Why Do Protein Architectures Have Boltzmann-Like Statistics?

Alexei V. Finkelstein, Azat Ya. Badretdinov, and Alexander M. Gutin
Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

ABSTRACT **A theoretical study has shown that the occurrence of various structural elements in stable folds of random copolymers is exponentially dependent on the own energy of the element. A similar occurrence-on-energy dependence is observed in globular proteins[1] from the level of amino acid conformations to the level of overall architectures. Thus, the structural features stabilized by many random sequences are typical of globular proteins while the features rarely observed in proteins are those which are stabilized by only a minor part of the random sequences.**
© 1995 Wiley-Liss, Inc.

Key words: **protein structure, stability, statistics, conformational temperature of protein statistics, melting temperature, random sequences**

INTRODUCTION

It is well known that low-energy elements occur more frequently than others in 3D structures of globular proteins,[1,2] and that the observed occurrence-on-energy dependence resembles Boltzmann statistics:

$$OCCURRENCE \sim \exp(-ENERGY/RT_*) \quad (1)$$

Here T_* is the "conformational temperature"[1] of protein statistics, which is close to room temperature, and R is the gas constant. These statistics were observed for different protein structure elements (Fig. 1): for the occurrence of various ϕ, ψ, χ angles;[1] for the distribution of residues between the surface and interior of globules,[3] as well as between different secondary structures and between different points within a secondary structure,[4,5] for the occurrence of *cis* and *trans* prolines,[6] of ion pairs,[7] of empty cavities[8] in protein globules, etc. This empirical relationship is never perfect, but is so common that it is even used to estimate interaction energies.[9]

A similar phenomenon is also observed at the level of overall protein architectures. The most common folding patterns[10–15] are those which have no "defects" such as, e.g., crossed loops (Fig. 2). The cost of these defects is 2–5 kcal/mol.[13] This cost could be covered by some lucky mutations,[15] but this is rarely done; this cost is small compared with the total energy of interactions in a protein globule (hundreds of kcal/mol)[16,17] and with the free energy reserve of protein stability under physiological conditions ($\approx$10 kcal/mol).[16] However, the defect cost is large in comparison with $RT_* \approx 0.6$ kcal/mol, and this suggests that Eq. (1) could account for their rare occurrence in protein structures.

Thus, Boltzmann's equation (1) can be used to connect, at least qualitatively, the occurrence of various structural elements in globular proteins with the energies of these elements.

However, the phenomenological similarity of protein and Boltzmann statistics is puzzling—that of protein cannot be a direct consequence of the other as their physical sense is hardly comparable.

Boltzmann statistics describe an equilibrium distribution of weakly interacting particles between their ground and excited states (for example, it describes the decrease of air density with elevation). This distribution is maintained by the equilibrium of *transitions* between the different states of each particle, and thus the number of particles in each of the states is proportional to the part of the time which one particle spends (in the long run) in each of the states depending on the energy of this state and the temperature of the medium.[18]

Hence, Boltzmann statistics of thermodynamic systems are applicable to an ensemble of particles as well as to each of them separately.

On the contrary, the observed Boltzmann-like form [Eq. (1)] of protein statistics is applicable *only to an ensemble* of protein structure elements but *not to each* of them separately.

For example, Leu-72 of sperm whale myoglobin is inside the globule,[19] and it is *never*, even in the long run, observed in any other position in any native sperm whale myoglobin molecule, irrespective of the fact that a quarter of all the leucines observed in proteins are located on the surface (cf. Fig. 1a).

Received November 29, 1994; revision accepted May 19, 1995.

Address reprint requests to Alexei Finkelstein, Institute of Protein Research, 142292, Pushchino, Moscow Region, Russian Federation.

Present address of A.M. Gutin: Department of Chemistry, Harvard University, Cambridge, MA 02138.

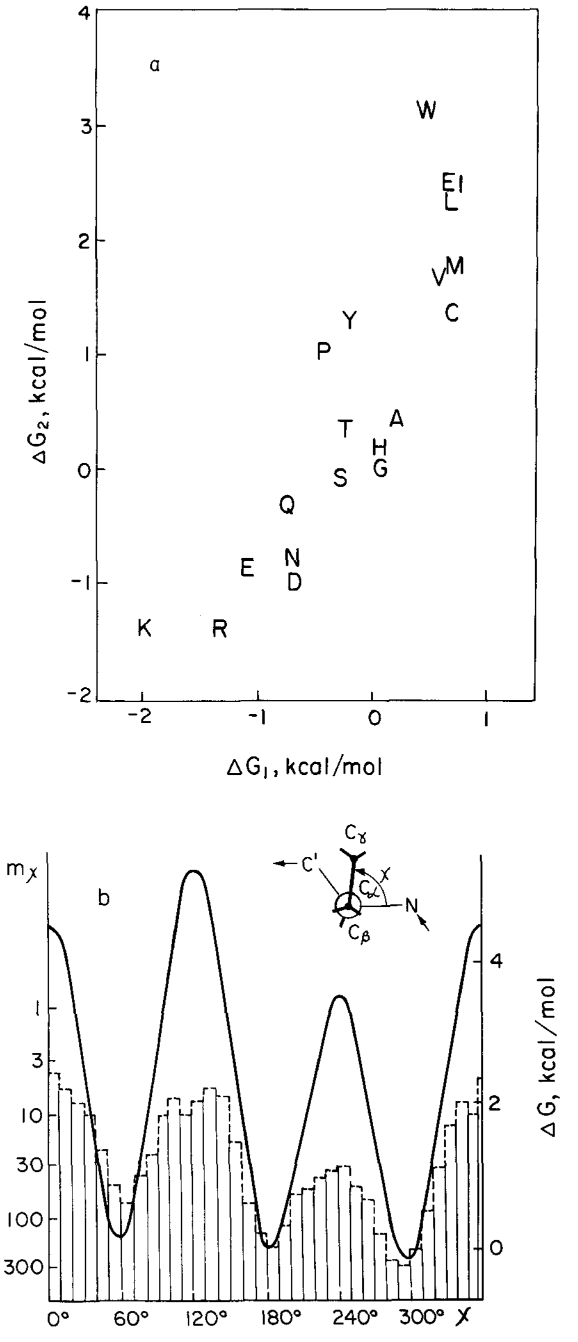

Fig. 1. (**a**) Correlation between transfer free energies.[3] $\Delta G_1 = -RT \ln f$: surface/interior transfer "free energy" from protein statistics. $RT = 0.6$ kcal/mol at 300 K. The partition coefficient $f = (N_s/\Sigma N_s)/(N_b/\Sigma N_b)$, N_s and N_b being the numbers of surface and buried residues of a given type. ΔG_2: experimental water/organic solvent transfer free energy. Amino acids are designated in the one-letter code. (**b**) Correlation between torsional energies. Broken line: $\Delta G_1 = -RT \ln(m_\chi)$: the "potential" from protein statistics; m_χ is the number of side chains with one C_γ atom which have $\chi_1 = \chi \pm 5°$ (statistical data are taken from McGregor et al.[35]). Solid line: ΔG_2: energy of the $C_\gamma H_3$ side chain computed with potentials of De Santis and Liquori.[36] ΔG_1 (180°) and ΔG_2(180°) are taken as zero. Both plots show a definite (though, not perfect) proportion between energy and the logarithm of occurrence. However, an estimate of conformational temperature T_*, which one can extract from these plots ($T_* \approx 350$–600°K) is rather rough, as (1) statistics in (**a**) are biased to definition of two states of a residue: "buried" and "surface"; (2) statistics in (**b**) are biased to refining programs of X-ray analysis.

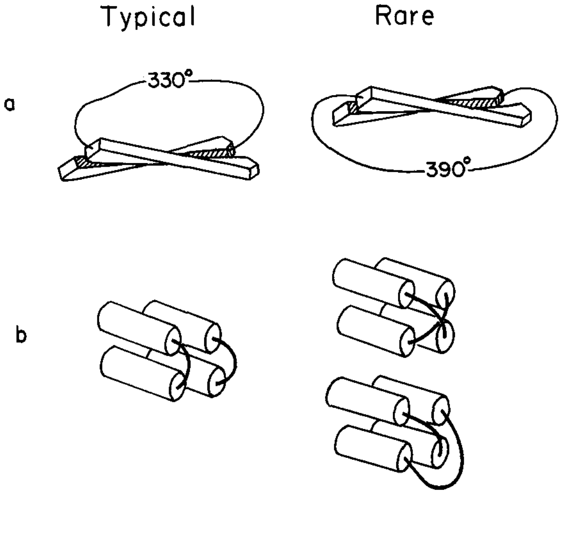

Fig. 2. Typical protein folding patterns in comparison with similar but rare ones. β-Strands are shown by arrows, α-helices by cylinders, loops by solid lines. (**a**) Right-handed connection between parallel β-strands is predominant;[37] it demands less loop bending than the left-handed one. (**b**) Crossing of loops is very rare; it either dehydrates a peptide group of a loop or demands additional loop bending. All these "defects" cost a few kcal/mol.[13]

Protein statistics deals only with the *ground* states of native protein globules (it is based on their X-ray or NMR structures and neglects the small thermal vibrations).[11,15] These statistics only tell how often a given state of elements (e.g., an interior position of leucines) is observed in proteins, while each element of a protein (each individual Leu) *is fixed* by many cooperative interactions in the globule,[20] and does not undergo any transitions from one state to another.

The question that arises is why do these powerful cooperative interactions usually fix a residue in the state corresponding to the minimum of its own, relatively weak potential (Fig. 3)?

The above consideration shows that the observed Boltzmann-like form [Eq. (1)] of protein statistics cannot be a simple consequence of Boltzmann statistics of separate amino acid residues; in particular, one cannot treat protein statistics as a Boltzmann distribution frozen by protein hardening, because this would mean that the same residue (e.g., Leu-72 of myoglobin) can be observed in different positions in different copies of the protein molecule.

The question that follows is why is the equation describing protein statistics so similar to the Boltzmann equation? And what then is the meaning of the "conformational temperature" T_* in protein statistics? (In any case, T_* cannot be the temperature of the medium: native protein folds do not change with temperature prior to denaturation.) Do these kinds of statistics, at the same conformational temperature, pertain only to small structural details (which has been established), or to overall protein folds as well?

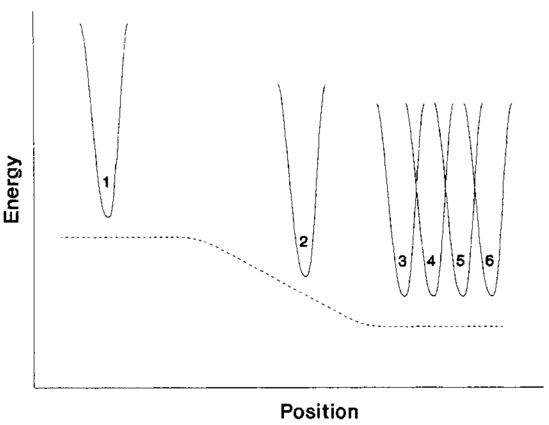

Fig. 3. A scheme which illustrates that strong cooperative interactions in a protein globule (shown as parabolas) usually fix a residue in the position which corresponds to the minimum of its own, relatively weak potential (shown as a dotted line).

To answer these questions, we compared the *numbers of random sequences stabilizing folds with structural elements of different energies.* We show that each high-energy element of a fold decreases the number of these sequences exponentially, according to Boltzmann's formula, with T_* being the characteristic freezing temperature for random amino acid chains. As a result, the low-energy elements are stabilized by many sequences, while the high-energy ones are stabilized by only a few.

THEORY

We assume that the choice of a protein native structure is determined by the stability of this structure.[13] It follows that a given fold can serve as a native one of a given chain only if its energy E is at the very bottom of the chain energy spectrum, below the energy level of any other fold of the chain. (The term "energy" is used here only for simplicity: actually, one should speak of mean force potentials, since the solvent-mediated hydrophobic and electrostatic forces are also considered.) How does the energy of one structural element affect the number of sequences which have this element within a fold corresponding to the very bottom of the chain energy spectrum?

Energy Spectra: Critical Energy and Critical Temperature

To answer this question we shall use the Random Energy Model[21] (REM). It has been shown that this simple model well describes the basic properties of energy spectra of heteropolymer globules,[22–26] which can be summarized as follows.

The form of the energy spectrum (see Fig. 5) is governed by the overall characteristics of the chain (such as the content of attracting and repulsing residues), which are close for a vast majority of long chains with random sequences. As a result, the energy spectra of most of the chains are also similar, though each fold obtains quite a different energy when one sequence is replaced by another.

In essence, the energies of sterically allowed compact folds of a randomly chosen sequence are distributed as if they acquire their energy according to the same statistical law and are independent of each other (these are the main assumptions of REM and they are approximately valid[24,25] for 3-D globular folds of heteropolymers). Namely, for a randomly chosen sequence, each fold acquires an energy E with a probability

$$p_E(E_0,\sigma^2) = (1/\sqrt{2\pi}\sigma)\exp[-(E-E_0)^2/2\sigma^2] \quad (2)$$

Here E_0 is the mean fold energy, and σ is the root mean square deviation of the fold energy from E_0. σ^2 is proportional to the number of interactions and thus to the number of chain links N. The probability p_E has a Gaussian form because the fold energy E summarizes many independent items (the energies of various contacts, bends, etc.).

The expected density of energy levels having energy E is

$$\overline{m}_E = Mp_E(E_0,\sigma^2) \quad (3)$$

where M is the number of different globular folds and is a large number, exponentially dependent on the number of chain links N.[24]

The energy of the lowest-energy fold of a random chain is usually close to a critical energy E_c determined by the condition that the expected number of spectrum lines with energy E_c or below is unity:

$$\overline{M}_{E_c} = \int_{-\infty}^{E_c} \overline{m}_E dE = 1 \quad (4)$$

The value of E_c can be readily estimated as M is an exponentially great number, and p_E [see Eq. (2)] changes with E also exponentially; thus,

$$E_c \approx E_0 - \sigma\sqrt{2 \ln M} \quad (5)$$

When $E<E_c$, $\overline{M}_E$ soon becomes much below 1 (as $\overline{m}_E$ rapidly changes with E); this means that most sequences have no energy levels below E_c.

When $E>E_c$, the average density of the energy spectrum, $\overline{m}_E$, soon becomes very large. Then the deviation of spectrum density from $\overline{m}_E$ is relatively small for a random sequence having $\sim\overline{m}_E \pm (\overline{m}_E)^{1/2}$ folds of energy $E<E_c$. Thus, its entropy is $S_E \approx R \ln(\overline{m}_E)$. According to conventional thermodynamics[18] the temperature is determined by the entropy-on-energy dependence: $T(E) \equiv (\partial \ln S_E/\partial E)^{-1}$. Thus, the rate of growth of $\overline{m}_E$ determines the temperature. The beginning of the energy spectrum gives a critical temperature $T_c = T(E_c)$ characteristic for domination of the lowest-energy folds:

$$RT_c = \sigma^2/(E_0-E_c) = \sigma/\sqrt{2\ln M} \qquad (6)$$

When the temperature falls below T_c, the lowest-energy folds are frozen out. Most heteropolymers freeze gradually,[24] but a small part of them forms the folds whose energies are below E_c by at least a few RT_c units, and they freeze by an "all-or-none" transition (like proteins) at temperature somewhat above T_c.[27,28] Thus, the value of T_c can be estimated from the "all-or-none" melting of these "protein-like" chains.

The temperature T_c depends on the properties of amino acid residues rather than on the chain length because σ^2, $\ln M$, and E_0-E_c are all proportional to the number of chain links.[24]

How Does the Energy of One Interaction Affect the Number of Fold-Stabilizing Sequences

Up to now, we considered (according to the strict REM) the thermodynamics of all the M globular folds. Let us consider further a model which has only one deviation from the conventional Random Energy Model. Namely, let us assume that all the folds can be divided into two groups: the M_1 folds of "Group 1" include a given structural element "1," and the M_0 folds of "Group 0" are without it, and that the REM is applicable to each of the groups separately. As a result of the presence of element "1," the groups have slightly different averaged properties: "Group 0" consists of folds with an E_0 mean energy and a σ^2 energy dispersion; "Group 1" consists of folds having a $E_1=E_0+\Delta_E$ mean energy and a $\sigma_1^2=\sigma^2+\Delta_\sigma$ energy dispersion.

This small deviation from the REM does not change the thermodynamics of random chains, but results in domination of folds from one or the other group among the lowest-energy folds of the chains.[30] To show this, let us estimate the expected density of energy levels corresponding to the folds of both groups at low $E-E_c$ energies.

For folds of Group 0, the expected number of spectrum lines at energy E is

$$\overline{m}_E^0 = M_0 p_E(E_0,\sigma^2) \approx (M_0/\sqrt{2\pi}\sigma)$$
$$\exp[-(E_c-E_0)^2/2\sigma^2]\exp[(E-E_c)/RT_c] \qquad (7)$$

This estimate follows from expansion of the exponent of Eq. (2) over a small difference $E-E_c$ (which is valid until $|E-E_c| < \sigma\approx\sqrt{N}\,RT_c$)[29] and the definition of critical temperature T_c given by Eq. (6) Taking into account also the definition of critical energy E_c given by Eqs. (3), (4), and (5), the expected number of obtained energy levels below energy E is

$$\overline{M}_E^0 = \int_{-\infty}^{E} \overline{m}_{E'}^0 dE' \approx (M_0/M)\exp[(E-E_c)/RT_c] \qquad (8)$$

The expected number of Group 1 folds having energy below E is

$$\overline{M}_E^1 = M_1\int_{-\infty}^{E} p_{E'},(E_0+\Delta_E,\sigma^2+\Delta_\sigma)dE'$$
$$\approx (M_1/M)\exp[(E-E_c)/RT_c]$$
$$\exp(-\Delta_E/RT_c+\Delta_\sigma/2R^2T_c^2) \qquad (9)$$

This estimate follows from the expansion of Eq. (2) over the small terms $E-E_c$, Δ_E, and Δ_σ, and the definitions of T_c and E_c [see Eqs. (3), (4), (5), (6)]; furthermore, we neglected the small term $\Delta_\sigma/\sigma^2\approx\Delta_\sigma/NT_c^2$ as compared with Δ_σ/T_c^2.

Let us consider the meaning of the ratio

$$\overline{M}_E^1/\overline{M}_E^0 = (M_1/M_0)\exp[-(\Delta_E-\Delta_\sigma/2RT_c)/RT_c] \qquad (10)$$

at low $E<E_c$ energies which are typical for stable folds.

Suppose that one has S random sequences, and $S>>1$. One can expect that they, altogether, have $\overline{M}_E^0\cdot S+\overline{M}_E^1\cdot S$ energy levels with the energy E or below, with $\overline{M}_E^0\cdot S$ of the levels corresponding to the folds of Group 0 and $\overline{M}_E^1\cdot S$ to the folds of Group 1. A sequence usually has at most only one energy level (that corresponding to its stable fold) at energies below E_c. Thus, when $E<E_c$, approximately $\overline{M}_E^0\cdot S$ of the S random sequences have their stable folds within Group 0, and $\overline{M}_E^1\cdot S$ of the sequences have them within Group 1.

This means that Eq. (10) gives the proportion of random sequences which stabilize folds of both groups and, being independent of the precise value of the threshold energy E (until $|E_c-E|<\sigma$), Eq. (10) is true also for that minority of "protein-like" random sequences which form abnormally stable folds whose energies are below E_c by a few RT_c units [see above; actually, even a considerable deviation of the threshold E from E_c results only in some modification of the temperature value used in Eq. (10)]. Thus, it is possible to approximate competition between different folds by the competition of each fold with the "stability threshold" E_c. When the energy of the fold is above E_c, it has many competitors and when below E_c it usually has none.

Equation (10), describing stable folds of random heteropolymers, resembles Eq. (1), the main empirical equation of protein statistics, and has the same physical meaning: both describe the statistical distribution of observable stable structures over the sequences and both have nothing to do with thermal fluctuations of the chains.

A few important consequences which follow from this equation are considered in the Discussion. Here it is sufficient to mention that the number of sequences stabilizing a given fold is exponentially dependent on the mean energetic advantage or disadvantage of the fold, no matter what structural detail causes them and that the temperature T_c which is included into Eq. (10), as T_* in Eq. (1), is the temperature typical for freezing of random heteropolymers.

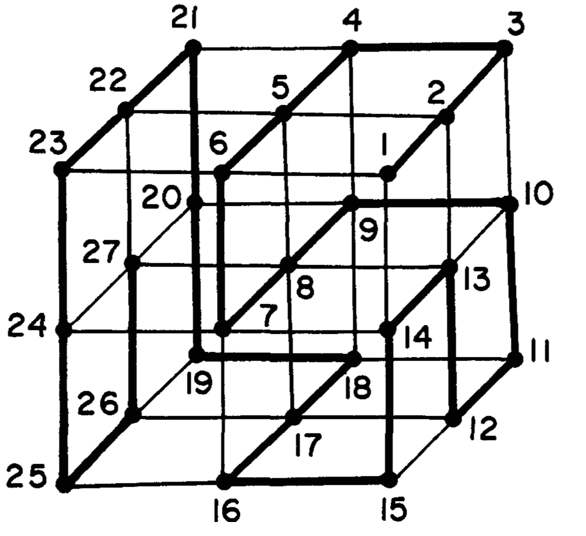

Fig. 4. One of the 103,346 compact folds of a self-avoiding chain of 27 links in a 3 × 3 × 3 fragment of the cubic lattice.[25] The fold energy E includes energies of contacts (like that formed by links 9 and 18) and bends (like link 11).

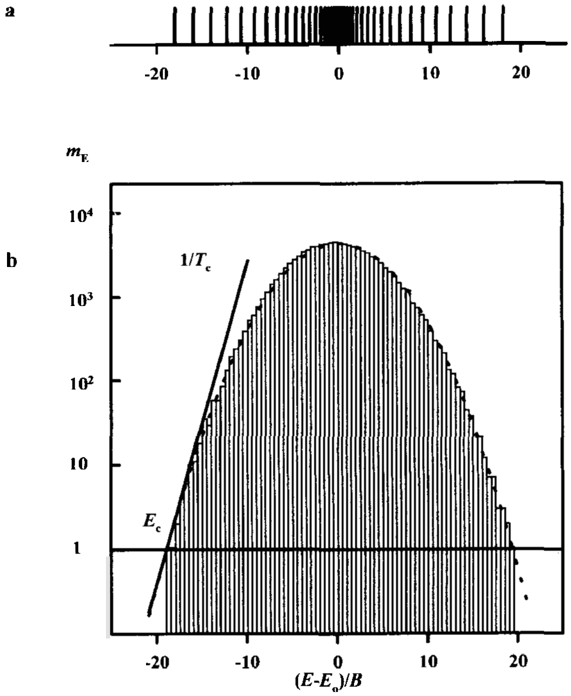

Fig. 5. Energy spectrum (**a**) and histogram of spectrum density (**b**) for one of the random sequencies. The spectrum contains 103,346 lines (only a few are drawn) corresponding to all the 103,346 compact folds of the chain shown in Figure 4. **E** is the energy of a fold, E_0 is the mean fold energy, m_E is the number of folds with energies in the range E±0.5. The average deviation of energy of one contact or bend from its mean value is taken as an energy unit. The logarithm of spectrum density can be approximated by a parabola (dotted line). The slope of this parabola at the point where the spectrum starts (i.e., where $m_E \approx 1$ and $E \approx E_c$, see the text) gives the value of $1/RT_c$. T_c is the critical temperature of freezing out of the low-energy folds. $RT_c = 1.25$ for the examined heteropolymers (on the average over 100 random sequences).

COMPUTER MODELING OF PROTEIN STATISTICS

Our analytical treatment is based on the Random Emergy Model which assumes that different folds acquire their energies irrespective of others. This assumption is not strictly correct as different folds can have some common link-to-link contacts, etc. It has been shown[24,25] that the energy spectra of random chains are, nevertheless, well described by the REM. However, this is valid, strictly speaking, only for gross properties of energy spectra, while our theory predicts a strong influence of small ($\Delta_E \sim RT_c$) energetic effects on the predominant conformations of heteropolymers.

Therefore, it is worthwhile to substantiate the obtained analytical results by a number of computer experiments. These were done using the simplest computer model of a protein chain[25] shown in Figure 4. This simple model is used only as *an example* to illustrate and test our *general* analytical theory.

The model chain consists of $N = 27$ links. We consider only the most compact folds of the chain as shown in Figure 4. Their number is $M = 103{,}346$.[25] They have the same number ($\kappa = 28$) of contacts between nonneighboring links; thus, the mean contact energy $\bar{\varepsilon}^c$ plays no role in our experiments, and we assume it as zero. The number of bends, ν, varies from 14 to 23. The number of all the possible positions of bends in the chain is $\nu_0 = 25$ (because two terminal links cannot be "bent"), while the number of all the possible link-to-link contacts is $\kappa_0 = 156$ (as only the even and odd links can be in contact, see Fig. 4).

The potential energies of the contacts, $\varepsilon^c_{i;j}$, and of the bends, ε^b_i (where i and j are the link numbers in the chain) were generated independently with Gaussian distributions

$$P_{\mathrm{cont}}(\varepsilon^c_{i;j}) = (1/\sqrt{2\pi}B)\exp[-(\varepsilon^c_{i;j})^2/2B^2]$$
$$P_{\mathrm{bend}}(\varepsilon^b_i) = (1/\sqrt{2\pi}B)\exp[-(\varepsilon^b_i - \bar{\varepsilon}^b)^2/2B^2] \quad (11)$$

Taking B as a unity in our computer experiments, we obtain all the energies in B units and all the temperatures in B/R units.

For each primary structure (i.e., for each set of all $\varepsilon^c_{i;j}$ and ε^b_i) we found the "native" conformation (one with the minimal energy) by an exhaustive enumeration of all the 103,346 chain conformations.

First, we randomly generated (taking $\bar{\varepsilon}^b = 0$) a hundred primary structures and computed their energy spectra (Fig. 5). The critical temperature T_c, calculated from the starting slopes of these spectra densities was 1.25.

We then examined the statistics of incorporation

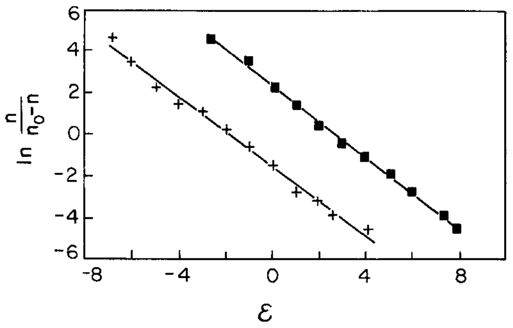

Fig. 6. Occurrence of a contact 9:18 (+) and of a bend 11 (■) in the "native" folds of random sequences. ε is the energy of the examined element, $n_0 = 100$ is the number of explored random sequences, n is the number of those which have the examined element in their native fold. A linear (Boltzmann-like) dependence of $\ln[n/(n_0 - n)]$ on ε is evident. The slope of the best-fitted line gives a value of $-1/RT_*$. The conformational temperature T_* obtained from these lines is virtually the same for the contact ($RT_* = 1.21$) and the bend ($RT_* = 1.16$), and is close to T_c($RT_c = 1.25$, see Fig. 5b).

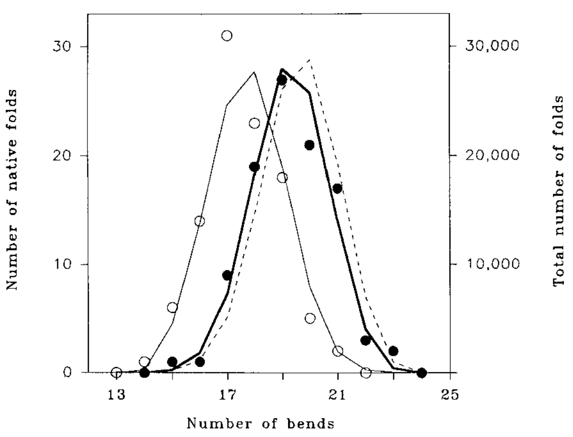

Fig. 7. The occurrence of "folding patterns" with a different number of bends among the native (lowest-energy) folds of the 100 random "protein chains." The investigated model is shown in Figure 4. The results are given for the mean bend energy $\bar{\varepsilon}^b = 0$ [●: computer experiment; bold line: theory, Eq. (10)] and $\bar{\varepsilon}^b = 1$ (○: computer experiment; thin line: theory). The distribution of all the possible 103,346 folds over the "folding patterns" with different numbers of bends is shown (dashed line) for comparison.

of arbitrarily chosen structural elements in native conformations of the model protein chains. Figure 6 shows the results for two elements: the contact between links 9 and 18, and the bend of link 11. For any fixed energy of this contact, $\varepsilon^c_{9:18}$ (or of the bend, ε^b_{11}), we generated $n_0 = 100$ sequences where the random values were assigned, according to Eq. (11), to all the other ε (again, we took $\bar{\varepsilon}^b = 0$). We found the number n of the sequences with the contact 9:18 (or the bend 11) in their native folds. It is noteworthy that the values of $\varepsilon^c_{9:18}$ (or of ε^b_{11}) determine the average energy difference between the folds with and without these elements, i.e., these values play the role of Δ_E in Eq. (10). Δ_σ does not depend on the Δ_E value as the different ε values are generated independently.

The results of a computer experiment (Fig. 6) clearly demonstrate the expected proportionality between $\ln[n/(n_0 - n)]$ and ε^c (or ε^b). The conformational temperature T_* which follows from the coefficient of this proportionality [$= -1/RT_*$, cf. Eqs. (1) and (10)], has virtually the same value for both ε^c and ε^b (Fig. 6) and is indeed very close to T_c, the critical temperature of chain freezing (Fig. 5b). Similar results (not shown) were obtained also when we took $\bar{\varepsilon}^b \neq 0$ and different B values for contacts and bends.

Finally, we carried out a computer simulation of the statistics of folding patterns. We divided all the 103,346 compact chain folds into 10 groups (imitating the "folding patterns") according to their number of bends (from $\nu = 14$ to $\nu = 23$). We generated a hundred random chains with mean bend energies $\bar{\varepsilon}^b = 0$ and a hundred with $\bar{\varepsilon}^b = +1$. This provided a predominance of less (when $\bar{\varepsilon}^b = +1$) and more (when $\bar{\varepsilon}^b = 0$) bent conformations in the stable chain folds. The dispersions of contact and bend energies were equal, as in Eq. (11). Then we found the native folds of these sequences and computed the number of representatives of each "folding pattern" among these native folds. The experimental results (Fig. 7) were compared with those calculated from Eq. (10) using M_ν, the computed numbers of all the possible folds with ν bends, the mean energies of these folds $E_\nu = \nu\bar{\varepsilon}^b$, and the dispersions $\sigma^2_\nu = \kappa(1 - \kappa/\kappa_0)B^2 + \nu(1 - \nu/\ \nu_0)B^2$. A good correlation of the computer experiment with the analytical theory is evident (Fig. 7).

It should be noted that the above results are obtained for the case of uncorrelated energies of different interactions.

To check the approximation which treats the energies of different interactions of a residue as independent values and to throw some light on those deviations from the strict Boltzmann-like statistics which are observed in proteins (Fig. 2), we investigated a case of strongly correlated interactions. Now the self-avoiding chain shown in Figure 4 consists only of "black" and "white" links, the contact energy being -1 for each "black–black" contact, $+1$ for each "white–white" contact, and 0 for each contact of "white" and "black" links. The numbers of "black–black," "white–black," and "white–white" contacts in the lowest-energy folds of these chains (Fig. 8) were calculated; when the chain had more than one lowest-energy fold, we chose one of them at random.

Now the energies of different contacts are extremely correlated, because the variety of residues is very small.[24] This leads to some deviation from the exponential dependence of occurrence of a structural element (a link-to-link contact) on its own energy, as the term Δ_E here includes not only this energy, but

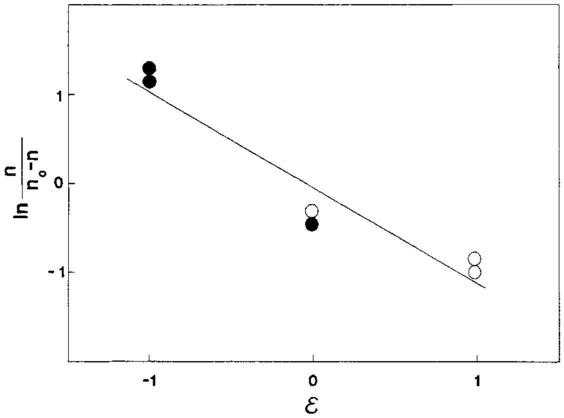

Fig. 8. Occurrence of different contacts in the lowest-energy folds of $n_0 = 100$ random sequences of "black" (attracting) and "white" (repulsing) beads. The points refer to "black–black," "black–white," and "white–white" contacts. The straight line corresponds to the best linear approximation. Its slope is approximately equal to that of the lines in Figure 6.

also the mean energies of interactions of this structural element with other links, and the term Δ_σ is no longer independent on the energy of the element. Nevertheless, a general tendency of the occurrence-on-energy dependence remains the same, and the T_* value from the average slope of this dependence remains close to the T_c value from the slope of the spectrum density logarithm in the low-energy part of the energy spectra of these chains (not shown).

DISCUSSION

Any structural element divides all the chain folds into two groups: the folds of one group include this element and the folds of the other do not. The "structural element" can be, for example, a contact between residues 9 and 18 or a bend in residue 11 or immersion of residue 72 into the globule or overcrossed loops (Fig. 2b), etc.

As concerns the sequences, two cases should be distinguished: when attention is focused on the amino acid residues which form the element and when it does not. In the second case (e.g., when overcrossed loops are examined), the folds are divided into those with crossed loops and those without, and all the possible sequences are considered. In the first case, only those chains are considered which include the residues in question in the named element. For example, if the immersion of Leu-72 is of interest, only the random sequences with Leu in position 72 are considered and the folds are divided into those where this Leu is immersed and where it is not. Being virtually independent of the Leu position in the chain, the obtained result for the immersion/nonimmersion of Leu-72 also shows the proportion of internal and surface leucines in the stable folds of random polypeptides.

The structural element influences the energies of those M_1 folds which include it. The average (over the sequences) energy difference Δ_E between the folds with and without this element is determined, first of all, by the (internal) energy of the element, although it is noteworthy that Δ_E can include also the averaged traces of the "external" interactions of the element with its environment (as in the example shown in Fig. 8).

Furthermore, the presence of a structural element changes the dispersion of energies of the folds which include it [term Δ_σ in Eq. (10)], because the residues involved in internal interactions have, on the average, less interactions with the remaining chain. It is noteworthy that the term $-\Delta_\sigma/2RT_c$ in Eq. (10) gives an energetic advantage to the elements with numerous interactions in the globule: a greater dispersion of these interaction energies increases Δ_σ and thus the chance that a randomly chosen sequence gives a sufficiently low total energy to the element in question.

The term $\Delta_E - \Delta_\sigma/2RT_c$ in Eq. (10) gives the expected energy of the element in the sequences containing it in their stable folds; taking this value together with the fraction of folds with this element, M_1/M, one gets an average "selective free energy" of the element "1" which determines its occurrence in the stable folds of random chains:

$$\Delta F^1_{\text{select}} = (\Delta_E - \Delta_\sigma/2RT_c) - RT_c \ln(M_1/M) \quad (12)$$

$$\frac{\textit{NATIVE FOLDS WITH "1"}}{\textit{NATIVE FOLDS WITHOUT "1"}} = \exp(-\Delta F^1_{\text{select}}/RT_c) \quad (13)$$

In essence, $\Delta F^1_{\text{select}}$ is the mean (averaged over possible surroundings) free energy of element "1" in the globule, and it is $\Delta F^1_{\text{select}}$ which is "visible" in the observable statistics of native folds.

Let us consider the terms contributing to $\Delta F^1_{\text{select}}$ in more detail taking, as an example, the simplest "element"—two residues at a given distance **r**. The statistics of interresidue distances in proteins is well investigated.[31] When the distance **r** is so short that two residues interact directly, the energy of this interaction mainly contributes to the dependence of $\Delta F^1_{\text{select}}$ or **r**, and thus governs the observed occurrence-to-distance relationship. However, when the distance **r** between the residues is too great for their direct interaction, the residues still can "interact" indirectly, through the intermediate or surrounding residues. Thus, at large distances **r**, these are the mediated interactions (rather than potentials of interresidue forces) that are visible in the statistics of interresidue distances observed in native structures of heteropolymers, including proteins, in the same way as indirect interactions of small molecules are visible in long-range correlation functions in liquids or solids.[32]

An answer can now be given to the question

"What is the fraction of random sequences that is able to stabilize a given element of protein structure?" The answer is that this fraction is directly proportional to the number of folds which include this element and is exponentially dependent on its mean energy [Eqs. (12) and (13)]. These equations show also that the same "conformational temperature," equal to the freezing temperature T_c, concerns any structural element, independently of whether it is a small detail or a motif of overall chain folding.

The greater the number of sequences stabilizing a structural element, the greater is the chance of its being observed.

This helps to understand the observed Boltzmann-like statistics of small elements of protein structures and also why the common folding motifs of globular proteins are those without "structural defects" like crossed loops (though these defects cost only 2–5 kcal/mol, while a sequence can change a fold energy by 50 kcal/mol).[17] The additional 2 or 5 kcal/mol are important only at low energies, when every additional kcal/mol is so hard to gain, and here they determine the expected result of competition between the folds: in particular, that a hundred or a thousand times smaller number of sequences can stabilize crossed loops than the sequences that can stabilize the noncrossed loops.

Equations (10), (12), and (13) show also that a greater number of individual folds within a folding pattern gives it a greater chance to be observed. In particular, this can be explained by the predominance of right-handed (less restricted) connections of parallel β-strands (Fig. 2a) over the more restricted left-handed ones: a random sequence will usually find its best fold within the group of less restricted loops, which includes more particular conformations.[30]

In other words, though the entropy of native protein is zero, a greater number of particular folds in a folding pattern simply gives a random sequence a greater chance to fit one of these folds, thus obtaining low energy.

CONCLUSION

Random heteropolymers demonstrate a Boltzmann-like distribution of various structural elements over their stable folds. The "temperature" in these statistics is the freezing temperature of the heteropolymers. The origin of these statistics is simply that the more sequences that provide the stability of folds with a given feature, the more often this feature can be observed. Since the chains of globular proteins (unlike, e.g., fibrous ones) resemble random heteropolymers,[33] this "multitude principle" must be valid for them as well. And, indeed, this distribution is observed at all protein structure levels.

This paper is concerned with the physical selection of protein structures and deals with the spontaneous consequences of only one restriction imposed on them: namely, that to exist the protein structure must be internally stable. Stability summarizes the energy contributions of many chain links; the links can change with mutations, and only this sum as a whole is under the control of selection. However, this control is imposed not only on proteins (RNA immediately comes to mind), and internal stability is not the only object of such a control. Similar limitations can be imposed by the absence of aggregation, by the requirement of fast folding, by the presence of some biological function,[34] etc. The key question of the current study: "What is the fraction of sequences that is able . . .?", posed both theoretically and experimentally, can elucidate these limitations and their spontaneous consequences for molecular structures.

ACKNOWLEDGMENTS

We thank E.I. Shakhnovich, O.B. Ptitsyn, J. Janin, and C. Chothia for valuable discussions. We acknowledge the financial support of the Russian Foundation for Fundamental Research (grant 93-04-6636). We are grateful to A. G. Raiher for editing the English text.

REFERENCES

1. Pohl, F.M. Empirical protein energy maps. Nature New Biol. 234:277–279, 1971.
2. Pohl, F.M. Statistical analysis of protein structures. In: "Protein Folding." Jaenicke, R. (ed.). Amsterdam: Elsevier/North-Holland Biomedical Press, 1980: 183–196.
3. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.
4. Finkelstein, A.V., Ptitsyn, O.B., Kozitsyn, S.A. Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structures with experiment. Biopolymers 16:497–524, 1977.
5. Serrano, L., Sancho, J., Hirshberg, M., Fersht, A.R. α-Helix stability in proteins. I. Empirical correlations concerning substitutions of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J. Mol. Biol. 227:544–559, 1992.
6. MacArthur, M.W., Thornton, J.M. Influence of proline residues on protein conformation. J. Mol. Biol. 218:397–412, 1991.
7. Bryant, S.H., Lawrence, C.E. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 9:108–119, 1991.
8. Rashin, A.A., Ionif, M., Honig, B. Internal cavities and buried waters in globular proteins. Biochemistry 25:3619–3625, 1986.
9. Miyazava, S., Jernigan, R.L. Estimation of effective inter-residue contact energies from crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
10. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature (London) 261:552–557, 1976.
11. Schulz, G.E., Schirmer, R.H. "Principles of Protein Structure." New York: Springer-Verlag, 1979.
12. Richardson, J.S. Anatomy and taxonomy of protein structures. Adv. Prot. Chem. 34:167–339, 1981.
13. Finkelstein, A.V., Ptitsyn, O.B. Why do globular proteins fit the limited set of folding patterns? Progr. Biophys. Mol. Biol. 50:171–190, 1987.
14. Chothia, C., Finkelstein, A.V. The classification and origins of protein folding patterns. Annu. Rev. Biochem. 59: 1007–1039, 1990.

15. Branden, C., Tooze, J. "Introduction to Protein Structure." New York: Garland, 1991.
16. Privalov, P.L. Stability of Proteins. Small globular proteins. Adv. Prot. Chem. 33:167–241, 1979.
17. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models. Implication for structure prediction. J. Mol. Biol. 177:787–818, 1984.
18. Landau, L.D., Lifshitz, E.M. "Statistical Physics." London: Pergamon, 1959.
19. Takano, T. Structure of myoglobin refined at 2.0 A resolution. J. Mol. Biol. 110:537–568, 1977.
20. Gelin, B.R., Karplus, M. Sidechain torsional potentials and motion of amino acids in proteins: Bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. U.S.A. 72:2002–2006, 1975.
21. Derrida, B. Random-energy model: An exactly solvable model of disordered systems. Phys. Rev. B 24:2613–2626, 1981.
22. Bryngelson, J.B., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U.S.A. 84:7524–7528, 1987.
23. Bryngelson, J.B., Wolynes, P.G. A smple statistical field-theory of heteropolymer collapse with application to protein folding. Biopolymers 30:177–188, 1990.
24. Shakhnovich, E.I., Gutin, A.M. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of replica approach. Biophys. Chem. 34:187–199, 1989.
25. Shakhnovich, E.I., Gutin, A.M. Enumeration of all compact conformations of copolymers with random sequence of links. J. Chem. Phys. 93:5967–5971, 1990.
26. Shakhnovich, E.I., Gutin, A.M. Implication of thermodynamics of protein folding for evolution of primary sequences. Nature (London) 346:773–775, 1990.
27. Shakhnovich, E.I., Gutin, A.M. Enginerring of stable and fast-fold sequences of model proteins. Proc. Natl. Acad. Sci. U.S.A. 90:7195–7198, 1993.
28. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya. Boltzmann-like statistics of protein architectures: origins and consequences. In: "Subcellular Biochemistry: Structure, Function and Protein Engineering," Vol. 32. Biswas, B.B., Roy, S. (eds.). New York: Plenum Press, 1995: 1–26.
29. Gutin, A.M., Badretdinov, A.Ya, Finkelstein, A.V. Why is the statistics of protein structures Boltzmann-like? Mol. Biol. (Russia), Engl. Transl. 26:94–102, 1992.
30. Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya. Why are the same protein folds used to perform different functions. FEBS Lett. 325:23–28, 1993.
31. Sippl, M.I., Jaritz, M. Predictive power of mean force potentials. In: "Protein Structure by Distance Analysis." Bohr, H., Brunak, S. (eds.). Amsterdam: IOS Press, 1993: 113–134.
32. Stanley, H.A. "Introduction to Phase Transitions and Critical Phenomena." Oxford: Clarendon Press, 1971.
33. Ptitsyn, O.B. Random sequences and protein folding. J. Mol. Struct. (Theochem.) 123:45–65, 1985.
34. Gregoret, L.M., Sauer, R.T. Additivity of mutant effects assessed by binomial mutagenesis. Proc. Natl. Acad. Sci. U.S.A. 90:4246–4250, 1993.
35. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198: 295–310, 1987.
36. De Santis, P., Liquori, A.M. Conformation of gramicidin S. Biopolymers 10:699–710, 1971.
37. Sternberg, M.J.E., Thornton, J.M. On the conformation of proteins: The handedness of β-strand–α-helix–β-strand unit. J. Mol. Biol. 105:367–382, 1976.