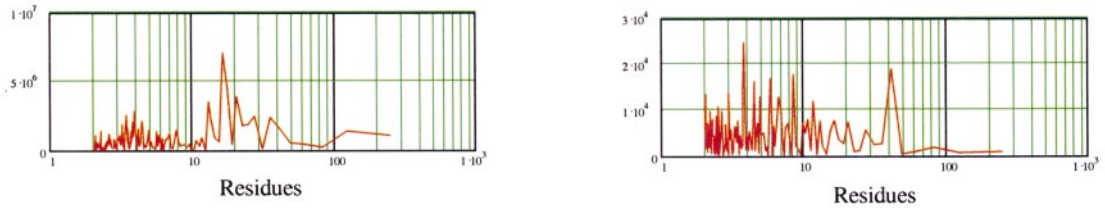
RADAR (Heger & Holm[13]) reveals the sequence data to contain 14 repeats of length 26 residues. The RADAR repeat results (not displayed herein) detect both parts of the α and β secondary units within their repeat length, but fail to delimit the repeat boundaries clearly and consistently. The data for both rASA and HΦ are shown in Figure 7(b) and (c); from these it is difficult to discern the repeating motif or the secondary structure



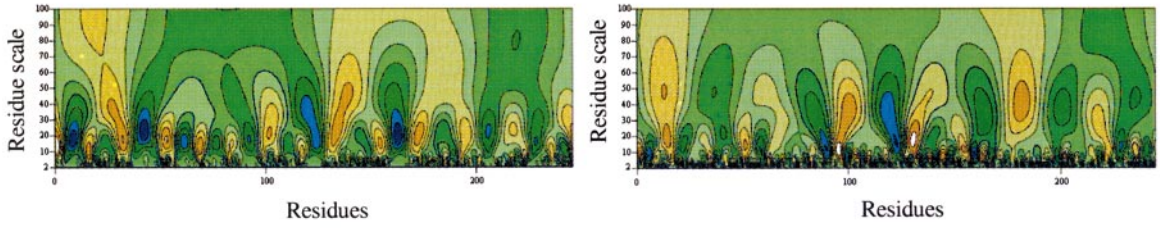
6(a) Rasmol cartoon of 1tim chain A.



6(b, c) rASA & HΦ Data



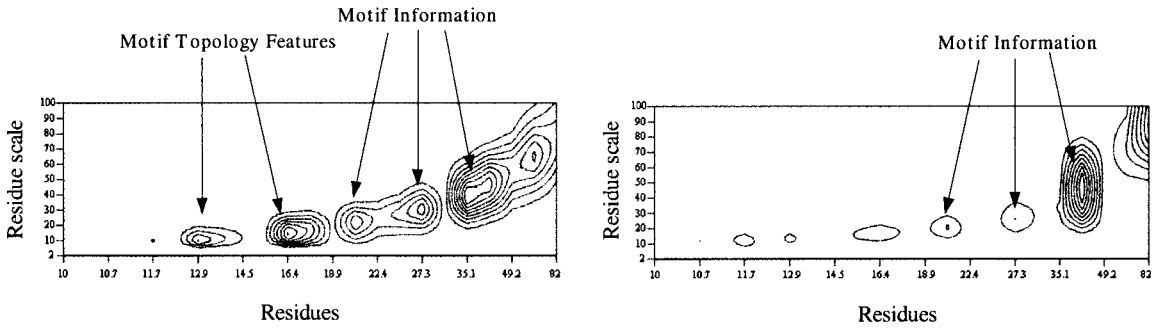
6(d, e) rASA & HΦ Fourier Spectra



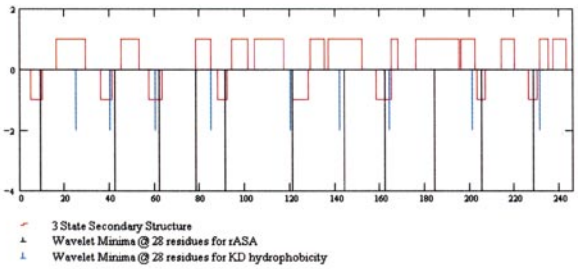
6(f, g) rASA & HΦ Scalograms

Figure 6 (*legend shown on page 356*)

components. Fourier spectra of the ASA data (Figure 7(d)) reveal a large spike at 3.6 residues, indicative of the periodicity associated with an α-helix. Spikes also occur at 14 and 29 residues, and these may represent the α-helix unit (average length 12.53 residues) and the repeating motif. The HΦ spectrum (Figure 7(e)) is less clear with little trace of the repeating motif at 29 residues and a



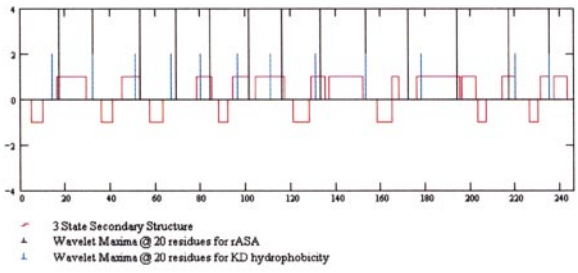
6(h,i) rASA & HΦ Fourier scale transforms



6(j) Wavelet maxima for rASA and HΦ at residue scale 28.

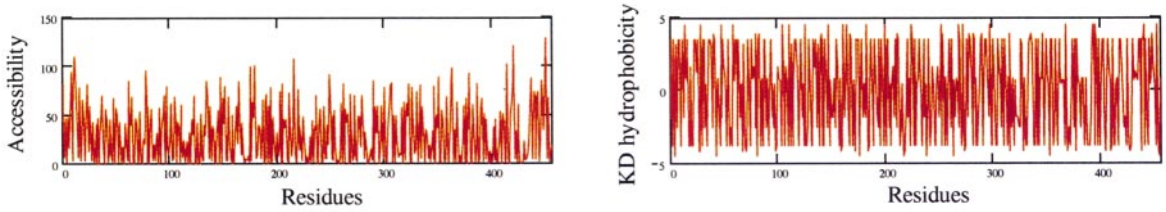
6(k) Rasmol representation of rASA maxima

6(l) Rasmol representation of HΦ maxima

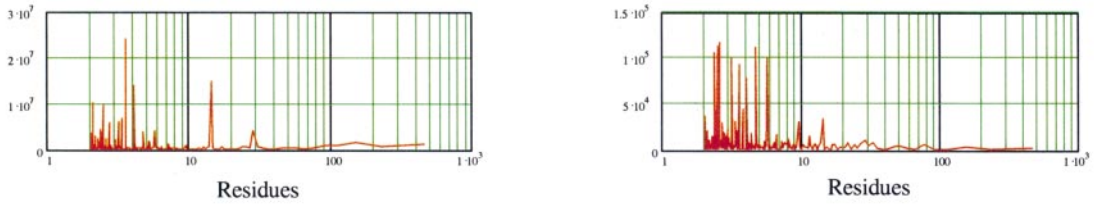


6(m) Wavelet minima for rASA and HΦ at residue scale 20.

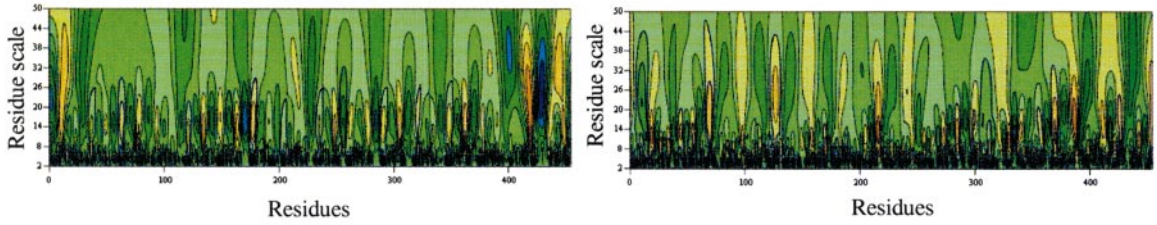
7(a) Rasmol cartoon of LRR ribonulease inhibitor



7(b,c) rASA & HΦ data



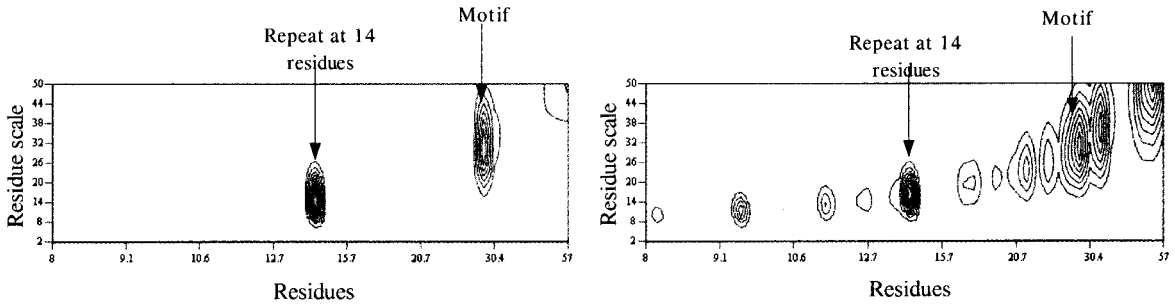
7(d,e) rASA & HΦ Fourier spectra



7(f,g) rASA & HΦ scalograms

Figure 7 (*legend shown on page 358*)

Figure 6. Fourier and wavelet analysis of the TIM barrel 1tim chain A. (a) Rasmol cartoon of 1tim chain A. (b) and (c) rASA and HΦ data. (d) and (e) rASA and HΦ Fourier spectra. (f) and (g) rASA and HΦ scalograms. (h) and (i) rASA and HΦ Fourier scale transforms. (j) Wavelet maxima for rASA and HΦ at residue scale 28. (k) Rasmol representation of rASA maxima. (l) Rasmol representation of HΦ maxima. (m) Wavelet minima for rASA and HΦ at residue scale 20.



7(h,i) rASA & HΦ Fourier scale transforms

Figure 7 (*legend shown on page358*)

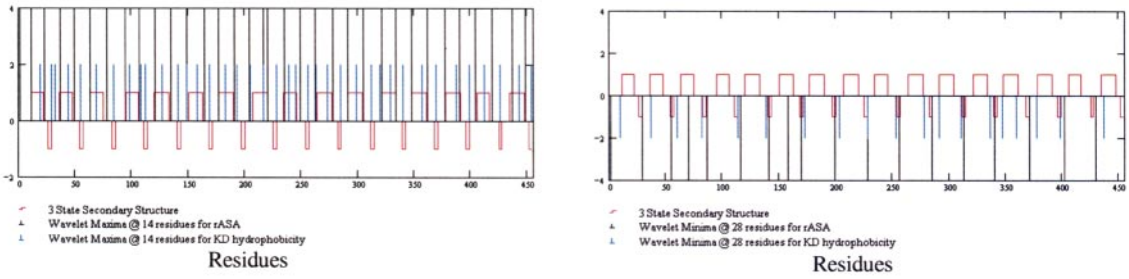
number of noisy high energy spikes residing in the range two to five residues, although an indication of the α-helix subunit is given by the small peak at 14 residues.

The rASA scalogram (Figure 7(f)) displays a dominant series of features for the entire length of the data at a scale centered at 14 residues, which are deemed to result from exposed residues which border the α-helix subunit. Additional large high amplitude features are exhibited at residue scales 40 to 100 at the beginning and ends of the scalogram, resulting from the ends of the polypeptide being exposed, as seen for the coiled coil 1d7m. Evidence of the repeating motif occurs throughout the entire length of the data at a residue scale of approximately 30 residues. Investigation of the HΦ scalogram (Figure 7(g)) reveals a series of features centered at approximately 30 residues, which are obvious over the first part of the data; in the second half large high period features are exhibited along with an intermittent range of features at a residue scale of approximately 14 residues, which

Table 3. Structural summary for 1dfj chain I

Motif	Motif range & length[a]	β	α
1	27-54, 28	26-28, VVR	36-47, EEHCKDIGSALR
2	55-83, 29	54-56, ELC	63-74, GDAGVHLVLQGL3
3	84-111, 28	83-85, KLS	95-105, GCGVLPSTLRS
4	112-140, 29	111-113, ELH	120-132, GDAGLRLLCEGLL
5	141-168, 28	140-142, KLQ	150-162, AASCEPLASVLRA
6	169-197, 29	168-170, ELT	177-189, GEAGARVLGQGLA
7	198-225, 28	197-199, TLR	207-219, PANCKDLCGIVAS
8	226-254, 29	225-227, ELD	234-245, GDAGIAELCPGL
9	255-282, 28	254-256, TLW	264-276, ASGCRDLCRVLQA
10	283-311, 29	282-284, ELS	291-303, GDEGARLLCESLL
11	312-339, 28	311-313, SLW	321-333, AACCQHVSLMLTQ
12	340-368, 29	339-341, ELQ	348-360, GDSGIQELCQALS
13	369-396, 28	368-370, VLC	378-390, NSGCSSLASLLLA
14	397-425, 29	396-398, ELD	406-416, DPGVLQLLGSL
15	426-452, 28	425-427, QLV	435-447, EEVEDRLQALEGS
Avg. length	28.33	3	12.53

[a]Numbered according to RCSB PDB.



7(j,k) .Extrema for HΦ and rASA at 14 residues (j) and 28 residues (k)

7(l) Rasmol representation of rASA wavelet maxima

7(m) Rasmol representation of HΦ wavelet maxima

7(n) Rasmol representation of REP server results, black residues indicate no motif .

Figure 7. Fourier and wavelet analysis of LRR ribonuclease inhibitor. (a) Rasmol cartoon of LRR ribonuclease inhibitor. (b) and (c) rASA and HΦ data. (d) and (e) rASA and HΦ Fourier spectra. (f) and (g) rASA and HΦ scalograms. (h) and (i) rASA and HΦ Fourier scale transforms. (j) and (k) Extrema for HΦ and rASA at 14 residues (j) and 28 residues (k). (l) Rasmol representation of rASA wavelet maxima. (m) Rasmol representation of HΦ wavelet maxima. (n) Rasmol representation of REP server results, black residues indicate no motif.

may correspond to the residues near the termini of the α-helix subunit. The Fourier scale transform for rASA (Figure 7(h)) displays the low frequency artifacts associated with the exposed protein ends and also the βα motif and α-helix subunit clearly. HΦ results are less defined with a number of low frequency features evident (as also seen in the corresponding scalogram) although evidence of the βα

motif and α-helix subunit are displayed at higher frequencies (Figure 7(i)).

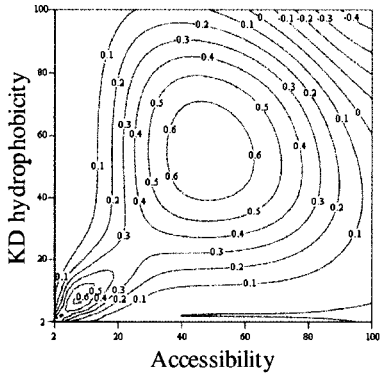
Further analysis of the wavelet data at a wavelet scale of 14 residues is shown by wavelet maxima for both data in Figure 7(j) along with three state secondary structure. These results show that rASA maxima are consistent detectors of α-helix termini with 15 of the 16 subunits delimited. HΦ maxima border the termini of ten α-helices but with numerous additional, apparently random, maxima. These results indicate that the regions surrounding the α-helix termini are locally the most exposed at this scale of investigation. Figure 7(k) depicts rASA and HΦ wavelet minima, at a scale corresponding to 30 residues. As might be expected for such a structure displaying obvious regular repeating motifs with few or no insertions or deletions, the rASA wavelet transform minima are excellent motif detectors. In comparison the HΦ results lack the accuracy of rASA but typically delimit the repeating motif to an average accuracy of 4.53 residues. Discrepancies arise in this case from both over and underdetection of the motif limits. To assist interpretation of these results cartoons of the wavelet data are displayed in Figure 7(l) and (m). Also displayed (Figure 7(n)) is a protein sequence motif result from the REP protein server[10] (available from URL: (http://www.embl-heidelberg.de/ ~ andrade) where black colored structure indicates that no motif has been found. It is observed that this method fails to accurately delimit the βα motif properly, often locating the motif start at the end of the α-helix subunit. Furthermore the REP server only predicts the location of 6 of the 15 repeating motifs. Visual comparison of rASA, HΦ and REP cartoons indicates that the wavelet based methods are the more consistent predictors in this case. Studies of other LRR structures indicate that both rASA and HΦ are good predictors of the βα repeat motif in these cases.

Discussion

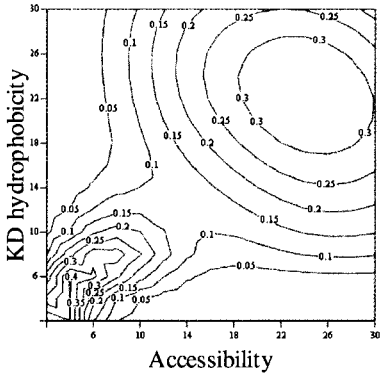
Protein repeats are important as they are very common and clearly reflect the evolutionary development of stable proteins. Here wavelet transform analysis for the detection and characterization of repeating motifs from rASA and HΦ data can be considered a success for all but the coiled coil of cortexillin I (using the HΦ data), although these data provide information on the location of interhelical salt bridges. Furthermore, by investigating high energy scales smaller than the motif scales it is possible to characterize the motif. For example wavelet maxima at a scale of 11 residues indicate the presence of hydrophilic β-turns in the propellor motif of 1hxn for both structural and sequence data. We also note a typically strong correlation between rASA and HΦ data. Using a novel 2D Pearson correlation measure this correlation can be quantified for each scale against all other scales. Details of this measure are given in the Appendix.

Figure 8 displays the 2D correlation measure for the four polypeptides tested; the topology of each correlation map differs, indicating that the relationship between rASA and HΦ data varies with structure. In this work correlations are considered statistically significant if $R > R_S$, where R_S is the significant correlation calculated at a confidence level of $P < 10^{-9}$ from statistical tables. Values of R_S for the test proteins are displayed in the Figure legend. For haemopexin (Figure 8(a)) there is a strong correlation between 5 and 15 residues indicative of the wavelet coefficients associated with β turns and hydrophilic β-strands evident in both data (see Figure 4(f) and (i)-(l)). At larger scales (40-60 residues) a second area of strong correlation ($R > 0.6$) corresponds to the scales where the motif is evident in both data. As might be expected the correlations displayed for cortexillin I are weaker than haemopexin (Figure 8(b)) with no correlations greater than R_S. The strongest correlation occurs at approximately three residues for both data and may correspond to the heptad repeats (period 3.5 residues) exhibiting some agreement with the small scale multi period structure of the HΦ data. The correlation measure for triosephosphate isomerase (Figure 8(c)) indicates that the data are well correlated ($R > 0.5$) with each other at all but the smallest scales (less than five residues). No peak is displayed at 28 residues (the average motif scale) although the correlation approaches 0.5 here; this is unsurprising considering the varying length of the motif (see Figure 6(f)-(i)). Indeed, all motif scales are well correlated, with R values tending to grow stronger at the larger scales; this may indicate that HΦ data may be useful for tertiary structure prediction for this type of protein. Further investigation is clearly warranted here. The ribonuclease inhibitor (Figure 8(d)) displays some correlation in the region 5-15 residues; this results from the wavelet coefficients that are associated with the α-helix termini in the protein structure. Little correlation is displayed at the motif scale (28 residues) and resulting from the inability of the HΦ data in some instances to detect the location of the motif, when compared to the more informative rASA data. At scales 40-60 residues a secondary correlation is evident, but the significance of this correlation in relation to the protein structure is doubtful. Therefore, any correlations displayed must be related to their wavelet scalograms and the protein structure to validate their biological significance. It should be noted that this correlation measure may prove useful for detecting structural repeats from sequence data alone, where a query sequence is aligned and then correlated with a sequence known to contain repeats. Additionally, for highly degenerate sequences with few indels this technique could be used to complement existing sequence-repeat detection methods.

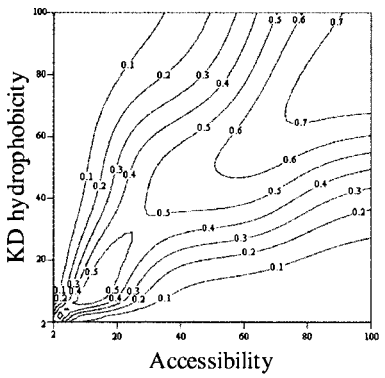
The results described in this paper indicate that HΦ data can provide useful information on repeating motifs and their topology, but with the caveat



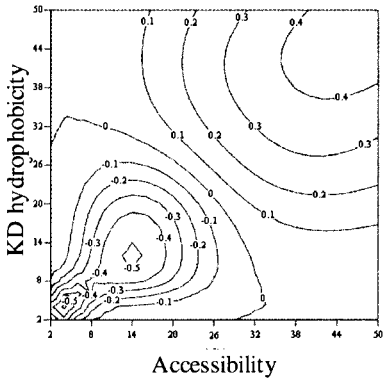
8(a) 1hxn $R_S = 0.35$



8(b) 1d7m chain A $R_S = 0.50$



8(c) 1tim chain A $R_S = 0.33$



8(d) 1dfj chain 1 $R_S = 0.25$

Figure 8. Pearson 2D correlation measure for rASA and HΦ. Axis units are residue scale. (a) 1hxn, $R_S = 0.35$; (b) 1d7m chain A, $R_S = 0.50$; (c) 1tim chain A, $R_S = 0.33$; (d) 1dfj chain, $R_S = 0.25$.

that wavelet transformations cannot easily detect insertions and deletions. However, this wavelet-based repeat detection method, with some refinement (see below), can be used to complement and validate other repeat/structure prediction methods, particularly when the target structure has low sequence similarity. The wavelet repeat method depends on the Fourier scale transform to detect the residue scale where the repeating motifs are displayed. This technique works best when the repeats are of similar lengths and there are no or few insertions and deletions. But for more complex repeats such as 1tim the motif scale can be less obvious, although using the wavelet scalogram, combined with the Fourier scale transform to resolve a suitable scale for further analysis, has proved successful. The Fourier scale transform also proved useful for finding an appropriate scale where insertions can be isolated from the motifs. This approach, however, is laborious and difficult and consequently awkward to automate, although the introduction of a multi-scale wavelet thresholding scheme† may provide a robust way of detecting varying length motif repeats. Such a method shrinks or eliminates coefficients below a certain threshold; a denoised signal, containing the repeat motifs, is then recovered by inverse transforming the wavelet coefficients. Finally, the location and type of repeats is unknown in most proteins, and consequently it becomes highly subjective to assess the quality of repeat detections for a given protein, from either sequence or structure. Thus, the next step in this study is to utilize purely structural data to identify where the repeating motifs reside

† D.L. Donoho & I.M. Johnstone: http://www-stat.stanford.edu/~donoho/Reports/index.html

with respect to insertions and deletions. Our wavelet based work has revealed that rASA carries insufficient information to predict the location and length of repeating motifs (e.g. protein secondary structure elements are not explicitly defined). It is hoped, however, that analysis of the protein 3D coordinates by SSAP (secondary structure alignment program; Orengo & Taylor[41]), combined with the wavelet techniques detailed here, will prove to be a suitable method to characterize repeating motifs from the structure. With a library of structural repeats it should be possible to improve the quality of repeat detection from sequence.

Acknowledgments

We thank David Jones, Sheena Radford and Adrian Shepherd for useful discussions. This work is supported by the BBSRC.

References

1. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A. (1998). Census of protein repeats. *J. Mol. Biol.* **293**, 151-160.
2. Heringa, J. (1998). Detection of internal repeats: how common are they. *Curr. Opin. Struct. Biol.* **3**, 338-345.
3. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all trends. *Trends Biochem. Sci.* **20**, 373-375.
4. Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E. & Sigler, P. B. (1996). Crystal structure of a G-protein beta gamma dimer at 2.1 Å resolution. *Nature,* **379**, 369-374.
5. Wilmanns, M., Priestle, J. P., Niermann, T. & Jansonius, J. N. (1992). Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 Å resolution. *J. Mol. Biol.* **223**, 477-507.
6. Papageorgiou, A. C., Shapiro, R. & Acharya, K. R. (1997). Molecular recognition of human angiogenin by placental ribonuclease inhibitor - an X-ray crystallographic study at 2.0 Å resolution. *EMBO J.* **16**, 5162-5177.
7. Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A. & Barford, D. (1999). The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell,* **96**, 99-110.
8. Leahy, D. J., Aukhil, I. & Erickson, H. P. (1996). 2.0 Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell,* **84**, 155-164.
9. Sliz, P., Engelmann, R., Hengstenberg, W. & Pai, E. F. (1997). The structure of enzyme IIAlactose from *Lactococcus lactis* reveals a new fold and points to possible interactions of a multicomponent system. *Structure,* **5**, 775-788.
10. Andrade, M. A., Ponting, C. P., Gibson, T. J. & Bork, P. (2000). Homology-based method for identification of protein repeats using statistical significance estimates. *J. Mol. Biol.* **298**, 521-537.
11. Pellegrini, M., Marcotte, E. M. & Yeates, T. O. (1999). A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins: Struct. Funct. Genet.* **35**, 440-446.
12. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
13. Heger, A. & Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Struct. Funct. Genet.* **41**, 224-237.
14. Bentley, P. M. & McDonnell, J. T. E. (1994). Wavelet transforms: an introduction. *IEE Electron. Comm. Eng. J.* **6**, 175-186.
15. Hirakawa, H., Muta, S. & Kuhara, S. (1999). The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics.* **4**, 141-148.
16. 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.
17. 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.
18. Lio, P. & Vannucci, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics,* **16**, 376-382.
19. Dodin, G., Vandergheynst, P., Levoir, P., Cordier, C. & Marcourt, L. (2000). Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theoret. Biol.* **206**, 323-326.
20. Altaiski, M., Mornev, O. & Polozov, R. (1996). Wavelet analysis of DNA sequences. *Genet. Anal.* **12**, 165-168.
21. Klevecz, R. R. (2000). Dynamic architecture of the yeast cell cycle uncovered by wavelet decomposition of expression microarray data. *Funct. Integr. Genom.* **1**, 186-192.
22. Tsonis, A. A., Kumar, P., Elsner, J. B. & Tsonis, P. A. (1996). Wavelet analysis of DNA sequences. *Phys. Rev. E.* **53**, 1828-1834.
23. Wako, H. & Blundell, T. L. (1994). 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.
24. Gao, W. & Li, B. L. (1993). Wavelet analysis of coherent structures at the atmosphere-forest interface. *J. Appl. Meteorol.* **32**, 1717-1725.
25. Addison, P. S., Murray, K. B. & Watson, J. N. (2001). Wavelet transform analysis of open channel wake flows. *J. Eng. Mech.* **127**, 58-70.
26. Mallat, S. G. & Hwang, H. L. (1992). Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory,* **38**, 617-643.
27. Hubbard, S. J. & Thornton, J. M. (1993). *Naccess. Computer Program.* Department of Biochemistry and Molecular Biology, University College London.
28. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
29. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132.
30. Jones, D. T., Tress, M., Bryson, K. & Hadley, C. (1999). Successful recognition of protein folds using threading methods biased by sequence similarity

and predicted secondary structure. *Proteins: Struct. Funct. Genet.* **37**, 104-111.
31. Faber, H. R., Groom, C. R., Baker, H. M., Morgan, W. T., Smith, A. & Baker, E. N. (1995). 1.8 Å crystal structure of the C-terminal domain of rabbit serum haemopexin. *Structure,* **3**, 551-559.
32. Burkhard, P., Kammerer, R. A., Steinmetz, M. O., Bourenkov, G. P. & Aebi, U. (2000). The coiled-coil trigger site of the rod domain of cortexillin I unveils a distinct network of interhelical and intrahelical salt bridges. *Struct. Fold. Des.* **8**, 223-230.
33. Fey, P. & Cox, E. C. (1999). Cortexillin I is required for development in polysphondylium. *Dev. Biol.* **212**, 414-424.
34. Crick, F. C. H. (1953). The packing of α-helices: simple coiled coils. *Acta Crystallog.* **6**, 689-697.
35. Lupas, A. (1997). Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* **7**, 388-393.
36. Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M. & Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA,* **92**, 8259-8263.
37. Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science,* **252**, 1162-1164.
38. Walshaw, J. & Woolfson, D. N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427-1450.
39. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C. & Wilson, I. A. (1976). Atomic coordinates for triose phosphate isomerase from chicken muscle. *Biochem. Biophys. Res. Commun.* **72**, 146-155.
40. Kobe, B. & Deisenhofer, J. (1995). A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature,* **374**, 183-186.
41. Orengo, C. A. & Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**, 617-635.

Appendix

Analyzing wavelets fall into two general categories, continuous[A1] and discrete.[A2] The choice of wavelet function used for analysis depends on a variety of factors including speed of computation, the shape of signal-specific features, the frequency resolution and the statistical analysis to be performed. The continuous and discrete transforms each have their own favorable properties. Due to its redundancy, or overspecification of the signal, the continuous transform is computationally expensive and strictly speaking does not lend itself well to statistical analysis.[A3–A5] However, it is superior in feature detection and localization due to its denser arrangement of temporal locations and spatial scales when compared with the sparse dyadic grid structure of the discrete transform. Consequently, continuous wavelets are the preferred method for resolving repeating protein motifs.

In order to be classified as a wavelet a function must have finite energy, and it must satisfy the following admissibility condition:

$$C_g = \int_{-\infty}^{\infty} \frac{|\hat{g}(\omega)|}{\omega} d\omega < \infty \qquad \text{(A1)}$$

where $|\hat{g}(\omega)|$ is the Fourier transform of $g(t)$, i.e. the wavelet must have no zero frequency component and hence have zero mean. C_g is known as the admissibility constant (which is equal to π for the Mexican hat wavelet) and is essential for energy calculations and the inverse wavelet transform.

Conversion of the wavelet a scale (which is a distance scale) to a related frequency requires knowledge of the passband center of the wavelet transform, which is given by:

$$\omega_0 = \frac{\int_{-0} \omega |\hat{g}(\omega)|^2 d\omega}{\int_{-0}^{\infty} |\hat{g}(\omega)|^2 d\omega} \qquad \text{(A2)}$$

For the Mexican hat wavelet with $a = 1$, ω_0 is equal to $\sqrt{5/2}$ rad/s (=1.58 rad/s), or in Hertz $f_c = 0.251$ Hz. The frequency, f, associated with any (other) a scale is $f = f_c/a$, or simply $f = 0.251/a$.

To determine the correlation between rASA and HΦ in wavelet space a novel 2D Pearson correlation measure is employed. Traditionally Pearson's correlation coefficient R is calculated for the entire data sets to give a global measure resulting in a single value. However, we propose to calculate Pearson's R value for all scales against each other, producing a correlation matrix. This

$$R(a,c) = \frac{\sum_b A(a,b)H(c,b) - \dfrac{\sum_b A(a,b)\sum_b H(c,b)}{N}}{\sqrt{\left(\left(\sum_b A(a,b)^2 - \dfrac{\left(\sum_b A(a,b)\right)^2}{N}\right)\left(\sum_b H(c,b)^2 - \dfrac{\left(\sum_b H(c,b)\right)^2}{N}\right)\right)}} \qquad \text{(A3)}$$

.

approach enables a more useful pseudo-local assessment of which scales in the data share common features. The correlation measure is defined as:
where $A(a,b)$ and $H(c,b)$ are the accessibility and hydrophobicty wavelet coefficients indexed by scales a and c and location b. A correlation of 1 or -1 means there is perfect positive or negative linear relationship between the data. A correlation of 0 means that there is no meaningful relationship. To test if the correlation measure is sensitive to the level dependant normalization of the wavelet coefficients, this work was repeated for the often used a^{-1} normalization.[A6] (an $a^{-1/2}$ normalization is used herein (see equation (1)), this results in all statistically significant correlations being preserved but with slightly lower correlation values (not shown herein).

References

A1. Grossman, A. & Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal*. **15**, 726-736.

A2. Daubechies, I. (1992). *Ten Lectures on Wavelets*. vol. 61 of CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, PA.

A3. Yamada, M. & Ohkitani, K. (1991). An identification of energy cascade in turbulence by orthonormal wavelet analysis. *Prog. Theor. Phys. (Japan),* **86(4)**, 799-815.

A4. Katul, G. G., Parlange, M. B. & Chu, C. R. (1994). Intermittency, local isotropy, and non-Gaussian statistics in atmospheric surface layer turbulence. *Phys. Fluids*, **6**, 2480-2492.

A5. Katul, G. G. & Parlange, M. B. (1995). Analysis of land surface heat fluxes using the orthonormal wavelet approach. *Water Resour. Res*. **31**, 2743-2749.

A6. Goupillard, P., Grossmann, A. & Morlet, J. (1984). Cycle-Octave and related transforms in seismic signal analysis. *Geoexploration*, **23**, 85-102.

Edited by G. von Heijne

(Received 20 September 2001; received in revised form 7 December 2001; accepted 10 December 2001)