Detection of protein secondary structures via the discrete wavelet transform

Jesús Pando
Department of Physics, DePaul University, Chicago, Illinois 60614, USA

Luke Sands
Department of Physics, DePaul University, Chicago, Illinois 60614, USA

Sean E. Shaheen
Department of Physics and Astronomy, University of Denver, Denver, Colorado 80208, USA

(Received 11 May 2009; revised manuscript received 14 October 2009; published 16 November 2009)

We subject the primary sequence of proteins gathered from the Structural Classification of Proteins (SCOP) database to a discrete wavelet transform (DWT) analysis to search for predictors of secondary structures. We use proteins with both alpha helices and beta sheets (the A/B, $A+B$ databases from SCOP). The amino acids composing the protein are converted to their hydrophobicity values using three hydrophobicity scales. Results prove to be independent of the scale used. Using a DWT multiresolution decomposition, each protein is coarse grained, in effect, creating snapshots of each protein at multiple scales. For each protein, a control data set is formed by generating random realizations that remove the positional informational in the sequence but still contain the same amino acid frequencies. Regions of salient hydrophobicity in the protein sequence are identified by comparing the transforms of the original sequence with those of the control set, at each resolution. We find significant matching between regions of salient hydrophobicity and the locations of secondary structure along the amino acid chains. We calculate the sensitivity, specificity, and Matthews correlation to quantify the agreement between the wavelet detected structures and the real protein. In addition we are able to distinguish between the morphologically different subsets, A/B and $A+B$. We also construct a correlation function based on the DWT that correlates quasilocalized structures at lengths in wavelet space. Through a similar comparison to the control data sets, features in this space-scale correlation are identified that show correspondence to the typical lengths of the secondary structures.

DOI: 10.1103/PhysRevE.80.051909 PACS number(s): 87.14.E−, 87.10.−e, 87.15.A−

I. INTRODUCTION

Despite advances made in both biophysical theory and computational power, a satisfactory solution to the protein folding problem remains out of reach [1]. The function of a protein is determined by its three-dimensional structure that is in turn determined by its amino acid sequence. However, this sequence-structure relationship is very complex and not fully understood. At present, there are two main approaches to try to resolve the problem. The first is based on direct calculation of the positions of the atoms in the protein using molecular dynamics or Monte Carlo simulations. Progress has been made in this approach as massively parallel computers and smart algorithms are applied to the problem [2]. The vast number of potential structures that need to be explored for a given protein make these direct calculation techniques difficult and computationally expensive. The second general approach is to use a knowledge-based algorithm for attempting to determine a protein's structure or function based on some measure of the similarity of its sequence to those of known proteins. These approaches require that learning sets be formed by training the algorithm on a large number of known sequences. The learning set is fed into a algorithm that then predicts the output state of a given amino acid within the chain. These methods rely on statistical probabilities derived from large sets of experimentally determined proteins, and they largely ignore the fundamental forces governing the dynamics of protein folding [3].

An approach that is seeing increased attention is the use of statistical analysis to uncover periodicities, correlations, or other implicit order in the one-dimensional amino acid sequence. For instance, Tiwari *et al.* have used Fourier methods to predict genomic sequences [4]. Evidence for nonrandomness in the primary sequence was found in 1996 using random walk techniques by Irbacket *et al.* [5]. Clustering of protein structures using hydrophobicity has been detected using Z-curve representations and fractal analysis [6]. Weiss and Herzel found hydrophobicity autocorrelation functions to be strongly oscillations and the α-helix propensity autocorrelation function to be monotonously decaying in a large set of nonhomologous protein sequences [7]. The wavelet transform has been used by Wen *et al.* [8] to search for functional similarity of proteins with low identity and by Pattini and Cerutti [9] to detect the presence of alpha helices in the protein secondary structure. Finally, Arneodo *et al.* [10] have used the wavelet in various ways, such as doing fractal analysis of DNA sequences.

Here we propose a wavelet-based approach that makes use of this technique's ability to detect multiscale features in a data series in order to give guidance in determining secondary protein structures given the primary sequence. The position and identity of an amino acid and its interaction with others in the protein chain and with the aqueous environment are the determining factors in the final configuration of the secondary structures. We expect that implicit structure is encoded in a given protein sequence as a result of its evolution toward a quickly folding, stable structure as well as physical constraints placed upon its three-dimensional structure, such as energetically favorable twist angles, inter-

action energies between given amino acids, and in particular the hydrophobicity of a given amino acid. We demonstrate here that the wavelet transform can be a powerful tool for uncovering this implicit information in a protein's sequence and correlating it with the secondary structure.

The approach we use here is to convert the amino acid sequence of a protein to a corresponding hydrophobicity sequence, since hydrophobic and hydrophilic interactions provide one of the strongest driving forces for the folding process. We do this by coarse graining the converted sequence via the discrete wavelet transform (DWT) and comparing the coarse grained signal to randomized sequences to detect areas of significance. We also define a space-scale correlation measure between the coarse grained identified structures that will yield important information about the size (scale) of secondary structures. The outline of the rest of the paper is as follows. In Sec. II we describe the discrete wavelet transform and how to use the resulting course grained information to yield useful statistical measures. Section III discusses the protein data selection criteria and conversion of the amino acid sequence into a hydrophobicity sequence. In Sec. IV we give our results, and Sec. V contains discussion and analysis.

II. DISCRETE WAVELET TRANSFORM

The discrete wavelet transform is ideal for studying data in which the extraction of information on both the scale and position of features is desired. A great deal of literature now exists on the DWT and its uses. Here we provide a conceptual description of the DWT and provide only the key expressions. For details see [11].

When a signal is wavelet decomposed, a localized smoothed and differenced set of coefficients is generated. The smooth coefficients are a locally coarse grained approximation to the signal, while the differenced coefficients capture the local fluctuations about a local mean. The differenced coefficients are stored, while the smoothed coefficients are passed on for further processing by the wavelet. The subsequent application of the wavelet transform on the smoothed signal produces another pair of smoothed and differenced set of coefficients which are now at half the resolution. There are half as many smoothed and differenced coefficients as in the previous case. The process continues until there are no more smooth coefficients to transform.

The smoothing and differencing is accomplished by passing the signal, $f(x)$, simultaneously through two filters, a high pass filter, $g(x)$ that gives the differenced set of coefficients, and a low pass filter, $h(x)$ that gives the smoothed set. Formally, the filtering process is a convolution between the signal and the filters. That is,

$$\phi_l(x) = f(x)h(x) = \int_l f(x)h(2x-l)dx$$

$$\psi_l(x) = f(x)g(x) = \int_l f(x)g(2x-l)dx, \tag{1}$$

where ϕ_l and ψ_l are the smoothed and difference coefficient, respectively. The subscript l indicates that the convolution is

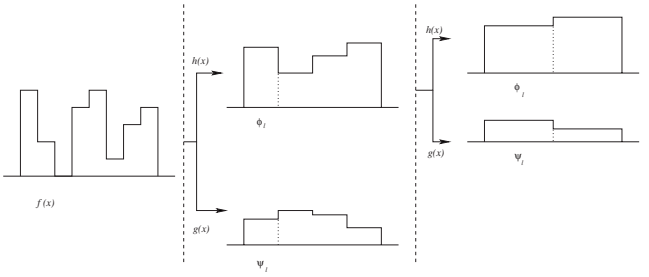

FIG. 1. Schematic representation of the DWT. The low pass filter, $h(x)$, produces a local approximation to signal, the high pass filter, $g(x)$, captures local fluctuations (after first vertical dashed line). The approximation can be further wavelet transform producing new local approximations and fluctuations at half the resolution (after second dashed line).

to be performed in a localized region l and the argument on the filter, $2x-l$, accounts for the size of the filter window. In total, there will be half as many ϕ_l and ψ_l as there were data points in the original signal. A complete reconstruction of the original data is possible by inverting the process. The process is shown schematically in Fig. 1.

One advantage of the wavelet transform is that filters at subsequent resolutions are simply stretched (or dilated) copies of the original filters. Labeling the resolution as j, we have that

$$h_{j,l}(x) = (2^j/L)^{1/2} h(2^j x/L - l), \tag{2}$$

and

$$g_{j,l}(x) = (2^j/L)^{1/2} g(2^j x/L - l), \tag{3}$$

where L is the size of the signal. In principle, the filters can be any function that obeys the admissibility condition

$$\int_R h(x)dx < \infty. \tag{4}$$

However, there is great advantage if the building block functions can be constructed so that they are orthogonal. In the late 80s, Daubechies found a construction for the these functions that accomplished this [12]. These functions do not admit a simple algebraic formula. However, Daubechies proved the simple recurrence relation for the so-called Daubechies 4 wavelet (higher order constructions are also possible),

$$h(r) = c_o h(2r) + c_1 h(2r-1) + c_2 h(2r-2) + c_3 h(2r-3),$$

$$g(r) = -c_o h(2r-1) + c_1 h(2r) - c_2 h(2r+1) + c_3 h(2r+2), \tag{5}$$

with initial values

$$h(0) = 0, \quad h(1) = \frac{1+\sqrt{3}}{2}, \quad h(2) = \frac{1-\sqrt{3}}{2}, \quad h(3) = 0 \tag{6}$$

that greatly facilitates their construction. The coefficients, c_n are

$$c_o = \frac{1+\sqrt{3}}{4}; \quad c_1 = \frac{3+\sqrt{3}}{4}; \quad c_2 = \frac{3-\sqrt{3}}{4}; \quad c_3 = \frac{1-\sqrt{3}}{4}. \tag{7}$$

These basic building blocks, $h(x)$ and $g(x)$, are translated across the signal to give the wavelet transform at the current scale. The transform at a different scale is achieved by dilating the filters and then translating across the signal. This process builds a cascade of smoothed and differenced signals to yield a multiresolution decomposition. Each tier in the cascade is at half the resolution of the previous tier. The smoothed signal is a localized average of the signal, while the differenced signal captures localized fluctuations.

Equations (5) and (6) imply that the convolution of the signal with the filters is now simple matrix multiplication. Furthermore, both the convolutions can be done simultaneously. That is, for a data vector,

$$\boldsymbol{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}_J$$

the smoothed and differenced coefficients are obtained by

$$\begin{pmatrix} \phi_0 \\ \psi_0 \\ \phi_1 \\ \psi_1 \\ \phi_2 \\ \psi_2 \\ \vdots \end{pmatrix}_{J-1} = \frac{1}{2} \begin{pmatrix} h_o & h_1 & h_2 & h_3 & & & \cdots \\ h_3 & -h_2 & h_1 & -h_0 & & & \cdots \\ & & h_o & h_1 & h_2 & h_3 & \cdots \\ & & h_3 & -h_2 & h_1 & -h_0 & \cdots \\ & & & & h_o & h_1 & \cdots \\ & & & & h_3 & -h_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \vdots \end{pmatrix}_J, \tag{8}$$

where the ϕ_l and ψ_l capture the local approximation and fluctuations of the original data at half the resolution. The subscripts $J-1$, l, give the resolution and position at which the data has been wavelet transformed. The original data are sampled at resolution J. The process then continues with the $\phi_{J-1,l}$ now serving as the data set.

The orthogonal relations for this construction are

$$\int h_{j,l}(x) h_{j,l'}(x) dx = \delta_{l,l'}, \tag{9}$$

$$\int g_{j,l}(x) g_{j',l'}(x) dx = \delta_{j,j'} \delta_{l,l'}, \tag{10}$$

$$\int h_{j,l}(x) g_{j',l'}(x) dx = 0, \quad \text{if } j' \geq j, \tag{11}$$

where the δ are the Kroneker deltas. These orthogonality relations ensure that information garnered via the DWT remains local. That is, the result of the decomposition at location l is independent of location l'. Unlike the Fourier basis which are not well localized, there is no leakage from one location to the next. When positional information is important, this property of the DWT is extremely useful.

Statistical properties

In addition to its multiresolution property, the DWT can be used to construct a variety of statistical measures for systems in which the position and scale information are relevant to uncovering structure [13,14]. In this work we concentrate on a correlation measure that measures the relationship between points displaced a distance in the wavelet scale space. As we will see, this measure reveals relevant structural information about the protein.

One of most important correlation measures is the two-point correlation, $\xi(r)$. The two-point correlation function is a measure of the excess probability for finding a neighbor a distance r away. In our case, the neighbors are the quasilocalized structures determined by the wavelet decomposition. Essentially, the two-point correlation function is a measure of the probability, dP, of finding a pair such that one object is in volume element dV_1 and the other object is in volume element dV_2. That is,

$$dP = \rho_o^2 [1 + \xi(r)] dV_1 dV_2, \tag{12}$$

where ρ_o is the mean density. One can interpret Eq. (12) as follows. For a Poisson distribution, the probability that two cells at separation r are both occupied is $\rho_o^2 dV_1 dV_2$. For clustering, the probability is modified by the term $[1+\xi(r)]$. Therefore, for a Poisson distribution, $\xi(r)=0$. If correlations exist, $\xi(r)>0$ and if data are anticorrelated then $-1<\xi<0$. Determining $\xi(r)$ gives clustering above (or below) what one would expect from a random distribution.

The orthogonality of both sets of wavelet coefficients gives great flexibility in constructing correlation measures that are analogs of $\xi(r)$. For example, the reconstructed distribution at each scale now contains structures that are coarse grained representations of the original sequence. These quasilocalized structures serve as objects localized at the different scales. Furthermore, the approximated signal is now free of some of the high frequency noise, and pronounced features arise from the signal. We can define correlation measures between the quasilocalized structures. For our purposes, we are especially interested in detecting correlations in the hydrophobicity content along a protein because a high degree of correlation could be a marker of secondary structure. To that end, we define the space-scale correlation first introduce by Feng *et al.* [15,16]

$$\chi(\Delta, j) = \frac{\langle \phi_{j,l} \phi_{j,l-l_0} \rangle}{\langle \phi_{j,l}^2 \rangle}. \tag{13}$$

The $\phi_{j,l}$ are the smoothed data found using Eq. (8) and $\Delta = l - l_o$ is the distance separating the $\phi_{j,l}$ in wavelet space at scale j. χ measures the relationship between points displaced a distance in the wavelet scale space. The coefficients $\phi_{j,l}$ are correlated with coefficients l_0 away in wavelet space. These calculations allow us to determine if there are any length scales that have significant correlations and to compare these to the lengths of alpha helices, beta strands, and beta sheets that have been found experimentally.

III. PROTEIN DATA SELECTION

We tested the above DWT algorithm on the Structural Classification of Proteins (SCOP) database, release 1.73 (http://scop.mrc-lmb.cam.ac.uk/scop/). The database contains over thirty thousand PDB entries organized by secondary structure type. For this work, we used proteins that contained both alpha helices and beta sheets. Within SCOP, these mixed secondary structure proteins come in two forms, A/B and $A+B$. The A/B proteins consist mainly of beta-alpha-beta units in which the alpha linkage allows the beta strands to be parallel. The $A+B$ proteins have distinct alpha and beta regions in which the beta strands are linked by short loop structures that force an antiparallel alignment between within the beta sheets. The two classes therefore have two differentiating characteristics: (i) the length scales on which the alpha and beta units are blended, and (ii) the sequence alignment within the beta sheets. We expect these structural difference to yield somewhat different results from this DWT analysis.

Of the 9735 proteins in the $A+B$ database, 3509 (about 36%) proteins were used for this analysis. The selection criteria we used was that the protein have only monomeric chains. Additional proteins were left off our analysis because they had missing amino acid residues within the primary structure. The three-dimensional structural descriptions were also subject to some error, which meant that the exact locations, lengths, and orientations of the secondary structure elements within a protein were not necessarily accurate. There were also apparent labeling errors of secondary structure position in some of the files that we used. In the A/B database, 3749 (about 34%) of the 10963 proteins were used using the same selection criteria as for the $A+B$ database.

IV. WAVELET STRUCTURE DETECTION

In order to extract information from a protein via the wavelet transform, its amino acid sequence must first be converted into a numeric signal. A common way to do this is to convert the sequence into its corresponding hydrophobicity. Hydrophobicity effects are the most influential factors in the folding process. In general, hydrophobic regions of the protein occur closer to the interior of the folded structure, away from the aqueous environment. Thus, the solvation environment of the protein provides a driving force for spatial separation and ordering of different sequence regions according to their hydrophobicity. Each amino acid can be characterized by its tendency to turn inward or outward and can be assigned a hydrophobicity value. There are several different hydrophobicity scales commonly in use among researchers in the field. The three scales used in this project are (Hopp-Woods—HW, Kyte-Doolittle—KD, and Engelman-Steitz—ES [17–19]). Hydrophobicity values are generally determined by the free energy difference ΔG between polar (water) and nonpolar environments for a given amino acid [20]. The three scales used here are among the most common. We note that the HW and KD scales have opposite direction. For the HW scale, a positive value indicates a hydrophilic amino acid, whereas for the KD scale a positive value represents a hydrophobic amino acid. The ES scale is based on how well a particular amino acid will enter a lipid bilayer from an aqueous environment. We will see that our results are not significantly affected by the choice of scale. Table I shows the hydrophobicity values for the three scales used in this work.

TABLE I. The three hydrophobicity scales used in this work.

Amino acid	Kyte-Doolittle	Hopp-Woods	Engelman-Steitz
ALA	1.8	−0.5	−1.6
ARG	−4.5	3.0	12.3
ASN	−3.5	0.2	4.8
ASP	−3.5	3.0	9.2
CYS	2.5	−1.0	−2.0
GLN	−3.5	0.2	4.1
GLU	−3.5	3.0	8.2
GLY	−0.4	0.0	−1.0
HIS	−3.2	−0.5	3.0
ILE	4.5	−1.8	−3.1
LEU	3.8	−1.8	−2.8
LYS	−3.9	3.0	8.8
MET	1.9	−1.3	−3.4
PHE	2.8	−2.5	−3.7
PRO	−1.6	0.0	0.2
SER	−0.8	0.3	−0.6
THR	−0.7	−0.4	−1.2
TRP	−0.9	−3.4	−1.9
TYR	−1.3	−2.3	0.7
VAL	4.2	−1.5	−2.6

The result of converting a protein's amino acid sequence to its hydrophobicity is a discrete, numeric signal. At the location of each amino acid in the sequence, we now have numeric value corresponding to the tendency of that acid to fold toward or away from its environment. We also create control data sets by randomly scrambling the order of each protein's amino acid sequence 500 times to create 500 randomized instantiations. The frequency of each kind of amino acid is kept the same in the control data sets as in the parent protein; only the ordering is altered. This is done to test the general assumption that the information that controls structure formation is contained not only in the frequency of the actual amino acids present in the chain, but also importantly in the specific ordering of these amino acids along the chain. Each randomized sequences is transformed into a hydrophobicity signal using the same hydrophobicity scale as the parent protein.

The proteins we analyze have primary sequence lengths ranging from 50 to 1000 amino acids. To use the wavelet transform on a signal of arbitrary length L, the sequence has to be lengthened to an integer power of two. Proteins are zero padded to bring the signal length up to the nearest integer power of two. Because the wavelet keeps information localized, zero padding affects only a few coefficients near the end of the original signal.

Analyses is conducted on the same set of proteins three times, each time using hydrophobicity values assigned by the

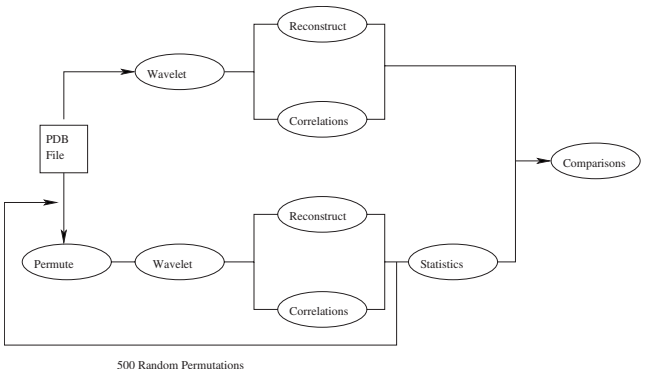

FIG. 2. A flow chart for the wavelet decomposition and statistical analysis for the SCOP proteins. Each protein is wavelet transformed after its amino acid sequence has been converted to a hydrophobicity scale. It is then reconstructed at four scales and the space-scale correlation determined. The amino acid sequence of that protein is also randomly rearranged and subjected to the same treatment (lower leg of figure). The results are averaged and compared to the real protein.

three different hydrophobicity scales. The wavelet decomposed signals are used to perform the multiscale reconstruction and space-scale correlations, and comparisons are made between the results of the actual protein and the control data. Figure 2 gives a schematic representation of the entire process. Detailed results are presented in the following section.

A. Hydrophobicity versus position

One of the primary objectives of this work is to demonstrate how wavelet analysis can be used to detect the location of secondary structures given just the amino acid sequence (as represented by their hydrophobic content) of a protein. We begin this analysis by using the multiresolution property of wavelets to investigate the coincidence of the location of secondary structures in the amino acid sequence with areas of high absolute hydrophobicity. The idea is somewhat analogous to locating genes in DNA in the sense that we were looking for localized structure embedded in a noisy background.

We proceed as shown schematically in Fig. 2. The amino acid sequence in each protein is converted to its corresponding hydrophobic value and wavelet reconstructed at different resolutions or scales. Each scale is a coarse grained representation of the previous scale. It is these multiresolution representations of the original protein that we subject to further analysis. In these kinds of analyses, it is common to set a threshold on the smooth coefficients so that any coefficient less than the threshold is set to zero. In effect, this denoises the data at multiple scales. Here we employ a different tactic. We keep all the coefficients and use the control data to set a threshold. Our hypothesis is that any signal outside this threshold is a marker for secondary structure. The procedure is performed separately on the $A+B$ and A/B data sets.

The coarse graining is accomplished using Eq. (8) (see also Fig. 1). A typical example of the resulting reconstructed hydrophobicity content is shown in Fig. 3. The figure shows the first four reconstructed levels for the protein $1RJF$. This

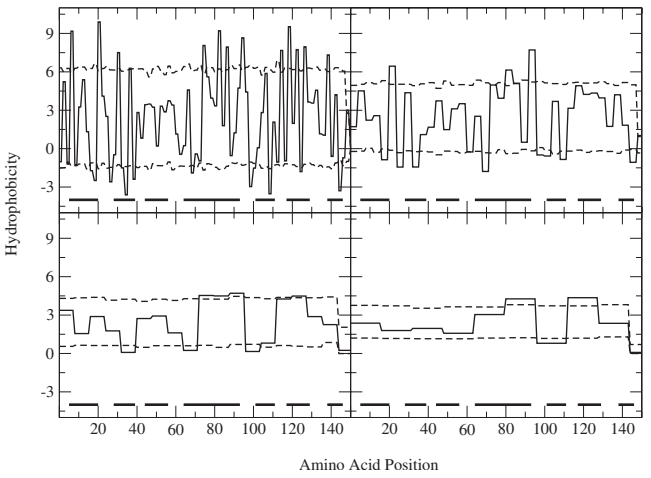

FIG. 3. The first four wavelet coarse graining using the *ES* scale for SCOP database protein, 1RJF. In each panel, the hydrophobicity content of the protein at a coarser resolution is presented as a function of amino acid position. The upper left panel is the first coarse graining, upper right the second, lower left the third, and lower right the fourth. The dotted lines are the $1-\sigma$ error bars determined from the 500 randomizations. The dark horizontal lines located at the bottom of the panels are the locations and lengths of α helices.

protein was randomly chosen among those proteins with relatively few alpha helices to avoid clutter in the figure. The actual locations of the alpha helices for this protein are also shown on the plot.

The threshold is set as the 1σ range obtained from the 500 random sequences for each protein and are shown as horizontal lines. The 1σ threshold implies that there is about a 1 in 3 chance of misidentifying an individual secondary structure (assuming a normal distribution). However, the chance of misidentifying all the secondary structures in a single protein is about $(1/3)^N$ where N is the number of structures. Thus the $1-\sigma$ threshold is a valid cutoff when considering the entirety of the protein. Each random sequence undergoes the same multiresolution process described above and is reconstructed at the same four resolutions. The coarse-grained hydrophobicity at each point along the sequence is found, and the coarse-grained value at that point is determined by averaging the 500 realizations.

Every sequence we analyzed contained some area of significance (hydrophobicity content beyond the threshold). Some alignments survived multiple reconstructions. This pointed favorably to a strong connection between hydrophobicity and secondary structure formation. Figure 3 shows multiple areas of significant hydrophobicity or hydrophilicity through all four passes of the wavelet.

To quantify the efficacy of our technique, we examine the correspondence between the positions of secondary structure (alpha helices and beta strands) as identified in the SCOP database and the results of the wavelet reconstructed sequence using binary classification. We begin by dividing the real protein sequence into regions that consist of either (a) a contiguous secondary structure or (b) unstructured residues. We define as true positive (T_p) a secondary structure region that contains at least one wavelet structure above the threshold anywhere within that region. If the secondary structure

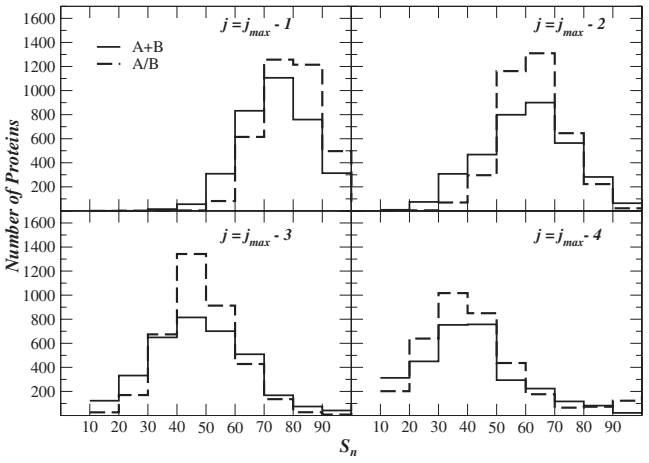

FIG. 4. The distribution of S_n for both the $A+B$ and AB data sets for the first four wavelet reconstruction using the Engelman-Steitz hydrophobicity scale. At the first resolution the average is 0.70 for the $A+B$ set, 0.74 for the A/B set.

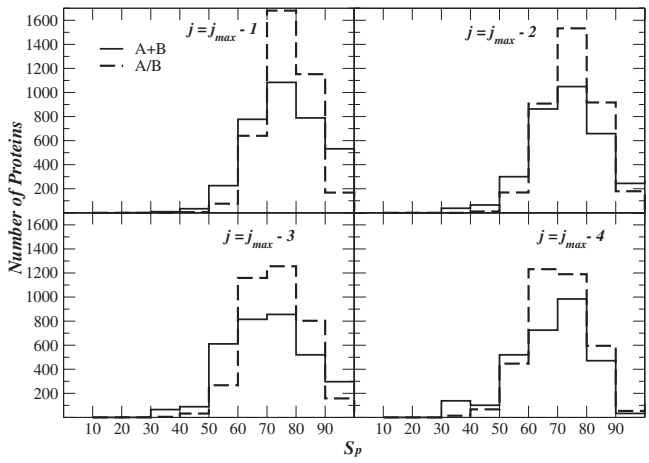

FIG. 5. Same as Fig. 4 but now for the quantity S_p. The average for the $A+B$ data subset is 0.72, and 0.77 for the AB set.

region contains no wavelet structures, it is counted as a false negative (F_n). Similarly, if a region with no secondary structures contains any wavelet structure above background, it is counted as false positive (F_p) and if a region with no secondary structures contains no wavelet structures above the threshold, it is a true negative (T_n). It is important to note that this analysis does not give the probability of an individual amino being a part of a secondary structure. Instead the reconstruction correlates regions of the sequence which the wavelet has detected as above (or below) the threshold with the existence of secondary structures somewhere within that region.

With these definitions we can define the sensitivity (S_n) and specificity (S_p) as

$$S_n = \frac{T_p}{T_p + F_n}, \tag{14}$$

$$S_p = \frac{T_p}{T_p + F_p}. \tag{15}$$

The sensitivity measures the proportion of regions that have been correctly identified as containing secondary structure, while specificity measures the proportion of predicted secondary structures that are real.

For each protein S_n and S_p were calculated for the first four resolutions. Figures 4 and 5 present the sensitivity and specificity, respectively, for both data sets. Plotted are the number of proteins versus either S_n or S_p for the ES hydrophobicity scale. The results for the other scales are tabulated in Table II and are not much different. For the $A+B$ data set, the average S_n for the data set was 0.70 at $j=j_{max}-1$, while it was 0.74 for the AB set. A comparison to other predictions will be made in Sec. IV D, but we note here that this on par with other methods. We prefer to show the entire distribution of either S_n or S_p rather than just report the average because it gives a more exhaustive measure of the effectiveness of our technique, and because, as we will discuss later, it allows for comparison between different morphological data sets.

The drop off in detection from the first pass of the wavelet, resolution $j_{max}-1$, to the second pass at resolution $j_{max}-2$ occurs because the quasilocalized structures during the second pass of the wavelet are on a scale twice as large as the first pass. The wavelength of the analyzing wavelet is too large to detect the smaller structures. Hence at the $j_{max}-2$ scale, only the larger structures are picked up. In Fig. 6, we show the distribution of the lengths of secondary structures for the SCOP A/B proteins. The distribution shows a narrow peak around four amino acids as the most common length for the secondary structures.

Figure 5 shows the specificity for both data sets for the first four reconstructions. The average for the $A+B$ data set is 0.72, while it is 0.77 for the A/B. As with the S_n results, our technique is doing a good job at recognizing regions containing secondary structure. Note however, that unlike the S_n

TABLE II. Tabulated are the sensitivity, S_n, specificity, S_p, and the correlation coefficient, M_c, for all three hydrophobicity scales and for both data sets.

		$A+B$			A/B		
Scale	j	S_n	S_p	M_c	S_n	S_p	M_c
ES	$j_{max}-1$	0.70	0.71	0.70	0.74	0.71	0.72
	$j_{max}-2$	0.55	0.69	0.61	0.57	0.70	0.63
	$j_{max}-3$	0.43	0.66	0.52	0.43	0.68	0.53
	$j_{max}-4$	0.33	0.57	0.42	0.37	0.64	0.47
HW	$j_{max}-1$	0.75	0.70	0.72	0.75	0.72	0.73
	$j_{max}-2$	0.55	0.70	0.62	0.54	0.70	0.61
	$j_{max}-3$	0.39	0.65	0.49	0.38	0.68	0.50
	$j_{max}-4$	0.33	0.59	0.43	0.27	0.63	0.40
KD	$j_{max}-1$	0.74	0.71	0.72	0.78	0.72	0.75
	$j_{max}-2$	0.54	0.70	0.61	0.57	0.70	0.63
	$j_{max}-3$	0.40	0.67	0.51	0.39	0.67	0.50
	$j_{max}-4$	0.35	0.61	0.44	0.29	0.64	0.42

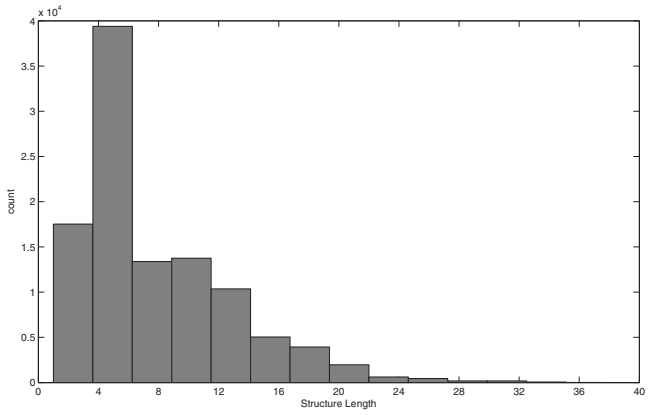

FIG. 6. Plot of the number of alpha helices and beta strands versus their length for the SCOP A/B data set. The distribution is peaked around four amino acids and drops relatively quickly after about 12 amino acids.

results, there is a smaller drop off as we decrease resolution. As we coarse grain the protein more, the number of true positives relative to false positives stays roughly the same. Thus at larger scales, fewer true positives are detected as are fewer false negatives. From Eq. (15) we see that this will cause little change is S_p.

Neither S_n or S_p alone provide a clear indicator of whether this technique is doing a good job detecting regions containing secondary structure. A better measure is to find the Matthews correlation coefficient, M_c [21]. This measure is commonly used in bioinformatics and can be shown to be approximately the geometric mean of S_n and S_p [22]. That is

$$M_c = \sqrt{S_p \cdot S_n}. \tag{16}$$

We calculated M_c for each protein in both data sets. The results are shown in Fig. 7, where we once again show the entire distribution rather than just the average. For $j=j_{max}-1$, the average for the correlation coefficient is $M_c=0.70$ for the $A+B$ set while $M_c=0.73$ for the AB set. The results for

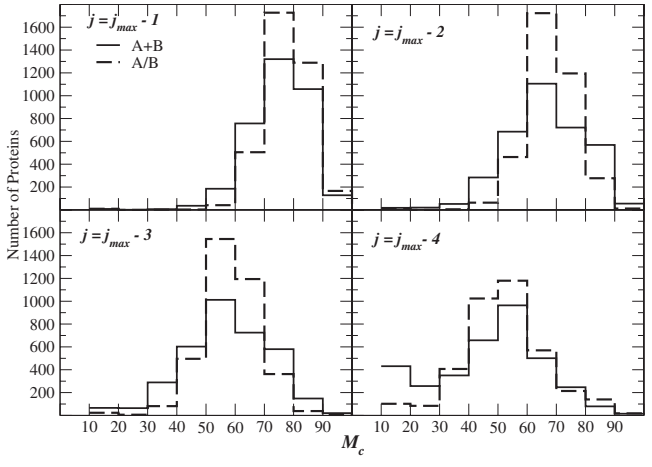

FIG. 7. The distribution of the Matthews correlation coefficient for both data sets at the first four resolutions. The average for the $A+B$ is 0.70, and it is 0.73 for the AB set at $j=j_{max}-1$.

the other resolutions and other hydrophobicity scale are listed in Table II. At least for the first two resolutions, this technique is able to recognize regions containing secondary structures with statistically valid accuracy.

As we hypothesized earlier and as indicated by Figs. 4 and 5, our analysis shows differences between the A/B and $A+B$ data subsets. To confirm the perceived difference, we performed a chi-square test comparing the match data results (i.e., Figs. 4 and 5) between the A/B and $A+B$ subsets. The test was done for each hydrophobicity scale individually and at four coarse grainings. All 12 comparisons (four resolutions, three hydrophobicity scales) showed essentially zero probability that the two distributions were drawn from the same underlying distribution. Our analysis has clearly detected the differences between these two subsets from just the amino acid sequence. We attribute the differences to the larger granularity between the alpha and beta units in the $A+B$ data than in the A/B proteins. Recall that in the $A+B$ proteins the alpha and beta regions appear in distinct areas of protein, while in the A/B proteins, these regions are in close proximity. The location of secondary structures should therefore be more evident (i.e., wavelet reconstruction above threshold) in the A/B set as these structures occupy longer contiguous regions than in the $A+B$ set.

B. Space-scale correlation

In the previous section we showed that the wavelet could pick up regions of significant hydrophobicity at various scales and that those regions corresponded to the location of secondary structures. However as pointed out in Sec. II, the wavelet allows for great flexibility in designing measures to detect structural features from a signal. We now demonstrate this by using the wavelet to detect typical length scales.

The coarse graining of the protein hydrophobicity signal accomplished by the DWT at different resolutions presents us with filtered versions of the original protein sequence data. The original amino acid sequence has been transformed into quasilocalized hydrophobic regions. We can now search for correlations between these regions to determine if there are salient or preserved length scales and to compare these areas to the lengths of alpha helices, beta strands, and beta sheets. In essence we are assuming that the formation of secondary structures are being at least partly governed by nonlocal effects. The nonlocality should be especially prevalent via spatial correlation functions once some of the high frequency signal has been smoothed. The smoothing is accomplished by the wavelet multiresolution decomposition. We now only need a spatial like correlation function to detect the nonlocality. The space-scale correlation defined in Eq. (13) gives us the appropriate tool to use.

Figure 8 shows a typical result for the protein 1*CDZ* in which we plot the correlation determined by Eq. (13) versus l_0 for four different resolutions of the wavelet. The error bars again represent 1σ for the 500 randomizations. The lengths of the actual alpha and beta structures in the protein are also indicated on the plot. A match occurs when the length of a secondary structure corresponds to the space-scale correlation being above the threshold at that same length. All the

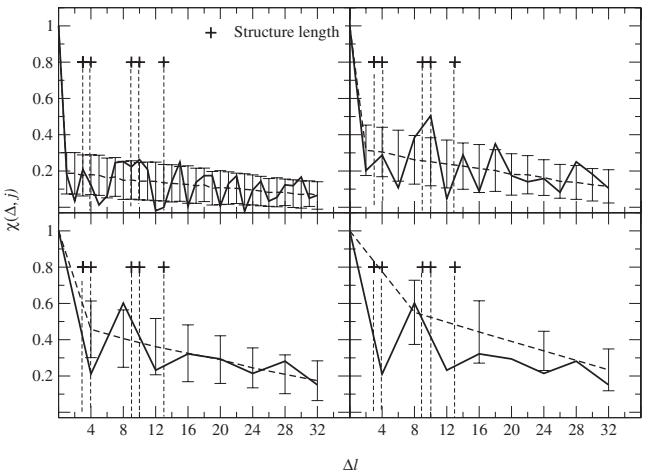

FIG. 8. Space-scale correlation at the first 4 resolutions, $j=6,5,4,3$, for the protein $1CDZ$ calculated using ES scale. The lengths of the secondary structures are shown as pluses. The solid line is the space-scale correlation. The dotted line is the average and $1-\sigma$ error bars calculated from the 500 random realizations. We drop a vertical line from the secondary structure length to the space-scale correlation. We count a match if at this location, $\chi(\Delta l, j)$ for the real protein is above the threshold. As expected, the random signal correlations quickly hit 0 (no correlation) and remain there for all lengths. The real data shows a great deal of fluctuation, with the fluctuations corresponding well to the lengths of the alpha or beta structures.

proteins revealed a match between lengths of secondary structures and lengths at which significant correlation occurred. Many of these significant correlation lengths can be explained as corresponding to the length of an alpha helix or a beta sheet, while other areas can be explained by a combination of structural elements. For example, the top left window of Fig. 8 shows a significant correlation at a length of roughly eight amino acids, and the lengths of the two alpha helices for this protein are nine and ten amino acids. A correlation, $\chi(\Delta l, j)$, above the threshold occurs at a length close to the lengths of actual structures. There remained areas of significant correlation that were not easily explained by the lengths of structural elements. However, the connection between lengths of secondary structure and the lengths of strong scale correlation of the wavelet coefficients becomes clear with statistical analysis. The matching of actual secondary structure to our space-scale correlation for all our data is shown in Figs. 9 and 10. We defined a match if the length of a structure (in this case an alpha helix or a beta strand) fell within one amino acid of a significant space-scale correlation $\chi(\Delta l, j)$ (i.e., above or below the threshold). All three hydrophobicity scales were examined to determine the number of proteins that showed strong matching of structure to hydrophobicity peaks. Figures 9 and 10 show the number of proteins versus matched structure percentage for all hydrophobicity indexes. The data show that more proteins had $\geq 90\%$ matching structures than any other percent-matched category for all three hydrophobicity scales and for both data sets. Using a $\geq 75\%$-matched criterion provides an even stronger correspondence between the space-scale correlations and the length scales of the secondary structures. In summary, for the

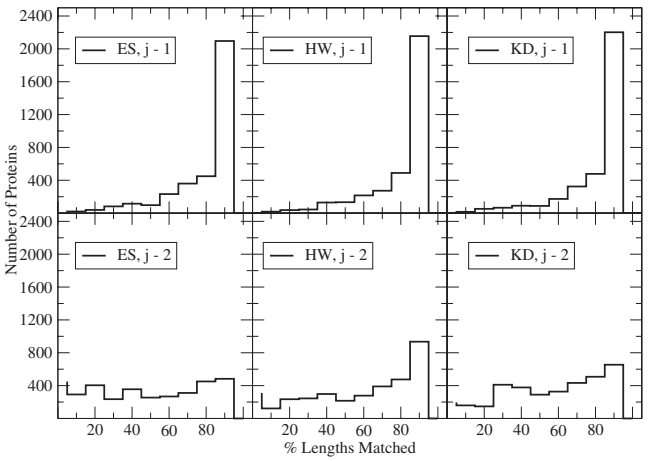

FIG. 9. The cumulative plot of the number of $A+B$ proteins versus percentage of structure lengths matched by the scale-scale correlation. The top row of panels shows the correlations after the first coarse graining, the second after the second coarse graining. For the first coarse graining, more proteins fall in the 90% range than any other. The distribution levels off after the second coarse graining as the wavelength or scale of the analyzing wavelet exceeds the lengths of many of the structures.

first wavelet reconstruction for the A/B set, the ES, HW, and KD hydrophobicity scales showed that 60%, 61.6%, and 58.9% had $\geq 90\%$ of their structures matched, respectively, and that 84.3%, 85.5%, and 84.6% had $\geq 75\%$ of their structures matched, respectively. For the A/B, the ES, HW, and KD hydrophobicity scales showed that 59.9%, 61.6%, and 62.9% had $\geq 90\%$ of their structures matched, respectively, and that 83%, 83.3%, and 85.8% had $\geq 75\%$ of their structures matched, respectively.

We note that the different hydrophobicity scales result in more uniform statistics here with the space-scale correlation analysis than with the previous position detection analysis. This is potentially explained by our ignoring the beta sheets (there only beta strands are counted) in the positional analysis of the previous section. That we see such consistent results here is encouraging and supports that the particular choice of hydrophobicity scale is not critical.

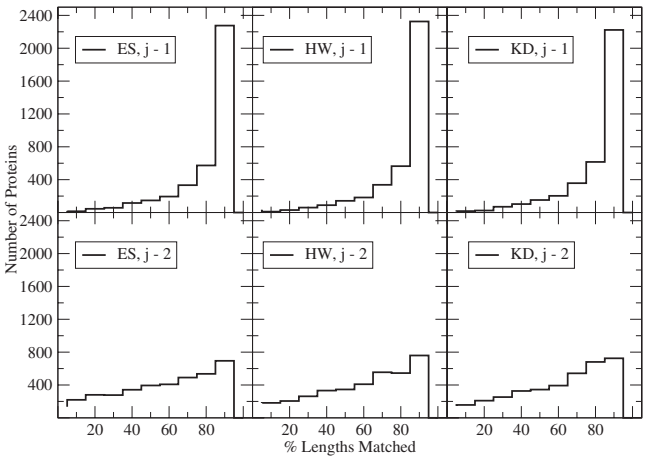

FIG. 10. Same as Fig. 9 for A/B proteins.

C. Comparison of hydrophobicity scales

As mentioned earlier, the calculations in this project were performed on three data sets, each set employing a different hydrophobicity scale. The results produced by these calculations were very consistent across all three sets. The wavelet decomposition revealed similar areas of significance for the three data sets despite initial signal variations caused by the three different hydrophobicity scales. The different measures used to define hydrophobicity in each of the three scales examined here do not critically affect the correlations detected by wavelet analysis.

D. Comparison to other methods

Direct comparison to other methods of secondary structure prediction is very difficult because our method relies on the wavelet's coarse graining property. In identifying the position of alpha helices, individual amino acids are replaced by quasilocalized structures corresponding approximately to a hydrophobicity average near the original amino acid. In the second part of our analysis, when we look for correlations in wavelet space, the length of any secondary structure element is what is detected.

The best predictors of secondary structure are around 80% according to the Evaluation of Automatic protein structure prediction (EVA) website [23]. Programs such as PHD [24], PSIPRED [25], or PROTEUS [26] use neural network methods and searches for homologous sequences with known structure to give a prediction of the secondary structure of each amino acid in a sequence (often along with a measure of the confidence level of the prediction). The wavelet-based technique described here is not meant to yield a specific prediction for each residue and thus cannot be compared directly to these techniques. As presented here, we have only examined the correlations between locations of significant hydrophobicity as discovered by wavelet analysis and the locations of real secondary structures in proteins of known structure. Applications and consequences of our approach are discussed in the following section.

V. DISCUSSION

Our goal was to find some meaningful pattern of information located in the primary structures of the proteins that we analyzed. Our results are encouraging, but must be treated as a preliminary effort. The approach of utilizing the discrete wavelet transform to unlock position and scale information by examining the hydrophobicity "signal" of these proteins is a relatively new idea and shows promise as technique that can reveal otherwise hidden structural information.

Our calculations suggest relationships involving both the locations of secondary structures and the length scales at which these structures emerge. We believe that this DWT approach sheds light on the scale dependence of secondary structure formation.

Specifically, we find that there exist correlations relating areas of strong hydrophobicity to physical structures within a protein. The number of secondary structures roughly matches the number of significant hydrophobicity peaks at certain scales. Most of the secondary structures in the proteins we examined had lengths that matched the length scales at which strong hydrophobicity correlations existed along the amino acid chain. These results prove to be essentially independent of the specific hydrophobicity scale used. Despite differences in the methods used to create these scales, our results suggest that the three scales are consistent with one another.

Significance levels were measured against randomized data that had the same number and kind of amino acids. This demonstrates that positional information in the amino acid sequence is critical in determining secondary structure formation, and wavelet analysis is able to distinguish different orderings at different length scales. Using a 1σ deviation from the randomized data provides a robust measure of significant measure of structure within the amino acid sequence that makes up the protein.

We suggest that are several possible avenues for the wavelet analysis presented here to find application in the future. First, the results of the positional analysis could be used to augment other techniques to improve their accuracy of secondary structure prediction even further. For instance, PROTEUS uses a combination of three different secondary structure predictors in concert with a neural network classifier and homologous pattern search [27]. Results of the wavelet technique could also be combined with this array of technique to improve the net accuracy even further. Second, the results of both the positional analysis and the space-scale analysis at different scales could be used to create a fingerprint or signature of a given protein. Such wavelet fingerprints could then be used to identify different families of proteins and to classify which family a given protein might be in. Third, the present technique could also be potentially useful in tertiary structure prediction. For instance, the wavelet may provide information about regions of a protein that are likely to undergo large folds and bends. Since the folding processes brings together residues that are far away in sequence space, we can consider that meaningful correlations will arise between the results of multiscale analysis of the one-dimensional (1D) sequence of a protein and measures of three-dimensional (3D) structure of the protein in its native state.

Lastly, this approach is very fast. We were able to produce the wavelet reconstructions and space-scale correlations for both the original proteins and their 500 randomizations for the approximately 3500 proteins in each data set in about an hour per set using a 2.0 GHz Power Mac G5.

Despite these positive results, some of the data remain a mystery. Many of the significant hydrophobic regions did not correspond to any secondary structure. It was not clear whether combinations of different secondary structures could account for these data. Further analysis on larger protein database sets is underway in order to better understand the potential of this technique.

[1] E. Shakhnovich, Chem. Rev. (Washington, D.C.) **106**, 1559 (2006).
[2] P. Bradley, K. M. S. Misura, and D. Baker, Science **309**, 1868 (2005).
[3] *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, edited by A. Baxevanis and B. F. Ouellette (Wiley, Hoboken, 2005).
[4] S. Tiwari, S. Ramachandran, A. Bhattacharva, S. Bhattacharva, and R. Ramaswamy, Comput. Appl. Biosci. **3**, 263 (1997).
[5] A. Irback, C. Peterson, and F. Potthast, Proc. Natl. Acad. Sci. U.S.A. **93**, 9533 (1996).
[6] Z. G. Yu, V. V. Anh, K. S. Lau, and L. Q. Zhou, Phys. Rev. E **73**, 031920 (2006).
[7] O. Weiss and H. Herzel, J. Theor. Biol. **190**, 341 (1998).
[8] Z. Wen, K. Wang, M. Li, F. Nie, and Y. Yang, Comput. Biol. Chem. **29**, 220 (2005).
[9] L. Pattini and S. Cerutti, Methods Inf. Med. **43**, 102 (2004).
[10] A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P. Graves, J. Muzy, and C. Thermes, Physica D **96**, 291 (1996).
[11] L. Fang and R. Thews, *Wavelets in Physics* (World Scientific, Singapore, 1998).
[12] I. Daubechies, *Ten Lectures on Wavelets* (SIAM, Philadelphia, 1992).
[13] J. Pando and L. Z. Fang, Phys. Rev. E **57**, 3593 (1998).
[14] J. Pando, L. Fang, P. Lipa, and M. Greiner, Astrophys. J. **496**, 9 (1998).
[15] L. Feng and L. Fang, ApJ **601**, 54 (2004).
[16] Y. G. Y. Chu and L. Fang, ApJ **610**, 51 (2004).
[17] T. P. Hopp and K. R. Woods, Mol. Immunol. **20**, 483 (1983).
[18] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[19] D. M. Engelman, T. A. Steitz, and A. Goldman, Annu. Rev. Biophys. Biophys. Chem. **15**, 321 (1986).
[20] J. Cornette, K. Cease, H. Margalit, J. Spouge, J. Berzofsky, and C. DeLisi, J. Mol. Biol. **195**, 659 (1987).
[21] B. Matthews, Biochim. Biophys. Acta **405**, 442 (1975).
[22] J. Gorodkin, S. L. Stricklin, and G. D. Stormo, Nucleic Acids Res. **29**, 2135 (2001).
[23] V. Eyrich, D. Przybylski, I. Koh, and B. Ros, http://cubic.bioc.columbia.edu/eva/doc/intro-sec.html (2007).
[24] B. Rost, G. Yachdav, and J. Liu, Nucleic Acids Res. **32**, W321 (2004).
[25] K. Bryson, L. McGuffin, R. L. Marsden, J. Ward, J. Sodhi, and D. Jones, Nucleic Acids Res. **33**, W36 (2005).
[26] S. Montgomerie, D. Wishart, S. Sundaraj, and W. Gallin, http://wks16338.biology.ualberta.ca/proteus/ (2008).
[27] S. Montgomerie, S. Sundaraj, W. Gallin, and D. Wishart, BMC Bioinformatics **7**, 301 (2006).