Algorithmic complexity of a protein

T. Gregory Dewey*
Department of Chemistry, University of Denver, Denver, Colorado 80208
(Received 8 March 1996)

The information contained in a protein's amino acid sequence dictates its three-dimensional structure. To quantitate the transfer of information that occurs in the protein folding process, the Kolmogorov information entropy or algorithmic complexity of the protein structure is investigated. The algorithmic complexity of an object provides a means of quantitating its information content. Recent results have indicated that the algorithmic complexity of microstates of certain statistical mechanical systems can be estimated from the thermodynamic entropy. In the present work, it is shown that the algorithmic complexity of a protein is given by its configurational entropy. Using this result, a quantitative estimate of the information content of a protein's structure is made and is compared to the information content of the sequence. Additionally, the mutual information between sequence and structure is determined. It is seen that virtually all the information contained in the protein structure is shared with the sequence. [S1063-651X(96)50707-5]

PACS number(s): 87.10.+e, 87.15.By, 89.70.+c

For most proteins, the information contained in the primary sequence is sufficient to dictate the three-dimensional structure. Although this succinct statement captures the essence of the protein folding problem, there are a number of manifestations of it and these have increasingly preoccupied the biochemical community [1]. Areas of investigation range from thermodynamic and kinetic aspects of folding to the computational ability to predict structure from sequence. Behind all these approaches lies the underlying assumption that the information content of the protein sequence is the source of the great structural specificity of a protein.

Despite this widespread interest, there has been little effort to quantify the information content either of protein sequences or of the folded state. In previous work, a large data base of protein sequences was analyzed to determine the Shannon information content [2]. It was shown that the Shannon entropy, S, was approximately 2.5 bits per amino acid. In the present work, we turn from the probabilistic analysis of protein sequences to the information content of the final folded state. To analyze an object such as the three-dimensional structure of a protein, the Kolmogorov information entropy, K, is employed. It has been shown for certain classes of problems that the Kolmogorov entropy or algorithmic complexity could be associated with the thermodynamic entropy [3–5]. The present work extends this result to polymer systems and allows the Kolmogorov information entropy of a protein to be estimated.

The Shannon information entropy of a message or sequence is related to the probability of receiving that message from an ensemble of messages. Paradoxically, in the Shannon formalism, once a message is received its information content is zero. Algorithmic complexity was a concept introduced by Kolmogorov and developed by a number of workers to avoid the probabilistic, ensemble interpretation of Shannon [6–8]. The algorithmic information is defined as the shortest binary string required to describe the object. Alternatively, it can also be defined as the shortest program required to produce the string with a universal computer. The algorithmic entropy is a property of the specific object under consideration rather than of the ensemble. There has been controversy in the literature regarding the relationship between information entropy and thermodynamic entropy [9]. These controversies are potentially resolved by a general definition of ''physical entropy'' as the sum of the uncertainty in the system, Shannon entropy, and what is known about the system, Kolmogorov entropy [4,5]. Zurek has also shown that for a Boltzmann gas, the algorithmic entropy can be estimated using the thermodynamic entropy, a result anticipated by Bennett [3]. In general, this is a good estimate for any statistical mechanical system whose configuration can be encoded with a binary sequence.

The protein folding problem can be represented as a Turing or universal computer problem. The data input, the protein sequence, is readily described using a binary string. This string can be encoded with 2.5 bits per amino acid as estimated from the Shannon entropy [2]. The output is the Cartesian coordinates of every atom in the protein and is also readily encoded in a binary string. The Kolmogorov entropy provides a means of describing the information content of this output string. Although it is difficult to calculate exactly, it is relatively easy to put upper bounds on the Kolmogorov entropy. Often these bounds are independent of the specific algorithm used to calculate them [5].

To determine the algorithmic complexity of an individual microstate of a polymer such as a protein, we encode its configuration. The simplest possible set of instructions for describing a protein is as follows: (a) list all the Φ and Ψ angular coordinates of the planar peptide linkages. (b) Order the list according to the occurrence of the coordinates along the sequence. This description captures the main features of the problem. Certainly rotameric states of side chains will be important. However, once the backbone configuration is set, sidechain rotamers are readily optimized computationally. Consequently, the main information content is found in the backbone configuration. The specification of the angular coordinates, Φ and Ψ, provides the secondary structure of the

*Fax: 303-871-2254. Electronic address: gdewey@du.edu

protein. The ordering or connectivity of the various Φ-Ψ pairs will then give the tertiary structure of the final folded state.

To specify the algorithmic complexity of part (a), the list of Φ-Ψ coordinates, an approach analogous to the one Zurek used for the Boltzmann gas, is followed [5]. A protein of N amino acid units is assumed to have peptide orientations distributed over a configurational space of volume $V=\Phi\Psi$. To specify a protein's secondary structure, a predetermined level of accuracy, ΔV, is required. With this accuracy, the location in configurational space of each peptide bond rotation can be described by a number whose size is $V/\Delta V$. If these are algorithmically random, the length of a program needed to encode the relevant angles for a single peptide bond is $\ln_2(V/\Delta V)$. The length of a program required to list all the angular coordinates is

$$K\approx N\ln_2(V/\Delta V)+O(\ln_2(N))+O(1). \tag{1}$$

For self-delimiting programs, the value of N will be known and the $O(\ln_2(N))$ correction can be eliminated [5]. Note that for the Boltzmann gas, a factor of $1/N$ appears in the first term of Eq. (1). This is a result of the indistinguishability of particles and does not appear in the protein problem because of the sequential ordering and amino acid composition. The configurational volume V is the volume available to a random coil and is often given the symbol z_{rc} [10]. It is given by

$$z_{rc}=\int_0^{2\pi}\int_0^{2\pi} e^{-\beta E(\Phi,\Psi)}d\Phi\, d\Psi, \tag{2}$$

where $E(\Phi,\Psi)$ is the internal energy associated with bond rotation and β is $1/kT$. The value of z_{rc} has been estimated as $4118^{\circ 2}$ [10,11]. Proteins are made up of secondary structural units that are defined in broad regions of Φ-Ψ space. Typically, these units are taken to be α helix, β sheet, β turn, and random coil. To determine the secondary structure in this configurational volume, one must know Φ and Ψ to an accuracy of $\pm 40^\circ$ [12]. Thus, a value of $1600^{\circ 2}$ has been used for ΔV. In Dill's notation, $\Delta V=z_g$ and $z=z_{rc}/z_g$. A correction is also added to the value of z to make it compatible with a cubic lattice model. Our first contribution to the Kolmogorov entropy is $K=N\ln_2 z$.

To estimate the contribution to the algorithmic complexity from the ordering of the list of coordinates [part (b)], the connectivity of the polymer must be considered. It is particularly convenient to consider a lattice model [10,13]. In this model, the polymer occupies N connected sites in a three-dimensional lattice of N_0 sites. To encode the polymer configuration as a binary string, we give an empty (or solvent) site a 0 and a filled (or polymer site) a 1. The polymer configuration is now represented by a binary string of length N_0 that contains N 1's. The algorithmic complexity of such a string can be estimated using a device that involves the lexicographic ordering of all possible strings [14].

To illustrate this approach, we first consider a system configuration in which the filled sites are randomly distributed in the lattice without a connectivity requirement. To specify a given string, we first compute all possible strings, print them in lexicographic order and find the address for the string of interest. The number of all possible strings is Ω and the addresses will run from 1 to Ω. Since an address numeral can be as large as Ω, a program must be able to represent this number. Because Ω will be a large number, the algorithmic complexity of the problem will be dominated by the mere representation of this number. The algorithmic complexity in this case will be at most $\ln_2\Omega$. For the present configuration of a random string of length N_0 with N 1's, one has [14]

$$K\leqslant \ln_2\binom{N_0}{N}+O(\ln_2 N)+O(1), \tag{3}$$

where again the $O(\ln_2 N)$ will not appear in a self-delimiting program. Also, note that as N gets large and approaches N_0, the information content approaches zero.

For the polymer problems, the calculation of Ω is more complicated. Using the Flory-Huggins approach [13,15,16], one successively ''builds'' the configuration in a stepwise fashion. There are N_0 sites available for the first polymer unit. In successive units, the probability of finding a vacant site to add the ith unit is given by a factor p_i and the number of configurations becomes

$$\Omega=N_0\prod_{i=1}^{N-1} p_i. \tag{4}$$

To account for excluded volume effects, Flory approximates p_i by

$$p_j=(N_0-j)/\{N_0-2j/q\}, \tag{5}$$

where q is the coordination number of the lattice. In the limit of large N, [13] $\Omega=a^{-N}$. The constant a depends on the lattice structure and is given by $a=(1-2/q)^{-(q/2-1)}$. For a cubic lattice, a is equal to 2.25. The contribution of the polymer connectivity to the algorithmic complexity is now given by $K\leqslant N\ln_2\Omega=N\ln_2(1/a)$.

Combining the results for encoding of a protein according to the directions in parts (a) and (b), the Kolmogorov entropy of a protein is given by

$$K\leqslant N\ln_2\left(\frac{z}{a}\right), \tag{6}$$

where we have assumed a self-delimiting program. This is essentially the thermodynamic configurational entropy for a protein and the value of this parameter has been discussed extensively by Dill [10]. For a cubic lattice model, z is estimated at 3.8 and $a=2.25$, giving $K\leqslant 0.77$ bits per amino acid. A more realistic estimate [10] that accounts for internal interactions between protein sidechains gives $z/a=1.4$ and $K\leqslant 0.49$. Thus, a program to compute the structure of a protein that is 100 amino acids long requires less than 49 binary digits. Also, it is seen that the Kolmogorov complexity is significantly less than the Shannon information content of the sequence, 2.5 bits per amino acid.

For protein structure prediction, a computer program is required that produces the protein's spatial coordinates given a specific sequence. Consequently, the algorithmic complexity of the protein itself is not as important as the shared or

mutual information between the structure and the sequence. But it would appear that we have determined two different quantities, the Shannon information entropy for sequences and the algorithmic complexity for structures. This problem is circumvented using Zurek's concept of physical entropy [4,5]. The physical entropy, S, is given by

$$S = K + I, \tag{7}$$

and is the sum of the information known about the system (Kolmogorov information), K, and the missing information (Shannon information), I.

To determine the shared physical entropy or information between the sequence (seq) and structure (str), we just calculate S(seq:str) from conditional entropies. The mutual information is given by [5]

$$S(\text{seq:str}) = S(\text{str}) - S(\text{str} \mid \text{seq}), \tag{8a}$$

$$= S(\text{seq}) - S(\text{seq} \mid \text{str}). \tag{8b}$$

S(str|seq) is the conditional entropy of a structure given a specific sequence. The conditional entropy is determined from the following relationship:

$$S(\text{str} \mid \text{seq}) = S(\text{str}) - S(\text{seq}) + S(\text{seq} \mid \text{str}). \tag{9}$$

S(str) is estimated by the Kolmogorov entropy, Eq. (6), and a value of approximately 0.5 bits per amino acid is assumed. For the value of S(seq), we use the Shannon entropy (2.5 bits per amino acid). The conditional probability of S(seq|str) can be estimated from exhaustive mutagenesis experiments where the amino acid sequence is altered and viable proteins are screened. Thus, for a given protein structure the number of possible sequences can be explored. A complete mutagenesis study of the protein coding domains of the HIV protease has been done [17]. This study produced all point mutations that could result in a folded protein. For all combinations of single mutations, one has 6.8×10^{60} possible protein sequences that fold to a given structure. From information theory, the number of possible sequences will be $2^{NI.}$ The information calculated from these experimental results is the conditional information related to the number of possible protein sequences given a specific protein structure. For this protein of 99 amino acids, this provides an estimate of 2.0 bits per amino acid for S(seq|str). This number is most likely an overestimate, as some combinations of single mutations may be incompatible.

With these various estimates and considerations, it is seen that S(str|seq) is approximately zero and the mutual information [Eq. (8)] is then given by S(seq:str)$\approx S$(str). This shows that all the information in the final structure is shared with that in the protein sequence. In information theory terminology, the protein folding process can be viewed as a noiseless communication channel, directly communicating the information from the sequence. It also suggests that outside environmental influences such as solvation or ionic strength do not add appreciably to the information content of the process, i.e., that the sequence itself is the main source of information.

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

[1] T. E. Creighton, *Protein Folding* (Freeman, New York, 1992).
[2] T. G. Dewey and B. J. Strait, Biophys. J. (to be published).
[3] C. H. Bennett, Int. J. Theor. Phys. **21**, 905 (1982).
[4] W. H. Zurek, Nature **341**, 119 (1989).
[5] W. H. Zurek, Phys. Rev. A **40**, 4731 (1989).
[6] A. N. Kolmogorov, Prob. Inf. Transmission **1**, 4 (1965).
[7] R. J. Solomonoff, Inform. Contr. **7**, 224 (1964).
[8] G. J. Chaitin, J. Assoc. Comput. Mach. **22**, 329 (1975).
[9] L. Brillouin, *Science and Information Theory* (Academic, New York, 1962).
[10] K. A. Dill, Biochemistry **24**, 1501 (1985).
[11] D. A. Brant, W. G. Miller, and P. J. Flory, J. Mol. Biol. **23**, 47 (1967).
[12] P. Y. Chou and G. D. Fasman, Annu. Rev. Biochem. **47**, 251 (1978).
[13] P. J. Flory, Proc. Natl. Acad. Sci. USA **79**, 4510 (1982).
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991), pp. 152 and 153.
[15] M. L. Huggins, J. Phys. Chem. **46**, 151 (1942).
[16] P. J. Flory, J. Chem. Phys. **10**, 51 (1942).
[17] D. D. Loeb, R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, and C. A. Hutchison III, Nature **340**, 397 (1989).