Algorithmic complexity and thermodynamics of sequence-structure relationships in proteins

T. Gregory Dewey*
Department of Chemistry and Biochemistry, University of Denver, Denver, Colorado 80208

(Received 28 April 1997)

The information contained in a protein's amino acid sequence dictates its three-dimensional structure. In this situation a frozen or embedded structure, the sequence, contains information that ultimately influences a thermodynamic entity, the protein structure. The interplay between information and thermodynamics is explored by considering the algorithmic complexity and Kolmogorov's universal probability of the sequence and of the structure. It is shown that the algorithmic complexity of a microstate of a polymer is given by its configurational entropy. Using this result and a lattice protein model, a quantitative estimate of the information contained in a protein's structure is made. This is compared to the information content of the sequence. The information content of the sequence is approximately 2.5 bits per amino acid, while the content in the structure is approximately 0.5 bits per amino acid. It is estimated that virtually all the information contained in the protein structure is shared with the sequence. A deeper connection can be made between the shared information content and the thermodynamic entropy governing the system. Using Kolmogorov's universal probability, it is possible to establish statistical-mechanical relationships for objects without resorting to a probabilistic ensemble formalism. This allows the thermodynamics of microstates of objects of known configurations to be determined. Using this formalism, the connection between sequence information and the structural thermodynamics of a protein can be made. This connection has strong implications for how protein sequences evolve over evolutionary time and demonstrates that this evolution is constrained by the thermodynamic evolution of the protein structure. [S1063-651X(97)03510-1]

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

I. INTRODUCTION

The fundamental premise of the protein folding problem is that the information contained in the protein sequence specifies the three-dimensional structure of the protein [1]. Although this premise is now supported by a wealth of experimental data, there have been few efforts to quantify the information content of the protein sequence (cf. [2]). Ideally, one could quantify the information content of the protein structure as well and determine the amount of information shared between sequence and structure. This shared or mutual information is an implicit component of programs involved with protein structure prediction and protein design. It is also of interest to ask how these information parameters evolve over evolutionary time. Is this evolution random, as suggested by Kimura's neutral theory of evolution [3], or does it follow specific dynamical laws? To approach these questions, the information content of both the protein sequence and the protein structure must be determined.

The information content of the sequence can be obtained by calculating the Shannon information entropy. The Shannon information entropy of the amino acid sequence reveals the smallest number of binary digits (bits) per amino acid that are needed for the most efficient coding of the sequence. This number can be estimated from the probability distribution of amino acids in a protein. Previous work in this laboratory suggests that this number is approximately 2–2.5 bits per amino acid [4]. This is a surprisingly small number considering that a uniform distribution of 20 amino acids would require 4.32 bits per amino acid $(\ln_2 20)$. This low information content indicates that sequences are not random and that some degree of correlation must exist within them. This nonrandomness is due in part to structural and thermodynamic constraints of the folded protein.

Recently, it was suggested that the information content of a protein's structure could be quantified using an information theoretical parameter known as the algorithmic complexity [5]. In previous work, it was shown that the algorithmic complexity of a protein is equal to its configurational thermodynamic entropy. The algorithmic complexity of an object is broadly defined as the length in bits of the shortest description for that object (cf. [6]). Alternatively, it is the length of the shortest program required to obtain the output. Kolmogorov devised this definition of the information content of an object to circumvent the probabilistic ensemble arguments used in defining the Shannon information. Shannon information has the paradoxical feature that information only exists when it can be described probabilistically as one possible ''message'' out of an ensemble of messages. Once the message is received, the probability of finding the message is unity and it no longer has an information content. Kolmogorov's definition of algorithmic complexity (sometimes referred to as Kolmogorov entropy) does not suffer from this problem and can be applied to individual objects such as the structure of phosphofructokinase. No reference need be made to an ensemble of proteins. Using the algorithmic complexity to estimate the information content of the protein from a lattice model, one finds that it contains approximately 0.5 bit per amino acid [5].

Shannon entropy and algorithmic complexity play complementary roles. Shannon entropy represents the information of the system that is not known and consequently uses a probabilistic ensemble treatment. Algorithmic com-

*FAX: (303) 871-2254. Electronic address: gdewey@du.edu plexity gives the known or measured information of the system. It is used for fully determined systems and probabilistic or ensemble arguments never enter. It is a remarkable relationship, discovered by Zurek [7,8], that the algorithmic complexity of a microstate in a statistical ensemble is equal to the thermodynamic entropy. Zurek has been able to derive statistical-mechanical relationships based on algorithmic complexity. Zurek established the relationship

$$S=K+I, \tag{1}$$

where S is the physical entropy, K is the algorithmic complexity, and I is the Shannon information entropy. Equation (1) says that the physical or thermodynamic entropy of a system is composed of two parts, that determined from the known information of the system K and that determined from the unknown or probabilistic information I. For a macroscopic system in which the microstates are unknown, $K=0$ and the entire entropy is due to the Shannon information, i.e., $S=I$. This result was established earlier by Jaynes [9,10]. As observations are made on a system, the information content shifts from I to K. If the position and momentum of all the particles of the system are known then $I=0$ and K represents the entire entropy of the system $(S=K)$.

In the present work we extend and generalize previous results that established the relationship between the algorithmic complexity and the thermodynamic entropy of a protein. In Sec. II it is shown that the algorithmic complexity of a polymer is given by its configurational entropy. Section III adapts this derivation to the specific case of a protein. The results of this section allow estimates of the information content of the structure of native, folded proteins. In Sec. IV these results are generalized to a classical many-body system. In this section it is shown that statistical-mechanics relationships can be derived using the Kolmogorov universal probability. This section is the algorithmic complexity counterpart to Jaynes's information theoretical development of statistical mechanics. It allows a formulation of statistical mechanics for systems in which a probabilistic approach is not needed. General relationships between the Kolmogorov universal probability and the classical partition function are established. In Sec. V the results of the previous sections are used to discuss the thermodynamic constraints on information transfer and dynamics in biological systems. It is seen that the evolution of protein sequence information is governed by thermodynamic laws. The paper is summarized in Sec. VI.

II. ALGORITHMIC COMPLEXITY OF A POLYMER

The algorithmic complexity of a polymer is given by the length of the shortest program required to describe it. Alternatively, one seeks the most compact signal, in bits, that can describe the object. At first, such definitions would appear to be very impractical as it would be very difficult to prove that any given program or signal is the shortest. In practice, however, this appears not to be a major constraint. Often very different and seemingly very inefficient programs will give essentially the same algorithmic complexity. This phenomenon is largely a result of the logarithmic nature of such problems.

An example of such an inefficient algorithm is the ''lexicographic'' trick [6]. This algorithm can be applied to any problem involving enumerations and will usually give the same result as more efficient algorithms. For a polymer, this algorithm would be to list all the Ω possible configurations of the polymer in lexicographic order, find the appropriate configuration for the microstate of interest, and print out that configuration. To perform this task, addresses must be given to each state so as to specify the location in the list of the microstate of interest. This address could be as high as Ω, so to execute this program the algorithmic complexity of representing the number Ω must be specified. Since Ω is such an enormous number, merely representing it dominates the information content. The smallest number of bits required to represent an integer is the logarithm in base 2 of that integer. Thus the algorithmic complexity of a polymer K is given by

$$K=\ln_2 \Omega. \tag{2}$$

This is the same result one expects for the thermodynamic entropy S and one has

$$K=\frac{S}{k\ \ln 2}, \tag{3}$$

where the natural logarithm is used in Eq. (3) and k is Boltzmann's constant.

This rather strange algorithm shows that the algorithmic complexity or information content of a protein will be its configurational entropy expressed in bits. In the remainder of this section and in Sec. III this result is established in a more physical and conceptually appealing manner. A polymer can be described by specifying the location of the monomeric units in space. To specify the spatial properties of a polymer, one must first divide the space in which it is embedded into discrete cells. These cells should be made large enough to encompass a monomeric unit, but small enough to avoid two units. If a cell is occupied with polymer, it is given a 1. If it is occupied with solvent, it received a 0. These cells are then numbered and the sequence of cell numbers that have 1's represent a specific microstate of a polymer. The sum of the logarithm of the addresses (or lattice coordinates) is the algorithmic complexity. For some situations, a spatial specification involving internal coordinates is required. For proteins, these internal coordinates are the $\Phi\Psi$ angles of the peptide linkages. These will be dealt with in Sec. III.

To describe a polymer, one then needs a list of addresses. The addresses are represented as integers whose value is given by the volume of the lattice V divided by the volume of the lattice cell ΔV. The lattice cell volume is given by $\Delta V \sim l^3$, where l is the bond length between polymeric units. On average an address can be represented by an integer $V/\Delta V$ and the algorithmic complexity of a polymer of n units is

$$K=n\ \ln_2\left(\frac{V}{\Delta V}\right). \tag{4}$$

Care must be taken to represent V in the most efficient manner [7]. Rather than representing the whole volume of the lattice, it is more efficient to represent the volume relative to an internal polymer point, such as the center of mass. The spacing between monomers in the polymer is at least V/n in such an ''internal'' coordinate system. For lattices of N sites, the total volume is $V=Nl^3$. Substitution into Eq. (4) gives the contribution from the sites with 1's in the lattice as $K=n\ \ln_2(N/n)$. If the contribution from the $N-n$ solvent sites is also considered, an analogous expression is obtained and the total contribution to the complexity is

$$K=n\ \ln_2\left(\frac{N}{n}\right)+(N-n)\ln_2\left(\frac{N}{N-n}\right). \tag{5}$$

Equation (5) is recognized as the entropy of mixing of an ideal gas rather than that of a polymer. The analogy between ideal gases and polymers has been made by Flory in his derivation of the configurational entropy of a polymer [11]. In the above derivation we failed to account for the connectivity of the polymer. This correction is readily achieved following the methods of Flory [11]. To introduce the connectivity of the polymer, the addresses of the monomeric units are listed in order of their connectivity. A site is chosen at random to initiate the polymer chain. It can fall on any site in the volume and therefore will have an address of order $V/\Delta V\sim Nl^3/l^3=N$. The second unit must be in a site adjoining the first one. The volume available to the second unit is the lattice coordination number q times the cell size $\Delta V=l^3$. This address will then be expressed by $V/\Delta V\sim ql^3/l^3=q$. For the third unit, one now has only $q-1$ sites that can be occupied. The available volume in this case is $V\sim(q-1)l^3(1-f)$, where f is the expectancy that a given cell adjacent to a previous one is unoccuppied (cf. [11]).

Proceeding in this manner, the algorithmic complexity of a polymer is given by

$$K_{\text{polymer}}=\sum_{i=1}^{n}\ \ln_2\left(\frac{V_i}{l^3}\right)$$

$$=\ln_2 n+\ln_2\ q+\cdots+\ln_2(q-1)(1-f_i) \tag{6a}$$

$$=\ln_2\left\{nq(q-1)^{n-2}\prod_{i=2}^{n}\ (1-f_i)\right\}. \tag{6b}$$

Following Flory [11], the site expectancy is approximated by

$$1-f_i=1-\overline{f_i}=\left(\frac{N-n}{N}\right), \tag{7}$$

where $\overline{f_i}$ is the average expectancy. The algorithmic complexity for the polymer is now given by

$$K_{\text{polymer}}=\ln_2(N)+(N-n)\ln_2\left(\frac{N}{N-n}\right)+(n-1)\ln_2\left(\frac{q-1}{e}\right). \tag{8}$$

Equation (8) is essentially the configurational entropy of a lattice polymer as derived by Flory [11]. It also gives the same result that would be obtained through the simpler algorithm based on Eqs. (2) and (3).

III. ALGORITHMIC COMPLEXITY OF A PROTEIN

Because proteins are compact polymers, the lattice can be made the size of the protein and solvent effects are eliminated. The complexity of a collapsed polymer is then given as $K_{\text{polymer}}\approx n\ \ln_2[(q-1)/a]$, where $a=e$. A more sophisticated analysis of the lattice excluded-volume effects [12] gives

$$a=\left(1-\frac{2}{q}\right)^{-(q/2-1)}. \tag{9}$$

In addition to a more accurate specification of the excluded-volume effect, it is important to specify the details of the configurational volume V and the size of the lattice cells ΔV required to specify a protein configuration. Essentially, both the connectivity and the secondary structural content of the protein are specified.

A protein has peptide orientations distributed over a configurational space of volume $V=\Phi\Psi$, where Φ and Ψ are the angles associated with the rotation of the planar peptide linkage. 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 $V/\Delta V$. The configurational volume V is the volume available to a random coil and is often given the symbol z_{rc} [13]. It is

$$z_{\text{rc}}=\int_0^{2\pi}\int_0^{2\pi}e^{-\beta E(\Phi,\Psi)}d\Phi\ d\Psi, \tag{10}$$

where $E(\Phi,\Psi)$ is the internal energy associated with bond rotation β is $1/kT$. The term z_{rc} replaces the factor $(q-1)l^3$ in the derivation of Sec. II. The value of z_{rc} has been estimated as 4118 deg^2 [14]. 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 an α helix, a β sheet, a β turn, and a random coil. To determine the secondary structure in this configurational volume, one must know Φ and Ψ to an accuracy of ± 40 deg [15]. Thus a value of 1600 deg^2 has been used for ΔV. In Dill's notation [13] $\Delta V=z_g$ and $z=z_{\text{rc}}/z_g$. The term z_g replaces the term l^3 used to specify ΔV in Sec. II. A correction is also added to the value of z to make it compatible with a cubic lattice model.

Combining the results for encoding of a protein, the Kolmogorov entropy of a protein is given by

$$K_{\text{protein}}\approx n\ \ln_2\left(\frac{z}{a}\right). \tag{11}$$

This is essentially the thermodynamic configurational entropy for a protein and the value of this parameter has been discussed extensively by Dill [13]. 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 [13] that accounts for internal interactions between protein side chains gives $z/a=1.4$ and $K\leqslant 0.49$. Thus a program to compute the structure of a protein that is 100 amino acid long requires less than 49 binary digits. Also, it is seen that the Kolmogorov complexity is significantly less than Shannon information content of the sequence, 2.5 bits per amino acid.

IV. KOLMOGOROV UNIVERSAL PROBABILITY AS A PARTITION FUNCTION

Using information theory, statistical mechanics can be recast not as a physical theory but rather as a theory based on statistical inference [9,10]. The appeal of such a fundamental shift is that certain physical assumptions, assumption not justified from mechanics, do not have to be made. However, statistical mechanics based on information theory still suffers from conceptual problems resulting from the need to extract probabilistic inferences from ensembles. The ‘‘maximum-entropy’’ method developed by Jaynes is a form of statistical inference that gives the optimal distribution when there is minimal knowledge about the system. But what about systems in which we have partial or even complete knowledge? How is entropy defined for them? What about individual objects or individual microstates of a system? How can entropy be defined for them?

Kolmogorov wrestled with similar problems when considering the information content or complexity of individual objects outside an ensemble. He defined the algorithmic complexity of an object to be the minimal length in binary code of a computer program required to describe the object [6]. This definition of complexity avoids all reference to probability distributions. This abstract construct from theoretical computer science can be directly related to the thermodynamic entropy.

Zurek established that the algorithmic complexity of a ‘‘typical’’ microstate of a Boltzmann gas is proportional to the thermodynamic entropy. He derived the Sackur-Tetrode equation from the algorithmic complexity of a microstate of the system [7]. In the present work we generalize these result by considering Kolmogorov’s universal probability rather than the algorithmic complexity. It is shown that for a classical statistical-mechanical system, the Kolmogorov probability can be related to the partition function. This relationship is deeper than the previous treatments because it allows for a consideration of different types of microstates. Only for a microstate of a microcanonical system can the algorithmic complexity and thermodynamic entropy be directly related to each other. This formulation of statistical mechanics allows one to define thermodynamic parameters for objects or microstates whose parameters are completely determined and need not be related to an ensemble. As such, it is ideally suited for discussing the thermodynamics of embedded, nonequilibrium structures, such as the sequence of a protein. Using the formalism developed in this section, the thermodynamics of protein sequences is discussed in Sec. V.

The Kolmogorov universal probability of an object is the probability that a program (in binary form) that describes the object can be generated by a random sequence of digits (cf. [6]). Thus, if a program is of length l, the probability of it being randomly generated is 2^{-l}. From a theoretical computer science prospective, the algorithmically most simple objects have the shortest programs (small l) and will be the most probable. The universal probability $P_{\mathcal{U}}$ is the sum of the probabilities of all randomly chosen programs that describe the object and is given by

$$P_{\mathcal{U}}=\sum_{i} 2^{-l_i}, \tag{12}$$

where the sum is over all programs. Since the algorithmic complexity is defined as the shortest program, its contribution will often dominate the sum and one frequently has [6]

$$P'_{\mathcal{U}}=2^{-K}, \tag{13}$$

where the prime designates the universal probability when K dominates the sum.

The universal probability of a point is phase space is now considered. All programs to specify this point must provide the position and momentum coordinates of each of the N particles. Following Zurek [7], the location of a given particle in position-momentum space is specified by covering the space with a grid of hypercubes with edge sizes of δp and δq. The edge sizes are adjusted so that only a single particle can fit into the hypercube. If an isotropic phase space is assumed, a single edge size can specify each of the respective three-dimensional coordinates. Each particle in the system will have a phase-space location specified by the numbers $p_x/\delta p$, $p_y/\delta p$, $p_z/\delta p$, $x/\delta q$, $y/\delta q$, and $z/\delta q$. The length of a program required to specify all the coordinates is

$$l_i=\sum_{j=1}^{N}\left[\ln_2\left(\frac{\mathbf{q}_j^{(i)}}{\delta q^{(i)}}\right)+\ln_2\left(\frac{\mathbf{p}_j^{(i)}}{\delta p^{(i)}}\right)\right], \tag{14}$$

where i specifies the coordinate system to be used and j identifies the N particles. The grid size $(\delta q\,\delta p)$ or hypercube may depend on the specific coordinate system under consideration. All programs are essentially the same, but some provide more convenient coordinate systems and consequently are more efficient.

The universal probability for this program is given by

$$P_{\mathcal{U}}=\sum_{i} 2^{-l_i}=\sum_{i}\left\{\prod_{j=1}^{N}\frac{\delta q\ \delta p}{\mathbf{q}_j^{(i)}\mathbf{p}_j^{(i)}}\right\}$$
$$=\sum_{i} h^N(v_1^{(i)}v_2^{(i)}\cdots v_N^{(i)})^{-1}, \tag{15}$$

where on the right hand side h is Planck’s constant and v_j is the phase-space volume of the jth particle. Following Tolman [16], we take the ‘‘natural’’ volume unit for the phase volume to be h, in accord with $h\sim\delta p\,\delta q$. The product of volumes in Eq. (15) is simply the phase-space volume for the system Ω. Even with the most efficient coordinate system, there can be many programs of the same efficiency. Because of the indistinguishability of particles, the various v_j’s can be interchanged without changing the basic description of the system. This results in $N!$ programs of the same efficiency. Thus Eq. (15) becomes

$$P_{\mathcal{U}}=\frac{N!\,h^N}{\Omega}. \tag{16}$$

If the system can be represented as a microcanonical ensemble then, on average, the phase-space volume will be given by

$$\langle\Omega\rangle=\int\ \delta(E-H(q,p))d^Nq\ d^Np, \tag{17}$$

where H is the Hamiltonian of the system and $d^N q\ d^N p$ is the volume element for the N spatial and N momentum coordinates. With Eqs. (16) and (17), the association between the Kolmogorov probability and the microcanonical partition function $Q(N,E,V)$ is made:

$$\langle P_{\mathcal{U}}^{-1}\rangle = Q(N,E,V) = \frac{1}{N!h^N}\int \delta(E-H(q,p))d^K q\ d^K p. \tag{18}$$

This result is somewhat more general than Zurek's, which simply states that $\langle K\rangle = S$.

Additional generalizations are possible by extending this description to canonical ensembles. Again it is seen that the universal probability is related to the partition function in a form similar to Eq. (18). To develop this relationship, we first consider a microcanonical system consisting of a reservoir or heat bath consisting of M particles and a sample in thermal contact that consists of N particles. The reservoir is such that $M \gg N$. It has a Hamiltonian H_R and the sample has the Hamiltonian H_S.

Because the reservoir is so large we ultimately would not wish to describe its algorithmic complexity along with that the sample. Nevertheless, the complexity of the entire system is given by

$$\begin{aligned} P_{\mathcal{U}}(R,S) &= \left\{\prod_{j=1}^{N} \frac{\delta q\ \delta p}{\mathbf{q}_j \mathbf{p}_j}\right\}\left\{\prod_{j=1}^{M} \frac{\delta q\ \delta p}{\mathbf{q}'_j \mathbf{p}'_j}\right\} \\ &= h^{N+M}(v_1 v_2 \cdots v_N)^{-1}(v'_1 v'_2 \cdots v'_M)^{-1}, \end{aligned} \tag{19}$$

where for simplicity only one of the possible permutations of phase space is considered and the primed and unprimed quantities are associated with the reservoir and sample, respectively. Using the integral representation of the phase-space volume, on average one has

$$\begin{aligned} \langle P_{\mathcal{U}}(R,S)^{-1}\rangle = \frac{1}{h^{N+M}}\int & \delta(E-H_R-H_S) \\ & \times d^M q' d^M p' d^N q\ d^N p, \end{aligned} \tag{20}$$

where we have assumed that the reservoir and the sample interact weakly, so that the combined systems can be described by a Hamiltonian that is the sum of the two, $H_R + H_S$. Performing the integration over the reservoir variables and representing the reservoir partition function by an entropy function, one obtains

$$\langle P_{\mathcal{U}}(R,S)^{-1}\rangle = \frac{1}{h^N}\int e^{S(E-H_S)}d^N q\ d^N p. \tag{21}$$

Because H_S is a small contribution to the reservoir energy, the entropy can be expanded to first order in a power series [17]

$$S(E-H_S) = S_R(E) - (\partial S/\partial H_S)H_S = S_R(E) - \beta H_S. \tag{22}$$

The reservoir entropy is a constant and can be moved outside of the integral in Eq. (21) giving

$$\langle P_{\mathcal{U}}(R,S)^{-1}\rangle = \frac{e^{S_R(E)}}{h^N}\int e^{-\beta H_S}d^N q\ d^N p. \tag{23}$$

Again, because the reservoir is so much larger than the sample, the product of reservoir volumes in Eq. (19) is given by

$$\left\{\prod_{j=1}^{M} \frac{\delta q\,\delta p}{\mathbf{q}'_j \mathbf{p}'_j}\right\} = e^{S_R(E)}. \tag{24}$$

Combining these results and again noting that for indistinguishable particles $N!$ programs of identical length can be generated by considering all permutations of the particles, one has

$$\begin{aligned} \langle P_{\mathcal{U}}(S)^{-1}\rangle &= \sum_i \left\{\prod_{j=1}^{N} \frac{\delta q\,\delta p}{\mathbf{q}_j^{(i)} \mathbf{p}_j^{(i)}}\right\}^{-1} \\ &= \frac{1}{h^N N!}\int e^{-\beta H_S}d^N q\ d^N p. \end{aligned} \tag{25}$$

Thus the Kolmogorov probability is related to the classical canonical partition function for this particular system. In this situation, the algorithmic complexity is now no longer related to the entropy, but rather to the Helmholtz free energy

$$\langle K\rangle = -\frac{\beta A}{\ln 2}. \tag{26}$$

The relationship between algorithmic complexity and thermodynamics depends on the specific ensemble under consideration. The Kolmogorov universal probability, on the other hand, is a more general quantity and is related to the appropriate partition functions. These results can be extended to the grand canonical partition function as well. This is outside the scope of the current paper.

V. THERMODYNAMICS OF SEQUENCE INFORMATION

Using the previous development, the relationship between information content and thermodynamics of protein sequences and protein structures can be explored. Sequence information is often construed as being independent of thermodynamics. As will be seen, this is not the case for protein sequences as they are tied to the thermodynamics of the structure via the shared information. Sequence-structure relationships are of particular importance in theoretical biology because they represent, on a molecular level, the connection between genotype and phenotype. This relationship is generally not thought to have a thermodynamic component, but in cases where the structure and stability of the phenotype are determined by thermodynamics, there will be a thermodynamic constraint on the genotype. A consequence of this is that the evolution of biological information is constrained by the second law of thermodynamics.

The shared Kolmogorov information or algorithmic complexity between two entities A and B is designated as $K(A{:}B)$. It can be related to the joint information $K(A,B)$ and to conditional information $K(A|B)$ and $K(B|A)$ and follows relationships similar to those for the Shannon information (cf. [6]). This shared information is easily visualized

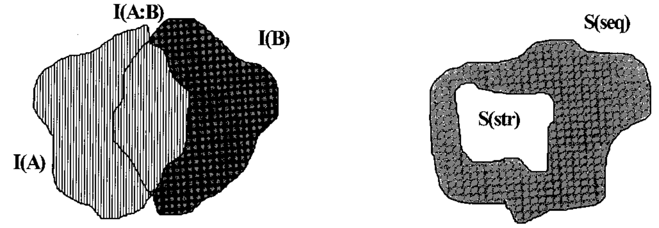

FIG. 1. Symbolic representation of information. (Left) Diagram showing the shared information $I(A{:}B)$ between signal A with information content $I(A)$ and signal B with information content $I(B)$. (Right) Shared information entropy between protein sequences $S(\mathrm{seq})$ and protein structures $S(\mathrm{str})$. Since all the information of the structure is contained in the sequence, $S(\mathrm{str})=S(\mathrm{str{:}seq})$.

diagramatically (see Fig. 1, left). The regions represent an abstract space of binary sequences that give programs to describe the object A or B. The joint information $K(A,B)$ is given by

$$K(A,B)=K(A)+K(B)-K(A{:}B). \tag{27}$$

The joint information also follows a relationship of the form $K(A,B)=K(A)+K(B|A)$, where $K(B|A)$ is the conditional entropy of B given A (it is the area in B that does not overlap with A). Using these relationships, the shared information is given by

$$K(A{:}B)=K(A)-K(A|B)=K(B)-K(B|A). \tag{28}$$

To determine the shared information between a protein's sequence (seq) and its structure (str), $K(\mathrm{seq{:}str})$ must be calculated from conditional entropies. The shared information is given by the counterpart of Eq. (28):

$$K(\mathrm{seq{:}str})=K(\mathrm{str})-K(\mathrm{str}|\mathrm{seq}) \tag{29a}$$

$$=K(\mathrm{seq})-K(\mathrm{seq}|\mathrm{str}). \tag{29b}$$

$K(\mathrm{str})$ has been estimated by the algorithmic complexity as described in Sec. III and a value of approximately 0.5 bits per amino acid is obtained. For the value of $K(\mathrm{seq})$, one can, in this instance, employ the close relationship between algorithmic complexity and the Shannon entropy (cf. [6]). Estimates for the Shannon information entropy put it in the range of 2.0–2.5 bits per amino acid [4]. The conditional information $K(\mathrm{seq}|\mathrm{str})$ and $K(\mathrm{str}|\mathrm{seq})$ is more difficult to estimate. Most physical evidence points to a single, native structure per sequence and this gives $K(\mathrm{str}|\mathrm{seq})=0$. Assuming this to be true for most sequences, one has $S(\mathrm{str{:}seq})=S(\mathrm{str})$. It is also possible to estimate $S(\mathrm{seq}|\mathrm{str})$ from mutagenesis experiments on a single sequence [5] or upon evolutionary changes in a given sequence [2]. Estimates from these methods are extremely variable, but typically are about 2.0 bits per amino acid for $K(\mathrm{seq}|\mathrm{str})$. These again recover the relationship $K(\mathrm{str{:}seq})\approx K(\mathrm{str})$.

These results suggest that all the information in the final structure is shared with that in the protein sequence. The abstract diagram for the information overlap in the protein folding problem is shown in Fig. 1 (right-hand side). In information theory terminology, the protein folding process is an interesting communication channel. It receives no information outside the sequence and therefore appears as a noiseless channel. Yet there is much more information contained in the sequence than is required by the structure. The biological significance of this additional sequence information is unclear at this time. Since parts of this sequence information could be changed without affecting the structure, the additional information could confer a robustness to mutation to the system. Such an effect would be in keeping with experience from mutagenesis studies.

The relationship of Eq. (29) provides a crucial link between the information content of the protein sequence and the thermodynamics of its structure. With the results of Sec. III, Eq. (29) can be rewritten as

$$\langle K(\mathrm{seq})\rangle-\langle K(\mathrm{seq}|\mathrm{str})\rangle=S(\mathrm{str})-S(\mathrm{str}|\mathrm{seq}). \tag{30}$$

This shows that the algorithmic complexity of the sequence can be related to the thermodynamics of the structure. This is an important relationship because it puts a thermodynamic constraint on the change of information during molecular evolution.

To see how the information dynamics parallels the physical evolution of the thermodynamic system, the Kolmogorov probabilities for the quantities in Eq. (30) are introduced in accord with the results of Sec. IV. In this specification, the protein structural parameters are associated with the system energy E and are given a Hamiltonian $H(q,p)$. The sequence parameters are observables $\mathbf{A}(q)$, which are a function only of position q in the protein and not of momentum. A vector $\mathbf{a}=(a_1,a_2,\ldots,a_N)$ is used to specify the identity of the amino acid at each position in a protein that is N amino acids long. Within the microcanonical description, the Kolmogorov probabilities are now given by

$$\langle P_{\mathcal{U}}(\mathrm{str})^{-1}\rangle=e^{S(E)}=\int\delta(E-H(\mathbf{q},\mathbf{p}))d^Np\;d^Nq, \tag{31a}$$

$$\langle P_{\mathcal{U}}(\mathrm{str}|\mathrm{seq})^{-1}\rangle=e^{S(E|a)}$$
$$=\int\delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))\delta(E-H(\mathbf{q},\mathbf{p}))d^Np\;d^Nq, \tag{31b}$$

$$\langle P_{\mathcal{U}}(\mathrm{seq})^{-1}\rangle=e^{K(\mathbf{a})}=\int\delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))d^N\mathbf{a}, \tag{31c}$$

$$\langle P_{\mathcal{U}}(\mathrm{seq}|\mathrm{str})^{-1}\rangle=e^{K(\mathbf{a}|E)}$$
$$=\int\delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))\delta(E-H(\mathbf{q},\mathbf{p}))d^N\mathbf{a}, \tag{31d}$$

where to simplify the notation factors of hN and $N!$ have been dropped and the structural integrals have been taken only to have the same number of degrees of freedom as the number of amino acids in the protein. None of these simplifications will have an impact on the following arguments.

An average or consensus sequence can be defined in two ways, one weighted with respect to protein structural stability and the other with respect to the conciseness of the sequence information or sequence complexity. The consensus sequence based on structure is

$$\langle \mathbf{a}\rangle_{\rm str}=\int \mathbf{A}(q)\,\delta(E-H(\mathbf{q},\mathbf{p}))d^Nq\;d^Np \Big/ \int \delta(E-H(\mathbf{q},\mathbf{p}))d^Nq\;d^Np \tag{32a}$$

$$=e^{-S(E)}\int \mathbf{a}\left\{\int \delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))\times\delta(E-H(\mathbf{q},\mathbf{p}))d^Nq\;d^Np\right\}d^N\mathbf{a} \tag{32b}$$

$$=\int e^{-S(E)+S(E|\mathbf{a})}\mathbf{a}\;d^N\mathbf{a}$$

$$=\int e^{-S(E:\mathbf{a})}\mathbf{a}\;d^N\mathbf{a} \tag{32c}$$

and the consensus sequence based on sequence complexity is

$$\langle \mathbf{a}\rangle_{\rm seq}=\int \mathbf{a}\;\delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))d^N\mathbf{a} \Big/ \int \delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))d^N\mathbf{a} \tag{33a}$$

$$=e^{-K(\mathbf{a})}\int \mathbf{A}(\mathbf{q})\left\{\int \delta(\mathbf{a}-\mathbf{A}(\mathbf{q}))\times\delta(E-H(\mathbf{q},\mathbf{p}))d^N\mathbf{a}\right\}d^N\mathbf{q}\;d^N\mathbf{p} \tag{33b}$$

$$=\int e^{-K(\mathbf{a})+K(\mathbf{a}|E)}\mathbf{A}(\mathbf{q})d^N\mathbf{q}\;d^N\mathbf{p}$$

$$=\int e^{-S(E:\mathbf{a})}\mathbf{A}(\mathbf{q})d^N\mathbf{q}\;d^N\mathbf{p}, \tag{33c}$$

where Eq. (30) is used in Eq. (32c). Equations (32) and (33) show that the probability density $e^{-S(E:\mathbf{a})}$ associated with each type of average is identical and is determined by the mutual entropy. As will be seen, this result is very significant for the time evolution of the system. Although the sequence and structural information are different, observables of the systems are determined from the mutual information. Thus sequence and structural information will not evolve independently, but rather will evolve with the mutual information.

The question to be addressed is whether the structural average $\langle \mathbf{a}\rangle_{\rm str}$ evolves with the same dynamics as the sequence complexity average $\langle \mathbf{a}\rangle_{\rm seq}$. If so, the evolution of the sequence information will mirror the entropic optimization of the structure. To approach this question, the time dependence of the following two correlation functions is considered:

$$\langle a_i(0)a_j(t)\rangle_{\rm str}=\int e^{-S(E:\mathbf{a})}a_i(0)a_j(t)d^N\mathbf{a}, \tag{34}$$

$$\langle a_i(0)a_j(t)\rangle_{\rm seq}=\int e^{-S(E:\mathbf{a})}A_i(0)A_j(t)d^N\mathbf{p}\;d^N\mathbf{q}, \tag{35}$$

where Eq. (34) shows sequence correlations based on structural thermodynamic considerations and Eq. (35) is based on sequence complexity. Presently, it is shown that Eqs. (34) and (35) have identical time dependences. The dynamics of Eqs. (34) and (35) can be revealed using a standard development of the fluctuation-dissipation theorem. The notation of Garrod [17] is followed where the time-dependent quantity $a_i(t,\mathbf{a}^0)$ is expanded about the equilibrium point $\mathbf{a}^0$ to first order in the thermodynamic force $\beta_k(\mathbf{a}^0)$. This gives

$$a_i(t,\mathbf{a}^0)=\sum_k F_{ik}(t)\beta_k(\mathbf{a}^0), \tag{36}$$

where $\beta_k(\mathbf{a}^0)=\partial S(E|\mathbf{a}^0)/\partial a_i$. The fluctuation-dissipation theorem states that [17]

$$F_{ik}(t)=-\langle a_i(0)a_k(t)\rangle_{\rm str}. \tag{37}$$

It is also possible to show that

$$F_{ik}(t)=-\langle a_i(0)a_k(t)\rangle_{\rm seq}. \tag{38}$$

This result shows that the evolution of sequences based on the complexity of the sequence is identical to that based on the thermodynamics of the structure. To obtain Eq. (38), Eq. (36) is substituted into Eq. (35), giving

$$\langle a_i(0)a_k(t)\rangle_{\rm seq}=\int e^{-S(E:\mathbf{a})}A_i(0)A_k(t)d^N\mathbf{p}\;d^N\mathbf{q} \tag{39a}$$

$$=\int e^{-S(E:\mathbf{a})}\delta(\mathbf{a}-A(q))a_i(0)a_k(t)d^N\mathbf{a}\;d^N\mathbf{p}\;d^N\mathbf{q} \tag{39b}$$

$$=\sum_j F_{jk}(t)\int e^{-S(E:\mathbf{a})}\delta(\mathbf{a}-A(q))a_i(0)\beta_j(\mathbf{a}^0)d^N\mathbf{a}\;d^N\mathbf{p}\;d^N\mathbf{q} \tag{39c}$$

$$=\sum_j F_{jk}(t)\int \delta(\mathbf{a}-A(q))a_i(0)\,\frac{\partial e^{-S(E:\mathbf{a})}}{\partial a_j}\,d^N\mathbf{a}\;d^N\mathbf{p}\;d^N\mathbf{q} \tag{39d}$$

$$=\sum_j F_{jk}(t)e^{K(\mathbf{a})}\int \delta(\mathbf{a}-A(q))a_i(0)\,\frac{\partial e^{-K(\mathbf{a}|E)}}{\partial a_j}\,d^N\mathbf{a}\;d^N\mathbf{p}\;d^N\mathbf{q} \tag{39e}$$

$$= \sum_j F_{jk}(t) e^{K(\mathbf{a})} \int A_i(0) \frac{\partial e^{-K(A|E)}}{\partial A_j} d^N\mathbf{p}\, d^N\mathbf{q} \tag{39f}$$

$$= -F_{ik}(t), \tag{39g}$$

where an integration by parts (cf. [17]) is used after Eq. (39d). This result demonstrates, within the approximations of the fluctuation-dissipation theorem, that the sequence complexity evolves with the same time course as the structural stability. Thus the sequence evolution is controlled by the thermodynamics of the structural evolution.

VI. SUMMARY

The sequence of a protein represents the ordering of amino acids along the protein chain. Because of the covalent nature of the bonding, these amino acids do not readily exchange in response to environmental changes in chemical potential. Thus the sequence represents a frozen or embedded structure that does not change during the lifetime of the protein. The nonrandomness of this sequence is a result of a variety of factors such as the thermodynamics of protein folding, the functionality of the protein and, perhaps, genetic factors associated with the DNA coding. The traditional ensemble formulation of statistical mechanics cannot be used to discuss the thermodynamics or evolution of such structures. Rather a formulation based on algorithmic complexity and Kolmogorov probability is required. Such a formulation provides a means to determine the thermodynamics of fixed, known objects. This formulation avoids the concept of probabilistic interpretations of microstates in an ensemble. Using this approach, it is seen that the sequence information and structural thermodynamics are linked by the shared or mutual information. Because this is a thermodynamic quantity, it is seen that protein sequences will evolve under the constraints of the thermodynamics of the structure.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant No. 1R15GM55910.

[1] C. B. Anfinsen, Science **181**, 223 (1973).

[2] H. P. Yockey, *Information Theory and Molecular Biology* (Cambridge University Press, Cambridge, 1992).

[3] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).

[4] B. J. Strait and T. G. Dewey, Biophys. J. **71**, 148 (1996).

[5] T. G. Dewey, Phys. Rev. E **54**, R39 (1996).

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[7] W. H. Zurek, Phys. Rev. A **40**, 4731 (1989).

[8] W. H. Zurek, Nature (London) **341**, 119 (1989).

[9] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[10] E. T. Jaynes, Phys. Rev. **108**, 171 (1957).

[11] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953), p. 495ff.

[12] P. J. Flory, Proc. Natl. Acad. Sci. USA **79**, 4510 (1982).

[13] K. A. Dill, Biochemistry **24**, 1501 (1985).

[14] D. A. Brant, W. G. Miller, and P. J. Flory, J. Mol. Biol. **23**, 47 (1967).

[15] P. Y. Chou and G. D. Fasman, Annu. Rev. Biochem. **47**, 251 (1978).

[16] R. C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, London, 1938), p. 170ff.

[17] C. Garrod, *Statistical Mechanics and Thermodynamics* (Oxford University Press, New York, 1995).