Statistical mechanics of protein sequences

T. Gregory Dewey
Department of Chemistry and Biochemistry, University of Denver, Denver, Colorado 80208
(Received 17 May 1999)

A statistical mechanical treatment of biological macromolecules is presented that includes the sequence information as an internal coordinate. Using a path integral representation, the canonical partition function can be represented as a product of a polymer configurational path integral and a sequence walk path integral. In most biological instances, the sequence composition influences the potential energy of intersubunit interaction. Consequently, the two path integrals are not separable, but rather ''interact'' via a sequence-dependent configurational potential. In proteins and RNA, the sequence walk occurs in dimensions greater than three and, therefore, will be an ideal ''polymer.'' The Markovian nature of this walk can be exploited to show that all the structural information is contained in the sequence. This latter effect is a result of the dimensionality of the sequence walk and is not necessarily a result of biological optimization of the system.
[S1063-651X(99)11910-X]

PACS number(s): 87.10.+e, 05.40.Fb, 87.15.Aa

I. INTRODUCTION

There has been ongoing interest in primitive theoretical models of protein folding. These models have been both analytic (cf. [1–3]) and computational (cf. [4–6], and [7]) and have focused on the minimal requirements that a polymer must have to fold like a protein. Typically, this work has focused on those properties that allow a polymer to have a distinct, well-separated ground-state energy and to be a maximally compact structure. Much of the impetus of this past work came from the development of the theoretical basis for heteropolymer freezing. This revealed the possibility of a phase transition between two compact globular polymeric states. One of these states has an exponentially large number of conformations while the other is characterized by a small number of low energy conformations. Heteropolymer freezing was originally couched in terms of a random energy model [1] and was based upon the analogy with spin glasses [8]. A more general model based on the replica method was subsequently developed [2]. Most recently, these different approaches were melded in a mathematically simplified model [3]. In this work, it was shown that an annealed heteropolymer model could yield results comparable to the random energy model and only require a simple averaging procedure. In the present work, additional properties of an annealed heteropolymer are explored using a path integral formulation.

Proteins in nature have other properties outside of those explored by the primitive models. Many of these properties arise from the interplay between sequence and structure. This sequence phenomenology can be grouped within three main observations. These are: sequence dictates structure, molecular evolution has a stochastic component, and sequence statistics have a Markovian nature. As increasingly sophisticated ''primitive models'' are explored, it is important to incorporate this broader phenomenology into the requirement for a minimal model of a proteinlike polymer.

The underlying premise of the protein folding problem is that sequence dictates structure. This premise has a long history in molecular biophysics and originated with the experimental observations of Anfinsen [9]. In ongoing work, the information content of protein sequences [10], of protein structures [11], and of the shared or mutual information between sequence and structure [12] has been estimated. It has been argued, using information complexity, that the information content of a protein sequence is directly proportional to its configurational entropy. It can also be shown that the information content of the structure is entirely contained in the information of the sequence. Thus, analysis from information theory is consistent with the experimental observation that the information contained in a protein sequence is sufficient to determine its structure.

The second feature of sequence-structure relationships is Kimura's observation of random neutral mutations. Kimura realized that protein sequences evolved at nearly a constant rate, independent of phylogeny. He attributed this to a stochastic evolutionary-neutral process of base substitution [13]. Thus, it appears that most proteins can sustain a significant amount of variation in their amino acid sequence without dramatically altering the structure. Since its inception, the neutral theory of evolution has sparked considerable debate. This theory was countered by a Darwinian selectionist point of view that requires an evolutionary advantage for mutations to become fixed in a population. Subsequent work from both camps have softened the stance on the ''arrow of time'' for evolution [14]. For our present needs, suffice it to say that there is a strong stochastic component to protein evolution.

The third observation on sequence-structure relationships comes from the modeling of sequence statistics. Markov and hidden Markov models (HMM) have been successfully applied to a number of problems of pattern recognition in protein sequences [15,16]. Such models have been used for multiple sequence alignment, modeling of secondary structure, consensus patterns in protein superfamilies and phylogenetic reconstruction. These models are based on first order Markov processes and are frequently successful in capturing the order within a family of sequences. These observations suggest that protein sequences will obey a statistical ''superposition principle'' that reflects their underlying Markovian nature.

In this paper, a path integral formulation for annealed heteropolymers is presented. This formalism is then used to

explore conditions that are consistent with the sequence phenomenology discussed above. In Sec. II, we present a heuristic derivation of the relationship of the Shannon information content of a polymer and its thermodynamic entropy. This section serves to introduce the notations and the information theoretical relationships used throughout the paper. Section III presents the configurational integral of the sequence-structure system using the sequence as internal coordinates of the system. In this section, the concept of sequence space is introduced and the sequence is viewed as a walk or ideal polymer in that space. A path integral representation of the canonical partition function is presented that views the system as one analogous to two separate, interacting polymers. In the present case, one of the polymers is the sequence walk while the second one is the actual polymer. This result is then used to derive general expressions for the entropy of the system and to show the contributions from sequence and polymer configuration (structure). In Sec. IV, the sequence-structure path integral is represented as a sequence path integral with an influence functional. A variational expansion of the path integral is developed and the Shannon information properties are established. In Sec. V, we show how to introduce biological constraints into the formalism. As will be seen, biological constraints effectively act as an external potential. Section VI presents a brief summary of these results.

II. INFORMATION CONTENT OF A POLYMER CONFIGURATION

In this heuristic section, the calculation of the Shannon information entropy for a configurational state of a polymer is presented. As will be seen, this quantity can be related to the thermodynamic entropy. Since the calculation of the Shannon entropy is central to later arguments, this section provides the groundwork for the paper. Following the canonical description of a polymer chain of N units with a set of bond vectors, $\{\mathbf{R}_i\}$, the probability of a given chain configuration is given by [17]

$$P(\{\mathbf{R}_i\}) = \frac{G(\{\mathbf{R}_i\})}{Q}, \tag{2.1}$$

where

$$G(\{\mathbf{R}_i\}) = \exp[-\beta W(\{\mathbf{R}_i\})] \prod_{j=1}^{N} \tau(\mathbf{R}_j), \tag{2.2}$$

and the canonical partition function is

$$Q = \int d\{\mathbf{R}_i\} G(\{\mathbf{R}_i\}). \tag{2.3}$$

The bond probability distribution function is given by $\tau(\mathbf{R}_j)$, and $W(\{\mathbf{R}_j\})$ is the potential energy.

Given a population of polymer configurations, the Shannon information entropy, I, is defined as

$$I = -\int d\{\mathbf{R}_j\} P(\{\mathbf{R}_j\}) \ln_2 P(\{\mathbf{R}_j\}) \tag{2.4}$$

$$= -\frac{1}{Q}\int d\{\mathbf{R}_j\} G(\{\mathbf{R}_j\}) \ln_2 G(\{\mathbf{R}_j\}) + \ln_2 Q. \tag{2.5}$$

The first term on the right-hand side of Eq. (2.5) represents the information content of encoding the walk of the polymer configuration. For an ideal polymer, it is the information content of a Markov chain. The second term is the length of the binary string required to specify the number of configurational states available to the system.

The thermodynamic entropy S for the polymer, can be determined from standard statistical mechanical relationships involving the Helmholtz free energy F:

$$Q = e^{-\beta F}, \tag{2.6}$$

$$S = -\frac{\partial F}{\partial T}. \tag{2.7}$$

With these results, one finds

$$S = -\frac{k}{Q}\int d\{\mathbf{R}_j\} G(\{\mathbf{R}_j\}) \ln(e^{-\beta W(\{\mathbf{R}_j\})}) + k \ln Q, \tag{2.8}$$

where the natural logarithm is now used. Using Gibbs' inequality [18], one obtains

$$S \geqslant kI \ln 2. \tag{2.9}$$

The difference in the Shannon information entropy and the thermodynamic entropy is a term that is proportional to the radius of gyration squared. In some applications, this will simply be a constant and I will be proportional to S.

III. PATH INTEGRAL FOR AN ANNEALED HETEROPOLYMER

A statistical mechanical model of biomolecular structure that formally incorporates sequence information into the configurational statistics of the polymer is presented in this section. The aim of this model is to establish the conditions sufficient to describe the three phenomenological features of sequence-structure relationships. To this end, we introduce sequence information as an internal coordinate to the polymer. This is analogous to a polymer whose units have discrete internal coordinates similar to an N-dimensional spin model. This puts sequence variables on an equal footing with spatial variables. Such a description has been employed previously in a heteropolymer, random energy model of a protein [3]. The simultaneous averaging over spatial and sequence coordinates gives the partition function for an annealed heteropolymer. This annealed polymer model had the attractive feature that the thermodynamic averages do not require the complicated replica averaging procedure found in other disordered systems. Such annealed polymer models show the properties required of primitive protein folding models.

To describe the configuration of such a polymer, one must consider the sequence (or spin) vector, $\mathbf{s}_i$, of each component along with the bond vectors, $\mathbf{r}_i$. The sequence vectors are unit vectors used to describe the chemical identity of the

polymeric unit. They are given by

$$s(1)=\{1,0,\ldots,0\},$$
$$s(2)=\{0,1,\ldots,0\},$$
$$\vdots$$
$$s(m)=\{0,0,\ldots,1\}, \tag{3.1}$$

where m is the length of the alphabet used to describe the sequence and each vector represents an individual amino acid (for proteins) or nucleotide (for RNA). For proteins, $m=20$, for the 20 different amino acids. For nucleic acids, $m=4$, representing the four different nucleotides. Note that this vector representation of sequence composition is different than the description of ''sequence space'' commonly used by evolutionary biologists [19,20]. Sequence space is a high-dimensional space where each possible sequence of an N unit polymer has its own dimension. Thus, for a protein of N units, sequence space will have 20^N dimensions. While this space is convenient for conceptualizing sequence evolution, it does not have practical utility for our present purpose. Instead, a biopolymer sequence is represented as a ''configuration'' of sequence vectors.

The sequence position vector for the jth unit within the polymer can be described in a similar manner as the spatial position vector:

$$\mathbf{S}_j=\sum_{i=1}^{j}\mathbf{s}_i. \tag{3.2}$$

The bond vector can also be represented in terms of position sequence vectors as $\mathbf{s}_j=\mathbf{S}_j-\mathbf{S}_{j-1}$. Using this set of spatial and internal coordinates, the probability of a configuration is given by

$$P(\{\mathbf{R}_i\},\{\mathbf{S}_i\})=G(\{\mathbf{R}_i\},\{\mathbf{S}_i\})/Q, \tag{3.3}$$

with the partition function given as

$$Q=\int d\{\mathbf{R}_i\}d\{\mathbf{S}_i\}G(\{\mathbf{R}_i\},\{\mathbf{S}_i\}). \tag{3.4}$$

The Greens function is defined as

$$G(\{\mathbf{R}_i\},\{\mathbf{S}_i\})=\prod_{i=1}^{N}\tau(\{\mathbf{R}_i\})\tau_s(\{\mathbf{S}_i\})\times e^{-\beta W(\{\mathbf{R}_i\})}e^{-\beta W'(\{\mathbf{R}_i\},\{\mathbf{S}_i\})}, \tag{3.5}$$

with the spatial bond probability distribution function [17]:

$$\tau(\mathbf{R}_i)=\tau(\mathbf{r}_i-\mathbf{r}_{i-1})=\left(\frac{3}{2\pi l}\right)^{3/2}\exp\left(\frac{-3\mathbf{R}_i^2}{2l}\right), \tag{3.6}$$

where l is the bond distance. The sequence ''bond'' probability function is assumed to be of the following form:

$$\tau(\mathbf{S}_i)=\tau_s(\mathbf{s}_i-\mathbf{s}_{i-1})=\left(\frac{m}{2\pi}\right)^{3/2}\exp\left(\frac{-m\mathbf{S}_i^2}{2}\right), \tag{3.7}$$

where the sequence ''bond distance'' is taken as unity. Following the treatment of discrete spatial coordinates, the sequence bond probability function is taken as a continuous Gaussian distribution. The potential terms in Eq. (3.5), $W(\{\mathbf{R}_i\})$ and $W'(\{\mathbf{R}_i\},\{\mathbf{S}_i\})$, are the potential of interactions between polymeric units that are independent and dependent on sequence, respectively. The present model differs from an ''Ising polymer'' in that there is no potential term such as $W''(\{\mathbf{S}_i\})$. This is because there is no physical basis for interactions between amino acids as compositional units. The sequential composition of the biopolymer influences successive units via its effect on the physical configuration, i.e., $W'(\{\mathbf{R}_i\},\{\mathbf{S}_i\})$ and not by any through space interactions as in the spin case.

Considering the discrete nature of the sequence walk, it may appear that Eq. (3.7) is a serious and highly specific assumption. Actually, when used in conjunction with the limiting procedure implicit in the path integral formulation, it provides a standard description of a discrete random walk ([21], see also [22]). One should also bear in mind that the sequence walk need not be self-avoiding. Thus, there is no need to incorporate excluded volume effects into the path integral for this walk.

Instead of working with the configurational probability distribution function, Eq. (3.3), the simpler end-to-end distribution function is considered. This is given by

$$P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)=G(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)/Q, \tag{3.8}$$

with the end-to-end partition function as

$$Q=\int d\mathbf{R}_0 d\mathbf{R}_N d\mathbf{S}_0 d\mathbf{S}_N G(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N). \tag{3.9}$$

Using a path integral representation for the end-to-end Greens function, one has

$$\begin{aligned}G(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)&=\int_{\mathbf{s}(0)=\mathbf{S}_0}^{\mathbf{s}(N)=\mathbf{S}_N}\mathcal{D}[\mathbf{s}(\tau')]\int_{\mathbf{r}(0)=\mathbf{R}_0}^{\mathbf{r}(N)=\mathbf{R}_N}\mathcal{D}[\mathbf{r}(\tau)]\\&\times\exp\left[-1/N\int_0^N\int_0^N d\tau d\tau'\{(2/m)\dot{\mathbf{s}}^2(\tau')\right.\\&\left.+(2/3l)\dot{\mathbf{r}}^2(\tau)\}\right]\\&\times\exp\left[-1/N\int_0^N\int_0^N d\tau d\tau'\{\beta V(\mathbf{r}(\tau))\right.\\&\left.+\beta V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))\}\right],\end{aligned} \tag{3.10}$$

where τ and τ' are the continuous space curves of the polymer configuration and the sequence walk, respectively, and the ''dot'' notation is the derivative with respect to these curves. The potentials $V(\mathbf{r}(\tau))$ and $V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))$ now represent the continuous versions of $W(\{\mathbf{R}_i\})$ and $W'(\{\mathbf{R}_i\},\{\mathbf{S}_i\})$.

Interestingly, the partition function, Eq. (3.9), can be viewed as representing two independent polymers that inter-

act through a potential. One polymer is the actual physical entity in three-dimensional space while the second ''virtual polymer'' is the *m*-dimensional walk in sequence space. When $V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))=0$, there are no interactions and the polymers are independent of each other. In this case, the partition functions are separable:

$$Q=Q_{\mathrm{str}}Q_{\mathrm{seq}}, \tag{3.11}$$

and the entropies and the Shannon information will be independent of each other:

$$S=S_{\mathrm{str}}+S_{\mathrm{seq}}, \tag{3.12}$$

$$I=I_{\mathrm{str}}+I_{\mathrm{seq}}. \tag{3.13}$$

In such a case, there is no shared information between sequence and structure. Note, however, that the sequence information makes a true contribution to the thermodynamic entropy. This thermodynamic contribution is loosely analogous to the residual entropy found is certain crystals when specific molecular configurations are frozen into the system. For considerations of molecular evolution, the important consequence of Eq. (3.12) is that the same thermodynamic pressures to maximize the entropy of the structure will also drive the sequence. In general, $V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))\neq 0$, and this situation is considered in the next section.

IV. SEQUENCE-STRUCTURE INFLUENCE FUNCTIONALS

For most biopolymers, Eqs. (3.11), (3.12), and (3.13) will not hold. The polymer-sequence interaction term ensures that there will be a shared Shannon information, as well as thermodynamic entropy, between sequence and structure. A detailed analysis using Eq. (3.9) would obviously be extremely difficult. Nevertheless, there are certain aspects of this problem which allow some conclusions to be drawn. The main consideration is the nature of the ''sequence polymer'' or walk. Because the dimensionality of the walk is four or higher for biopolymers of interest, the resulting ''sequence polymer'' will be ideal (cf. [22,23]). This is because the probability of a polymer folding back on itself in dimensions greater than three is virtually nonexistent. Therefore, there will be no excluded volume effect. Consequently, the end-to-end probability function can be represented as an ideal polymer, i.e., a Gaussian distribution. Because of this effect, perturbation techniques could be used to treat the potential energy term.

One could develop cluster expansions of the potential using the polymer perturbation approaches described previously [24]. However, in the present work, a variational approach using influence functionals is taken. The Green's function, Eq. (3.10), can be represented in a manner similar to Feynman's influence functionals [25]. Feynman developed the influence functional to describe the interactions of a microscopic system with a heat bath. In the present case, the polymer path integral equivalent of the influence functional represents the interaction of two ''polymers.'' The system consists of the ''virtual polymer,'' given by the sequence walk, interacting with potential $V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))$ with the actual polymer.

The end-to-end Greens function can be represented as

$$G(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)=\int_{\mathbf{s}(0)=\mathbf{S}_0}^{\mathbf{s}(N)=\mathbf{S}_N}\Phi(\mathbf{R}_0,\mathbf{R}_N,\mathbf{s}(\tau'))\times\exp\left[-\int_0^N d\tau'\{(2/m)\dot{\mathbf{s}}^2(\tau')\}\right]\mathcal{D}[\mathbf{s}(\tau')], \tag{4.1}$$

with the influence functional defined as

$$\Phi(\mathbf{R}_0,\mathbf{R}_N,\mathbf{s}(\tau'))=\int_{\mathbf{r}(0)=\mathbf{R}_0}^{\mathbf{r}(N)=\mathbf{R}_N}\mathcal{D}[\mathbf{r}(\tau)]\times\exp\left[-\int_0^N d\tau\{(2/3l)\dot{\mathbf{r}}^2(\tau)\}\right]\times\exp\left[-1/N\int_0^N\int_0^N d\tau d\tau'\times\{\beta V(\mathbf{r}(\tau))+\beta V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))\}\right]. \tag{4.2}$$

The form of the Greens function, Eq. (4.1), is indicative of the sequence walk being driven by a random external force. In this analogy, the force is derived from the physical interunit potentials in the polymer. This formulation is in keeping with Kimura's observation of a stochastic component to sequence evolution. In this case, structurally neutral changes drive an apparent random change in sequences.

The sequence walk described above occurs in a space with dimensionality greater than three. This is the critical dimension for random walks. Walks of dimensionality greater than or equal to four will be ideal random walks [22,23]. While the potential term $V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))$ will provide some strong constraints on possible sequences, these constraints will essentially be potential spikes in a high dimensional space. The sequence walk can avoid forbidden regions without significantly altering its end-to-end statistics. This is a consequence of the high dimensionality of the space.

With these considerations, it is anticipated that the path integral in Eq. (4.1) will result in a Gaussian distribution of the end-to-end sequence vector. The sequence walk can be viewed as an extremely ''high temperature'' walk that will not be strongly influenced by external potentials. This situation is ideal for the application of variational methods [26]. In such an approach, the influence functional will take the following form:

$$\Phi(\mathbf{R}_0,\mathbf{R}_N,\mathbf{s}(\tau'))=\int_{\mathbf{r}(0)=\mathbf{R}_0}^{\mathbf{r}(N)=\mathbf{R}_N}\mathcal{D}[\mathbf{r}(\tau)]\times\exp\left[-\int_0^N d\tau\{(2/3l)\dot{\mathbf{r}}^{2(\tau)}\}\right]\times\exp\left[-\int_0^N d\tau\{\beta V(\mathbf{r}(\tau))+\beta V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\}\right], \tag{4.3}$$

where the trial potential $V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))$ is a potential taken over some average path of the sequence walk. For the present formal development, we need not specify the specific form needed to optimize this potential.

The sequence end-to-end probability function is given by

$$P(S_0,S_N,N)=\int P(R_0,R_N,S_0,S_N,N)dR_0dR_N. \tag{4.4}$$

A perturbative expansion of the potential term gives

$$\langle V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle = \int \int_{\mathbf{r}(0)=\mathbf{R}_0}^{\mathbf{r}(N)=\mathbf{R}_N} \mathcal{D}[\mathbf{r}(\tau)]\beta V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau)) \times \exp\left[-\int_0^N d\tau\{(2/3l)\dot{\mathbf{r}}^2(\tau)+\beta V(\mathbf{r}(\tau))\}\right] d\mathbf{R}_0 d\mathbf{R}_N . \tag{4.5}$$

Because the high-dimensional sequence walk will be ideal, a first order perturbative expansion can be considered [24]. This gives an end-to-end probability function as

$$P(\mathbf{S}_0,\mathbf{S}_N,N) = \frac{e^{-\beta\langle V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle}}{Q} \int_{\mathbf{s}(0)=\mathbf{S}_0}^{\mathbf{s}(N)=\mathbf{S}_N} \times \exp\left[-\int_0^N d\tau'\{(2/m)\dot{\mathbf{s}}^2(\tau')\}\right]\mathcal{D}[\mathbf{s}(\tau')]. \tag{4.6}$$

This distribution function is essentially a random walk path integral multiplied by a potential-specific amplitude term. This function shows the Markovian nature of the sequence walks. Since Eq. (4.6) will obey a Markovian superposition principle, it establishes the connection with the phenomenology of sequence modeling. This result suggests that Markov and hidden Markov behavior of protein sequences will occur when the perturbative expansion in Eq. (4.5) is valid.

The final phenomenological feature to be explored is the relationship between sequence and structural information. If sequence predicts structure, then all of the information contained in the structure is shared with the sequence. The information of the sequence-structure system is given by [12,27]

$$I(R,S) = I(S) + I(R) - I(R{:}S), \tag{4.7}$$

where $I(R,S)$ is the joint information content of the sequence-structure system, $I(S)$ and $I(R)$ are the information contained in sequence and structure, respectively, and $I(R{:}S)$ is the mutual or shared information between sequence and structure. Observations on proteins suggest that $I(R{:}S) = I(S)$. This is equivalent to

$$I(R,S) = I(S). \tag{4.8}$$

The joint Shannon information based on the end-to-end joint distribution function is given by

$$I(R,S) = \int d\mathbf{R}_0 d\mathbf{R}_N d\mathbf{S}_0 d\mathbf{S}_N \{P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N) \times \ln P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)\}, \tag{4.9}$$

and the sequence Shannon information is

$$I(S) = \int d\mathbf{S}_0 d\mathbf{S}_N \{P(\mathbf{S}_0,\mathbf{S}_N,N)\ln P(\mathbf{S}_0,\mathbf{S}_N,N)\}. \tag{4.10}$$

Condition (4.8) is satisfied when

$$\int d\mathbf{R}_0 d\mathbf{R}_N \{P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)\ln P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)\} = \left\{\int P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N) d\mathbf{R}_0 d\mathbf{R}_N\right\} \times \ln\left\{\int P(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N) d\mathbf{R}_0 d\mathbf{R}_N\right\}. \tag{4.11}$$

The use of Eq. (4.6) implies that $\exp\langle -\beta V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle \approx \langle \exp -\beta V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle$. This is equivalent to $\langle V'^N(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle = \langle V'(\overline{\mathbf{s}}(\tau'),\mathbf{r}(\tau))\rangle^N$. Consequently, the equality in Eq. (4.11) can be established by a series expansion. Thus, the third phenomenological feature, sequence dictates structure, is handled by the present model.

The role of sequence information in determining the folded three-dimensional structure of the polymer has historically been considered a salient feature of proteins [9]. It is natural to presume that this feature goes hand in hand with the biological function of the folded state of the protein. While there are doubtless biological driving forces favoring protein sequences that fold into unique structures, this phenomena is not necessarily a result of biological selection. As shown in this work, any polymer with a composition of more than three distinguishable chemical units can have the property of sequence dictating structure. This is a consequence of the high dimensionality of the sequence walk. In the present work, we considered a variational treatment of the polymer statistics that leads to the condition that all the structural information is contained within or shared with the sequence information. Our goal in doing this is not to explicitly solve the configurational partition function. Rather, the formal result is that within the assumptions leading to an accurate variational solution, sequence will dictate structure. Thus, there are physical conditions that can lead to the biologically important observations on protein sequence-structure relationships.

V. INTRODUCING BIOLOGICAL CONSTRAINTS INTO THE FORMALISM

The results of the previous section show that sequence can dictate structure in biopolymers when certain physical conditions are met. This physical explanation is counter to the traditional view that protein evolution was driven by biological structural and functional constraints. In this traditional view, biological pressures forced proteins to be more than collapsed heteropolymers, but rather to assume highly specific three-dimensional structures. To achieve structures with specific functions, proteins consisting of a unique amino acid sequence are required. As a result of this evolution, the information inherent in the sequence is carried over to the structure.

Given the physical model of biological sequences presented in the previous section, the natural question is, how do biological pressures influence sequence walks? At first, it might appear that because of the high dimensionality of the walk, the statistics of protein sequences will be unaffected by external influences. This is true for parameters such as the radius of gyration or end-to-end probability. The vastness of sequence space allows significant constraints to be put on the

system without altering the nature of the sequence walk. Biological constraints placed by evolutionary pressure can be viewed as external to the sequence-structure system. These constraints are fixed by the nature of the organism, the environment, and the individual phenotypes that are being expressed by a given individual. They will rarely be altered once the protein has been optimized through evolution. In this regard, they are similar to systems with ''permanent entanglements.'' Such systems occur when external fields or size constraints are placed on the system. Such a system may be a polymer in a confined space. Similarly, if the system is prepared in such a manner where there are irreversible linkages, as in crosslinked polymers, then the system has fixed constraints.

The specification of a biologically functional macromolecule can yield these sorts of constraints in sequence space. For instance, a DNA binding protein must be positively charged to associate with the negatively charged DNA. In this case, negatively charged amino acids must be effectively excluded from such a protein. This condition will essentially form a wall in sequence space, that the sequence walk cannot penetrate for biological reasons. A second example can be found in enzymes. Typically, there are two or three amino acids at the active site of an enzyme that are inviolate. Altering these amino acids will destroy all enzymatic activity. This condition serves as an external biological constraint that crosslinks or ties down the sequence space trajectory to specific locations. The need for biological activity for a macromolecule creates a complex and varied set of constraints for the sequence space trajectory. Yet, because sequence space is of such a high dimensionality, these constraints will have little effect on the overall statistics of the trajectory.

To incorporate external biological constraints into the formalism, we then follow an approach similar to that used to treat crosslinked polymers [21]. Taking Eq. (3.4) as a starting point, these constraints are included as

$$Q=\int d\{\mathbf{R}_i\}d\{\mathbf{S}_i\}\,\delta(k-K(\{\mathbf{R}_i,\mathbf{S}_i\}))G(\{\mathbf{R}_i\},\{\mathbf{S}_i\}), \tag{5.1}$$

where $K(\{\mathbf{R}_i,\mathbf{S}_i\})$ is a mathematical description of the constraint, and k is the particular value of the constraint in the system. For the case of an enzyme, $K(\{\mathbf{R}_i,\mathbf{S}_i\})$ may represent the specification of the structural and sequence characteristics that the system must have for the enzyme to have a catalytic activity at a biologically viable level. It might include the sequence and structural requirements for the enzyme to have some minimal steady state kinetic property. One can return to the Gaussian limit of the δ function and use the limiting process that defines the functional integral. Since $K(\{\mathbf{R}_i,\mathbf{S}_i\})$ will have a complicated form, this approach will recover a Gaussian integration over a complicated function in $\{\mathbf{R}_i,\mathbf{S}_i\}$.

A conceptually and mathematically more appealing approach is to use the Dirichlet δ function to represent the constraint [21]. The constraint defines regions of configuration and sequence space that cannot be occupied because of loss of biological activity. This definition also allows one to return to the path integral representation. The region that can be occupied is defined as $\mathcal{V}$. The constraint is now defined as

$$\theta(\{\mathbf{r},\mathbf{s}\})=\begin{cases}1 & \{\mathbf{r},\mathbf{s}\}\in\mathcal{V}\\ 0 & \text{otherwise.}\end{cases} \tag{5.2}$$

The Green's function with constraints is now given by

$$G(\mathbf{R}_0,\mathbf{R}_N,\mathbf{S}_0,\mathbf{S}_N,N)=\int_{\mathbf{s}(0)=\mathbf{S}_0}^{\mathbf{s}(N)=\mathbf{S}_N}\mathcal{D}[\mathbf{s}(\tau')]\int_{\mathbf{r}(0)=\mathbf{R}_0}^{\mathbf{r}(N)=\mathbf{R}_N}\mathcal{D}[\mathbf{r}(\tau)]\exp\left[-1/N\int_0^N\int_0^N d\tau d\tau'\{(2/m)\dot{\mathbf{s}}^2(\tau')+(2/3l)\dot{\mathbf{r}}^2(\tau)\}\right]$$
$$\times\exp\left[-1/N\int_0^N\int_0^N d\tau d\tau'\{\beta V(\mathbf{r}(\tau))+\beta V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))\}\right]\prod_{\tau_{\rm seq}}^{N}\prod_{\tau_{\rm str}}^{N}\theta[\mathbf{r}(\tau_{\rm str}),\mathbf{s}(\tau_{\rm seq})]. \tag{5.3}$$

Treating the Dirichlet δ function as a logarithm of exponentials, one recovers an infinite barrier potential function. It goes to infinity outside of $\mathcal{V}$ and is 0 inside of $\mathcal{V}$. The constraint modifies the sequence-structure potential term in the path integral, $\beta V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))$, giving an effective interaction:

$$\int_0^N\int_0^N d\tau d\tau'\,\beta V'_{\rm eff}(\mathbf{s}(\tau'),\mathbf{r}(\tau))$$
$$=\int_0^N\int_0^N d\tau d\tau'\left\{\beta V'(\mathbf{s}(\tau'),\mathbf{r}(\tau))+\frac{1}{N}\ln\theta[\mathbf{r}(\tau_{\rm str}),\mathbf{s}(\tau_{\rm seq})]\right\}, \tag{5.4}$$

where the $\ln\theta[\mathbf{r}(\tau_{\rm str}),\mathbf{s}(\tau_{\rm seq})]$ term provides an infinite well potential.

These arguments show that biological constraints will introduce a correction to the sequence-structure potential term. Yet the arguments of the previous section still hold. The sequence must wind around infinite potential spikes in its space, but nevertheless will preserve its Markovian statistics as a result of the high dimensionality of the space. Thus, considerations of external biological constraints do not effect the main conclusion of this work. Structure is dictated by sequence as a result of the high dimensionality of sequence space.

VI. SUMMARY

In this work, a path integral formulation of annealed polymers is presented. As in many applications of path integrals,

the power of the technique lies more in the formal development than in practical computations. In the present case, it is seen that the sequence walk and the polymer configuration can be treated as separate, but interacting ''polymers.'' This formulation is analogous to Feynman's influence functionals that have been used to treat microscopic systems coupled to heat baths. The heat bath provides a random external force perturbing the system. In the present case, it is argued that because of its high dimensionality, the sequence walk will behave as an ideal ''polymer.'' The influence functional of the structure on the sequence is analogous to a random external field. This provides the theoretical underpinning for Kimura's neutral theory of evolution where there is a strong stochastic component to sequence evolution. Because of the ideal nature of the sequence walk, the ''coupling'' of the structure to the sequence can be treated in a variational fashion. This insures the Markovian nature of the resulting Green's function and the phenomenological connection with the Markov and hidden Markov models of sequence statistics. Finally, when the variational problem is treated in the ''high temperature'' limit, the Shannon information of the structure is contained in the sequence. Thus, the variational approximation presented in this work provides a model of an annealed heteropolymer that is consistent with a broad range of observed protein sequence behavior. The present formalism can then be used to develop a physical theory of molecular evolution. In particular, it provides a statistical mechanical framework for describing rugged adaptive landscapes. Secondly, it can provide a physical interpretation for the transition probabilities obtained from statistical analysis of sequences with Markov models. This will allow, at least formally, a connection between sequence statistics and structure.

ACKNOWLEDGMENT

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

[1] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[2] E. Shakhnovich and A. Gutin, Biophys. Chem. **34**, 187 (1989).
[3] V.S. Pande, A.Y. Grosberg, and T. Tanaka, Biophys. J. **73**, 3192 (1997).
[4] E. Shahknovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[5] H.S. Chan and K.A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[6] V.S. Pande, A.Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. USA **91**, 12 976 (1994).
[7] J.N. Onuchic, P.G. Wolynes, Z. Luthey-Schulten, and N.D. Socci, Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).
[8] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).
[9] C.B. Anfinsen, Science **181**, 223 (1973).
[10] B.J. Strait and T.G. Dewey, Biophys. J. **71**, 148 (1996).
[11] T.G. Dewey, Phys. Rev. E **54**, R39 (1996).
[12] T.G. Dewey, Phys. Rev. E **56**, 4545 (1997).
[13] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, England, 1982).
[14] T. Ohta, BioEssays **18**, 673 (1996); M. Kreitman, *ibid.* **18**, 678 (1996).
[15] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, MA,1998), pp. 9–11.
[16] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, England, 1998).
[17] K.F. Freed, R*enormalization Group Theory of Macromolecules* (John Wiley & Sons, New York, 1987).
[18] A. Isihara, *Statistical Physics* (Academic Press, New York, 1971), pp. 36–40.
[19] M. Eigen, *Steps Toward Life: A Perspective on Evolution* (Oxford University Press, Oxford, England, 1992).
[20] S.A. Kauffman, *The Origins of Order* (Oxford University Press, New York, 1993).
[21] K.F. Freed, Adv. Chem. Phys. **22**, 1 (1972).
[22] R.J. Rubin, J. Chem. Phys. **20**, 1940 (1952); **21**, 2073 (1953).
[23] P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979), pp. 45 and 46.
[24] H. Yamakawa, *Modern Theory of Polymer Solutions* (Harper & Row, New York, 1971).
[25] R.P. Feynman and A.R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, 1965), pp. 321–351.
[26] R.P. Feynman, *Statistical Mechanics, A Set of Lectures* (W.A. Benjamin, Inc., Reading, PA, 1972), pp. 72–96.
[27] T.M. Cover and J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991), p. 20.