Journal of Statistical Physics, Vol. 93, Nos. 3/4, 1998

Transformational Homologies in Amino Acid Sequences Suggest Memberships in Protein Families

Arnold J. Mandell,[1] Karen A. Selz,[1] and Michael F. Shlesinger[2]

Received February 3, 1998; final May 26, 1998

The 20 amino acid monomers composing polymeric proteins are encoded using their individual properties relatable to thermodynamic potentials such as aqueous partial specific volumes, aqueous molar volumes, and free energies of transfer from hydrocarbon to water solvents. These principally hydrophobic solvation, "hydrophobicity"-derived free energies are minimized in protein folding as well as protein–protein and peptide–receptor interactions. Sequential patterns in the one-dimensional distribution of these energies, reflected in dominant wavelengths of amino acid hydrophobicity, and the locations of singular hierarchical, secondary and supersecondary structures are elucidated by orthogonal decomposition and eigenfunction construction followed by continuous wavelet and all poles, maximum entropy power spectral transformations. The resulting graphs discriminate among examples of structural families of proteins.

KEY WORDS: protein physics; wavelets; hydrophbic free energy eigenfunctions; maximum entropy power spectra.

1. INTRODUCTION

We are honored to contribute to a volume dedicated to the recognition of Leo Kadanoff's broad range of important contributions to science in general and mathematical and theoretical physics in particular.(1)

We address an open problem in biophysics concerning what can be conjectured about a three dimensional protein's family membership from its one dimensional amino acid sequence. Contributions in this area are of increasing value in the present era of survey research such as the Human Genome Project and its resulting welter of cDNA expressible proteins with

[1] The Cielo Institute, Emory University, Asheville, North Carolina 28804, and Department of Mathematics, Florida Atlantic University, Boca Raton, Florida 33431.

[2] Physical Sciences Division, Office of Naval Research, Arlington, Virginia 22217.

0022-4715/98/1100-0673$15.00/0

known amino acid sequences but unknown structure and behavior. These unidentified proteins do not arise, as in the past, from within a particular program of biochemical research, but rather emerge without hints as to their categorical anatomical or functional family memberships. The years often required for the crystallization of even one of these unknown proteins for structural characterization make molecular anatomic techniques impractical as routine procedures.

One approach to the current flood of unidentified proteins has involved protein sequence data base development and searches by scientists in a relatively new field called protein informatics. These procedures are encouraged by the belief that "...the first and best clue to (protein) function..." is the specific amino acid sequence homology between these new proteins and those of known function.[2] For example, evolutionary distances between proteins are computed from the Hamming distances between their amino acid sequences. However, even in common, related proteins, sequence homology is quite low. For example, myoglobin, hemoglobin and the light-gathering phycocyanin of algae have as little as 16% amino acid sequence homology in pairwise comparisons, while members of this globin family manifest identical tertiary fold structure and chemical function.[3] More complicated yet nearly identical families of tertiary structures such as α/β barrels and functions such as the serine proteinases have no specific amino acid sequence homology[4] and represent a growing number of examples of divergent amino acid sequences in proteins associated with convergent structures and functions.

As the number and kinds of uncharacterized proteins grows, it is now being acknowledged by protein chemists that searches for homologies will require new systems representing "...a twilight zone of more and more distant similarity."[5] Our program exploits physically meaningful numerical representations of the twenty different amino acids in protein sequences, referable to individual amino acid thermodynamic potentials, which tend toward minimization in protein folding, protein-protein interactions and peptide ligand-receptor binding. Familial patterns in these real number sequences are sought by orthogonal decomposition followed by continuous wavelet and maximum entropy power spectral transformations.[6–8]

These procedures have yielded characteristic patterns consisting of signatory amino acid wavelengths and sequential distributions of localized, hierarchical (also called secondary and supersecondary) structures that successfully predict peptide drug-receptor interactions, suggest a statistical mechanical mechanism[8–10] and indicate memberships in categorical families of protein structure.[6, 7] In spite of many instances of nonsequential connectivities between singular secondary and supersecondary structures in protein domain formation composing tertiary structures, this system of

transformational analyses of one dimensional sequences has proven of value in their preliminary classification. The discrimination between representatives of *globular* and *polyprotein* families are examined here as examples.

2. REPRESENTATION OF AMINO ACIDS BY REAL NUMBERS

Proteins are polypeptide polymers consisting of unbranched chains of amino acid monomers strung together by identical covalent peptide bonds. Each of the genetic code specified 20 amino acids that compose these polymers has a unique side chain. It is these side chains that give to each monomer its unique physical and chemical properties and it is the cooperative influence of these properties that determine the tertiary structure and function of polymeric proteins. Descriptions of these properties often group the monomers into an "apples versus oranges verus bananas," qualitative categorical system. Using their diverse chemical compositions and properties, one can discriminate among amino acids side chains containing aromatic rings, hydroxy groups, acidic groups, basic groups, nucleophiles, electrophiles, imino rings and simple aliphatic chains.

Another approach to classification involves quantifiable properties that can order the entire group of twenty amino acids along a particular physicochemical dimension of varying continuity. Because proteins achieve their physiologically relevant structure and function via noncovalent bonding mechanisms in water, the hydrodynamic properties common to all amino acids side chains such as diffusivity, frictional coefficients, aqueous cavity surface area, aqueous molar volume and partial specific volume, *PSV*, have been of particular interest.[11] Intuitively, the latter at physiological temperature is related (inversely) at constant volume and pressure to density, an intensive equilibrium thermodynamic property and it correlates with other measures relatable to chemical potentials (see below).

With respect to *PSV*, we consider a binary solution of molecules of water, N_1, and molecules of a particular amino acid, N_2, having a total volume computed (at fixed temperature and pressure) from the partials of their molecular volumes, $V = N_1\{(\partial V/\partial N_1)(\equiv v_1)\} + N_2\{(\partial V/\partial N_2)(\equiv v_2)\}$. Dividing both sides by V and letting n stand for their respective concentrations, $1 = n_1 v_1 + n_2 v_2$, where the *rhs* is the sum of fractional volumes of the components. Letting ρ_1 and ρ_2 represent their respective densities and ρ the density of the binary solution, $\rho = n_1 v_1 \rho_1 + n_2 v_2 \rho_2$, where the *rhs* is the sum of the fractional densities. This can be rearranged, $\rho = \rho_1 + (\rho_2 - \rho_1) n_2 v_2$ and letting $w_2 = (n_2 v_2 \rho_2)/\rho$ the weight fraction of the amino acid and letting $c_2 = w_2 \rho$, the concentration of amino acid in grams/ml, we see the relationship between partial specific volume and the density, $\rho = \rho_1 + (1 - (\rho_1/\rho_2))\, c_2$.

The range of observed density increments, $(\rho - \rho_1)/c_2$, with the addition of particular amino acids to water reflects the differences in their *PSV* among amino acids. Amino acids restructure their surrounding water (generating various amounts of aqueous "hydrogen bond cage strain") and this is reflected in both their v_i and ρ_i. For concreteness, a set of values of experimentally determined *PSV*'s in gm/ml for the twenty essential (required in the diet) amino acids, but here range normalized and therefore dimensionless, are:[12] $A = 0.5541$, $R = 0.2852$, $N = 0.1311$, $D = 0.0000$, $C = 0.1705$, $Q = 0.3115$, $E = 0.2098$, $G = 0.1738$, $H = 0.2983$, $I = 1.000$, $L = 1.000$, $K = 0.6885$, $M = 0.5443$, $F = 0.6393$, $P = 0.5869$, $S = 0.1115$, $T = 0.3607$, $W = 0.5081$, $Y = 0.4360$, $V = 0.8787$. *PSV* is an example of a thermodynamically relevant, quantifiable property of all amino acids which can be ordered along a single dimension.

Water molecules have a much stronger attraction for each other than they do for hydrocarbons and this property has been called the "hydrophobic effect."[13,14] A nonpolar, uncharged amino acid side chain, unable to hydrogen bond with water, results in volume expanding hydration shells composed of clathrate structures, hydrogen bond cages of many molecular layers of water in a process also called hydrophobic hydration[16,17,15] which is particularly prominent in the *PSV*'s of amino acids with purely hydrophobic side chains, *L*, *I* and *V* above. These cages of structured water are of high heat capacity and low entropy, and both properties are sensitive to changes in temperature.[18] Water solvation of hydrophobic residues is associated with calorimetrically measurable increases in heat capacity at constant pressure, $\Delta C_p > 0$, positive enthalpy, $\Delta H > 0$, and negative entropy, $\Delta S < 0$, which compete as Second Law governed countervailing influences on the Gibbs free energy, ΔG.[19] It is this energy which is minimized when hydrophobic groups "pack" together during protein folding as well as protein–protein and peptide–receptor interactions.

Polar, uncharged side chains may hydrogen bond with just one molecule of water, but otherwise undergo the same kind of hydrophobic hydration as the non-polar side chains. With respect to side chains with charged residues, the high dielectric constant of water minimizes the energy of their electrostatic fields, such that even in these side chains in water, the hydrophobic effect is prominent. These polar groups have small hydration heat capacities and very large but ill-defined hydration enthalpies and entropies. These findings have led to the conclusion that hydrophobic hydration is by far the largest contributor to heat capacity changes with amino acid solvation.[20]

Another and closely related set of quantities that differ among amino acid side chains in water is their hydrophobic free energies, *H* in kcal/mol,

which correlates with PSV ($R^2 = 0.671$, $P = 0.0184$) as well as with molar volume in aqueous solution ($R^2 = 0.723$, $P = 0.00475$) (both computed from published tables of values[11]). H of an amino acid is derived from its "free energy of transfer" from a hydrocarbon solvent to water computed as a chemical potential, μ, from its relative equilibrium concentrations, X_i, in the two media.[21] In addition, computable proportionality coefficients relate μ and solvent accessible surface area, SAS, (reduced when the hydrophobic groups of amino acid side chains aggregate in aqueous protein folding or peptide ligand-protein interactions) to values for ΔC_p, ΔH and ΔG.[22] Internal polymeric arrangements leading to the burial of hydrophobic surfaces, "hydrophobic packing," reduce SAS accompanied by $\Delta C_p < 0$ and "enthalpy-entropy compensation" such that ΔG tends to remain small.[19, 23]

If X_{HC} and X_W are the equilibrium distributions of an amino acid between the hydrocarbon and water solvents in mole fraction units and μ_{HC} and μ_W are the corresponding standard chemical potentials, the amino acid's *hydrophobic free energy* $\equiv$ *hydrophobicity* $\equiv H = \mu_W - \mu_{HC} = -RT \ln(X_{HC}/X_W)$ in kcal/mol, with the reference value for glycine, G, set equal to 0.[24] The set of experimental results at physiological temperature and pressure used in these studies yield the following set of amino acid hydrophobic free energy values, H, in kcal/mol: G, $Q = 0.0$; S, $T = 0.07$; $N = 0.09$; $D = 0.66$; $E = 0.67$; $R = 0.85$; A, $H = 0.87$; $C = 1.52$; $K = 1.65$; $M = 1.67$; $V = 1.87$; $L = 2.17$; $Y = 2.76$; $P = 2.77$; $F = 2.87$; $I = 3.15$; $W = 3.77$.[24, 25] Different experimental conditions, solvents and computational schemes have led to other sets of estimates of H_i and their corrections.[26–30]

We are interested in the sequential patterns of these hydrophobic free energies because their minimization is generally considered to be the major influence in the polymeric self organization, "folding" into characteristic tertiary structure via hydrophobic contacts, hydrophobic packing and reduction of SAS of the hydrophobic side chains of peptides in water.[31] Whereas the total amount of energy involved in protein hydrophobic aqueous solvation is large, the conjugate variations in ΔH and ΔS in normal physiological processes, enthalpy-entropy compensation, make the net variations in free energy, ΔG, relatively small. In a classic computation using x-ray crystallographic data of an n long protein and the above Tanford values for the $H_{i,\, i=1 \cdots n}$, Chothia[32] estimated that the hydrophobic contribution to the free energy in the unfolding of a representative protein (lysozyme) is large, approximating 340 kcal/mol. For comparison, a change in conformation in a representative two state protein such as hemoglobin, involving a transition between folded states, has ΔG's in the vicinity of 3–5 kcal/mol.

3. SEQUENTIAL PATTERNS IN AMINO ACID HYDROPHOBICITY

Although intuition would suggest that one dimensional patterns in sequential values of H_i in proteins would have little to contribute to guesses about the three dimensional folded geometry of hydrophobic free energy minimizing contacts between their hydrophobic side chains, the results of our empirical studies have indicated otherwise.[6,7] In addition, recent studies exploiting nine orders of magnitude in folding time in a set of non-homologous single domain proteins found that the chain length normalized, average sequence separation between contacting residues in the folded state, varied directly with folding time.[33] "Local guidance by contacting groups" appears important for the protein to find its native state.[34] As an example, two of a protein's polypeptide segments with almost identical rotation numbers in H (the average number of sequential amino acids required to go from low to high to low again in side chain hydrophobicity values) have been shown to selectively aggregate resulting in "supersecondary" structures in protein domains called "hydrophobic zippers."[35,36]

Guidance by hydrophobic free energy minimizing arrangements among sequential patterns of hydrophobic groups can also be viewed in the context of a long range (>20 angstrom) hydrogen bond cage strain minimizing "attractive force" between hydrophobic surfaces,[37] and has been measured directly (in *newtons*) using atomic force microscopy.[38] It is for these and other reasons that the cooperative three dimensional spatial arrangements assumed by one dimensional peptide sequences in aqueous solution have been generally understood to be dominated by the influence of sequential locations and hydrophobic properties of amino acid side chains.[39,40]

We treat the amino acid hydrophobic free energy sequence of an n-amino acid long protein as a function, $H(i)$. Previously, we have found that expanding $H(i)$ in a Fourier series required many terms to converge to a close representation of the data, and the relative shortness of the polypeptide chains introduced numerical and end effects.[41,42] It is for these reasons that in our current work we have chosen two systems of transformation which are without the sensitivity to finite length, mode multiplicity and other constraints of Fourier or related transformations and/or exploit "best" basis functions which arise from the amino acid sequence data itself. They are: (1) The continuous Morlet wavelet transformations, $W_{a,b}(H_n)$, of protein hydrophobicity series, allowing for the multiresolution study of serial hydrophobic amplitudes with respect to a range of characteristic sizes (the scale referred to as dilations, dilates, inverse

sequential hydrophobic frequencies or wave numbers) and their sequence locations, with a Heisenberg-like, reciprocal trade-off in the relative precision of sequence location versus dilation measurements,[43] using a family of derivatives of the Gaussian as "mother wavelets."[44] (2) The linear decomposition of order M lagged, autocovariance matrices, C_M, of the amino acid hydrophobic free energy series, H_n. From the set of ordered eigenvalues, $\{\nu_j\}_{j=1}^{M}$, of these C_M, the corresponding set of eigenvectors $X_{j=1\cdots M}$ are computed and then serially convolved with H_n to form a set of hydrophobic free energy eigenfunctions, $\psi_{j=1\cdots M}$.[45, 46] The dominant mode(s), hydrophobic free energy in not necessarily integer units of amino acid wavelengths, inverse sequential frequencies, ω_i^{-1}, of each ψ_j is then determined by their "all poles," maximum entropy power spectral transformation, $h(\omega)$.[47]

We compare the transformations of two families of proteins manifesting two kinds of physically established global hydrophobic free energy mode structures. The first are folded *globular proteins* (*globular* means compact, densely folded, quasi-spherical with minimal surface yet soluble in aqueous medium) with the symmetry of organization across size: primary sequences, secondary modules composed of the primary sequences and supersecondary arrangements of secondary modules constituting tertiary structure. This hierarchical protein structural pattern was first described over a half-century ago by Kai Linderstrøm-Lang.[48] As will be evident, some of the continuous wavelet graphs of globular proteins look very much like the continuous wavelet transformations of multiscale correlated Brownian motion[49] and/or hierarchical irrational rotations on the torus. The second group of proteins examined here are called *polyproteins* because after synthesis they tend to remain unfolded and are catalytically separated by proteolytic enzymes into several smaller proteins.

Globular proteins sharing similar mode structures throughout their length, aggregate into localized repeating cooperative structures across several scales of correlation which can be called "folding." *Polyproteins* appear to delay folding and, therefore, remain more extended and accessible to enzymatically catalyzed separation. Contributing to the vulnerability of *polyproteins* to proteolytic segmentation is their more varied composition and more irregular distribution of hydrophobic modes. In addition, the relations between the rotation numbers of the *polyprotein's* modes are more incommensurate than those of *globular proteins*. These factors would appear to make polyprotein segments less mode matched and less "hydrophobic zippering" prone. The *globular proteins* examined are human *hemoglobin* A, and human *prealbumin* and the *polyproteins* are human *pro-opiomelanocortin* and the human AIDS-related core protein, *GAG1*.

4. BRIEF DESCRIPTIONS OF THE COMPUTATIONS

4.1. Continuous Morlet Wavelet Transformations

Continuous Morlet wavelet analysis of the proteins' $H_{i,\,i=1\ldots n}$ series, using a wavelet transform, $W(a, b)$, generally, consists of decomposing it into translated $W(n) \rightarrow W(n-b)$ and scaled $W(n) \rightarrow W(n/a)$ ("scale" is in terms of the radian sequential frequency or radian wavenumber of a trigonometric function) versions of the real valued mother wavelet, w. w is a waveform with an average value of zero ($\int_{-\infty}^{\infty} w(n)\,dn = 0$), of finite duration, arbitrary regularity and symmetry, and which is composed with hydrophobic value data series, $H_1, H_2, \ldots, H_n$, as $W(a, b) = (1/\sqrt{a}) \sum_0^n H(n)\, w((n-b)/a)\, dn$. For w we chose a member of the family of continuous, symmetric, infinitely regular and differentiable, modulated Gaussian Morlet wavelets:[44] $w(x) = (1/\sqrt{2\pi}) \exp(-x^2/2) \exp(2\pi i f x)$. Note that the Morlet wavelet when used in the context of real data analysis is called the pseudoMorlet wavelet because an admissibility condition (having a zero mean) is violated, $\int_{-\infty}^{\infty} w(x)\,dn = \exp((-2\pi^2 f^2) \neq 0$, but for large f the integral can be made arbitrarily small.

In our graphs of the moduli of the continuous Morlet wavelet transformations, the x-axis indexes the location along the one dimensional amino acid sequence. The y-axis indicates the relative dilation of $w(x)$ (composed with H_n) in dilate divisions, dd, with the shortest mother wavelet being $\approx 1/0.5 = 2$ amino acids at the bottom and $1/0 \rightarrow \infty$ amino acids at the top. For example, location in a dilate space with 64 dilate divisions, $dd_{1\ldots 64}$, can be approximated as mother wavelet wavelengths, ϖ, in amino acids, aa, as $\varpi \approx 1/[0.5 - (dd * 0.5/64)]$. The modular amplitudes of the wavelet transformations are grey scale shaded with relative maxima being more white and relative minima being more black. These absolute amplitudes in each of the 64 dilate ranges were normalized to 100% ("coloration by scale"). This choice of "by scale" versus "cross scale" color coding of modular amplitudes does not portray the relative dominance of structures across all dilate ranges (which results in the loss of wavelet structural detail), but rather outlines the relative amplitudes of modular patterns and their locations at each dilate range.[50]

Assuming the protein structural organization first suggested by Linderstrøm-Lang[48] and assuming 64 dilate divisions in the wavelet graph, we may ask for some or all of the following kinds of information from the Morlet wavelet transformations: (1) At relatively small scales, we want to know the sequence locations and fundamental inverse sequential hydrophobic frequencies or wave numbers of the protein's characteristic secondary structures. For examples, α-helices contain from 3.2 to 3.7 amino

acids per hydrophobic free energy rotation ($\approx$24 to 30 *dd*), while β-strands have rotation numbers which may range from 2.2 to 2.6 amino acids ($\approx$5 to 15 *dd*). (2) At intermediate scales we want to assess the characteristic sequence sizes and locations of singular, hierarchical, secondary structures. For examples, although there is considerable variability, individual helices in helical bundles generally average in the range of 7 to 15 residues in length ($\approx$48 to 55 *dd*) and β-strands in sheets or barrels may range from 4 to 8 residues ($\approx$32 to 45 *dd*). (3) At the next largest scale, we sometimes exploit the multiresolution capacity of $W(a, b)$ to locate another kind of sequence singularity[51] characteristic of the multiscale, hydrophobicity sequence content of the longer and shorter loops (called "random coils") which serve as transitions between more dilate localized secondary modules of helices or sheets. "Random coils" (which are widely acknowledged to be neither random nor coils) range, generally from 2 to 16 residues (2 to 56 *dd*) or longer. (4) The modular maxima at the largest scales (60 to $\geqslant$62 *dd*) are relatively long hierarchical, hydrophobic domains of 40 to 50 residues or larger.

The global hydrophobic domains can also be observed as the sequence lengths between relative maxima or minima when the $H_{i,\, i=1\cdots n}$ is smoothed by iterative, nearest neighbor averaging, a standard technique in protein biology which results in what is called a "hydropathy plot."[26] We have found that these descriptions, taken together and to a first approximation, appear to be useful in discriminating among structural families of proteins.[6, 7]

4.2. Construction and Mode Indexing of the Proteins' Leading Hydrophobic Free Energy Eigenfunctions

The transformed sequence of n length of the protein, $H_1, H_2, \ldots, H_n$, is decomposed into orthogonal functions. From the M lagged vectors (M chosen such that the first eigenfunction, ψ_1, see below, minimizes least squares errors when fit to the protein's hydropathy plot, a polynomial like pattern as the 13 $\times$ iteratively applied three point moving average of H_n[8]) $V_1 = (H_1, H_2, \ldots, H_M)$, $V_2 = (H_2, H_3, \ldots, H_{M+1}), \ldots, V_{n-M+1} = (H_{n-M+1}, H_{N-M+2}, \ldots, H_n)$, with $M = 14$ to 16, an $M \times M$ covariance matrix, C_M, is formed in which $K = n - M + 1$:

$$C_M = \frac{1}{K}\begin{pmatrix} \sum_{j=1}^{K} H_j H_j & \sum_{j=1}^{K} H_j H_{j+1} & \cdots & \sum_{j=1}^{K} H_j H_{j+M-1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{j=1}^{K} H_{j+M-1} H_j & \sum_{j=1}^{K} H_{j+M-1} H_{j+1} & \cdots & \sum_{i=1}^{K} H_{j+M-1} H_{j+M-1} \end{pmatrix}$$

We compute the eigenvalues, $\{v_j\}_{j=1}^{M}$, and the associated eigenvectors, $X_j(l)$, of C_M, where j goes from 1 to M and labels the eigenvector and l also ranges from 1 to M and refers to the lth component of the eigenvector $X_j(l)$. $\{v_j\}_{j=1}^{M}$, are ordered from largest to smallest as are the $X_j \geqslant 0$. The leading X_j are then used as multiplicative "weights" to transform the $H_1, H_2, \ldots, H_n$ into M eigenfunctions, $\psi_j(l)$, where $j = 1 \cdots M$ labels the eigenvector and $l = 1 \cdots n - M$ indexes its lth component. The $\psi(l)$, for $l - k + 1 > 0$, are given by $\psi_j(l) = \sum_{k=1}^{M} X_j(k)\, H_{l-k+1}$, where H_l is the first numerically transformed amino acid in the sequence. In words, the convolution of each of the leading eigenvectors with the hydropathy series is carried out by computing the sums of the scalar products of the M length eigenvector with an M length of the hydrophobic series to produce a point in the eigenfunction; this process is translated down the data series by one step and repeated to generate each of the sequential points of the eigenfunction that corresponds to its ordered eigenvalue associated eigenvector in the computation

In effect, C_M results from a scan for hydrophobic free energy modes across a range of autocovariance/correlation lengths from 1 to M. Because C_M is by definition real, symmetric and normal, its $\{v_j\}_{j=1}^{M}$ are real, nonnegative and distinct, and its associated X_j constitute natural bases for orthonormal projections on $H_1, \ldots, H_n$. The set of ψ_j constitute the orthonormally decomposed sequences of moving average values, each of which has been weighted with respect to the hydrophobic variation attributable to the jth of the $\{v_j\}_{j=1}^{M}$ ordered X_j.

The ψ_j of each protein's sequence of H, were transformed into their dominant inverse sequential amino acid hydrophobic frequencies, ω^{-1}, using an "all poles," maximum entropy power spectra, $h(\omega) = (\sigma^2/2\pi)(1/|1 + a_1 \exp(-i\omega + \cdots + a_k \exp(-ik\omega)|^2)$, which will peak (have poles) at the zeros of the denominator. The Fourier coefficients, a_i, are calculated from a small set of k known autocorrelations, chosen so that the entropy of the spectral estimate, $\mathscr{H} = \int \ln h(\omega)\, d\omega$, is maximal. This happens because beyond the limited information of a small set of autocorrelation-matched Fourier coefficients, the "correlation matching property,"[47] the process is extended into a Gaussian process such that $\mathscr{H}$ is maximized.[52] It is known that the Gaussian function maximizes the entropy under the constraints of a finite number of second order correlations. In these studies $k \leqslant 8$ for protein eigenfunctions of sequence lengths of several hundred amino acids to avoid "splitting" $h(\omega)$ into spurious modes. $h(\omega)$ is much like an autoregressive, maximum likelihood spectral estimate in that it is not model-dependent, but is derived directly from the data of ψ_j and behaves like a filter that yields its one or two leading complex poles of discrete hydrophobic variational frequency in the hydrophobic free energy eigenfunction.

5. RANDOM VERSUS DETERMINISTIC SIMULATIONS OF PROTEIN WAVELET HIERARCHIES

5.1. A Fractional Brownian Series

Lila Gatlin[53] applied the information theoretic ideas she developed in content analyses of adult's versus children's texts to a Neo-Darwinian theory of peptide sequence evolution which involved random point-wise nucleotide mutations with selective trends based on advantageous macrostructure. She hypothesized that the conserved macrostructure would be seen in the one dimensional code as increased serial dependencies and therefore reflected in increased autocorrelations. From this point of view, the Linderstrøm-Lang hierarchical protein can in some ways be approximated by a correlated random process.

Generally, a fractional Brownian series is a Gaussian, zero-mean, nonstationary stochastic process, $B_h(n)$, indexed by a scalar parameter, h, such that $0 < h < 1$ and $h = 0.5$ is a series composed of uncorrelated Brownian increments. $B_h(n)$'s covariance structure indicates series location dependence, and can be written $B_h(n)\, B_h(n+i) = (\sigma^2/2)(|n|^{2h} + |i|^{2h} - |n-i|^{2h})$, making its (nonstationary) variance $\mathrm{var}(B_h(n)) = \sigma^2 |n|^{2h}$, with a power law *average spectrum*, $\mathscr{S}_{B_h}(\omega) = \sigma^2/|\omega|^{2h+1}$, characteristic of hierarchical ("$1/f$ like") processes.[54] Though globally nonstationary, this form of fractional Brownian series has stationary increments, such that the probability properties are dependent only on the lag variable $|n-i|$, making the incremental process "self-similar." These two properties, nonstationary sequence location dependence and self-similar scale dependence, have been implicated in fractal theories of protein structure,[55, 56] and make this kind of sequential structure well suited to wavelet analysis with simultaneous representation in both the sequence location and scale dimensions. The synthesis of fractional Brownian motion using inverse wavelet transformations has been described.[57]

Figure 1A represents a specific fractional Brownian series, $V_h(n)$, and its continuous Morlet wavelet transformation. The wavelet moduli demonstrate statistical scaling behavior. Generally, if the scale of the sequence partition size, n, is changed by factor r, rn, then the increments, $\Delta V_h(n)$, are changed by a factor r^h. As noted above, in this process, the expectation of the increments, $\langle \Delta V_n(rn) \rangle^2 \approx r^{2h} \langle \Delta V_h(n)^2 \rangle$ in which $0 < h < 1$, with $h = 0.5$ representing the Brownian motion of independent random events. Choosing a sample size of 500, corresponding to a common protein length, the series portrayed results from a highly correlated, "persistent," series, exponent $h = 0.9$, obtained using a standard recursive generating algorithm called *random midpoint displacement*, which simulates *fractal Brownian* series with correlations at all scales.[58, 59]

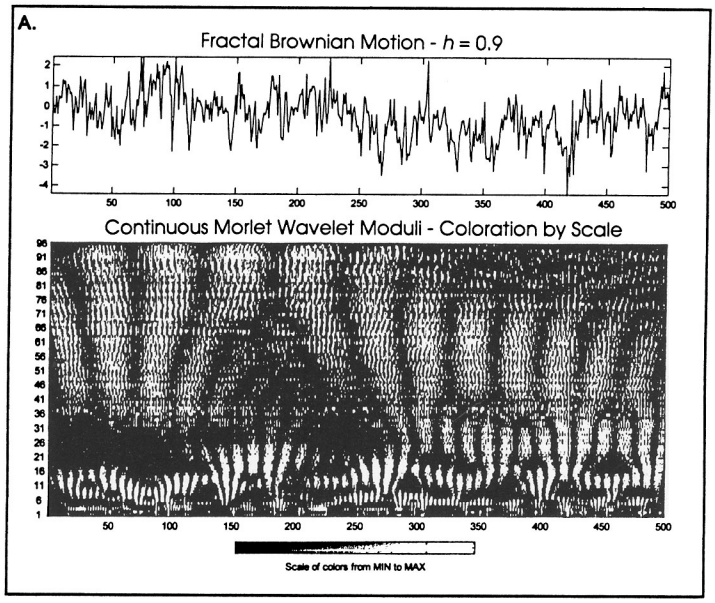

Fig. 1. (A) Graphs from a Fractal Brownian sequence of increments (generated by the random midpoint displacement method with scalar parameter, $h = 0.9$) and their continuous Morlet wavelet transformation shaded for modular amplitudes, normalized at each scale and plotted along the sequence location, x, and dilate dimension, y. Increased correlations between sequential values results in higher modular amplitudes across dilates similar to those seen in wavelet transformations of hydrophobic free energy series of globular proteins (Figs. 2A and 2B). See text. (B) Graphs of the sum of an amplitude scaled sine series truncated at five low valued Fibonacci inverse frequencies. Its continuous Morlet wavelet transformation demonstrates hierarchically and translationally symmetric ranks of modes in dilate space, not unlike the characteristic of continuous wavelet transformations of the hydrophobicity sequences of globular proteins. See text.

As preparation for interpreting the graphs of the Morlet wavelets of the H_n sequences of proteins, we attend to the Morlet wavelengths and locations of some of the maximal moduli in the wavelet graph in Fig. 1A, computed as $\varpi \approx 1/[0.5 - (dd * 0.5/96)]$. At dilation divisions, 2 to 6 *dd*, we see modular maxima at mother wavelet wavelengths (in units of series length) of $\varpi = 2.04$ to 2.13 intermittently throughout the series. At series locations, 1 to 135, we see in addition, maximal moduli in the dilate range of 7 to 16 *dd*, implicating wavelengths $\varpi = 2.16$ to 2.40, as well as hierarchical arrangements of larger variations riding on still larger ones of 36 to 92 *dd*, wavelengths range from $\varpi = 3.2$ to 48.0. Figure 1A demonstrates the expected vertical lengthening and intensification at the middle and larger scales, consistent with the increased cross scale correlations compared with wavelet studies at $h = 0.5$ (not shown).

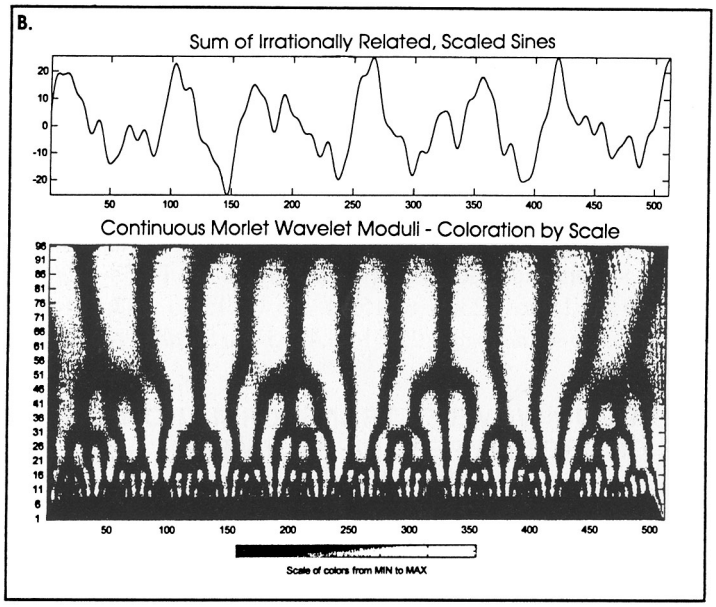

Fig. 1. (*Continued*)

5.2. Hierarchical Rotations of Incommensurate Modes

A hierarchical continuous wavelet graph also results from a deterministic model of hydrophobic free energy rotation numbers (the average number of sequential amino acids required to go from low to high to low again in side chain hydrophobic values). Independence of their rotation numbers might be speculated to reduce the probability of "hydrophobic zipper" coupling between a protein's peptide segments. This property of a set of rotation numbers has been found to be dependent upon their number theoretic character, specifically how well they can be approximated by rationals.[60] It has been proven that the "more irrational" sequential mode ratios are, the less likely they are to participate in mode matched binding.[61] The most irrational set of numbers are called the *nobel numbers*,[62] which are approximated by a Fibonacci series, $\omega_{i+1} = \omega_i + \omega_{i-1}$, with the ratio of successive terms converging on $\omega_{i+1}/\omega_i = (1+\sqrt{5})/2 = 1.618...$ (or its inverse, $1/1.618... = 0.618...$). Given the finite number of fundamental hydrophobic modes found in peptide transmitters, receptor and other proteins,[41] the hydrophobic mode *golden* amino acid wavelength sequence could be approximated by 2, 3, 5, 8, 13...

Figure 1B is the Morlet continuous wavelet transformation of the sum of an amplitude scaled sine series using a short set of Fibonacci numbers

as amplitude scaled inverse frequencies: $y(i) = (2 * \sin((1/2) * i) + (3 * \sin((1/3) * i) + (5 * \sin((1/5) * i) + (8 * \sin((1/8) * i) + (13 * \sin((1/13) * i))$, $i = 1 \cdots 500$ residue pseudo-protein sequence. In contrast with the more inhomogeneous and statistically hierarchical wavelet graph of the $h = 0.9$ correlated fractal Brownian sequence of Fig. 1A, the five inverse-frequencies, amplitude scaled sequence seen in Fig. 1B are more clearly demarcated and regularly localized in both dilate and sequence location space on the wavelet plane. It demonstrates a sequence of translationally and hierarchically symmetric modes computed using $\varpi \approx 1/[0.5 - (dd * 0.5/96)]$: 7 to 10 $dd = \varpi = 2.15$ to 2.23; 7 to 15 $dd = \varpi = 2.15$ to 2.37; 7 to 27 $dd = \varpi = 2.15$ to 2.78; 7 to 41 $dd = \varpi = 2.15$ to 3.49; 57 to 94 $dd = \varpi = 4.97$ to 96.

6. CONTINUOUS WAVELET TRANSFORMATIONS OF TWO KINDS OF PROTEIN PEPTIDE POLYMERS

6.1. Two Globular Proteins

The amino acid sequences of the proteins were obtained from SWISS-PROT Protein Sequence Data Bank, Release, 32.0, September, 1995 (Med. Biochem. Department, University of Geneva, Switzerland, Amos Bairoch) and transformed into amino acid equivalent hydrophobic free energy values, $H_{i=1\cdots n}$, as indicated above.[24, 25]

Figure 2A are graphs of the H_n series of the α chain monomer of the human hemoglobin tetramer, *hemoglobin A* and its continuous wavelet transformation. The levels of hierarchical organization of this typical Linderstrøm-Lang globular protein are qualitatively portrayed. We remind ourselves that the values along the x-axis indicate position in the sequence and the numbers along the y-axis of the wavelet plot represent increasing dilation (longer wavelength, slower frequency of hydrophobic free energy variation along the amino acid sequence) with the range of the moduli scaled relative to the range at each of 64 dilate divisions, *dd*, and not across all *dd*. The mother wavelet wavelengths were relative to 64 *dd*'s as $\varpi \approx 1/[0.5 - (dd * 0.5/64)]$ *aa*.

Our attention is drawn to the modular maxima of the α-helical regions in dilate space since *hemoglobin A* has ≈80% α helical content, as established in many x-ray crystallographic studies and others using optical rotary dispersion and most easily seen in "ribbon diagrams" as graphical summaries of physical data relevant to tertiary structure.[63] We note the intermittent appearance of ≈2 *aa* modular prominences between helix connectivities and of turns at $dd = 9$ to 13, $\varpi = 2.32$ to 2.50 *aa*, but the most regular and prominent of the sequential structures in this graph are the modular maxima of the helical regions in dilate space which cluster in the

vicinity of $dd = 28$ to 30, $\varpi = 3.56$ to 3.79 *aa*. With reference to the known x-ray crystallographic data, the first helical segment at sequence positions 4 to 35 is evidenced in two modular maximal patches, whereas the second in sequence and second shortest helix, 37 to 42 is seen as one. The third helical segment at sequence position 53 to 71 is well marked in helical dilate space, but the shortest at 76 to 79 and the next one at 81 to 89 are more diffuse. The sixth helical segment at sequence location 96 to 112 is also not well marked, but the large seventh helical segment at 119 to 136, like the first and largest one at sequence location 4 to 35, is manifested in two patches of modular maxima. The globin fold does not involve helical segments in bundles, so that the modular hierarchy above the helical locations are not as well defined, as has been the case in other helix barrel bundle containing proteins we have studied.[6,7]

Cross scale connectivities in shorter and longer aperiodic loops, "cross scale singularities"[51] from $dd = 5$ or 9 to 45 or higher, $\varpi \approx 2$ to 7 *aa* appear intermittently throughout the graph. The three large regions of hierarchical modular maxima centered at sequence positions ≈ 27, 64 and 108 with $dd = 53$ to 63, $\varpi = 11.6$ to 42.7 correspond to the local relative maximum, minimum and maximum respectively of their iteratively smoothed, nearest neighbor averaged hydrophobic sequences. This finding is consistent with the suggestion (and a hydropathy plot of its H_n) that the entire polypeptide chain at largest scales consists of two hydrophobic domains.

In the interpretation of this and other wavelet graphs of the hydrophobic free energy sequences of proteins, at least two caveats require consideration: (1) Since the wavelet transform is the scalar product of the function with the mother wavelet (as dilated and translated), the amplitudes of the modular maxima in addition to their sequential correlation are also a function of the average hydrophobicity of the data segment. It is this second factor rather than the fit of the mother wavelet scale that accounts for the differences in the modular densities of the wavelet structures representing helices in this graph. For example, the average hydrophobicity of the clearly demarcated modular maxima of the initial α-helical segment, sequence location 4 to 16, PADKTNVKAAWGKV, is 1.33 kcal/mol contrasted with the more poorly marked helical segment with modular maxima at sequence location 81 to 89, ALSDLHAHK, with an average hydrophobicity per residue of 1.02 kcal/mol. A similar continuous wavelet structure with more uniform sequential hydrophobicity and more symmetric wavelet transform graph has been elucidated for another globin protein, *myoglobin*.[6,7] (2) Maximal modular densities widen and become more vague with upward vertical distance from these helical patches, partially explained by the dependence on dilate scale of the correlation length

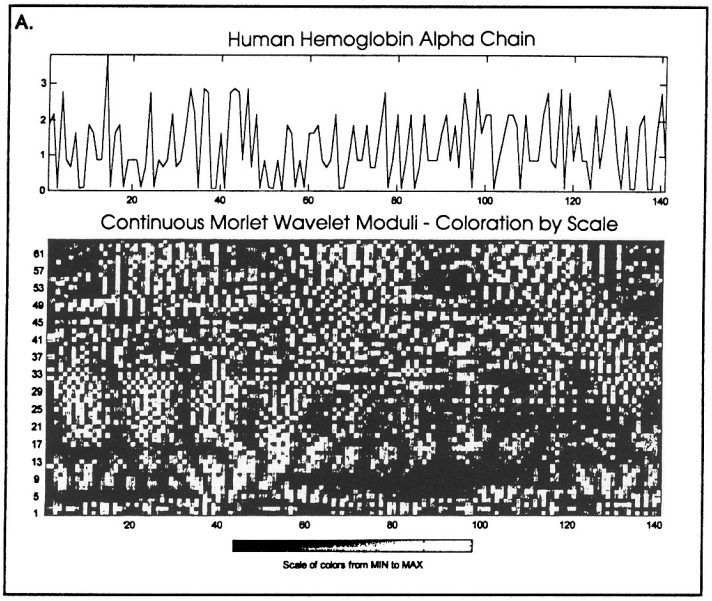

Fig. 2. (A) The hydrophobic free energy series, $H_{i=1,\ldots,n}$ of human hemoglobin *A* α-chain and its continuous Morlet wavelet transformation. The secondary and supersecondary levels of organization of this typical Linderstrøm-Lang globular protein are qualitatively portrayed across the sequence. Notice the sequential repetitions of the hierarchical structures in the vicinity of dilate divisions, $dd = 27$–30, regions of modular maxima corresponding to 3.46–3.79 amino acid α-helical wavelengths. See text for details. (B) A graph of the $H_{i=1,\ldots,n}$ series and wavelet transformation of hierarchical globular protein, *prealbumin*. *Prealbumin* is dominated by the repeating hierarchical hydrophobic mode structures of β-strands and their aggregate as four stranded β-sheets: $dd = 5$–13, $\varpi = 2.17$–2.50 amino acids of the fast frequency hydrophobic variations composing the β-strands. Also prominant are the $dd = 29$–43, $\varpi = 3.65$–6.09 amino acids representing the repeating β-strand lengths which compose the β-sheets. The relevance of this transformation of one dimensional information to tertiary structure is facilitated by the x-ray crystallographic finding that sequentially encoded β-strands are generally adjacent in the domain pairs of four stranded β-sheets. See text for details.

of the analyzing wavelet. In $W(a, b)$, the effective correlation length of the mother wavelet w at location n_0 varies with scale a, such that $|n_0 - b| \leqslant \sqrt{\text{var}_w}\, a$, where var_w is the variance of the mother wavelet, w. The diffuse and widening modular densities above the helical patches confound hierarchical quasi-regular supersecondary and global hydrophobic domains with what is known as the scale dependent expanding "cone of influence"of the mother wavelet with vertical distance from the index point n_0.[51] One way to separate these influences is the linear decomposition and construction of ψ_i (as above) before continuous wavelet transformation.[7]

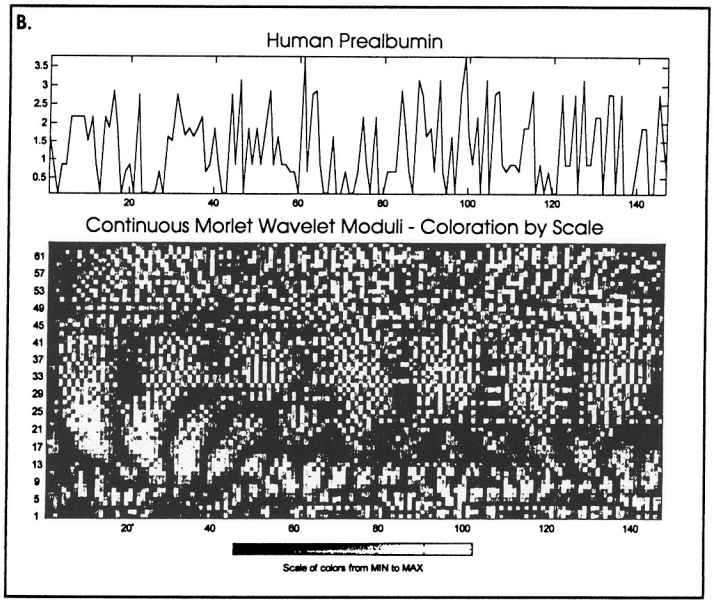

Fig. 2. (*Continued*)

Figure 2B is a graph of the H_n series and wavelet transformation of the globular protein, *prealbumin*, also called *transthyretin*. From x-ray crystallographic studies, it is known that *prealbumin* is dominated by β-strands with the supersecondary structure resembling a "sandwich" of two four stranded β-sheets, with both individual β-strands and their loop connectivities somewhat variable in length.[63] It is also known that sequentially encoded β-strands are generally adjacent in the domain pair of four stranded β-sheets. Two regions of dilate space are of particular interest in the global sequential structure. The first, $dd \approx 5$ to 13, $\varpi \approx 2.17$ to 2.50 *aa*, captures the characteristic hydrophobic mode composition within β-strands, which repeats somewhat irregularly throughout the sequence. The second, $dd \approx 29$ to 43, $\varpi \approx 3.65$ to 6.09 repeats rather regularly throughout the sequence and represents the characteristic lengths of this protein's β-strands which make up the two four stranded β-sheets described above. This interpretation is consistent with this protein's known domination ($\approx 75\%$) by β-strands. The somewhat ill-defined maximum modular bands in the $dd \approx 48$ to 52, $\varpi \approx 8$ to 9 region above and roughly between the β-strand lengths in sequence location, and in the vicinity of $dd = 1$ to 5, $\varpi \approx 2$ may represent the two different loop lengths that connect the β-strands characteristic of the so called *Greek Key* configuration of this protein.[63] In the

dilate region of the global sequential hierarchical hydrophobic domains, $dd = 53$ to 63, $\varpi = 11.6$ to 42.7, quasi-asymptotic smoothing of the hydrophobicity sequence demonstrates a fit to the somewhat diffuse modular maxima with sequence location centered at 18 a local relative minimum, at 37 a local relative maximum, at 73 a local relative minimum, and at 110 a local relative maximum.

Note that the wavelet graphs of both *prealbumin* and *hemoglobin A* manifest crude homogeneities and symmetries composed of translational, semi-invariant, repeating structures in the form of β-strands and α-helices respectively, consistent with the wavelet transformations of the hydrophobicity sequences of other globular proteins that have been studied in this way.[6, 7] There are, in addition, many proteins which combine, in some order, semi-invariant patterns of alternating and repeating structures of both β-strands and α-helices. The *polyproteins* are without this translational

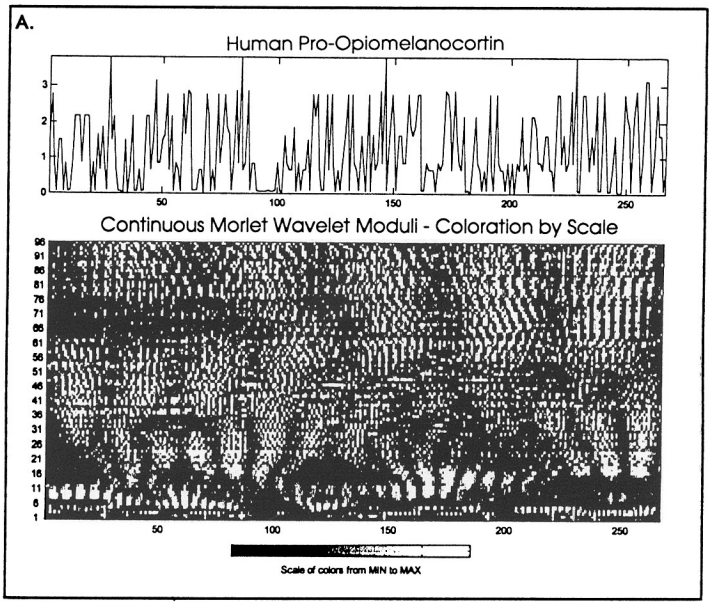

Fig. 3. (A) The $H_{i=1\ldots n}$ series of a representative polyprotein, *pro-opiomelanocortin*, which is enzymatically decomposed into several peptide hormones. The graph of its continuous wavelet transformation, in contrast with the wavelet graphs of globular proteins, demonstrate the irregular appearance and disappearance of a variety of modes along their sequences. Notable is the absence of repeating modular maxima characteristic of α-helices or β-strands. See text for details. (B) The $H_{i=1\ldots n}$ series and its continuous wavelet transformation of a viral polyprotein sequence, the core protein product of the human AIDS related retroviral *GAG1* gene. In contrast with the globular proteins, it shares the sequential variety of the polyprotein, *pro-opiomelanocortin*. See text.

homogeneity and consistency, and the graphic analysis of the continuous wavelet transformation of their H_n sequences help identify these differences in the global properties of proteins from their primary sequences.

6.2. Two Polyproteins

Polyproteins are polypeptide chains that are composed of the sequences of two or more proteins. We chose to compare them with globular proteins because, whereas their size ($\geqslant 50$ residues) is sufficient to have the propensity to fold and remain folded,[64] their subsequence processing requires that they remain sufficiently unfolded to be available for enzymatic decomposition into their final functional forms by trypsin-like proteolytic enzymes acting at basic cleavage sites. Specific proteases cleave polyproteins into their component polypeptides and/or proteins. We speculated that the translationally invariant repeating or alternating modular wavelet structure characteristic of globular proteins would be absent in polyproteins. We examined this proposition using two typical polyproteins.

The polyprotein, *pro-opiomelanocortin*, is synthesized in both the anterior and intermediate lobes of the pituitary and, depending upon location, is broken down into several different polypeptide hormones.[65]

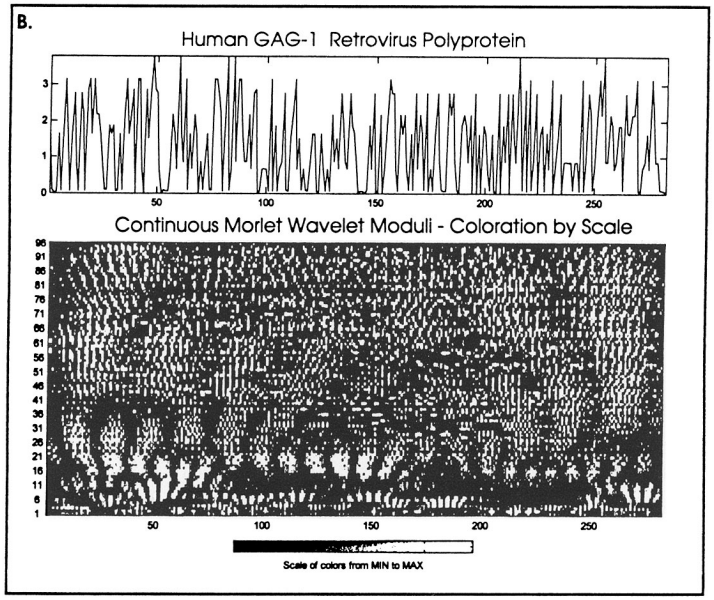

Fig. 3. (*Continued*)

Polyproteins are also common in viruses, such as those of polio and AIDS. The human AIDS related retroviral core protein product of the *GAG1* gene is proteolytically cleaved to yield several proteins with different nucleotide binding properties. The graphs of their continuous wavelet transformations as in Fig. 3A and Fig. 3B demonstrate the irregular appearance and disappearance of a variety of dilate modes along their sequences.

The multiplicity of peptide hormones in the sequence of *pro-opiomelanocortin* have been located and characterized.[65] As noted, depending upon the organ of synthesis, *pro-opiomelanocortin* is broken up in a variety of ways. This evidences itself in the wavelet graphs as the appearance and disappearance of hierarchical wavelet structures along the sequence and the resulting absence of the translational symmetry of globular proteins. We recall that with 96 dilate divisions, $\varpi \approx 1/[0.5 - (dd * 0.5/96)]$.

We describe a few examples of the peptide segments in *pro-opiomelanocortin.* The *signal sequence* is at locations 1 to 26 and demonstrates modular maxima in the vicinity of $dd = 8.5$, ($\varpi \approx 2.25$ *aa*), $dd = 36$ ($\varpi \approx 3.2$ *aa*) and $dd = 59$ ($\varpi \approx 5.2$ *aa*) which was confirmed using the peaks of the $h(\omega)$ of its H_n. The next sequence, the *n-terminal* segment, sequence locations 27 to 102, is dominated by six sequential clusters of maximal moduli centered at $dd = 21$, $\varpi \approx 2.5$ *aa* suggesting a sequential set of β-strands characteristic of β-sheet structures. In addition there are three vertical bands that bracket the α-helical region, $dd = 36$ to 61, $\varpi = 3.2$ to 5.5 *aa*. *γ-melanotropin*, sequence locations 77 to 87, a part of the *n-terminal* segment, demonstrates this β-strand and the $h(\omega)$, as well as the lateral band at $dd = 61$ in the $W(a, b)$ graph, indicates the presence of a $\varpi \approx 5.5$ *aa* in this regions. *β-endorphin*, a source of endogenous opiates, appears uniquely at sequence locations 237 to 267 in the form of three vertical clusters of maximum moduli, ranging in wavelength from $dd = 16$, $\varpi = 2.4$ *aa* to (the last one) reaching to $dd = 52$, $\varpi = 4.4$ *aa*, the latter reported previously as the characteristic mode of the brain's opiate peptides.[8] The *adrenocorticotropin hormone, ACTH*, is at sequence locations 138 to 176 and is dominated by hydrophobic modes at $dd = 6$ to 16, $\varpi = 2.25$ to 2.4 *aa* as is *α-melantropin*, sequence locations 138 to 150. At largest dilations, $dd = 77$ to 92, the wavelet graph demonstrates five clusters of maximum moduli and quasi-asymptotic smoothing of the hydrophobicity sequence showed their correspondence in sequence locations with two minima intercalated among three maxima inscribing three global sequential hydrophobic domains. The lack of translational symmetry in the *GAG-1* polyprotein and the appearance and disappearance of modes of several sizes and locations along the sequence are seen in Fig. 3B.

The global hierarchical plan of the Linderstrøm-Lang globular proteins, as described and evidenced above in the continuous wavelet transformation

of their hydrophobic free energy sequences, is missing in both of these polyproteins. It appears that the presence (*globular*) or absence (*polyprotein*) of a sequence that manifests a fundamental hydrophobic frequency embedded in a pattern of secondary structures can be postulated from the translationally invariant, repeating, maximum modular hydrophobic mode structures found in the graphical patterns of the continuous wavelet transformation of a protein's one dimensional hydrophobic sequence.

The following section employs a different method with which to discriminate protein families. It characterizes protein polypeptide sequences in terms of the"all poles," maximum entropy power spectral peaks of the dominant orthogonal eigenfunctions constructed from their H_n's.

7. EIGENFUNCTION MODE VARIETY IN GLOBULAR AND POLYPROTEIN POLYPEPIDES

To briefly review from above, following hydrophobic free energy transformations of the protein's amino acid sequences, lagged autocovariance matrices of order eight yield sets of ordered eigenvectors, such that their vectorial convolution with the original hydrophobic free energy sequence results in orthogonal hydrophobic free energy (eigen)functions, ψ_i. The eigenvector, X_1, associated with the largest eigenvalue $\{v_j\}_{j=1}$ of the covariance matrix of a hydrophobic sequence composed with the hydrophobic sequence generates ψ_1. The $X_j(l)$ that follow in statistical weight in the linear decomposition are independent and dominated by different wavelengths. Maximum entropy power spectral transformations, $h(\omega)$, are then used to index the four leading eigenfunctions with respect to their inverse-frequencies, ω^{-1}, in amino acid sequential hydrophobicities, identifying the protein's hydrophobic free energy modes. We use these techniques to further examine our conjecture that in contrast with the anticipated greater mode variety, discontinuities and co-primeness of the polyproteins, the globular proteins would be dominated by more repetitious, continuous and less coprime ω^{-1}'s, facilitating their self hydrophobic zipper-like mode coupling.

Figure 4 (left side) contains graphs of the $\psi_{i=1\ldots4}$ of *pro-opio-melanocortin*, a representative polyprotein, and (right side) their associated $h(\omega)$. The $\psi_{i=1\ldots4}$ demonstrate five different ω^{-1}'s, four of which approximate the first four terms of a Fibonacci-like sequence: $\omega^{-1}=2$, 3.21, 4.95, 8.76. The $h(\omega)$ of ψ_1 yield $\omega^{-1}=27.11$ which reflects the average length of the multiple of peptide product size composing the post-proteolytic pattern of seven peptide hormones and their signal sequence. Figure 5 (left side) are graphs of $\psi_{i=1\ldots4}$ of *hemoglobin A*, a generic globular protein, dominated by α-helical ω^{-1}'s. The $h(\omega)$ of $\psi_1=3.44$ *aa*

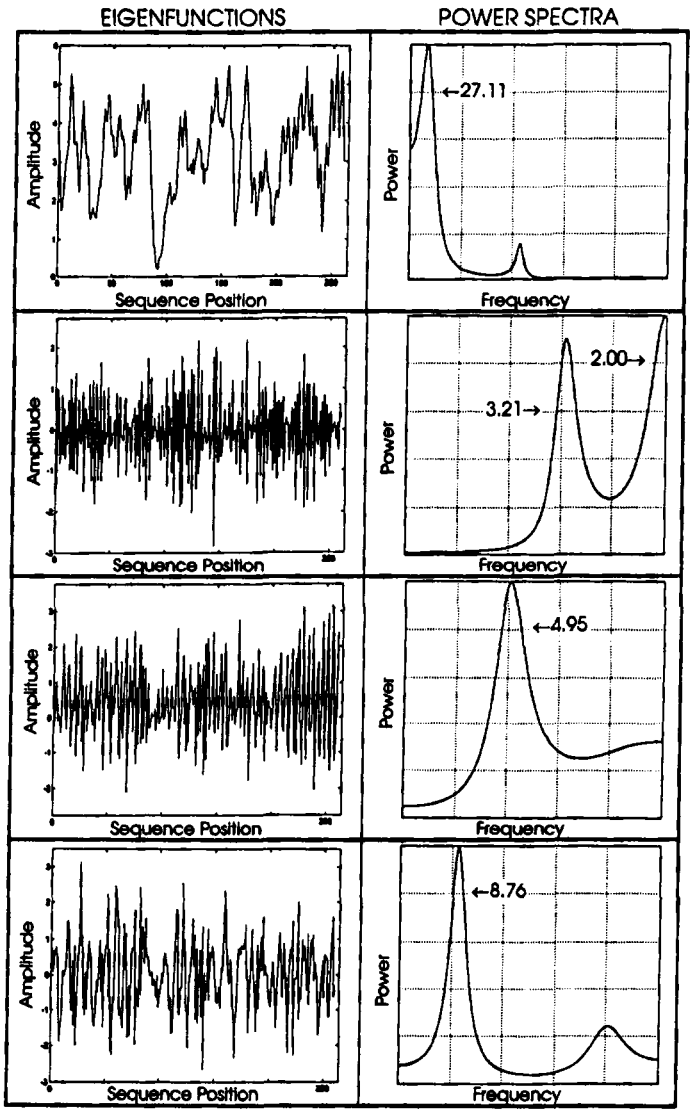

Fig. 4. Left column contains graphs of the first four eigenfunctions of *pro-opiomelanocortin*, a representative polyprotein, and right column their associated maximum entropy power spectra with their inverse frequencies labeled. See text. The first four eigenfunctions demonstrate five different modes, four of which approximate the first four terms in the Fibonacci sequence. The leading eigenfunction mode of $\omega^{-1} = 27.11$ amino acids reflects an approximate average length of the seven final peptide products. The polyprotein mode structure is more varied than that of the globular protein and demonstrates a series of rotation number modes that tended to be number theoretically "more incommensurate" in ratio. See text for details.

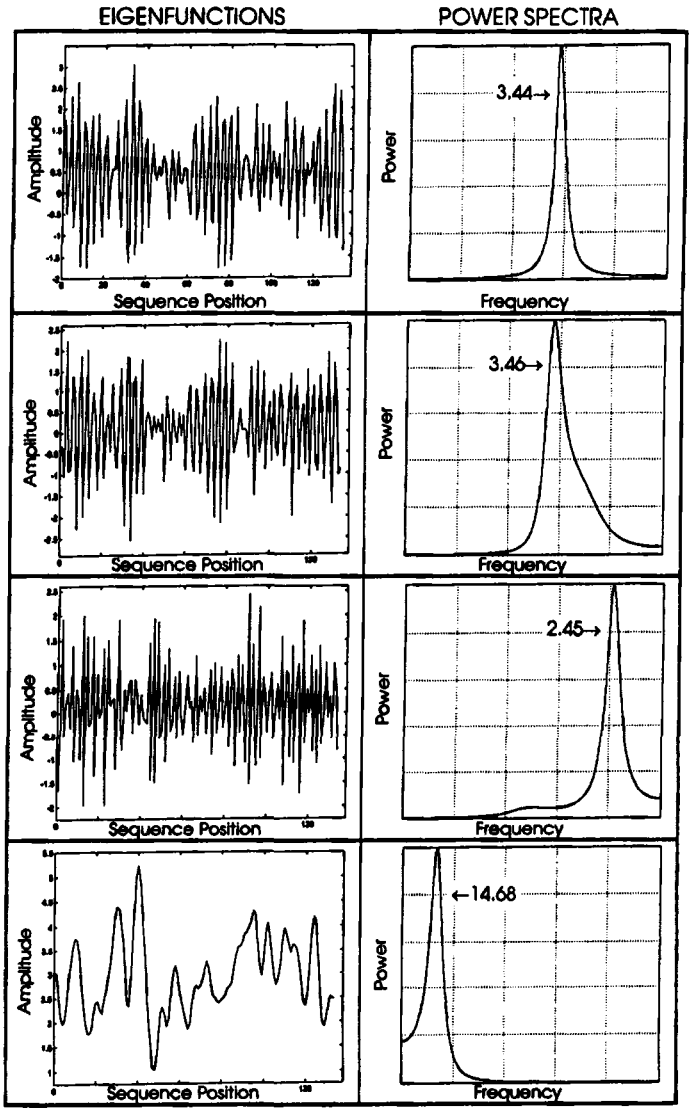

Fig. 5. Left column are graphs of the first four eigenfunctions of hemoglobin A, a generic globular protein with $\approx 80\%$ α-helix. This is reflected (right column) in the dominance of the α-helical inverse frequencies, $\omega^{-1} = 3.44$ *aa*, $\omega^{-1} = 3.46$ amino acids. The β-strand inter-helical connective loops manifest in ψ_3 as $\omega^{-1} = 2.45$ amino acids and an approximate representation of a typical helical barrel length is found in ψ_4 as $\omega^{-1} = 14.68$ amino acids. See text for details.

and the $h(\omega)$ of $\psi_2 = 3.46$ *aa*. The β-strand connective loops are seen in ψ_3 with $h(\omega) = 2.45$ *aa* and a not atypical helical barrel length in ψ_4 with $h(\omega) = 14.68$ *aa*. As we anticipated, the polyprotein mode structure was more varied than that of the globular protein and demonstrated ratios of hydrophobic free energy rotation numbers that were number theoretically "more incommensurate." We interpret this difference as contributing, along with its sequential hydrophobic transience and heterogeneity to the polyprotein's greater resistance to internal mode matched, "hydrophobic zippering."

ACKNOWLEDGMENTS

This work was supported by the National Institute of Mental Health, Small Business Innovation Research Grant R43 MH58026-01. The authors express appreciation for the helpful suggestions of a reviewer.

REFERENCES

1. L. P. Kadanoff, *From Order to Chaos; Essays: Critical, Chaotic and Otherwise* (World Scientific Press, Singapore, 1993).
2. D. Gershon, *Nature* **389**:418 (1997).
3. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **136**:225 (1980).
4. C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, N.Y., 1991).
5. D. Searls, *Nature* **389**:417 (1997).
6. A. J. Mandell, K. A. Selz, and M. F. Shlesinger, *Physica A* **244**:254 (1997).
7. A. J. Mandell, K. A. Selz, and M. F. Shlesinger, *Proc. International School of Physics; Enrico Fermi Course. CXXXIV*, F. Mallamace and H. E. Stanley, ed. (I.O.S. Press, Amsterdam, 1997), pp. 175–192.
8. A. J. Mandell, K. A. Selz, and M. F. Shlesinger, *Proc. Natl. Acad. Sci.* **94**:13576 (1997).
9. A. J. Mandell, Michael J. Owens, Karen A. Selz, W. Neal Morgan, Michael F. Shlesinger, and Charles B. Nemeroff, *Biopolymers* **46**:89 (1998).
10. E. A. Di Marzio and A. J. Mandell, *J. Chem. Physics* **107**:5510 (1997).
11. T. E. Creighton, *Proteins, Structures and Molecular Properties* (Freeman, N.Y., 1984).
12. A. A. Zamyatnin, *Prog. Biophys. Mol. Biol.* **24**:109 (1972).
13. W. Kauzmann, *Adv. Prot. Chem.* **14**:1 (1959).
14. C. Tanford, *Proc. Natl. Acad. Sci.* **76**:4175 (1979).
15. F. H. Stillinger, *Science* **209**:451 (1980).
16. W. Blokzijl and J. B. Engberts, *Angew. Chem. Int. Edn. Engl.* **32**:1545 (1993).
17. C. Y. Lee, J. A. McCammon, and P. J. Rossky, *J. Chem. Phys* **80**:4448 (1984).
18. J. T. Edsall, *J. Am. Chem. Soc.* **57**:1506 (1935).
19. J. M. Sturtevant, *Proc. Natl. Acad. Sci.* **74**:2236 (1977).
20. P. L. Privalov and N. N. Khechinashvili, *J. Mol. Biol.* **86**:665 (1974).
21. J. A. Reynolds, D. B. Gilbert, and C. Tanford, *Proc. Natl. Acad. Sci.* **71**:2925 (1974).
22. B. K. Lee and F. M. Richards, *J. Mol. Biol.* **55**:379 (1971).
23. R. Lumry and S. Rajender, *Biopolymers* **9**:1125 (1970).
24. Y. Nozaki and C. Tanford, *J. Biol. Chem.* **246**:2211 (1971).

25. P. Manavalen and P. K. Ponnuswamy, *Nature* **275**:673 (1978).
26. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**:105 (1982).
27. S. H. White and R. E. Jacobs, *Biophys. J.* **57**:911 (1990).
28. M. Degli Esposti, M. Crimi, and G. Venturoli, *Eur. J. Biochem.* **190**:207 (1990).
29. K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig, *Science* **252**:106, 1991.
30. J. M. Sanches-Ruiz, *Eur. Biophys. J.* **24**:261 (1996).
31. J. Janin, *Prog. Biophys. Molec. Biol.* **64**:145 (1995).
32. C. Chothia, *Nature* **254**:304 (1975).
33. K. W. Plaxco, K. T. Simons, and D. Baker, *J. Mol. Biol.* **277**:985 (1998).
34. H. S. Chan, *Nature* **392**:761 (1998).
35. A. M. Gronenborn and G. M. Clore, *Science* **263**:536 (1994).
36. K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci.* **90**:1942 (1993).
37. R. M. Pashley, P. M. McGuiggan, B. W. Ninham, and D. F. Evans, *Science* **229**:1088 (1985).
38. J. Israelachvili and H. Wennerstrom, *Nature* **379**:219 (1996).
39. C. Chothia, *Ann. Rev. Biochem.* **53**:537 (1984).
40. G. D. Rose, *Nature* 272:586 (1978).
41. A. J. Mandell, *Ann. Rev. Pharmacol. Toxicol.* **24**:237 (1984).
42. A. J. Mandell, P. V. Russo, and B. Blomgren, *Ann. Rev. N.Y. Acad. Sci.* **504**:88 (1987).
43. I. Daubechies, A. Grossman, and Y. Meyer, *J. Math. Phys.* **27**:1271 (1986).
44. A. Grossmann and J. Morlet, *SIAM J. Math. Anal.* **15**:723 (1984).
45. D. S. Broomhead and G. P. King, *Physica D* **20**:217 (1986).
46. D. S. Broomhead, R. Jones, and G. P. King, *J. Phys. A* **20**:L563 (1987).
47. R. N. Madan, *Maximum Entropy and Bayesian Methods* (Kluwer Academic, Netherlands, 1993).
48. K. Linderstrøm-Lang and J. Schellman, *Enzymes* **1**:443 (1959).
49. A. Arneodo, G. Grasseau, and M. Holschneider, *Phys. Rev. Lett.* **61**:2281 (1988).
50. M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software* (A. K. Peters, Wellesley, MA, 1994), pp. 103–150.
51. M. Farge, *Ann. Rev. Fl. Mech.* **24**:395 (1992).
52. J. P. Burg, *Geophysics* **37**:375 (1972).
53. L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, New York, 1972).
54. A. M. Yaglom, *Correlation Theory of Stationary and Related Random Functions* (Springer-Verlag, New York, 1986).
55. G. C. Wagner, J. T. Colvin, J. P. Allen, and H. J. Stapleton, *J. Am. Chem. Soc.* **107**:5591 (1985).
56. S. Alexander, J. Bernasconi, W. Schneider, and R. Orbach, *Rev. Mod. Physics* **53**:175 (1981).
57. S. G. Mallat, *IEEE Trans. on Pattern Anal. Machine Intell.* **PAMI-II**:674 (1989).
58. A. Fournier, D. Fussell, and L. Carpenter, *Commun. ACM* **25**:371 (1982).
59. R. F. Voss, *NATO ASI Series* **F17**, R. A. Earnshaw, ed. (Springer-Verlag, New York, 1985).
60. J. C. Yoccoz, *Ann. Sci. de l'ENS* **17**:333 (1984).
61. M. R. Herman, *Asterisque* **1**:103 (1983).
62. G. H. Hardy and E. M. Wright, *Introduction to the Theory of Numbers*, 5th ed. (Clarendon, London, 1979).
63. J. S. Richardson, *Adv. Prot. Chem.* **34**:167 (1981).
64. J. Janin, *Bull. Inst. Pasteur* **77**:337 (1979).
65. J. Douglass, O. Civelli, and E. Herbert, *Annu. Rev. Biochem.* **53**:698 (1984).