Multifractals and decoded walks: Applications to protein sequence correlations

Bonnie J. Strait and T. Gregory Dewey*
Department of Chemistry, University of Denver, Denver, Colorado 80208

(Received 8 June 1995)

Correlations occur in protein and nucleic acid sequences that have important implications for the evolution and stability of these macromolecules. A number of fractal analyses of sequence data have been developed that decode sequence information into a random walk. Alternatively, a generalized box-counting analysis of decoded sequences can be used to establish multifractal properties. In this work, the connection between these two seemingly disparate approaches is established. This connection is exploited to investigate correlations in protein sequences. A comparison is made between the hydrophilicity profile, a composition parameter, and the solvent accessibility, a parameter reflecting the final folded structure. It is seen that the hydrophilicity of proteins sequences are distributed as a multiplicative binomial process. The solvent accessibility, on the other hand, is a more complicated binary process with one-step memory. The evolutionary benefits of this latter process are seen by calculating the redundancy as determined from the information dimension.

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

There has been considerable recent interest in the statistical properties of nucleic acid and protein sequences [cf. 1,2]. Information on correlations within these sequences has a direct bearing on evolutionary mechanisms and on the thermodynamic stability of biomacromolecules. In addition to conventional statistical approaches [cf. 3,4], a number of fractal methods have been developed to investigate sequence correlations [5–10]. In one of these methods, encoded sequence information is mapped into a random walk problem. The sequence may be decoded according to composition, as is done in the DNA walk problems [5–7] or according to a specific chemical or physical property of the monomeric unit, as done in the "bridge analysis" of protein sequences [8]. Drift can occur in the decoded walk as a result of the overall compositional bias of the sequence. The "bridge analysis" is an algorithmic device that compensates for this and allows correlations to be observed. Deviations of the decoded walk from random behavior provides evidence for long-range correlations.

In a seemingly very different approach, the decoded data sequence can be analyzed with a generalized box-counting procedure to obtain a multifractal spectra [10]. In box-counting algorithms, a curve is covered with boxes of a fixed size and the moment distribution of the density is determined. The scaling of the moments of the distribution with box size provides an infinite number of fractal dimensions. This infinity of dimensions gives a multifractal spectrum. The breadth of the multifractal spectrum provides evidence of an underlying hierarchical structure. The goal of this work is to show the correspondence between these two methods. Additionally, correlations in the hydrophilicity of a number of protein sequences are examined and compared with those for a protein structural property, the solvent accessibility. It is seen that the hydrophilic residues in 16 different proteins are distributed as a multiplicative binomial process. The solvent accessibility that is dictated by the final folded structure shows a more complicated correlation and can be modeled as a 2×2 P process with one-step memory [11]. The implications of these differences are discussed.

In the "bridge analysis," a protein sequence is decoded by a numerical correspondence between each amino acid and a physical property associated with it. This correspondence provides a decoded sequence, $\{\xi_1,\xi_2,\ldots,\xi_L\}$, where ξ_i is a numerical value associated with the amino acid in the ith position along the sequence and L is the length of the protein sequence. Often ξ_i takes on values of ±1 depending on the chemical composition of the unit [5,8]. In a previous application, the hydrophilic, Coulombic, and hydrogen bonding properties of amino acid sequences were separately decoded [8]. A trajectory can be mapped using the decoded sequence, and for a one-dimensional mapping this is given by

$$x(l)=\sum_{i=1}^{l}\xi_i \ . \tag{1}$$

Walks defined in this fashion will show strong drift as a result of composition. This effect can obscure correlations, and attempts have been made to compensate for this [cf. 12]. Here we consider a drift correction known as the "bridge analysis" [8]. In this analysis the reduced trajectory is considered:

$$y(l)=x(l)-(l/L)x(L) \ . \tag{2}$$

This trajectory $y(l)$ will start and return to the origin and form a "Brownian bridge."

Because trajectories of individual proteins are noisy, it is more practical to consider ensemble averages. An en-

*Author to whom all correspondence should be addressed. FAX: (303)-871-2254. Electronic address: gdewey@cair.du.edu semble averaged squared displacement $\langle z^2(l) \rangle$ is defined as

$$\langle z^2(l) \rangle = \left\langle \frac{y^2(l)}{L(\xi - \xi^{-})^2} \right\rangle , \tag{3}$$

where the brackets represent averages over many proteins and the bars represent an average within a protein sequence. For example, $\bar{\xi} = (1/L)\sum_{i=1}^{L} \xi_i$. The term $L(\xi - \bar{\xi})^2$ eliminates the L dependence and corrects for different lengths and variances between proteins. The mean squared trajectory follows a scaling law: $\langle z^2(l) \rangle \sim l^{2\alpha_w}$, where the exponent α_w will equal $\frac{1}{2}$ for a random walk. When α_w is greater than $\frac{1}{2}$, the walk demonstrates persistence and α_w less than $\frac{1}{2}$ indicate antipersistence. Correlations in protein sequences were seen with α_w being 0.520 for walks based on hydrophilic and on hydrogen bonding, and an α_w of 0.470 for walks based on static charge distributions.

A very different analysis has been used to demonstrate correlations in amino acid solvent accessibilities in individual protein structures [10]. In this analysis, one starts with a decoded sequence as before. A function $Z_q(l)$ is defined to examine the qth moments of the sequence:

$$Z_q(l) = \sum_{j=1}^{L/l} x_j^q(l) , \tag{4}$$

where j labels individual sequences of length l within the complete protein sequence. There will be a total of L/l of these sequences for a given protein of length L. Using a scaling ansatz, $Z_q(l) \sim l^{-\tau(q)}$, where $\tau(q)$ is a generalized exponent and is related to a generalized fractal dimension, D_q by $\tau(q) = (q-1)D_q$. Using the box-counting method described previously [10], $\tau(q)$ can be calculated for an individual decoded protein sequence. Frequently, the Legendre transformation properties of $\tau(q)$ are employed to represent the multifractal spectrum. These properties define two functions, $f(q)$ and $\alpha(q)$, that are related to $\tau(q)$ by $\tau(q) = f(q) - q\alpha(q)$. Multifractal spectra, $f(q)$ versus $\alpha(q)$, for the protein concanavalin A are shown in Fig. 1. The spectrum for the decoded solvent accessibilities is seen to be much broader than for the hydrophilicity. Both the these spectra are broader than one obtained from a random sequence of numbers of the same length. These results show that both the solvent accessibility data and the hydrophilicity show nonrandom correlations. However, these two parameters are correlated in a different manner.

The difference between the Brownian bridge and the multifractal approach is that multifractals are concerned with the moment distribution of many short trajectories that make up a protein sequence. The Brownian bridge focuses on the root mean displacement of a single protein trajectory. Because such a trajectory is too short to generate good statistics, an average over many proteins must be considered. However, given a protein of a long enough sequence there is no reason to anticipate a scaling law that would differ from those for the ensemble average. In the multifractal approach, the sum of trajectories generated from a given sequence can be used as our ensemble. If protein sequences are ergodic in the information theory sense, then a correspondence between the two approaches can be made. Because of the finite length of a protein sequence, the number of trajectories decreases with trajectory size as L/l and the multifractal sum must be normalized accordingly. The scaling of $\langle y^2 \rangle$ is related to the second moment of $Z_q(l)$ by $\langle y^2 \rangle \sim Z_2(l)/(L/l)^2 \sim l^{2-\tau(2)}$, where a fractal dimension of 1 for the support, i.e., the linear sequence, is implicitly assumed. This provides a relationship between the walk exponent α_w and one of the multifractal exponents:

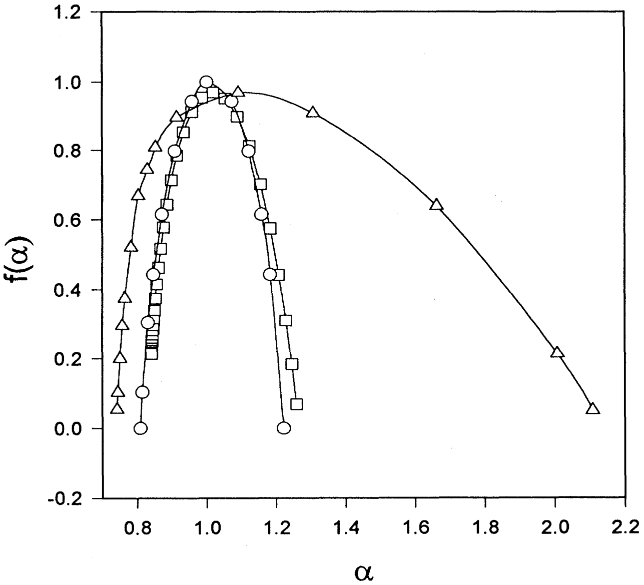

FIG. 1. Multifractal spectra of concanavalin A: □, spectrum determined from the hydrophilicity profile; ○, theoretical curve for a multiplicative binomial process used to fit hydrophilicity curve; △, spectrum determined from the solvent accessibility profile.

$$\alpha_w = 1 - \frac{\tau(2)}{2} . \tag{5}$$

Thus, persistence is seen for $\tau(2) < 1$ and antipersistence occurs when $\tau(2) > 1$. Multifractal spectra (f versus α) obtained to date from decoded protein sequence can be accurately fit using three parameters, $f_{\max}$, $\alpha_{\min}$, $\alpha_{\max}$. For linear sequence problems, $f_{\max}$ is fixed at unity. Consequently, the multifractal approach provides one more parameter than the Brownian bridge. Of course, Brownian bridges could be constructed from higher order mean displacements to generate different scaling exponents. These exponents can then be related to the multifractal spectra by equations similar to Eq. (5).

The multifractal spectrum of the hydrophilicity profile of 16 different proteins were obtained. The hydrophilicity is an empirical index that describes the affinity for water of an amino acid sidechain [13]. It takes on dimensionless values from -5 to 5 and has commonly been used in protein structure prediction programs. Figure 2 shows the hydrophilicity profile (or decoded sequence) for

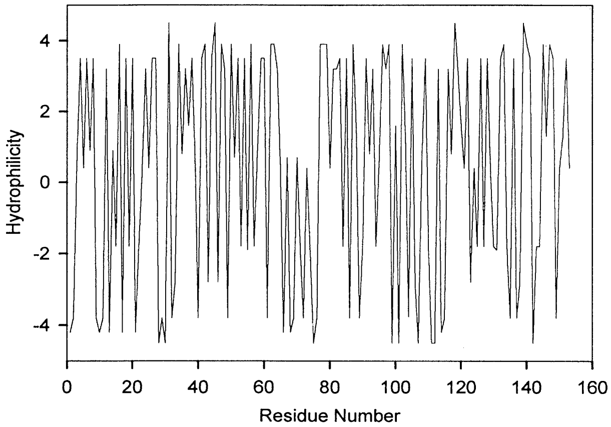

FIG. 2. Hydrophilicity profile of the protein myoglobin (153 amino acid residues). The empirical hydrophilicity index for each amino acid residue is plotted versus the position along the protein sequence.

the protein myoglobin. Figure 1 shows the multifractal spectrum determined from such profiles. For the 16 proteins investigated, the multifractal spectra could be accurately fit (see Fig. 1) using a multiplicative binomial process. This model provides a relationship for $\tau(q)$ (cf. [14]),

$$p^q+(1-p)^q=2^{-\tau(q)} , \tag{6}$$

where p is the probability of finding a hydrophilic residue and the factor of 2 results from treating the problem as a binary process. Using Eq. (6) a multifractal spectrum can be determined with the following relationships for $\alpha_{\min}$ and $\alpha_{\max}$.

$$\alpha_{\min}=-\frac{\ln(1-p)}{\ln 2} , \tag{7a}$$

$$\alpha_{\max}=-\frac{\ln p}{\ln 2} . \tag{7b}$$

The value of p is obtained using the $\alpha_{\min}$ and $\alpha_{\max}$ values determined from the intercepts of the multifractal spectrum. If the process is truly a random multiplicative one, the values of p determined separately from Eqs. (7a) and (7b) should agree. Table I shows the experimental values of $\alpha_{\min}$ and $\alpha_{\max}$ and the corresponding $p_{\min}$ and $p_{\max}$. As can be seen, these values are very close, although often $p_{\min}$ has slightly higher values then $p_{\max}$. From a comparison of data analyzed in forward and reverse sequences, it is estimated that there is a 10–20 % error in $\alpha_{\max}$ and a significantly lower error in $\alpha_{\min}$. Given these considerations, the multifractal spectrum is well represented by a binomial multiplicative model. Figure 1 shows a representative example of such a fit, where p_{ave} is used to generate the curve for the model.

Using Eqs. (5) and (6), the walk exponent α_w for individual proteins is determined, and these are shown in Table I. As is seen in the table, these values center around 0.52. The average is 0.517±0.005 for all 16 proteins. This is in excellent agreement with the value of 0.520±0.005, determined previously from the bridge analysis of hydrophilicity of an ensemble of proteins [8]. Thus, the correspondence between the average of walks within a protein to ensemble average walks is justified in this instance. The ergodicity of protein sequences have important implications for information theory approaches to molecular evolution. It is also interesting to note that all multiplicative binomial processes will show persistence, as $\tau(q)$ can only vary from 0 to 1 in these cases.

These results differ substantially from those obtained when profiles of the solvent accessibility are analyzed. Using a "ball rolling" algorithm, 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. The sequence of solvent accessibilities is a reflection of the geometry of the final, folded state of the protein. Thus, it is a very different parameter from hydrophilicity that is based on composition. The multifractal analysis of this data has been presented previously [10], and these results are summarized in Table I. For this decoded parameter, the multifractal spectra are considerably wider than the hydrophilicity spectra and cannot be accurately represented by a multiplicative binomial process. To accurately fit the experimental spectra, a 2×2 P model [11] with one-step memory was used. This model employs Feigenbaum scaling factors $\sigma_p(00)$, $\sigma_p(10)$, $\sigma_p(11)$, $\sigma_p(01)$ defined as

$$\sigma_p(ij)=\frac{P(ij)}{P(j)} , \tag{8}$$

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 for all the units gives the probability for a specific configuration or sequence. For our present case, 0 could be associated with a hydrophilic reside and 1 is associated with a hydrophobic residue. These scaling factors are governed by two conservation equations,

$$\sigma_p(00)+\sigma_p(10)=1 , \tag{9a}$$

$$\sigma_p(01)+\sigma_p(11)=1 . \tag{9b}$$

The generalized exponent is given by

$$\tau(q)=\frac{-\ln\lambda(q)}{\ln 2} , \tag{10}$$

where

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

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

TABLE I. Multifractal and walk parameters of protein sequences.

	Hydrophilicity				Solvent Accessibility		
Protein (length)	$\alpha_{\min}$	$\alpha_{\max}$	$p_{\min}$ ($p_{\max}$)	α_w	$\alpha_{\min}$	$\alpha_{\max}$	α_w
Ferrodoxin(54)	0.88	1.30	0.46(0.41)	0.51	0.88	1.12	0.51
Ferrodoxin(98)	0.80	1.33	0.43(0.40)	0.52	0.76	1.53	0.53
Cytochrome *C*(103)	0.82	1.28	0.43(0.41)	0.52	0.79	1.47	0.53
Parvalbumin(107)	0.85	1.33	0.45(0.40)	0.52	0.87	1.36	0.52
Myoglobin(153)	0.80	1.24	0.42(0.42)	0.52	0.73	1.52	0.54
Plastocyanin(99)	0.85	1.19	0.45(0.44)	0.51	0.82	1.51	0.53
α-lytic protease(198)	0.86	1.26	0.45(0.42)	0.51	0.75	1.73	0.55
Concanavalin *A*(237)	0.84	1.26	0.44(0.42)	0.51	0.73	2.12	0.57
Acid Proteinase(330)	0.90	1.18	0.46(0.44)	0.51	0.71	2.13	0.57
Flavodoxin(138)	0.81	1.36	0.43(0.39)	0.52	0.75	2.49	0.59
Adenylate kinase(194)	0.79	1.32	0.42(0.40)	0.52	0.71	2.04	0.57
Carboxypeptidase *A*(307)	0.79	1.29	0.42(0.41)	0.52	0.70	2.88	0.60
Papain *D*(212)	0.81	1.35	0.43(0.39)	0.52	0.70	2.46	0.59
Actinidin(218)	0.84	1.31	0.44(0.40)	0.52	0.70	2.40	0.59
Carbonic anhydrase *B*(261)	0.81	1.34	0.43(0.40)	0.52	0.75	2.53	0.59
Thermolysin(316)	0.82	1.34	0.43(0.39)	0.52	0.70	2.93	0.60

$$\alpha_{\min}=\frac{\ln[\sigma_p(00)]}{-\ln(2)}\ , \tag{12a}$$

$$\alpha_{\max}=\frac{\ln[\sigma_p(11)]}{-\ln(2)}\ . \tag{12b}$$

Again, the third independent parameter $f_{\max}$ is fixed at unity as a result of having a binary process on a linear sequence. With Eqs. (9) and (12), the scaling factors are determined from the intercepts of experimental multifractal spectra. Using Eqs. (5), (10), and (11), the walk exponent is determined. For the 16 proteins studied, α_w had an average value of 0.56 ± 0.03. While the walk exponents for the solvent accessibility and the hydrophilicity are very similar, the corresponding multifractal spectra are quite different. This is because the walk dimension is determined from a single point on the spectra. A minimal representation of the hydrophilicity data requires a multiplicative binomial process while the solvent accessibility is not as simple. In this latter case, a binary multiplicative process with one-step memory must be employed.

It is interesting to calculate the Shannon or information entropy. For the binomial process, one has $S=-\sum p_i \ln p_i$, while the one-step memory process gives $S=-\sum_{ij}\sigma_{ij}\ln\sigma_{ij}$. The redundancy R of the sequence is defined as

$$R=1-\frac{S}{S_0}\ , \tag{13}$$

where $S_0=\ln\Omega$, with Ω equal to 2 for the binomial process (singlet code) and 4 for the one-step memory process (doublet code). For the protein concanavalin A the redundancy of the hydrophilicity sequence is 0.014, while the solvent accessibility has a redundancy of 0.216. In general, doublet code gives larger redundancies than a single code [2], yet this case shows much larger differences than those observed in linguistic texts. A similar large difference in redundancy can also be observed by considering the information dimension of the two processes. The Shannon entropy can be defined using the information fractal dimension [14,15] and allows a comparison of different multiplicative processes. The Shannon entropy is given by $S=-\alpha(1)\ln 2$, where $\alpha(1)=-[d\tau(q)/dq]|_{q=1}$. The redundancy associated with the solvent accessibility data of concanavalin A as determined from the 2×2 P model [Eqs. (10) and (11)] is 0.98, while that determined for the hydrophilicity from Eq. (6) is identical to that determined above 0.014. Thus, the solvent accessibility shows a much higher redundancy than it would have if it were generated by a simple multiplicative binomial process. In general, high redundancy of a code makes it more resistant to errors. These results suggests that a structural parameter, the solvent accessibility, is more resistant to error than a composition parameter, the hydrophilicity. There may be an evolutionary advantage to having a high redundancy associated with the folding process.

This work was supported in part by Grant No. 1R15GM51019 from the NIH.

[1] R. F. Doolittle, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequence, Methods in Enzymology* (Academic, New York, 1990), Vol. 183.

[2] M. V. Volkenstein, *Physical Approaches to Molecular Evolution* (Springer-Verlag, Berlin, 1994).

[3] S. Karlin, P. Bucher, V. Brendel, and S. F. Altschul, Annu. Rev. Biophys. Biophys. Chem. **20**, 175 (1991).

[4] S. H. White, Annu. Rev. Biophys. Biophys. Chem. **23**, 407 (1994).

[5] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin,

F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

[6] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2675 (1993).

[7] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[8] V. S. Pande, A.Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. USA **91**, 12 972 (1994).

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

[10] J. S. Balafas and T. G. Dewey, Phys. Rev. E **52**, 880 (1995).

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

[12] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1691 (1994).

[13] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).

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

[15] R. C. Hilborn, *Chaos and Nonlinear Dynamics* (Oxford University Press, New York, 1994), pp. 367–481.