Multifractal analysis of solvent accessibilities in proteins

James S. Balafas and T. Gregory Dewey*
Department of Chemistry, University of Denver, Denver, Colorado 80208†

(Received 24 February 1995)

The solvent accessibilities of amino acid side chains in a protein can be determined computationally from x-ray crystallographic data. The sequential profile of these accessibilities shows a seemingly random variation. A generalized box-counting analysis of such profiles shows multifractal behavior. Multifractal spectra obtained for a variety of proteins are broader than a corresponding random array, indicating an underlying hierarchical structure to the proteins. The multifractal parameters are used to extract an underlying binary multiplicative process with one-step memory. Similar binary processes occur in the generation of helix-coil sequences in biopolymers. Computer simulations were performed that generated misfolded proteins. The misfolded proteins have narrower multifractal spectra than the properly folded ones. Thus, this multifractal analysis can be used as a diagnostic tool in assessing proper folding in structure prediction algorithms.

PACS number(s): 87.10.+e, 87.15.By, 61.43.Hv, 05.90.+m

I. INTRODUCTION

Fractal geometry provides a mathematical formalism for describing complex spatial and dynamical structures [1,2]. It has been used to investigate such complex phenomena as dielectric breakdown, turbulence, and diffusion-limited aggregation. The fractal approach allows a unified characterization and comparison of diverse structures. Because of the underlying complexity of their structure, it is natural to develop fractal descriptors of proteins. To date, such applications have been limited. Protein structure has been characterized by two fractal descriptors, the fractal dimension of the protein backbone and the fractal surface dimension. Proteins appear to follow "universal" laws with regard to these descriptors. A large data set of proteins shows that the dimension of the contour of the protein backbone is slightly less than 3 [3]. Thus, proteins behave statistically as collapsed polymers. The fractal dimension of the accessible surface of a protein has been investigated in a number of studies [3–6]. Typically, the value for the surface fractal dimension is approximately 2.2. Protein surfaces are slightly more convoluted than a smooth spherical surface, which would have a dimension of 2. Thus, the fractal description of a protein is consistent with the common view of proteins as compact structures with smooth surfaces [7]. The utility of the approach lies in the quantization of these features. Recently, these gross morphological factors have been related to the kinetics of hydrogen isotope exchange [8,9]. This fractal theory of the kinetics of small molecule-protein interactions demonstrates the role of the fractal surface dimension of the protein.

Despite the success of the fractal formalism, by the mid 1980s the limitations of the approach had become apparent. Interest shifted from characterizing a set of points in a complex geometric structure to characterizing complex probability densities. Such problems required a generalization of fractal concepts, and the multifractal formalism was developed (for reviews see [10–12]). Density distributions are characterized by an infinity of fractal dimensions, hence the terminology multifractal. The multifractal formalism has been used as a descriptor for a variety of physical and chemical phenomena, such as diffusion-limited aggregation [13], percolating clusters [14], energy dissipation in fully developed turbulent flows [15], configuration of Ising spins at critical points [16], and the characterization of strange attractors [17]. Recently the multifractal approach has been extended to the description of helix-coil transitions in biopolymers [18,19].

In the present work, we apply the multifractal formalism to the analysis of a structural parameter in proteins known as the solvent accessibility. Using a "ball rolling" algorithm [7], it is possible to determine from x-ray structures the exposed surface area of each amino acid residue in the protein sequence. Typically, a probe of the radius of a water molecule is used. The fractional solvent accessibility is the exposed surface area divided by the area of a fully exposed amino acid as it would appear in the middle of a tripeptide. In Fig. 1, the solvent accessibilities of the amino acid side chains are shown as a function of position along the protein backbone. Two questions will be addressed with regard to this sequential display of data. First, is there any correlation in this seemingly random pattern? Second, if there are correlations, what are the implications for the structure of the protein? There is a wide variety of methods for analyzing correlations in "noisy data." Previously, a Hurst or range analysis was used to demonstrate long-range correlation among Debye-Waller factors at different positions along a protein chain [20]. In the present work, we use the multifractal formalism to analyze solvent accessibility profiles. One reason for choosing this particular ap-

*Author to whom correspondence should be addressed.

†FAX: (303)-871-2254; electronic address: gdewey@cair.du.edu

proach is that the multifractal spectrum is a reflection of a hierarchical structure. Ultimately, one would hope to relate the multifractal properties of proteins to other hierarchical models, such as the conformational substates of Frauenfelder, Sligar, and Wolynes [21]. As will be seen, the multifractal spectrum provides a signature for correlations in properly folded proteins. Thus, it provides a diagnostic for assessing protein folding algorithms.

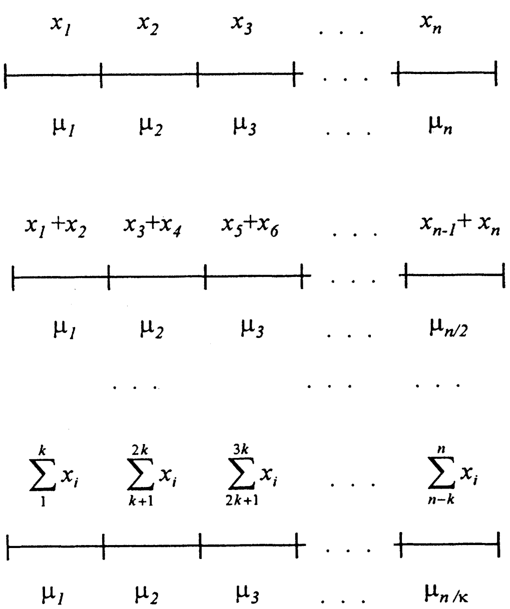
In Sec. II, the algorithm for calculating multifractal spectra of the solvent accessibilities of protein amino acid side chains is presented. This algorithm is essentially a "box-counting" method and is common in the multifractal literature [22]. In this section, the multifractal parameters are defined, and technical details regarding the calculation are presented. Section III presents the multifractal spectrum for 18 proteins. These results show the inherent multifractal nature of the solvent accessibilities. Results for "misfolded" proteins are also presented. The multifractal spectra can distinguish proper from misfolded structures. Empirical evidence shows that a properly folded protein will have maximized the width of its multifractal spectrum. In Sec. IV, it is shown that protein multifractal spectra can be represented by a binary multiplicative model with one-step memory. The formalism for this model is presented, and it is used to analyze the protein spectra. Finally, in Sec. V, the meaning of multifractal behavior in the context of protein structure is discussed.

II. ALGORITHM FOR CALCULATING MULTIFRACTAL SPECTRA

A generalized box-counting method is used to analyze sequential data (as in Fig. 1) on the solvent accessibilities of protein side chains. This method is used to determine the "generalized" dimensions associated with the shape of the profile. These generalized dimensions provide information on the hierarchical nature of the "noise" in Fig. 1. To proceed with the details of the method, consider the solvent accessibilities as a sequence of length n,

$$\{x_i : i=1,\ldots,n\}\ , \tag{1}$$

where n is the number of amino acids that constitute the protein, i.e., the length of the protein. The solvent accessibilities are ordered along a linear array according to their respective positions in the amino acid sequence. This sequence is "covered" with boxes of a defined length and the value(s) of the accessibilities within a given box is assigned to the box. Initially, one starts with boxes of size 1, i.e., it covers a single residue. The value of the accessibility of that residue is assigned to its respective box. In the second iteration, the box size is doubled, so one box covers two residues. The sum of the two accessibilities covered is now assigned to the respective box. The procedure is repeated with increasing box sizes. The following illustrations shows how the box-counting algorithm was applied to a linear sequence for boxes of size $\delta=1$, 2, and k:


x_1 x_2 x_3 . . . x_n
μ_1 μ_2 μ_3 . . . μ_n
x_1+x_2 x_3+x_4 x_5+x_6 . . . $x_{n-1}+x_n$
μ_1 μ_2 μ_3 . . . $\mu_{n/2}$
. . . .
$\sum_1^k x_i$ $\sum_{k+1}^{2k} x_i$ $\sum_{2k+1}^{3k} x_i$ . . . $\sum_{n-k}^{n} x_i$
μ_1 μ_2 μ_3 . . . $\mu_{n/k}$

where x_i is the value of the solvent accessibility for the ith residue along the chain and μ_j is the sum in the jth box,

$$\mu_j(\delta)=\sum_{i=1}^{\delta} x_{i+(j-1)\delta}\ . \tag{2}$$

In our algorithm, boxes of size $\delta=5$ to 34 were used.

A function $Z_q(\delta)$ is now defined that provides the qth moment of the measure,

$$Z_q(\delta)=\sum_{j}^{n/\delta} \mu_j^q(\delta)\ . \tag{3}$$

The scaling ansatz is made such that [2]

$$Z_q(\delta)=\lim_{\delta\to 0} \delta^{-\tau(q)}\ , \tag{4}$$

where $\tau(q)$ is known as the mass exponent. Since q is a discrete variable, the limiting behavior of the mass exponent was obtained from the initial slope of a $\ln[Z_q(\delta)]$ versus $\ln(\delta)$ plot using a linear least squares fit. This procedure is represented as

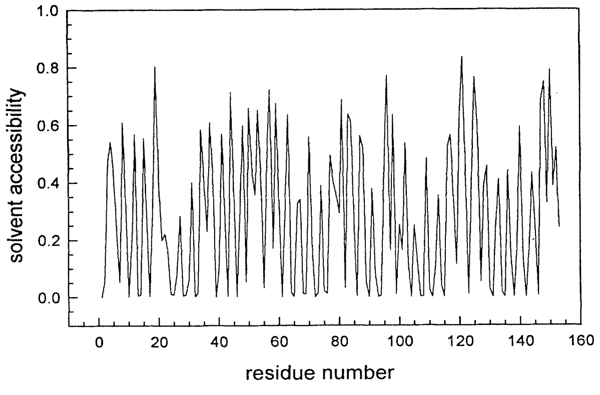


FIG. 1. Solvent accessibility as a function of residue number for the protein myoglobin (153 residues).

$$\tau(q) = -\frac{\ln[Z_q(\delta)]}{\ln(\delta)} \ . \tag{5}$$

The singularities of the measure are characterized by the Lipschitz-Hölder exponent α. This parameter is related to the mass exponent τ according to the relation

$$\alpha(q) = -\frac{d}{dq}\tau(q) \ . \tag{6}$$

Substitution of Eq. (5) into Eq. (6) yields

$$\alpha(q) = -\frac{\sum_j \mu_j^q \ln(\mu_j)}{Z_q(\delta)\ln(\delta)} \ . \tag{7}$$

Again, the limiting behavior of the discrete variable q requires that the Lipschitz-Hölder exponent be obtained from the initial slope of a plot of $\sum_j \mu_j^q \ln(\mu_j)/Z_q(\delta)$ vs $\ln(\delta)$. Now, the multifractal spectrum $f(\alpha)$ versus α can be calculated according to

$$f(\alpha) = q\alpha(q) + \tau(q) \ , \tag{8}$$

where Feder's convention has been used in Eq. (8) [2].

The function $Z_q(\delta)$ is analogous in structure to the partition function of statistical mechanics (cf. [10]). In this analogy the multifractal parameters become "generalized" thermodynamic functions. This correspondence is based on the Legendre transformation properties of $\tau(q)$. Thus, q is a generalized temperature, τ is the generalized free energy, α is the generalized energy, and f is the generalized entropy. In this formal analogy, the multifractal spectrum represents a relationship between the generalized energy and the entropy of the problem.

As seen in the previous presentation, a spectrum is generated by calculating $Z_q(\delta)$ at a fixed q value and varying δ. From the two linear regressions, $f(\alpha)$ and α are determined. The entire spectrum is generated by varying q. Both positive and negative integer values of q were used. For a given box size δ, there may not be sufficient residues in the last box to complete the sequence. To circumvent this problem, an initial segment of the sequence was appended to the sequence to fill the last box. Thus, the algorithm entailed a wraparound method (periodic boundary conditions). As a check on the error introduced by this procedure, periodic boundary conditions were applied for sequences taken in both the forward (amino→carboxy) and reverse (carboxy→amino) direction. This changes the content of the appended sequence. The wraparound algorithm produced essentially the same spectra, regardless of the direction of the sequence. A typical comparison of forward and reverse sequence spectra is shown in Fig. 2.

All multifractal spectra were generated from the x-ray crystallographic data using coordinate files obtained from the Protein Data Bank at Brookhaven National Laboratory [23]. The fractional solvent accessibilities of the amino acid side chains were determined computationally using the computer program ACCESS [24]. Initial slopes for determining τ and α [Eqs. (5) and (7)] were determined from linear regressions.

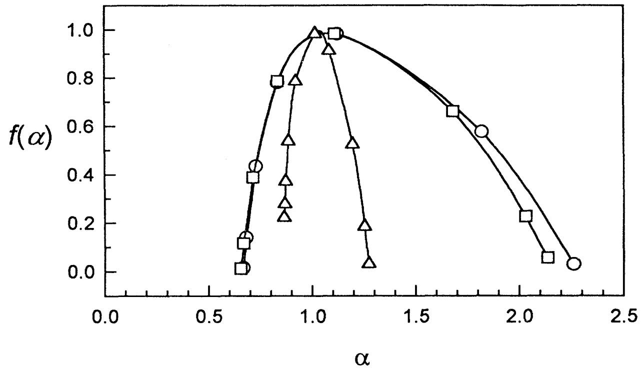


FIG. 2. Multifractal spectra of hexokinase using the wraparound algorithm: forward sequence (○) and reversed sequence (□). The multifractal spectrum for data derived from a random number sequence having a length identical to hexokinase (374 residues) is also shown (△). Note the narrowness of the random number spectrum as compared to that of the protein.

III. RESULTS

The multifractal spectra of 18 proteins were obtained using the algorithm in Sec. II. The proteins analyzed belong to five standard classes [25]: α [Fig. 3(A)], β [Fig. 3(B)], α, β alternate [Fig. 4(A)], α, β segregate [Fig. 4(B)],

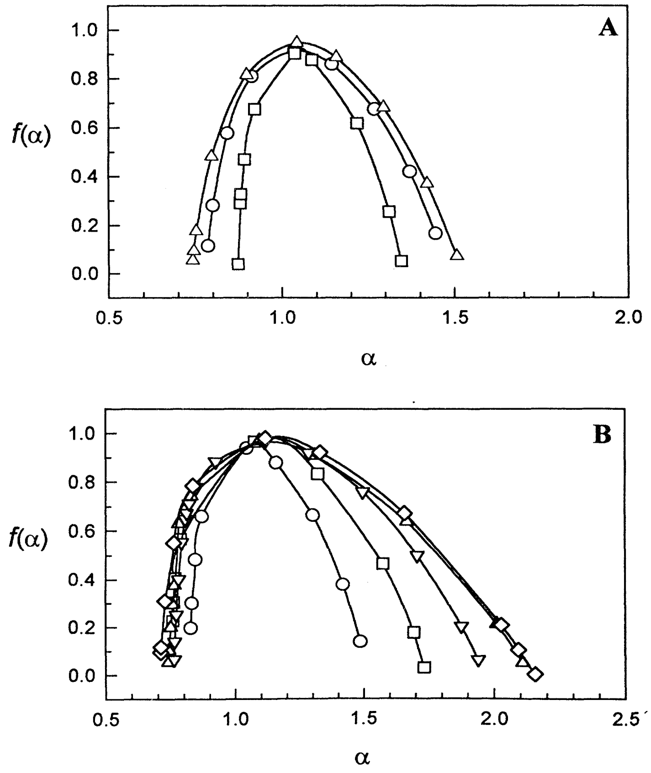


FIG. 3. (A) Multifractal spectra of solvent accessibilities for the α class: ○, cytochrome C; □, Ca-binding parvalbumin; △, myoglobin. (B) β class: ○, plastocyanin; □, α-lytic protease; △, concanavalin A; ▽, elastase; ◇, acid proteinase.

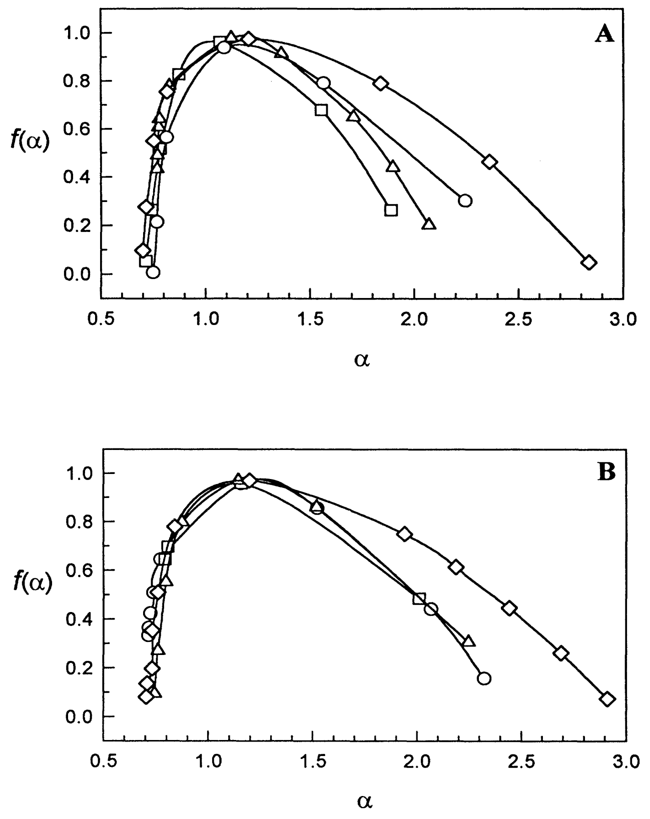


FIG. 4. (A) Multifractal spectra of solvent accessibilities for the class α, β alternate: ○, flavodoxin; □, adenylate kinase; △, arabinose-binding protein; ◇, carboxypeptidase *A*. (B) α, β segregate: ○, papain *D*; □, actinidin; △, carbonic anhydrase *B*; ◇, thermolysin.

and FeS (spectra not shown) proteins. In all cases, a simple convex spectrum is obtained that can be characterized by its intercepts $\alpha_{\min}$ and $\alpha_{\max}$. Note that all spectra have a maximum at $f=1$, reflecting the dimensionality of the support. The values of $\alpha_{\min}$ and $\alpha_{\max}$ are given in Table I for all the proteins investigated. The region to the left of the maximum is determined by the positive moments of the distribution and is dominated by high probability sequences of accessibility values. The region to the right of the maximum is due to the negative moments and, consequently, is dominated by low probability sequences. As a comparison, the spectrum for a random sequence was generated using a sequence of random numbers. A typical comparison of the protein spectrum and a random number spectrum is shown in Fig. 2. It was found in all cases that random number sequences had much narrower spectra than the protein fractional accessibility sequences of corresponding length. Thus, the breadth of the proteins multifractal spectrum can be attributed to nonrandom effects within the data sequence.

After considering the comparison to random sequences, the multifractal behavior of sequences generated from improperly folded proteins is treated. An important problem in biochemistry is the development of algorithms to predict protein structure from sequence information (cf. [26,27]). As competing algorithms are developed, it is important to be able to assess their relative performance on protein sequences whose structures are not known. The multifractal approach potentially provides a diagnostic tool for assessing the performance of such algorithms. Computer simulations were performed that generated improperly folded proteins. The proteins were misfolded in the manner of Novotny, Bruccoleri, and Karplus [28]. Two proteins having a com-

TABLE I. Summary of the results of the multifractal analysis. Note that $\sigma_p(01)$ and $\sigma_p(10)$ can be obtained from the conservation equations, Eq. (11).

Class	Protein	Length	$\alpha_{\min}$	$\alpha_{\max}$	$\sigma_p(00)$	$\sigma_p(11)$
FeS	Ferrodoxin	54	0.88	1.12	0.543	0.460
	Ferrodoxin	98	0.76	1.53	0.590	0.346
α	Cytochrome *C*	103	0.79	1.47	0.578	0.361
	Ca-binding Parvalbumin	107	0.87	1.36	0.547	0.390
	Myoglobin	153	0.73	1.52	0.603	0.349
β	Plastocyanin	99	0.82	1.51	0.566	0.351
	α-lytic protease	198	0.75	1.73	0.595	0.301
	Convanavalin *A*	237	0.73	2.12	0.603	0.230
	Acid proteinase	330	0.71	2.13	0.611	0.228
α,β alternate	Flavodoxin	138	0.75	2.49	0.595	0.178
	Adenylate kinase	194	0.71	2.04	0.611	0.243
	Carboxypeptidase *A*	307	0.7	2.88	0.616	0.136
α,β segregate	Papain *D*	212	0.7	2.46	0.616	0.182
	Actinidin	218	0.7	2.4	0.616	0.189
	Carbonic anhydrase *B*	261	0.75	2.53	0.595	0.173
	Thermolysin	316	0.7	2.93	0.616	0.131

TABLE II. Comparison of multifractal parameters for native and misfolded proteins. The last column lists $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$.

Protein	$\alpha_{\min}$	$\alpha_{\max}$	$\Delta\alpha$
Elastase	0.76	1.96	1.20
Misfold (concanavalin *A* frame)	0.78	1.71	0.93
Concanavalin *A*	0.74	2.14	1.40
Misfold (elastase frame)	0.76	1.72	0.96
Carboxypeptidase *A*	0.7	2.88	2.2
Misfold (thermolysin frame)	0.7	2.06	1.4
Adenylate kinase	0.71	2.04	1.33
Misfold (α-lytic protease frame)	0.71	1.76	1.05
Flavodoxin	0.75	2.49	1.74
Misfold (myoglobin frame)	0.75	1.70	0.95
Carboxypeptidase *A*	0.7	2.88	2.2
Misfold (arabinose-binding protein frame)	0.74	1.96	1.22
Arabinose-binding protein	0.77	2.21	1.44
Misfold (carboxypeptidase *A* frame)	0.70	2.26	1.56
Ferredoxin	0.76	1.53	0.77
Misfold (plastocyanin frame)	0.83	1.40	0.57
Plastocyanin	0.82	1.51	0.69
Misfold (ferredoxin frame)	0.81	1.42	0.61

mensurate number of amino acid residues but dissimilar structures were selected. The amino acid side chains of one protein were then placed in sequence on the backbone of the other, and vice versa. The incorrectly folded structures were then energy minimized using the geometry optimization algorithm of the computer program HYPERCHEM (Hypercube, Inc.), which implements an AMBER (Assisted Model Building and Energy Refinement) force field. The following protein pairs were used: plastocyanin (99 a.a.) (a.a. denotes amino acid) and ferredoxin (98 a.a.), carboxypeptidase *A* (307 a.a.) and arabinose-binding protein (306 a.a.), and elastase (240 a.a.) and concanavalin *A* (237 a.a.). When the lengths were incommensurate the longer sequence was truncated. The multifractal parameters $\alpha_{\min}$ and $\alpha_{\max}$, for these pairs are shown in Table II. The following proteins were also misfolded: papain *D* (212 a.a.), adenylate kinase (194 a.a.), flavodoxin (138 a.a.), and again carboxypeptidase *A*. These sequences were placed on larger frames and then truncated. The frames used belong to the following proteins: actinidin (218 a.a.), α-lytic protease (198 a.a.), myoglobin (153 a.a.), and thermolysin (316 a.a.), respectively.

The notable feature of the multifractal spectra of the misfolded structures when compared to the spectra of the properly folded proteins is that the misfolded spectra are narrower. This is well illustrated in Fig. 5, where the spectra of the elastase-concanavalin *A* swap-pair is exhibited. The reduced width ($\Delta\alpha$) for misfolded proteins is also seen in Table II. Visualization of the improperly folded proteins via computer graphics reveals that the structures are not closely packed, containing many open regions. These open spaces destroy the alternating regions of high and low solvent accessibilities of the sequence found in properly folded proteins. They introduce more randomness and less correlation into the sequence data. This effect is discussed further in Sec. V.

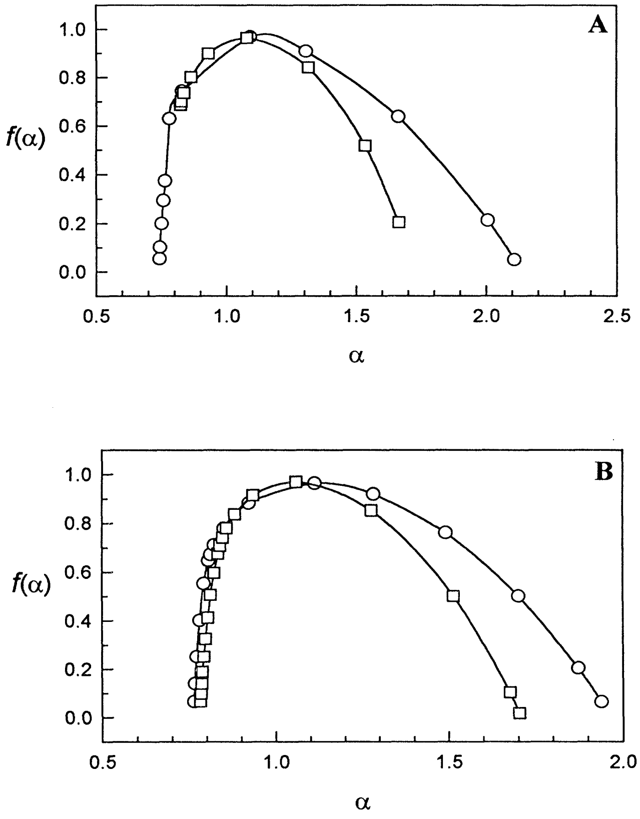


FIG. 5. Multifractal spectra of solvent accessibilities of improperly folded proteins and the corresponding properly folded structure. (A) ○, concanavalin *A*; □, misfolded concanavalin *A* on an elastase frame. (B) ○, elastase; □, misfolded structure elastase on a concanavalin *A* frame.

The narrowing of the spectrum is an observation that exists for all the misfolded proteins except one, arabinose-binding protein. In this case, the protein has a large binding cleft [25] and is thus a rather open structure. Improperly folding the arabinose-binding protein sequence on the carboxypeptidase *A* frame produced a well-packed structure. Therefore, the multifractal spectrum of arabinose-binding protein is narrower than that of the misfolded structure.

IV. RANDOM MULTIPLICATIVE MODEL

Although the multifractal spectra in Figs. 3 and 4 have a range of widths, they all can be fit by a single model. This model involves a binary random multiplicative process with one-step memory. In this section, the procedure for extracting a random multiplicative process from the multifractal spectrum is presented. The development of Chhabra, Jensen, and Sreenivasen [29] is followed. Figure 6 illustrates a binary random multiplicative process. It is a simple model in which units in the sequence can exist in one of two states. For heuristic purposes, the context of a protein chain is considered. Proceeding down the length of the chain, each amino

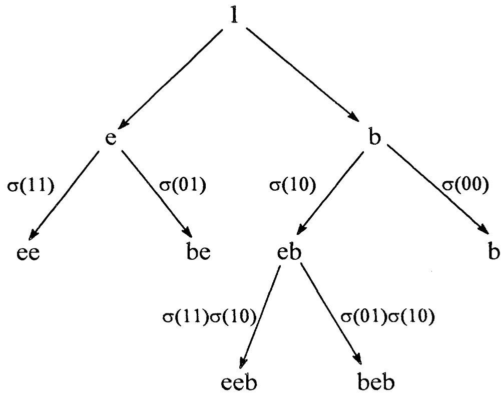


FIG. 6. Schematic showing a portion of a binary tree that corresponds to a random multiplicative process. Sequences of exposed (e) or buried (b) units are generated by a random process. Probabilities of a given step are determined by the preceding unit. The Feigenbaum scaling functions $\sigma_p(ij)$ are associated with the indicated steps in the tree. See text for details.

acid is assumed to be in one of two states, exposed (e) or buried (b). The configuration is part of a binary tree containing all possible combinations of e and b sequences. If the probability of an e or b is determined by the preceding unit, then the model has one-step memory. These binary trees are hierarchical structures, as the early branches dictate the region that the sequence will end up in.

The multifractality of such binary trees has been extensively investigated. In general, such trees give multifractal spectra that can be characterized by three independent parameters, $\alpha_{\min}$, $\alpha_{\max}$, and $f_{\max}$. These parameters can be derived from the probabilities associated with each branch of the tree (cf. [29]). The connection between these probabilities and the multifractal spectrum is briefly outlined here, where a model known as the 2X2 P model is adopted. The addition of each unit on the tree has a probability associated with it. This probability is incorporated into a parameter known as the Feigenbaum scaling factor. For a binary model with one-step memory, there are four such scaling factors (see Fig. 6): $\sigma_p(00)$, $\sigma_p(10)$, $\sigma_p(11)$, and $\sigma_p(01)$. These scaling factors are defined as

$$\sigma_p(ij)=\frac{P(ij)}{P(j)}\ , \tag{9}$$

where $P(ij)$ is the probability of an i unit following a j unit and $P(j)$ is the probability of a j unit existing. The product of the scaling factors associated with each branch gives the probability for a specific configuration or sequence. Probability must be conserved at every splitting of a branch, so two conservation equations exist. They are

$$\sigma_p(00)+\sigma_p(10)=1\ , \tag{10a}$$

$$\sigma_p(01)+\sigma_p(11)=1\ . \tag{10b}$$

The problem of calculating the probability distribution of binary trees can be represented as an eigenvalue problem [30]. Binary trees with one-step memory are conveniently handled using a transfer matrix method [29]. The transfer matrix $\mathbf{T}$ is defined as

$$\mathbf{T}=\begin{bmatrix}\sigma_p^q(00) & \sigma_p^q(01)\\ \sigma_p^q(10) & \sigma_p^q(11)\end{bmatrix}\ . \tag{11}$$

The eigenvalues of this matrix, $\lambda(q)$, lead directly to the mass exponent

$$\tau(q)=\frac{-\ln\lambda(q)}{\ln 2}\ , \tag{12}$$

where the factor of 2 in Eq. (12) is a result of the binary process. Once Eq. (12) is solved, Eqs. (6) and (8) lead directly to the multifractal spectrum. The eigenvalues are determined explicitly from the quadratic equation

$$\lambda^2(q)-\lambda(q)\mathrm{Tr}[\mathbf{T}]+\det[\mathbf{T}]=0\ . \tag{13}$$

This gives

$$\lambda(q)=\frac{\sigma_p^q(00)+\sigma_p^q(11)}{2}+\left\{\frac{[\sigma_p^q(00)+\sigma_p^q(11)]^2}{2}+\sigma_p^q(01)\sigma_p^q(10)\right\}^{1/2}\ . \tag{14}$$

Using the above results, the scaling functions can be extracted from the multifractal spectrum. At the extrema of the spectra ($f=0$), one can assign

$$\alpha_{\min}=\frac{\ln[\sigma_p(00)]}{-\ln(2)}\ , \tag{15a}$$

$$\alpha_{\max}=\frac{\ln[\sigma_p(11)]}{-\ln(2)}\ . \tag{15b}$$

The conservation equations [Eq. (10)] are then used to determine the other parameters. The third independent parameter $f_{\max}$ is fixed at unity as a result of having a binary process on a linear sequence. In Table I, the values determined for the two independent scaling factors are given for all the proteins studied.

This procedure has fixed three points (the two extrema and the maximum) on the simple, convex curves in Figs. 3 and 4. Because the curves vary smoothly, the intervening regions are accurately fit. Thus, this simple empirical model can accurately fit all the protein data. This, of course, does not prove that the protein data result from a binary multiplicative process. However, it shows that the underlying phenomenon does not require a complicated model to describe its multifractal behavior. To test the validity of the multiplicative model, one must provide an interpretation of the scaling parameters from a physical model. For instance, it was previously shown that the Bragg-Zimm model of an alpha helix can be mapped into a binary multiplicative model with one-step memory [19]. Using these results, one obtains physically unrealistic parameters for the cooperativity parameter of the alpha helix. Not too surprisingly, the helix-coil model of an alpha helix is seen not to be applicable to proteins. In the next section, alternate models are discussed that may lead to a clearer physical and less empirical picture of the observed multifractal spectrum.

V. DISCUSSION

In this work, it was shown that a protein structural parameter, the solvent accessibility of amino acid side chains, is distributed in a multifractal fashion along the length of polymer. It was also seen that this distribution is consistent with that generated by a random multiplicative process with one-step memory. To understand the implications of the observed multifractality, one must address the question of the very nature of multifractals. Indeed, this has been the topic of recent research (cf. [31]) and there is, at this time, no single answer to this question. However, two phenomena are known to give rise to multifractal behavior. They are random multiplicative processes and processes that involve the convolution of two probability functions. As will be argued, both of these effects may be occurring in proteins.

The multifractal formalism can be used to characterize certain distributions of random variables. To distinguish between multifractal behavior and other, more common statistical behavior, one must focus on the breadth of the distribution function. For instance, in many random additive processes, the central limit theorem can be applied and a Gaussian distribution is obtained. If one observes the probability function $\langle P \rangle$ that is an average over the independent random variables of the system, it is seen to have a narrow distribution whose moments obey $\langle P^q \rangle = a_q \langle P \rangle^q$, where a_q is some nonsingular function in q. Multifractal distributions on the other hand cannot be characterized by such simple moment relationships. This signifies a loss of scaling for the system, that is, the system no longer has a characteristic length. This results in an extremely broad distribution. For multifractals the moments are such that $\langle P^q \rangle \sim \langle P \rangle^{\tau(q)}$, where $\tau(q)$ is the function defined in Eqs. (3) and (4).

Multifractal moment distributions arise in problems involving random multiplicative processes. For such processes an averaged parameter $\langle A \rangle$ is characterized by the product given by

$$\langle A \rangle = \prod_i A_i p_i \ , \tag{16}$$

where the product is over the i random variables of the systems. There are two major differences between the behavior of multiplicative and additive random processes. For multiplicative processes, a rare event (small p_i) can dominate the distribution, while for additive processes rare events have very little impact. Also, in multiplicative processes short range correlations can have a strong impact on the product in Eq. (16). Additive processes are insensitive to short range correlations because correlated pairs are often distributed as a single random variable.

Multifractal behavior can also arise in cases where a distribution results from a convolution of two distributions. This situation arises in problems of random walks on random structures [32,33]. If one has two relatively narrow distributions, say $P(r,t)$ and $\Phi(l,r)$, the convolution can result in a broad logarithmic distribution. In the random walk problem, t is time, r is a Euclidean distance, and l is a "chemical" or polymer contour distance. The multifractal spectrum is generated by the convolution integral

$$\langle P^q \rangle = \int_0^\infty \Phi(l,r) P^q(r,t) dr = \langle P \rangle^{\tau(q)} \ . \tag{17}$$

Such convolution problems present a direct analytic theory of multifractal behavior.

Both of the above forms of multifractality may be occurring in proteins. As discussed in Sec. IV, it has previously been shown that the Bragg-Zimm model of the alpha helix exhibits multifractal behavior [18,19]. Such order-disorder models are adaptations of the one-dimensional Ising model and consider only nearest neighbor interactions. The partition function for such a model is obtained by summing all possible sequences. In most helix-coil models, these sequences can be generated by a simple binary process. Because only nearest neighbors are considered when adding a unit to the sequence, the process is binary with one-step memory. This model can be mapped into the 2X2 P model used in Sec. IV to analyze the experimental multifractal spectrum.

Such simple helix-coil models are clearly inadequate for describing protein structure and thermodynamics. Nevertheless, they can serve as a starting point for other statistical models. Recent work has made the analogy between proteins and spin glasses [34–36]. Helix-coil models of the alpha helix are a form of the Ising spin model. Helix-coil models can be used to predict secondary structure by introducing mathematical devices that introduce long range and disordering effects. Similar problems arise in the field of spin glasses. These effects are modeled in a spin Hamiltonian by introducing spin-spin coupling factors that are treated as random variables. Wolynes and co-workers have exploited these analogies to model protein structure and folding [34–36]. Such statistical models of proteins employ the replica technique developed by Edwards (cf. [37]). This method results in a convolution of a probability function with the partition function. Thus, it will have the features of Eq. (17). Therefore, if one utilizes the spin glass analogy, multifractality in proteins can arise from two sources. First, the Ising problem, even in one dimension, has an intrinsic fractal nature [38,39]. This can be put in the context of a random multiplicative process and the multifractal behavior is readily demonstrated (see [18,19] for biopolymer applications). In addition to this effect, the averaging of coupling constants introduces a convolution of probability functions, a second potential source of multifractal behavior. The influence of this convolution effect on the multifractal behavior of spin glass models is currently under investigation.

Finally, we turn to the intriguing question of why misfolded structures show narrower multifractal spectra than properly folded structures. Again, considering the folding to represent a random multiplicative process, rare events will have a profound influence on the distribution. A properly folded protein will have rare (or nonrandom) combinations of exposed residues. Thus, it will have unusual sequences of solvent accessibilities. A misfolded protein, on the other hand, appears closer to a random sequence. In plots of $f(\alpha)$ versus α, the region to the left of the maximum corresponds to positive moments ($q>0$). This region is dominated by common events or sequences. The region to the right of the maximum is influenced by negative moments ($q<0$). It is dominated by rare events or sequences. The misfolded spectrum shows small changes in the left-hand region while large changes are observed in the right-hand region with the misfold appearing closer to a random sequence than the native protein. Thus, the multifractal spectrum shows great sensitivity to correlated packing of side chains as reflected in this right-hand region.

The multifractal behavior of the solvent accessibilities potentially provides a useful tool for addressing a number of problems in protein structure. First, it can provide a diagnostic test for comparing protein folding algorithms for sequences whose structures are unknown. It is anticipated that the best algorithm will generate the widest multifractal spectrum. This approach can also be used to compare two structures solved for the same x-ray diffraction data or for comparing NMR structures to x-ray ones. The correct structure would be expected to have a broader spectrum than the incorrect one. Finally, it may also be used in assessing whether a given sequence is likely to fold into a native structure. Since the hydrophobicity of a sidegroup is strongly correlated with its solvent accessibility, one would anticipate hydrophobicity sequence data to behave in a similar fashion as the accessibility data. If a multifractal spectrum from the hydrophobicity sequence data appears as a random spectrum does, it is unlikely that such a sequence would fold into a well-defined structure. Such issue are currently under investigation.

ACKNOWLEDGMENTS

We acknowledge the assistance of Dr. Bruce Bowler for help with the ACCESS program and for useful discussions. This work was supported in part by Grant No. 1R15GM51019 from the NIH.

[1] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Academic, New York, 1983).

[2] J. Feder, *Fractals* (Plenum, New York, 1988), pp. 66–103.

[3] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).

[4] P. Pfiefer, U. Welz, and H. Wipperman, Chem. Phys. Lett. **113**, 535 (1985).

[5] B. A. Fedorov, B. B. Fedorov, and P. W. Schmidt, J. Chem. Phys. **99**, 4076 (1993).

[6] M. Lewis and D. C. Rees, Science **230**, 1163 (1985).

[7] F. M. Richards, Ann. Rev. Biophys. Bioeng. **6**, 151 (1977).

[8] T. G. Dewey, Proc. Natl. Acad. Sci. USA **91**, 12 101 (1994).

[9] T. G. Dewey, in *Dynamics in Small Confining System II*, edited by J. M. Drake, J. Klafter, R. Kopelman, and D. D. Awschalon, MRS Symposia Proceedings No. 366 (Materials Research Society, Pittsburgh, in press).

[10] H. E. Stanley and P. Meakin, Nature **335**, 405 (1988).

[11] T. Tél, Z. Naturforsch. Teil A **43**, 1154 (1988).

[12] L. Pietronero, C. Evertsz, and A. P. Siebesma, in *Stochastic Processes in Physics and Engineering*, edited by S. Albeverio *et al.* (Reidel, Dordrecht, 1988), pp. 253–278.

[13] T. A. Witten, Jr. and L. M. Sander, Phys. Rev. Lett. **47**, 1400 (1981).

[14] L. de Arcangelis, S. Redner, and A. Coniglio, Phys. Rev. B **31**, 4725 (1985).

[15] I. Procaccia, J. Stat. Phys. **36**, 649 (1984).

[16] K. G. Wilson, Sci. Am. **241**, 158 (1979).

[17] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).

[18] T. G. Dewey, in *Fractals in the Natural and Applied Sciences, IFIP Transactions*, edited by M. M. Novak (North-Holland, Amsterdam, 1994), Vol. A-41, pp. 89–100.

[19] T. G. Dewey, Fractals (to be published).

[20] T. G. Dewey, Fractals **1**, 179 (1993).

[21] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).

[22] L. V. Meisel, M. Johnson, and P. J. Cote, Phys. Rev. A **45**, 6989 (1992).

[23] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).

[24] B. Lee and F. M. Richards, J. Mol. Biol. **55**, 379 (1971).

[25] J. S. Richardson, Adv. Protein Chem. **34**, 167 (1981).

[26] J. S. Fetrow, and S. H. Bryant, Biotechnol. **11**, 479 (1993).

[27] T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, and J. M. Thorton, Nature **326**, 347 (1987).

[28] J. Novotný, R. Bruccoleri, and M. Karplus, J. Mol. Biol. **177**, 787 (1984).

[29] A. B. Chhabra, R. V. Jensen, and K. R. Sreenivasan, Phys. Rev. A **40**, 4593 (1989).

[30] M. J. Feigenbaum, I. Procaccia, and T. Tel, Phys. Rev. A **39**, 5359 (1989).

[31] H. E. Stanley, in *Fractals and Disordered Systems*, edited by A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1991), pp. 1–49.

[32] S. Havlin and A. Bunde, Phys. D **38**, 184 (1989).

[33] A. Bunde, S. Havlin, and H. E. Roman, Phys. Rev. A **42**, 6274-6277 (1990).

[34] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. **84**, 7524 (1987).

[35] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).

[36] P. G. Wolynes, in *Spin Glasses and Biology*, edited by D. L. Stein (World Scientific, Singapore, 1992), pp. 225–259.

[37] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).

[38] P. Bak and R. Bruinsma, Phys. Rev. Lett. **49**, 249 (1982).

[39] R. Bruinsma and P. Bak, Phys. Rev. B **27**, 5824 (1983).