Energy Distribution of the Compact States of a Peptide Chain

W. John Wilbur* and Jun S. Liu†

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, and Department of Statistics, Harvard University, Cambridge, Massachusetts 02138

Received October 28, 1993; Revised Manuscript Received February 7, 1994®

ABSTRACT: We assume that the energy of contact between residues of a peptide chain is governed by an interaction matrix and derive a number of relationships between this interaction matrix and the energy spectrum over the compact states of the peptide chain. If the random energy model (REM) with a fixed number of contacts is assumed, the energy spectrum for the compact states of a peptide is known to be Gaussian. This leads to clear relations between the Hamiltonian, the energy spectrum, and the probability of a random peptide folding to a native state. While these developments are of great theoretical interest, it is evident that structural predictions for real proteins require a more detailed Hamiltonian which distinguishes the different types of residue–residue contacts. Here we consider a Hamiltonian which takes the form of an energy matrix and which explicitly defines the energy of the different types of residue–residue contacts. Statistical conditions are discussed for the contact sets of the compact states which again lead to a Gaussian energy distribution as a limiting form for large numbers of contacts. As for the REM, a simple relation exists between the energy matrix and the resulting energy spectrum. This in turn leads to predictions relating the energy matrix and the probability of a native state, and we show how such predictions may be extended to the case where the number of contacts is variable over the set of compact states. We further give the form of the energy matrix that will maximize the probability of a native state when the individual interaction energies obey certain plausible constraints. While these results may be regarded as exact for the limiting Gaussian distributions, we discuss the approximate nature of the results in realistic cases.

I. Introduction

One of the keys to understanding the protein-folding process and predicting tertiary structure from the primary sequence of a peptide is the Hamiltonian describing the energy of interaction between individual amino acid residues. In one of its simplest forms known as the REM, the interaction energy between two contacting residues is assumed to be obtained by randomly sampling from a Gaussian distribution. Such sampling is repeated for each contact in each potential structure to determine the energy spectrum for the set of compact structures. This approach was introduced by Derrida[1] for the study of spin glasses, first applied to proteins by Bryngelson and Wolynes[2] and adapted by Shakhnovich and Gutin.[3,4] If the number of contacts for the compact structures is assumed to be a constant, the REM leads to a Gaussian distribution for the energy spectrum and to predictions of the probability that a random sequence of residues will have a unique native state.[5] Both the energy distribution and the probability of folding bear a simple relation to the variance of the distribution from which random contact energies are sampled. This allows one to manipulate the system in a predictable manner by changing the Hamiltonian. The energy spectrum may be altered and the conditions for the existence of a native structure may be made more or less stringent by altering the variance of the distribution of contact energies.

While the REM allows one to answer some important theoretical questions, as Chan and Dill[6] have pointed out, the investigation of how a particular sequence produces a particular native structure requires in some sense a more refined analysis. One approach to accomplish this is the introduction of a Hamiltonian for contact interactions which takes the form of an energy interaction matrix. In the HP lattice model, introduced by Lau and Dill,[7] a very simple 0–1 matrix is used. On the other hand, several investigators[8,9] have used a statistical analysis of contact frequencies in known crystal structures to assign "statistical" contact free energies, providing an interaction matrix. Because there is still much to be learned about the true Hamiltonian describing contact energies, it is an important question how the properties of an interaction matrix may affect the energy spectrum and in turn the probability of folding for a peptide sequence. It is the aim of this work to investigate this question and to see in what manner the simple relations found in the REM may also hold for this more complex model.

The paper is organized as follows. In section II we note that when all compact states are assumed to have the same number of contacts and these contacts are assumed to be randomly and independently selected, the central limit theorem applies with a resultant Gaussian energy distribution in close analogy to the REM. However, in real proteins the pairs of contacts involving a common residue are not independent. Because such dependencies have a simple form, the variance of the energy distribution may be explicitly calculated. Further, because the dependencies are limited, we are able to prove a central limit theorem showing the limiting energy distribution is still Gaussian when the number of contacts becomes large provided the number of contacts and the number of dependencies between contacts are fixed over all compact structures. When the energy distribution is Gaussian, the probability of a unique native fold is an increasing function of its variance. This implies certain simple relations between the probability of a unique native fold and the energy matrix. In section III we establish conditions under which these relations continue to hold even when the number of contacts is not a constant over the set of compact structures. Finally, in section IV we show how under constraints the energy matrix may be chosen to maximize certain variances of random contact energies. We deem this important because, on the one hand, these variances

* To whom correspondence should be sent at the National Library of Medicine, Building 38A, Room 8S806, 8600 Rockville Pike, Bethesda, MD 20894; Phone 301-496-2475.
† Department of Statistics, Harvard University.
® Abstract published in *Advance ACS Abstracts*, March 15, 1994.

0024-9297/94/2227-2432$04.50/0 © 1994 American Chemical Society

relate to the nature of the energy distribution and the probability of folding to a unique native state, while, on the other hand, information may exist or become available which constrains the energy matrix.

II. A Gaussian Distribution

To carry out energy calculations it is necessary to have definite models of sequences and structures. The set of sequences is just the set of all linear strings of letters of some fixed length from the alphabet of amino acids. The set of structures is somewhat less obvious. There seems to be general agreement among researchers in the area,[6,10] however, that the energy of conformation of a sequence in a particular compact structure may be characterized by the contacts which occur between residues which are not adjacent along the linear string of residues. A model for the compact structures should be as realistic as possible, but not so complicated that calculations become intractable. A key element in our discussion is the observation that the contact sets for different compact structures of a peptide chain have only a random relation to each other.[3,10,11] Thus in some sense the different compact folds of a peptide chain may be considered to be statistically independent structures. One manner in which this might be expressed is what we shall call the independence model of contacts:

Independence Model: The contact sets for different compact structures behave as random independent samples from the set of all possible contacts for a sequence. (1)

This assumption for three-dimensional structures finds support from several different sources. First, for three dimensions the number of nonlocal contacts exceeds the number of local contacts in a structure.[10] Second, theoretical calculations based on random heteropolymers suggest that different low-energy states for such a heteropolymer will have very few contacts in common.[3,4] Finally, actually enumeration of all the compact conformations of a 27-mer on a $3 \times 3 \times 3$ lattice confirms the theoretical calculations.[11]

While these observations have been used to support the application of the random energy model (REM) of Derrida[1] to proteins,[10] we point out that (1) is essentially equivalent to the REM in the case when the number of contacts in a compact structure is sufficiently large. Let the interaction energies of different pairs of residues be governed by a symmetric energy matrix $R = (e_{ij})$ and let $\{p_i\}_{i=1}^n$ represent the probabilities of the n different letters of the alphabet along some sequence, seq. Then a random contact along the sequence will have energy e_{ij} with probability p_ip_j. Let μ and σ denote the mean and standard deviation of this energy distribution. Then if str_c represents a randomly chosen structure with c contacts from the set of all structures and if $E(\mathrm{seq},\mathrm{str}_c)$ represents the energy of the sequence seq when in the configuration str_c, we may conclude that

$$\frac{E(\mathrm{seq},\mathrm{str}_c) - c\mu}{c^{1/2}\sigma} \tag{2}$$

tends to the unit normal distribution as c becomes large. This is a consequence of (1) and the central limit theorem for a sequence of independent identically distributed random variables.[12]

While (1) may in many circumstances be an adequate model on which to base energy calculations, there are systematic effects due to the dependence of contacts which involve a common residue. Note that for a single contact between randomly determined residues we have

$$\mu = \sum_{i,j=1}^{n} p_ip_je_{ij} \tag{3}$$

and

$$\sigma^2 = \sum_{i,j=1}^{n} (e_{ij} - \mu)^2 p_ip_j \tag{4}$$

Now for any residue type j we may write

$$\mu_j = \sum_{i=1}^{n} p_ie_{ij} \tag{5}$$

$$\sigma_j^2 = \sum_{i=1}^{n} (e_{ij} - \mu_j)^2 p_i \tag{6}$$

We next define

$$\sigma_b^2 = \sum_{j=1}^{n} (\mu_j - \mu)^2 p_j \tag{7}$$

It is then not difficult to show that

$$\sigma^2 = \sum_{i=1}^{n} \sigma_j^2 p_j + \sigma_b^2 \tag{8}$$

Now let us suppose the residues i and k both contact the residue j. It is then a relatively simple calculation to show that

$$\sum_{i,j,k=1}^{n} (e_{ij} - \mu)(e_{jk} - \mu)p_ip_jp_k = \sigma_b^2 \tag{9}$$

This reveals what the covariance is between dependent contacts.

To discuss the energy of a compact structure it will be convenient to use the terminology of graph theory. Let a compact structure be represented as a graph $G = (V,L)$ where the set of vertices, V, of the graph is the set of residues in the structure and the set of edges, L, is the set of contacts between nonadjacent residues. Then for a particular contact $l \in L$ we will write the energy associated with l as e_l. The mean and the variance of e_l are, of course, given by (3) and (4). The total energy of the structure G may then be written as

$$E = \sum_{l \in L} e_l \tag{10}$$

If we consider each vertex to be occupied by a residue sampled according to the distribution $\{p_i\}_{i=1}^n$, we obtain a distribution of total energies with mean $|L|\mu$. Note that $|L|$ is the number of contacts denoted by c in (2). The variance of this distribution may also be computed. For each vertex $v \in V$ let d_v denote the degree of v, i.e., the number of residues that contact v. Then we have the following theorem.

Theorem A. For a fixed graph G and residues assigned randomly to vertices according to the probability distribution $\{p_i\}_{i=1}^n$, $\mathrm{var}(E) = |L|\sigma^2 + (\sum_{v \in V} d_v^2 - 2|L|)\sigma_b^2$.

Proof:

$$\mathrm{var}(E) = \sum_{l \in L} \mathrm{var}(e_l) + \sum_{l_1 \neq l_2} \mathrm{cov}(e_{l_1}, e_{l_2})$$
$$= |L|\sigma^2 + \sum_{l_1 \sim l_2} \mathrm{cov}(e_{l_1}, e_{l_2}) \quad (11)$$

where $l_1 \sim l_2$ implies that the two edges have one vertex in common. The covariance for such a pair of edges is σ_b^2 as noted in (9). For a given vertex, v, with degree d_v, the number of pairs of edges with v in common is $d_v(d_v - 1)/2$. Thus the total number of dependent edge pairs is $\sum_{v \in V} d_v(d_v - 1)/2$. Now since $\sum_{v \in V} d_v = 2|L|$, the result follows. QED.

Because the dependencies are a local phenomenon in a structure, as the number of contacts becomes large, the distribution of the total energy approaches the Gaussian. This is a consequence of a result of Stein.[13]

Theorem B. Let $d = \max_{v \in V} d_v$ be bounded by d_0. Then the central limit theorem holds for E as $|L| \to \infty$.

Proof: Let $W = (E - |L|\mu)/(\mathrm{var}(E))^{1/2}$. Then by Corollary X.2 of Stein (1986, p 110),

$$|P(W \leq w_0) - \Phi(w_0)| \leq 2S_1^{1/2}/\mathrm{var}(E) + (1.252S_2^{1/2})/\mathrm{var}(E)^{3/4} \quad (12)$$

where

$$S_1 = E[\sum_{l_1} \sum_{l_2 \sim l_1} \{e_{l_1}e_{l_2} - E(e_{l_1}e_{l_2})\}]^2 \quad (13)$$

and

$$S_2 = E[\sum_{l_1} \{|e_{l_1}|(\sum_{l_2 \sim l_1} e_{l_2})^2\}] \quad (14)$$

and Φ is the cumulative distribution function of the standard normal distribution.

Here we want to compute explicit limits for S_1 and S_2. Since the maximal degree for the vertices is bounded by d_0, the maximum number of edges correlated with a given edge e_l is at most $2d_0$. Hence there are at most $2d_0|L|$ terms in the sum S_1. Now for a term $e_{l_1}e_{l_2}$ with $l_1 \sim l_2$, the maximum number of correlated terms of the form $e_{l_3}e_{l_4}$ with $l_3 \sim l_4$ is $2 \times (3d_0) \times (2d_0) = 12d_0^2$. Since

$$\mathrm{cov}(e_{l_1}e_{l_2}, e_{l_3}e_{l_4}) \leq \sqrt{\mathrm{var}(e_{l_1}e_{l_2})}\sqrt{\mathrm{var}(e_{l_3}e_{l_4})} \leq \beta_4\sigma^4 \quad (15)$$

where $\beta_4 = \max(\mathrm{var}\{e_{l_1}e_{l_2}\})/\sigma^4$, it follows that

$$S_1 \leq (2d_0|L|)(12d_0^2)\beta_4\sigma^4 = 24d_0^3|L|\beta_4\sigma^4 \quad (16)$$

As for S_2, we have

$$S_2 \leq (2d_0)^2|L|\beta_3\sigma^3 \quad (17)$$

where $\beta_3 = \max\{E(e_{l_1}e_{l_2}e_{l_3})|/\sigma^3$. Since $\mathrm{var}(E) \geq |L|\sigma^2$, we conclude that

$$|P(W \leq w_0) - \Phi(w_0)| \leq 2\sqrt{6d_0^3\beta_4}/|L|^{1/2} + 2.504d_0\beta_3^{1/2}/|L|^{1/4} \quad (18)$$

It follows that the normal approximation is valid as $|L| \to \infty$. QED.

This result suggests a modification of (1).

Local Dependence Model: The contact sets for different compact structures behave as random independent samples from the set of all possible contacts for a sequence, except for the dependency that exists between those pairs of contacts involving a common residue. (19)

Both the independence and the local dependence models as here formulated are based on the assumption that the contacts in different compact conformations in a three-dimensional molecule behave as random and independent samples of the set of all possible contacts. While there is evidence in support of this approach,[3,4,10,11] there is also evidence that excluded volume and chain constraints induce a dependency between the different contacts in a compact conformation. Direct simulations on a cubic lattice have revealed the existence of significant geometric constraints, e.g., the blocking of one contact by another.[14] Two questions seem appropriate in regard to the existence of such steric correlations. First, are they of sufficient prevalence to have a significant impact on the overall energy distribution of a peptide? Second, does a central limit theorem still hold when they are accounted for in a model? To the first of these questions we do not know the answer. To the second, we believe that a central limit theorem will still hold because steric constraints are generally limited to the local volume surrounding the contacts producing them. However, the larger the volume in which contact energies have dependencies, the less useful a central limit theorem becomes because the size of the molecule becomes insufficient to observe the limiting effects. Thus it is important to realize that the models we have defined in (1) and (19) are only an approximation to the constraints that operate in the folding of real peptide chains.

While we can only conclude that the energy distribution of the compact states of a peptide is Gaussian in the limit of large numbers of contacts, the large numbers of contacts in actual proteins suggest the approximation may be quite good. Of course, even a seemingly good approximation may not give useful data in the extreme tails of the distribution. However, for the bulk of the energy spectrum the Gaussian approximation would seem to be an adequate description. In this regard we note that the distribution of energies for the contact graph of a real protein appears Gaussian when a statistically derived energy model is used, and random sequences are assigned.[9] This corresponds to the local dependence model with fixed graph, G, as discussed here. Our primary interest in the remainder of the paper is with the folding question and how folding may be affected by the energy distribution. Thus tail approximations are of key importance, and we consider them further in the next section.

III. Mixtures

If we make the assumption that the tail approximations that arise from the central limit theorem as applied to peptide molecules are adequate to study the folding question, then we have a model which in many ways is equivalent to the REM. Thus we shall begin this section assuming the REM. This will enable us to establish several results with rigorous proofs. We will then consider in some detail the validity of the results obtained for the case of a discrete interaction matrix.

Let STR denote the set of compact structures possible for a sequence seq and let $\Lambda = \{E_1, E_2, ..., E_N\}$ denote the associated set of energies. Let us assume that E_j is the lowest energy in the set. Then following Shakhnovich and Gutin,[5] we say that seq folds if and only if

$$e^{-E_j/(kT)}/Z \geq p_{cut} \tag{20}$$

where the partition function, Z, is defined in the usual manner. The value of p_{cut} is not well defined. Shakhnovich and Gutin[5] suggest the value of 0.99 as a reasonable cutoff to define a sequence with a native structure. Clearly the value must be larger than 0.5 so that to a folder corresponds a unique structure, yet sufficiently less than 1.0 so that it is satisfied by a reasonable number of sequences at reasonable temperatures. Beyond these considerations the exact value is somewhat arbitrary.

Now if the energy distribution of the compact states of a peptide chain is Gaussian, then a simple consequence is that the probability of folding to a unique structure increases as the variance of this distribution increases. Let us assume that the REM with a fixed number of contacts c is applicable to all structures in STR and a given sequence seq. Then we note that the difference in energy between different structures or states is directly proportional to σ, the standard deviation of the contact energy. From (20) it then follows that the probability of folding is an increasing function of σ (in this regard see also the results of Shakhnovich and Gutin[5]). This statement is also true under somewhat more general conditions in which the number of contacts is not required to be a constant. On the basis of the folding criterion (20), we have the following result.

Theorem C. Assume the REM is applicable and let Δc represent the maximum difference in number of contacts between any two structures in STR. Then if

$$|\Delta c\mu| \leq kT \ln\left(\frac{p_{cut}}{1-p_{cut}}\right) \tag{21}$$

is satisfied, the probability of folding is a nondecreasing function of σ.

Proof: For a given σ, let the energies corresponding to seq in the conformations of STR be written as $\Lambda^\sigma = \{E_1^\sigma, E_2^\sigma, ..., E_N^\sigma\}$, where each E_i^σ is assumed to be the sum of c_i pair interaction energies. Because the independence model is assumed, each E_i^σ may be regarded as a normal random variable with mean $c_i\mu$ and variance $c_i\sigma^2$. If E_j^σ is the lowest energy in Λ^σ, then the probability of folding is defined as

$$P_\sigma(\text{fold}) = P(e^{-E_j^\sigma/(kT)}/Z^\sigma \geq p_{cut}) \tag{22}$$

We introduce a sample space Ω consisting of all possible peptide chain energy assignments. Then each E_i^σ may be regarded as a function from Ω to the real line, $E_i^\sigma(\omega)$. We may write $E_i^\sigma(\omega) = c_i\mu + c_i^{1/2}\sigma\xi_i(\omega)$, where the ξ_i's are i.i.d. normal random variables with mean zero and variance one. We want to prove that the set

$$A_\sigma = \{\omega | e^{-E_j^\sigma(\omega)/(kT)}/Z^\sigma(\omega) \geq p_{cut}\} \tag{23}$$

increases as σ increases where again we assume that $E_j^\sigma(\omega)$ is the lowest energy in the set corresponding to ω and σ.

Let us consider a particular σ' and any realization $\omega \in A_{\sigma'}$. Since $Z \geq e^{-E_j/(kT)} + e^{-E_i/(kT)}$ for any $i \neq j$, we have

$$kT \ln\left(\frac{p_{cut}}{1-p_{cut}}\right) \leq E_i^{\sigma'}(\omega) - E_j^{\sigma'}(\omega) = (c_i - c_j)\mu + \{c_i^{1/2}\xi_i(\omega) - c_j^{1/2}\xi_j(\omega)\}\sigma' \tag{24}$$

Because of condition (21), we see that $c_j^{1/2}\xi_j(\omega) \leq c_i^{1/2}\xi_i(\omega)$ for all i and any $\omega \in A_{\sigma'}$. Hence for any other $\sigma'' > \sigma'$ and the same $\omega \in A_{\sigma'}$,

$$E_i^{\sigma''}(\omega) - E_j^{\sigma''}(\omega) \geq E_i^{\sigma'}(\omega) - E_j^{\sigma'}(\omega) \geq 0 \tag{25}$$

That is, $E_j^{\sigma''}(\omega)$ is still the smallest energy and the differences between it and any other E_i increase as the variance increases. We have proved that for any $\omega \in A_{\sigma'}$, it is also true that $\omega \in A_{\sigma''}$ whenever $\sigma'' > \sigma'$. Thus the conclusion of the theorem is established. QED.

Theorem C has a corollary that is of interest.

Corollary 1. If the REM is applicable, then without restriction on the numbers of contacts in structures, the probability of folding converges to one as the variance σ^2 goes to infinity.

Proof: Let us assume the same sample space Ω and representation of $E_i^\sigma(\omega)$ in terms of c_i, σ, and $\xi_i(\omega)$. Then for almost all ω there must be some j such that $c_j^{1/2}\xi_j(\omega) < c_i^{1/2}\xi_i(\omega)$ for any $i \neq j$. It follows that for almost all ω there is some j such that for any $i \neq j$, $E_j^\sigma(\omega) - E_i^\sigma(\omega)$ goes to negative infinity as σ^2 goes to infinity. Because there are only a finite number of structures, the result follows.

Example. Here we give an example showing that the condition (21) of Theorem C cannot be eliminated. Let $N = 2$. Then

$$P(\text{fold}) = P(|E_1 - E_2| \geq B) \tag{26}$$

where $B = kT \log[p_{cut}/(1 - p_{cut})]$. With simple algebraic manipulations we find that

$$P(|E_1-E_2|\geq B) = 1 - P(|E_1-E_2|<B) = 1 - \left\{\Phi\left(\frac{B-\Delta c\mu}{(c_1+c_2)^{1/2}\sigma}\right) - \Phi\left(\frac{-B-\Delta c\mu}{(c_1+c_2)^{1/2}\sigma}\right)\right\} \tag{27}$$

Evidently if condition (21) fails the two arguments for Φ are both nonzero and of the same sign and as a consequence as $\sigma \to 0$ the probability of folding must go to one. Since by Corollary 1 the probability of folding must also go to one as $\sigma \to \infty$, it is evident that the probability of folding cannot be a nondecreasing function of σ. On the other hand, if (21) is satisfied, the arguments must both be nonzero and of the same sign or one argument may be zero. In any of these cases the probability of folding is a strictly increasing function of σ.

Case of a Discrete Interaction Matrix. We turn now to the question whether Theorem C might also hold when energies arise from a discrete interaction matrix. To begin we analyze the proof of Theorem C. Let the average contact energy, μ, and the number of contacts, c, be fixed and let σ' and σ'' denote particular variances of contact energies satisfying

$$\sigma' < \sigma'' \tag{28}$$

Denote the cumulative energy distributions corresponding to σ' and σ'' by $F_{\sigma'}$ and $F_{\sigma''}$, respectively. To each cumulative distribution there is a centered distribution with mean zero defined by the relationship

$$FC_\sigma(E) = F_\sigma(E + c\mu) \tag{29}$$

The two centered distributions corresponding to σ' and σ'' are related by

$$FC_{\sigma'}(E) = FC_{\sigma''}(f_{\sigma'\sigma''}(E)) \tag{30}$$

where under the conditions of Theorem C the transformation $f_{\sigma'\sigma''}$ takes the particularly simple form

$$f_{\sigma'\sigma''}(E) = \frac{\sigma''}{\sigma'}E \tag{31}$$

The two important properties of $f_{\sigma'\sigma''}$ that we need in the proof of Theorem C are that $f_{\sigma'\sigma''}$ is an expanding mapping and is independent of the number of contacts, c. By an expanding mapping we mean that its derivative is everywhere greater than or equal to one so that energy differences always increase under the mapping.

We apply the two properties of $f_{\sigma'\sigma''}$ just described to see how they yield a proof of Theorem C. First we note that just as in the proof of Theorem C we may derive the relation

$$kT \ln\left(\frac{p_{\text{cut}}}{1-p_{\text{cut}}}\right) \le E_i^{\sigma'}(\omega) - E_j^{\sigma'}(\omega) \tag{32}$$

which is a part of (24). This coupled with (21) then yields the relation

$$(E_i^{\sigma'}(\omega) - c_i\mu) - (E_j^{\sigma'}(\omega) - c_j\mu) = d' \ge 0 \tag{33}$$

We proceed to apply the transformation $f_{\sigma'\sigma''}$ to (33) to obtain the relation

$$(E_i^{\sigma''}(\omega) - c_i\mu) - (E_j^{\sigma''}(\omega) - c_j\mu) = d'' \ge d' \tag{34}$$

which follows from the expanding nature of the mapping and the fact that it applies independent of contact number. But (33) and (34) together easily imply (25) and the theorem follows.

Examination of the argument just given reveals that the only place the Gaussian nature of the energy distributions is used is in (31) to obtain the fact that f expands the energy scale uniformly. Now an f satisfying (30) is guaranteed for any two continuous distributions. The fact that it must apply to a range of values for c we believe need not be too restrictive provided the variation in contact number is small relative to the number of contacts. Theorem C or an approximation to Theorem C will follow if f simply expands the energy scale whether uniformly or not.

Let us now assume the independence model of contacts as in (1). From the discussion of the previous section it is evident that in the ideal limit the energy distribution would be Gaussian and Theorem C would apply. Presumably at some point in the process of taking the limit dictated by the central limit theorem the approximation will become adequate to ensure that f is expanding and an approximate version of Theorem C will follow.

To illustrate the kind of behavior we envision, we examined the binomial distribution. This is a discrete distribution similar in construction to the discrete energy distributions that we are concerned with yet simple enough that exact calculations can reveal the behavior in the extreme tails of the distribution. A central limit theorem applies but extreme tail approximations can be poor. To compare the energy scaling for different distribution functions, we define a quantity we term the *density* for a discrete distribution. The relationship to the continuous case we have been discussing is that an expanding transformation corresponds to a transformation that decreases the density of a distribution. Let $\{p_n\}_{n=1}^{\infty}$ denote successive values in a tail of the distribution. Then for any positive integer, k, we set

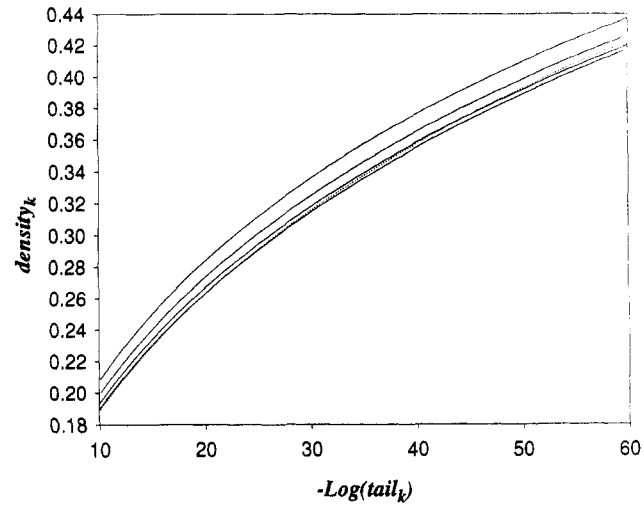


Figure 1. Relative energy densities in the extreme tails of several different binomial distributions, all with an N of 1600. The top curve is for a p of 0.75 followed successively below by the solid curves corresponding to 0.7, 0.65, and 0.6. This progression of decreasing density corresponds to increasing σ. The dotted curve is for a p of 0.5 and shows that the progression is not perfect.

$$\text{tail}_k = \sum_{n=k}^{\infty} p_n$$

$$\text{density}_k = p_k/\text{tail}_k \tag{35}$$

To picture the density of energies of a distribution we graph density$_k$ against–log(tail$_k$), and two such graphs for different distributions allow the relative scaling of energy densities to be compared. Figure 1 is such a comparison of the extreme lower tails of the binomial distribution with an N of 1600 and several different choices for p corresponding to different variances. It is evident from the figure that, while not perfect, there is yet a strong tendency for higher variance to produce a lower energy density. This is true even though the value tail$_{1050}$ of the curve for a p of 0.75 (corresponding to an abscissa of about 17 on the graph) has a normal approximation with continuity correction that is too small by an order of magnitude.

Another way in which the variance of a binomial distribution may be increased is to simply scale up the energy difference between success and failure. This, however, will give perfect density scaling and requires no analysis. We may conclude that at least for the binomial distribution there is a strong tendency to see the kind of energy density scaling required by Theorem C. While Theorem C and its corollary are stated for the REM, the statistical success of threading energies in matching sequences with their native structures suggest a discreteness and specificity in the contact energies. Thus we believe the question of Theorem C's applicability to the discrete energy distributions arising from an interaction matrix is important. Further elucidation of this question is needed. We suggest it will depend on finding methods of examining the extreme tails (20 or 30 standard deviations from the mean) of discrete energy distributions arising from a complex interaction matrix.

If Theorem C is assumed applicable to a molecule with a discrete interaction matrix, the question naturally arises as to the limit placed on Δc in realistic cases. Let us use the value of 0.99 for p_{cut} suggested by Shakhnovich and Gutin.[5] Then the right side of (21) is $4.6kT$. We have estimated $|\mu|$ to be less than $0.01kT$ using contact energies based on protein threading[9] and a typical chain compo-

Table 1. Number of Residues and Contacts for Two Groups of Monomers[a]

	chain length	contact number
1ALC	122	1698
3RN3	124	1805
2LZT	129	2026
2APR	325	6176
4APE	330	6366
2LIV	344	6300

[a] Date from Stephen Bryant (personal communication) and contact numbers computed by him using a threading model.[9] Within each group of three molecules, the chain length is comparable and the contact numbers have a variation less than 400.

sition. This yields a limit of about 500 on Δc. Using the same threading model used to estimate contact energies,[9] Bryant (personal communication) has supplied the numbers of contacts for several peptide monomers from the PDB. These are contained in Table 1. The variation in contact number within groups of similar chain length is less than 400. This suggests that folding may take place under conditions making Theorem C relevant.

IV. Maximal Variance under Energy Constraints

We have seen that under appropriate conditions the problem of maximizing the probability of folding is equivalent to the conceptually simpler problem of maximizing the variance of the energy distribution. If the independence model is assumed, one is in turn concerned with the maximization of σ^2 for a random contact. While the maximization of σ^2 is not meaningful in general, it does become a meaningful problem if the possible interaction energies are bounded. As a convenient way to bound interaction energies, we shall assume that μ is fixed in value. With this assumption we may bound the interaction energy distribution in two different ways which allow a statement in regard to the maximum value of σ^2.

Theorem D. If μ is constant and $e_{ij} \leq 0$ for each contact type i,j, then σ^2 achieves its maximum when all but one of the e_{ij} are zero.

Proof: Suppose that in order to achieve the maximum only m of the e_{ij} may be nonzero and let us rename these contact energies as $\{r_i\}_{i=1}^{m}$. If r_k stands for e_{ij}, let q_k equal p_ip_j if $i = j$ or $2p_ip_j$ if $i < j$. Then we may set

$$\sum_{i=1}^{m} q_i = \gamma \leq 1 \tag{36}$$

and note that

$$\sum_{i=1}^{m} q_i r_i = \mu \tag{37}$$

Let us suppose $m > 1$. We seek an extremal value of

$$\sigma^2 = \sum_{i=1}^{m} q_i(r_i - \mu)^2 + (1 - \gamma)\mu^2 \tag{38}$$

with the constraint specified by (37). This constraint may be removed by solving (37) for r_m

$$r_m = \frac{1}{q_m}(\mu - \sum_{i=1}^{m-1} q_i r_i) \tag{39}$$

and substituting the result into (38). When this is done an interior extremal value must satisfy the relations

$$\frac{\partial \sigma^2}{\partial r_i} = 2q_i(r_i - \mu) + 2q_m(r_m - \mu)\left(\frac{-q_i}{q_m}\right) = 0, \quad 1 \leq i < m \tag{40}$$

But this is only possible if for each $i < m$, r_i is equal to r_m. This does indeed correspond to an extremal value of σ^2, namely, its unique minimum of zero. The maximum could only occur on the boundary of the set. From the hypothesis of the theorem and the relation (39) it is evident that the boundary of the set of possible points is just the set of points for which some one of the r_i is zero. It follows that we may remove some one of the r_i from our set and reduce m by one without impeding our search for the maximum. The argument may then be repeated. This may be continued until $m = 1$. QED.

Corollary 1. Under the hypothesis of Theorem D, if for some fixed i, $p_i \leq p_j$ for all $j \neq i$, then σ^2 achieves a maximum of $\mu^2(1/p_i^2 - 1)$ when $e_{ii} = \mu/p_i^2$ and all other e_{ij} are zero.

Proof: This result is obtained by direct comparison of the different possible candidates for the maximum of σ^2 as described in Theorem D.

A similar proof to that for Theorem D may be used to establish the following theorem.

Theorem E. If μ is constant and $\alpha_{ij} \leq e_{ij} \leq \beta_{ij}$ for each contact type i,j, then σ^2 achieves its maximum when for all but one of the e_{ij}, $e_{ij} = \alpha_{ij}$ or $e_{ij} = \beta_{ij}$.

To deal with the case of the local dependence model we have the following result.

Theorem F. Suppose μ is constant and $e_{ij} \leq 0$ for each contact type i,j and further suppose that for some fixed i, $p_i \leq p_j$ for all $j \neq i$. Then σ_b^2 achieves a maximum of $\mu^2(1/p_i - 1)$ when $e_{ii} = \mu/p_i^2$ and all other e_{ij} are zero.

Proof: We first recall the definition of σ_b^2 from (7) and note the constraint

$$\mu = \sum_{j=1}^{n} \mu_j p_j \tag{41}$$

on the μ_j. We then treat the μ_j as though they were independent except for the relation (41). This allows us to mimic the proof of Theorem D to conclude that σ_b^2 achieves its maximum when all but one of the μ_j are zero. As in Corollary 1 to Theorem D, a direct calculation shows that this maximum is $\mu^2(1/p_i - 1)$ and that it is achieved when $\mu_i = \mu/p_i$ and all other μ_j are zero. But this latter condition is fulfilled exactly when $e_{ii} = \mu/p_i^2$ and all other e_{ij} are zero. QED.

Example. To illustrate some of our results we shall consider the case of a 27-mer where the compact states have been taken to be all conformations possible on a cubic lattice three residues on an edge. Contacts are counted between each pair of nodes which are adjacent to each other in the x, y, or z direction but not adjacent on the chain. Simple enumeration shows that c is 28. Further there must be at least 38 pairs of dependent contacts. This latter observation comes from the fact that Hamiltonian paths on the $3 \times 3 \times 3$ cube must begin and end at either vertices (corners) or the centers of faces.[11] The smallest number of dependent pairs, namely, 38, occur when such a path both begins and ends at vertices. Thus we can write

$$\sigma^2(E) = 28\sigma^2 + 76\sigma_b^2 \tag{42}$$

where we understand that the number 76 is a slight underestimate in some of the cases. Let us consider the HP lattice model and assume that the H–H interaction

energy is −1 while the H–P and P–P interactions are zero. Denote the probability of an H by p_1 and assume that $p_1 \leq 1/2$ is satisfied. Then by Theorems D and F the model is so formulated as to maximize both σ^2 and σ_b^2 among models with the same mean interaction energy and with positive interaction energies excluded. Under these circumstances it is not difficult to show that

$$76\sigma_b^2/(28\sigma^2 + 76\sigma_b^2) = 76p_1/(28 + 104p_1) \qquad (43)$$

If p_1 is 1/3 it is evident from (43) that at least 40% of the variance of the energy spectrum is due to σ_b^2, i.e., to the dependence between contacts with a common residue. If p_1 is increased to 1/2 this fraction rises to 47.5% or almost half of the variance. Because variances are maximized, the HP model thus formulated is a candidate to maximize the probability of folding to a unique native structure. Note that here all structures considered have exactly 28 contacts; thus we might conclude that the probability of folding is maximized provided the tail approximations obtained from Theorem B are sufficiently accurate. While this seems unlikely, the argument given in section III, which suggests that the density of energy states in the tail of the distribution scales inversely with the variance of a contact energy, we believe is applicable. From this we conclude that maximum variance is likely to confer at least a near-maximum folding probability. Such a conclusion may relate to real peptides. When one notes that a statistically derived energy matrix has a dominant "hydrophobic" component,[9] our results seem to support the view that hydrophobicity not only is responsible for the collapse of a peptide chain to a compact state but also plays a significant role in its assumption of a unique native state.[6]

For an analysis of the statistical properties of chain foldings on a cubic lattice we refer the reader to Chan and Dill.[14] Likewise a detailed analysis of the thermodynamics of the folding transition applicable to two-letter models such as HP may be found in Sfatos, Gutin, and Shakhnovich.[16]

Acknowledgment. The authors would like to thank Stephen Bryant and the referees for reading the paper and making a number of helpful suggestions.

References and Notes

(1) Derrida, B. *Phys. Rev.* **1981**, *B24*, 2613.
(2) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524.
(3) Shakhnovich, E. I.; Gutin, A. M. *Biophys. Chem.* **1989**, *34*, 187.
(4) Shakhnovich, E. I.; Gutin, A. M. *J. Phys.* **1989**, *A22*, 1647.
(5) Shakhnovich, E. I.; Gutin, A. M. *Nature* **1990**, *346*, 773.
(6) Chan, H. S.; Dill, K. A. *Phys. Today* February 1993, 24–32.
(7) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986.
(8) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534.
(9) Bryant, S.; Lawrence, C. E. *Proteins* **1993**, *15*, 92.
(10) Karplus, M.; Shakhnovich, E. I. *Protein Folding: Theoretical Studies of Thermodynamics and Dynamics*; Creighton, T., Ed.; W. Freeman: New York, 1992; pp 127–195.
(11) Shakhnovich, E. I.; Gutin, A. M. *J. Chem. Phys.* **1990**, *93* (8), 5967.
(12) Larson, H. J. *Introduction to Probability Theory and Statistical Inference*, 3rd ed.; John Wiley & Sons: New York, 1982.
(13) Stein, C. *Approximate Computation of Expectations*; IMS: Hayward, CA, 1986.
(14) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1990**, *92*, 3118.
(15) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. C. Protein Data Bank. In *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Bonn, Chester, Cambridge, 1987; p 107.
(16) Sfatos, C. D.; Gutin, A. M.; Shakhnovich, E. I. *Phys. Rev. E* **1993**, *48*(1), 465–475.