Multifractals, Encoded Walks and the Ergodicity of Protein Sequences

T. Gregory Dewey and Bonnie J. Strait, Department of Chemistry, University of Denver, Denver, CO 80208, (303)-871-3100, FAX (303)-871-2254, email gdewey@cair.du.edu

Abstract

A variety of statistical methods have been developed to explore correlations in protein and nucleic acid sequences. Such correlations have important implications for the evolution and stability of these macromolecules. Recently, a number of fractal analyses of sequence data have been developed. These analyses have considerable appeal as they are extremely sensitive to long range correlations and to hierarchical structures. One such analysis decodes sequence information into a random walk and the statistics of the resulting random walk is investigated. Anomalous scaling of such walks has been interpreted as indicative of a fractal structure. 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. An ensemble consisting of a comprehensive data set of representative protein sequences is analyzed to establish the ergodicity of protein sequences. The implications of this ergodicity for information theoretical approaches to protein structure prediction is explored.

1: Introduction

The statistical analysis of sequence data has generated ongoing interest. There has been a number of studies on the statistical properties of nucleic acid and protein sequence [for reviews see 1, 2] and such information has important implications for both the evolution and the thermodynamic stability of biomacromolecules. In addition to conventional statistical approaches [cf. 3, 4], a number of fractal methods have been recently developed to investigate sequence correlations [5-10]. In one of these methods, sequence information is mapped into a random walk problem. The encoded walk results from assigning a specific numerical value and spatial direction to a property of the members in the sequence. For instance, in DNA problems it is common to give purines a +1 step on a 1-dimensional lattice and pyrimidines a -1 step [5-7]. Similar walks have been studied in protein sequences and have been based on specific chemical or physical property of the monomeric unit [8]. The resulting trajectories of these encoded walks can be analyzed as diffusion problems. Deviations of the encoded walk from random behavior provides evidence for long-range correlations. Significant controversy has surrounded encoded walks for DNA sequences, particularly with regard to correlations in non-coding or "junk" DNA [cf. 11].

In a seemingly very different approach, the encoded data sequence can be analyzed with a generalized box counting procedure to obtain a multifractal spectra [10]. The multifractal formalism was developed in an attempt to generalize fractal concepts to encompass complicated probability densities (for reviews see [12-14]). Density distributions can be characterized through their moment distributions and can give rise to an infinite number 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 [15], percolating clusters [16], energy dissipation in fully developed turbulent flows [17], configuration of Ising spins at critical points [18], and the characterization of strange attractors [19]. Recently the multifractal approach has been extended to the description of helix-coil transitions in biopolymers [20, 21] and to the analysis of the solvent accessibility profile of proteins [10]. It was seen that the multifractal spectrum of the solvent accessibility provides a signature for correlations in properly folded proteins.

In this work we relate scaling laws determined from trajectories of the encoded walk to generalized fractal dimensions determined from the multifractal analysis (Section 2). These results are exploited to examine correlations in the hydrophilicity of a large data base of protein sequences (Section 3). The multifractal approach allows correlations within individual protein sequences to be accurately determined. This provides a significant advantage over the analysis of encoded walks as the trajectory of an individual protein is too "noisy" to obtain good statistics. Consequently, this latter approach has been restricted to the study of "ensembles" of many protein sequences. Using the multifractal approach we examine the ergodicity of protein sequence correlation (Section 3). For a comprehensive, non-redundant data set of protein sequences it is seen that sequence correlations within individual proteins are the same as correlations within the ensemble of sequence. By establishing the ergodicity (in the information theory sense), the information content of protein sequence can be statistically related to the number of probable sequences. From a k-tuplet analysis of the information content (Section 4), it is seen that the number of probable sequences is significantly smaller than the number of possible sequences.

2: Encoded Walks and Multifractals

An encoded walk is generated from a protein sequence using a numerical correspondence between the each amino acid and a physical property associated with it. This correspondence provides a sequence of numbers, $\{\xi_1, \xi_2, \ldots \xi_L\}$, 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 $\underline{+}1$ 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 encoded [8]. A trajectory is generated from encoded sequence. In the case of a one-dimensional mapping, this is given by:

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

Walks defined in this manner will usually show strong drift as a result of the overall composition of the sequence. This effect can obscure correlations and attempts have been made to compensate for it [cf. 22]. 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, (*l=0*), and return, (*l=L*), to the origin, regardless of composition. Typically, the trajectory is symmetric about the midpoint, $l = L/2$. At the midpoint of the trajectory, the walk will have its largest excursion from the origin and this point represents the mean squared displacement. A construct of this kind is known as the "Brownian bridge."

Trajectories of individual proteins have a limited number of data points and, therefore, appear "noisy". For practical analysis of the walk statistics, ensemble averages must be considered. An ensemble averaged squared displacement $\langle z^2(l) \rangle$ is defined as:

$$\langle z^2(l) \rangle = \left\langle y^2(l) \Big/ L\overline{(\xi - \bar{\xi})^2} \right\rangle \tag{3}$$

where the brackets represent averages over a large data set of different proteins and the bars represents an average within a protein sequence. For example, $\bar{\xi} = (1/L)\sum_{i=1}^{L} \xi_i$. The term $L\overline{(\xi - \bar{\xi})^2}$ eliminates the L dependence and corrects for different lengths and variances between proteins. The mean squared trajectory follows a scaling law:

$$\left\langle z^2(l)\right\rangle \sim l^{2\alpha_w} \qquad (4)$$

where the exponent α_w will equal 1/2 for a random walk. When α_w is greater than 1/2, the walk demonstrates persistence and α_w less than 1/2 indicates anti-persistence. Correlations in proteins sequences were seen with α_w being 0.520 for walks based on hydrophilic properties or on hydrogen bonding and an α_w of 0.470 was determined for walks based on static charge distributions [8].

The multifractal approach [10] provides a method for detecting correlations within sequences of individual proteins. It has been used to explore correlations in the solvent accessibility profile and the hydrophilicity profile of a number of proteins [11]. In this analysis, one again starts with an encoding of the sequence as in the walk problem. However, in previous applications ξ_i did not take on values of $\underline{+}1$, rather they were assigned the continuous values associated either with the fractional solvent accessibility [23] or with the hydrophilicity index [24]. A generalized box-counting method is used to analyze the sequential data. In this approach, the sequential array is covered by boxes of length, *l*, and the trajectory of each box is calculated. The procedure is repeated with increasing box sizes and the dependence of the trajectory on box size is found. In the multifractal approach one is not concerned with the displacement, *x(l)*, alone, but with all moments of the displacement, $x^q(l)$. These moments are used to determine the "generalized" dimensions associated with the shape of data profile. These generalized dimensions provide information on the hierarchical nature of the data set.

A "partition function", $Z_q(l)$, is defined to examine the qth moments of the sequence and it is given by:

$$Z_q(l) = \sum_{j=1}^{L/l} x_j^q(l) \qquad (5)$$

where j labels individual boxes or sequences of length l within the complete protein sequence of length, L. There will be a total of L/l of these boxes. 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: $\tau(q) = (q-1)D_q$. The generalized exponent is obtained from the initial slope of a $\log(Z_q(l))$ versus $\log(l)$ plot using a linear least squares fit.

It has been demonstrated that $\tau(q)$ follows a Legendre transformation and two new function *f(q)* and α*(q)* can be defined. 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 and gives *q* as a generalized temperature, τ as the generalized free energy, α as the generalized energy and *f* as the generalized entropy. Often one sees a multifractal spectrum as a plot of *f* versus α. In this formal analogy, the multifractal spectrum represents a relationship between the generalized energy and the entropy of the problem.

The function, *a,* (not to be confused with a_w) is determined from the relation:

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

With Eq. (5), this gives:

$$\alpha(q) = \frac{\sum_j x_j^q(l)\ln\{x_j(l)\}}{Z_q(l)\ln(l)} \tag{7}$$

The function a is obtained from the initial slope of a plot of $\sum_j x_j^q \ln(x_j) \big/ Z_q(l)$ vs. $\ln(l)$. 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]. Commonly, one sees multifractal spectra represented either as an *f*(*a*) versus α plot or as a D_q versus *q* plot. These are merely different parameteric ways of representing the same information.

In keeping with the above discussion, a single point on the spectrum is generated by calculating $Z_q(l)$ at a fixed *q* value and varying *l*. From the two linear regressions, *f*(α) and α are determined. The entire spectrum is generated by varying *q*. Both positive and negative integer values of *q* were used. Figure 1 shows an example of two data sets that can be analyzed with the multifractal formalism, the hydrophilicity index and the solvent accessibility for concanavalin

A. The corresponding multifractal spectra are shown in Figure 2. Although to the eye the profiles in Figure 1 may appear similar, the multifractal analysis reveals that they are correlated in very different ways. The spectrum for the solvent accessibilities is seen to be much broader than for the hydrophilicity. Both of 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 non-random correlations. However, these two parameters are correlated in a different manner. In previous work [25], it was shown that hydrophilicity profiles could be modeled as a simple multiplicative random process. The solvent accessibility requires a slightly more complicated model, a binary model with one-step memory.

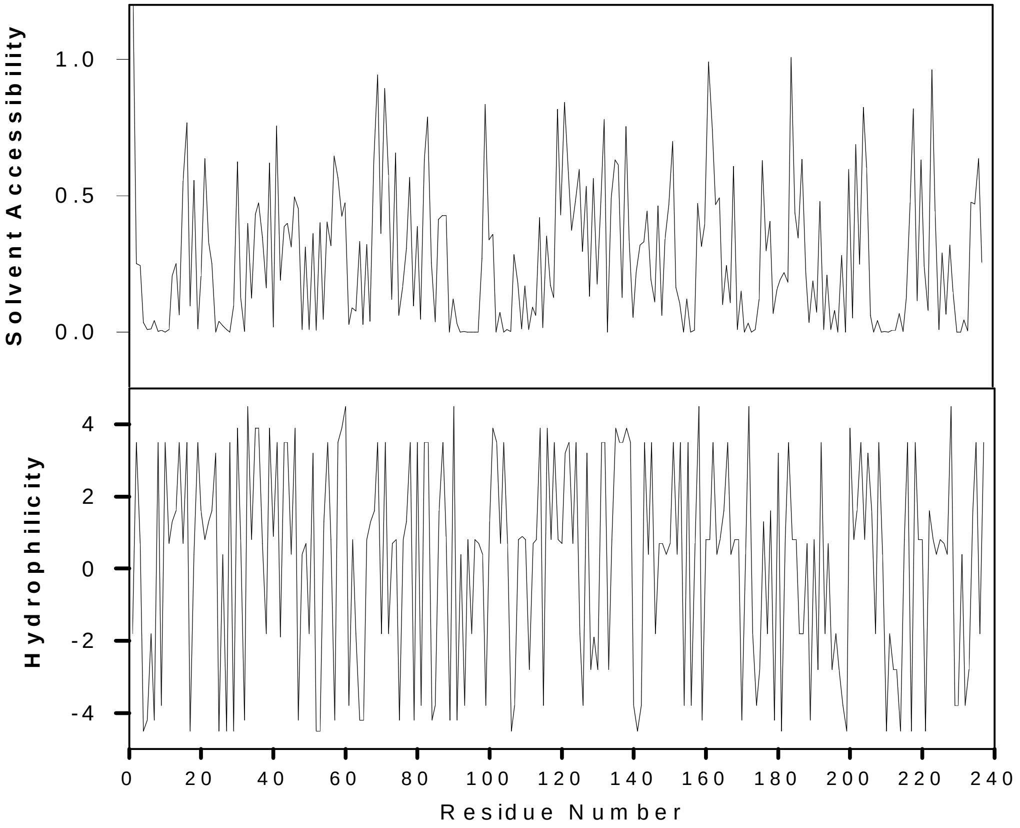

Figure 1. Profile for concanavalin A for the hydrophilicity (top) and the solvent accessibility (bottom). Hydrophilicity was determined as descibed in [24] and the determination of solvent accessibility from X-ray structures is described in [10]. Both profiles show similar

variations along the sequence. However, correlations with each profile is statistically very different.

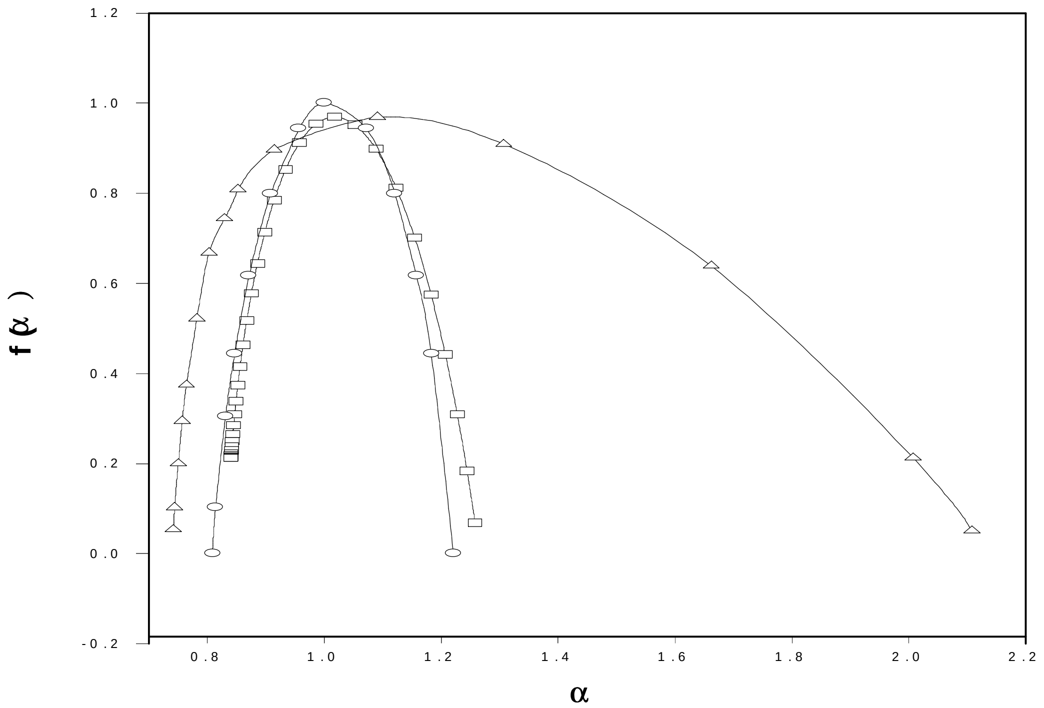

Figure 2. Multifractal spectra of concanavalin A: □, spectrum determined from the hydrophilicity profile, ❍, theoretical curve for a multiplicative binomial process used to fit hydrophilicity curve, Δ spectrum determined from the solvent accessibility profile. The broader spectra indicate stronger deviations from random behavior.

At first it may appear as if the multifractal approach is quite different from the encoded walks. However, they are actually closely related. The main difference in the two approaches is that multifractals are concerned with the moment distribution of many short trajectories that make up a protein sequence while the encoded walks focus on the root mean displacement of a single trajectory from an ensemble of protein sequences. Because the encoded walks rely on a single trajectory, an average over many proteins must be considered to generate good statistics. However, given a protein of a long enough sequence one could in principle obtain a scaling law (Eq. 4) for a single protein rather than for an ensemble average. In the multifractal approach, the sum of trajectories generated from a given sequence replaces the ensemble average of the encoded walk. If protein sequences are ergodic in the information theory sense, then the statistics within an individual protein sequence will be the same as those within the ensemble. In such cases, a correspondence between the two approaches can be made.

Because of the finite length of a protein sequence, the number of trajectories in the multifractal approach decreases with trajectory size as L/l. Consequently, to relate the multifractal sum to the mean displacement of the encoded walks, one must normalized the sum. The scaling of $\langle y^2 \rangle$ is related to the second moment of $Z_q(l)$ by:

$$\langle y^2 \rangle \sim Z_2(l) / (L/l)^2 \sim l^{2-\tau(2)} \qquad (9)$$

where a fractal dimension of one for the support, i.e., the linear sequence, is implicitly assumed. Using Eq. 4 and 9, one obtains a relationship between the walk exponent, α_w, and one of the multifractal exponents, $\tau(2)$:

$$\alpha_w = 1 - \frac{\tau(2)}{2} \qquad (10)$$

Thus, the result of the encoded walk trajectory is related to a single point in the multifractal spectra. One could, of course, examine higher order moments of the average displacement of the trajectory and establish a relationship such as: $\langle z^q(l) \rangle \sim l^{q\alpha_w(q)}$. These higher moments are related to the appropriate multifractal exponent by:

$$\alpha_w = 1 - \frac{\tau(q)}{q} \qquad (11)$$

From this it is seen that the multifractal approach has a direct correspondence with the encoded walks and that, except for negative moments, it provides the same information. The multifractal approach is the method of choice because it can generate scaling exponents for single protein sequences.

3: Ergodicity of Protein Sequences

The multifractal spectrum of the hydrophilicity profile (see Figure 2 for an example) of a large data set of proteins could be accurately fit using a multiplicative binomial process [25]. This model provides the following relationship for $\tau(q)$ [cf. 26]:

$$p^q + (1-p)^q = 2^{-\tau(q)} \qquad (12)$$

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. 12, a multifractal spectrum can be determined with the following relationships for $\alpha_{\min}$ and $\alpha_{\max}$:

$$\alpha_{\min} = -\frac{\ln(1-p)}{\ln 2} \tag{13a}$$

$$\alpha_{\max} = -\frac{\ln p}{\ln 2} \tag{13b}$$

The value of p is obtained using the $\alpha_{\min}$ and $\alpha_{\max}$ values determined from the intercepts of the multifractal spectrum. Most multifractal spectrum are simple concave curves that are well specified using three points, $\alpha_{\min}$, $\alpha_{\max}$, and $f_{\max}$. For a linear array, $f_{\max}$ is fixed at 1. The dimension, $\tau(2)$, is directly determined from the data with the aid of Eq. 5. Using Eq. 10, the α_w associated with the corresponding encoded walk is found. The data set shows persistence when $\tau(2) < 1$ and anti-persistence when $\tau(q) > 1$. Binomial multiplicative processes will always show persistence as $\tau(2)$ must always be less than 1 (see Eq. 12).

In previous work the average α_w was determined to be 0.517 ± 0.005 for 16 different proteins [25]. This is in excellent agreement with the value of 0.520±0.005 determined from the bridge analysis of hydrophilicity of an ensemble of proteins [8]. The correspondence between the average of walks within a protein to ensemble average walks suggests that protein sequences are ergodic. The ergodicity of protein sequences have important implications for information theory approaches to molecular evolution and sequence statistics as all the central theorems of information theory assume ergodicity in the signal or message.

To actually prove ergodicity is a difficult task and, at best, one can hope to establish consistency or inconsistency with an assumed ergodicity. To this end, a large, representative data set of protein sequences was selected. These were based on two algorithms [27] whose goals were to reduce redundancy in the set while maximizing coverage. Using these algorithms, representative sets of 155 (Set I) and 190 (Set II) different proteins sequences were examined. The multifractal spectrum derived from the hydrophilicity index for each protein was determined and the value of α_w for the corresponding encoded walk was calculated using Eq. 10. Representative values for individual proteins are shown in Table I. Additionally, the average values of the entire data sets (I and II) are given. The statistics of the ensemble of proteins in the data set were determined by concatenating the entire data set into a single sequence and determining the multifractal spectrum. The value and error of $\tau(2)$ was determined from a linear regression of the $\log(Z_q(l))$ versus $\log(l)$ plot for $q=2$. These values are shown in Table I. As can be seen, the mean of the individual proteins is well within a standard deviation of the value for the ensemble of proteins. It establishes the hydrophilicity as an ergodic parameter for these data sets.

Table I. Walk Exponents Derived from Multifractal Spectra

Set	Size	α_W
Concatenated Set I	30,309 A. A.	0.5156 ± 0.0008
Concatenated Set II	37,101 A. A.	0.5123 ± 0.0007
Average of Individuals, Set I	155 Proteins	0.517 ± 0.007
Average of Individuals, Set II	190 Proteins	0.517 ± 0.006
Representative Proteins	Size	α_W
Cytochrome B562	106 A.A.	0.523
Hemaglutinin	328 A.A.	0.521
Aconitase	754 A.A.	0.516
Tryptophan Synthase	248 A.A.	0.512

4: Implications for Information Theory

The number of possible proteins sequences of length N is 20^N. For a protein of only 100 amino acids the number of possible sequences is astronomical. Much has been written about the size of this number [cf. 28] and its implication for molecular evolution. Yet from an information theory perspective this number is not all that significant. An analogy can be made with the information content of languages. Certainly there are astronomical numbers of letter sequences that one can generate, but this has nothing to do with the structure or information content of the language. A basic theorem in information theory relates the most probable number of messages, Ω, of length N to the information entropy, I, by [cf. 29]:

$$\Omega = 2^{NI} \tag{14}$$

This theorem was derived for ergodic messages. The information entropy can be estimated from the distribution of k-tuplets in a sequence. This has been done for languages [30] and for DNA sequences [2, 31, 32] but we are unaware of any work on protein sequences. At the lowest order approximation, the entropy is defined as:

$$I = -\sum_{k=1}^{20} p_k \ln_2 p_k \qquad (15)$$

where p_k is the probability of finding the kth amino acid and base 2 is used to represent the entropy in terms of bits. At the lowest order of approximation, I_0, each amino acid is equally probable and $I_0 = \ln_2 20 = 4.3219$. The first order approximation accounts for the non-uniform amino acid composition and gives $I_1 = 4.177$for data set II. Higher order approximations can be determined from the k-tuplet distribution using [30]:

$$I_m = -\sum_{k=1}^{20} \sum_{s} p_k\, p(k|s) \ln p(k|s) \qquad (16)$$

where s represents a sequence that is m-1 units long and the inner sum is over all possible sequences of this length. The conditional probability, $p(k|s)$, is the probability of a kth amino acid following an s sequence. Doublet and triplet frequencies give $I_2 = 4.161$ and $I_3 = 3.988$, respectively. Higher order entropies drop precipitously. This is a result of the limited text. The number of possible k-tuplets exceeds the number in the text and gives the appearance of an unusually low information content. If a large enough, non-redundant text existed it would be possible to accurately determine the higher order entropies and these would decrease until the true information entropy was reached. Because we were not able to obtain this limit, the value of 3.988 can be considered to be an upper limit to the information entropy. (Interestingly, estimates of the information content of nucleic acid sequences (from an early and limited data base [31]) found an I of 1.94. This gives the not too surprising result that two bases do not quite have enough information to code an amino acid while 3 bases have an excess. This allows for redundancy in the genetic code. A Huffman encoding of the genetic code could be more efficient but would require codons consisting of singlets and doublets)

With this crude estimate of the upper bound for the information entropy, Eq. 14 can be used to calculate the most probable number of protein sequences. First, it is recast into a more convenient form:

$$\Omega = 20^{0.2314\,NI} \tag{17}$$

Using I_3 in Eq. 17, it is seen that for a protein with $N = 100$, Ω is at most 20^{92}, over 10 orders of magnitude less than the number of possible sequences. This indicates that there are significant regions of "sequence space" that have little or no probability of being populated. Nevertheless, sequence space is still vast and the number of proteins that have existed throughout evolution is still minute compared with the number of "most probable" proteins. One must bear in mind that the number of probable protein sequences is an upper estimate. Although this number could be considerably lower, from estimates of information content of nucleic acid one would not anticipate it approaching the number of proteins visited during evolution.

Proteins have been described as "slightly edited random polymers" [cf. 8]. Because sequence space is so vast, it is unlikely that a primordial soup generating random sequences of proteins could explore major regions of this space during the history of the earth. The number of proteins that could have existed during our history is difficult to estimate and these numbers have ranged from 10^{35}-10^{48} for a protein of 100 units [28]. Despite this wide range, these estimates are still minute compared with the number of possible sequences. This suggests that the conditions for forming a protein from a random polymer cannot be all that stringent. Yet the concept of a "slightly edited random polymer" is extremely vague. This work serves to better define this condition. A "most probable" protein sequence of 100 units is approximately 10^{-10} as likely as a random protein sequence, suggesting that considerable editing has taken place. Yet even with this, sequence space for "most probable sequences" is still much larger than the space that has been explored to date, again suggesting that finding a protein in this space is not too stringent a condition.

Acknowledgment

This work was supported in part by a grant #1R15GM51019 from the NIH.

References

1. R. F. Doolittle, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequence. Meth. Enz.* **183** (Academic Press, Inc. 1990).
2. M. V. Volkenstein, *Physical Approaches to Molecular Evolution* (Springer-Verlag, 1994).

3. S. Karlin, P. Bucher, V. Brendel, S.F. Altschul, *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175-203 (1991).

4. S. H. White, *Annu. Rev. Biophys. Biophys. Chem.* **23**, 407-439 (1994).

5. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley, *Nature* **356**, 168-170 (1992).

6. S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, M. Simons, *Biophys. J.* **65**, 2673-2679 (1993).

7. R. F. Voss, *Phys. Rev. Lett.* **68**, 3805-3808 (1992).

8. V. S. Pande, A.Y. Grosberg, T. Tanaka, *Proc. Natl. Acad. Sci. USA* **91**, 12972-12975 (1994).

9. T. G. Dewey, *Fractals* **1**, 179-189 (1993).

10. J. S. Balafas, T. G. Dewey, *Phys. Rev. E* **52**, 880-887 (1995).

11. H. E. Stanley, S. V. Buldyrev, A.L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadink, C.-K. Peng, M. Simons, *Physica A* **205**, 214-253.

12. H. E. Stanley, and P. Meakin, *Nature* **335**, 405-409 (1988).

13. T. Tél, *Z. Naturforsch.* **43a**, 1154-1174 (1988).

14. L. Pietronero, C. Evertsz, and A. P. Siebesma, in *Stochastic Processes in Physics and Engineering*, eds. Albeverio, S. et al. (D. Reidel Publishing Co., 1988) pp. 253-278.

15. T. A. Witten, Jr., and L. M. Sander, *Phys. Rev. Lett.* **47**, 1400-1403 (1981).

16. L. de Arcangelis, S. Redner, and A. Coniglio, *Phys. Rev.* **B 31**, 4725-4727 (1985).

17. I. Procaccia, *J. Stat. Phys.* **36**, 649-665 (1984).

18. K. G. Wilson, *Sci. Am.* **241**, 158 (1979).

19. T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B.I. Shraiman, *Phys. Rev.* **A 33**, 1141-1151 (1986).

20. T. G. Dewey, in *Fractals in the Natural and Applied Sciences, IFIP Transactions* **A-41** ed. Novak, M. M. (North-Holland, 1994) pp. 89-100.

21. T. G. Dewey, *Fractals* **3**, 9-22 (1995).

22. C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, *Phys. Rev. E* **49**, 1685-1689 (1994).

23. F. M. Richards, *Ann. Rev. Biophys. Bioeng.* **6**, 151-176 (1977).

24. J. Kyte, R. F. Doolittle, *J. Mol. Biol.* **157**, 105-132 (1982).

25. B. Strait, T. G. Dewey, *Phys. Rev. E* (accepted) (1995).

26. J. Feder, *Fractals* (Plenum Press, 1988) pp. 66-103.

27. U. Hobohm, M. Scharf, R. Schneider, C. Sander, *Protein Science* **1**, 409-417.

28. S. A. Kauffman, *The Origins of Order* (Oxford University Press, NY, 1993) pp. 20-22; 121-172.

29. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, NY, 1991) pp. 50-59.

30. C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1962) pp. 7-33.

31. L. L. Gatlin, *Information Theory and the Living System* (Columbia University Press, NY, 1972) pp. 47-72.

32. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley, *Phys. Rev. Lett.* **73**, 3169-3172.