PROTEINS: Structure, Function, and Genetics 23:151–162 (1995)

Perfect Temperature for Protein Structure Prediction and Folding

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

ABSTRACT **We have investigated the influence of the "noise" of inevitable errors in energetic parameters on protein structure prediction. Because of this noise, only a part of all the interactions operating in a protein chain can be taken into account, and therefore a search for the energy minimum becomes inadequate for protein structure prediction. One can rather rely on statistical mechanics: a calculation carried out at a temperature T_c somewhat below that of protein melting gives the best possible, though always approximate prediction. The early stages of protein folding also "take into account" only a part of all the interactions; consequently, the same temperature T_c is favorable for the self-organization of native-like intermediates in protein folding.**

Key words: protein structure, prediction, self-organization, stability, melting temperature.

INTRODUCTION

The folding process and 3D structure prediction have a common feature: they both have to operate on the basis of *only some part* of the interactions which determine the native protein structure.

In the case of protein folding, it is known that some fluctuating secondary structures and then some globular states appear at early stages of in vitro self-organization.[1–7] These early intermediates demonstrate some native-like features (in particular, secondary structure content and location). The intermediates are formed by hydrogen bonds and hydrophobic interactions much earlier than other forces (such as specific van der Waals interactions responsible for tight packing) begin to influence the molecular structure. Being responsible for the stability of native protein structures,[8] these "late" interactions are rather powerful. The question arises: Why do the "late" forces preserve some features of the intermediates formed only by the "early" forces? Is this a consequence of a special concordance of the "early" and "late" interactions?

A similar question arises in protein, RNA, etc. structure prediction.

Any prediction is based on some energetic parameters, and any set of parameters can give only some approximation of actual interactions. Thus, a prediction is always based on *some part* of the actual interactions, while the other part remains unknown and can be only roughly estimated. As a result, the fold which has the lowest "calculated" energy can be very different from that actually having the lowest energy. It has been shown that the native fold is among those few having a low calculated energy,[9,10] but it seldom has the lowest calculated energy. Thus, the question arises: How many folds of a low calculated energy form a set of "promising" candidates for the role of the native structure? And the main question: How can the predictions be more or less successful, at least in some cases,[11,12] if they ignore a considerable part of the actual interactions?

Here we should explain the usage of the terms "energy" and "number of folds" in this paper. (1) The term "energy" is used here only for simplicity; strictly speaking, the term "mean force potential" must be used, because the hydrophobic and electrostatic forces are solvent-mediated ones. (2) It is assumed that a chain has a finite number of folds. This is always the case for polymer models which use 3-D lattices, like Flory's model; in continuous space, the "folds" can be compared with the local energy minima.

THEORY

We start with the problem of 3D structure prediction and reformulate the above questions in terms of energy spectra. This allows us to use some ideas of the Random Energy Model, or REM[13] which proves to be effective in analysis of general properties of heteropolymers, including proteins.[14–22]

REM accounts for the properties of energy spectra of heteropolymers. It assumes that the chains have no definite regularities in their sequences (such as the overall periodicities which are not typical for globular protein, though present in fibrous ones). Strictly speaking, REM has been proved to be valid

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.

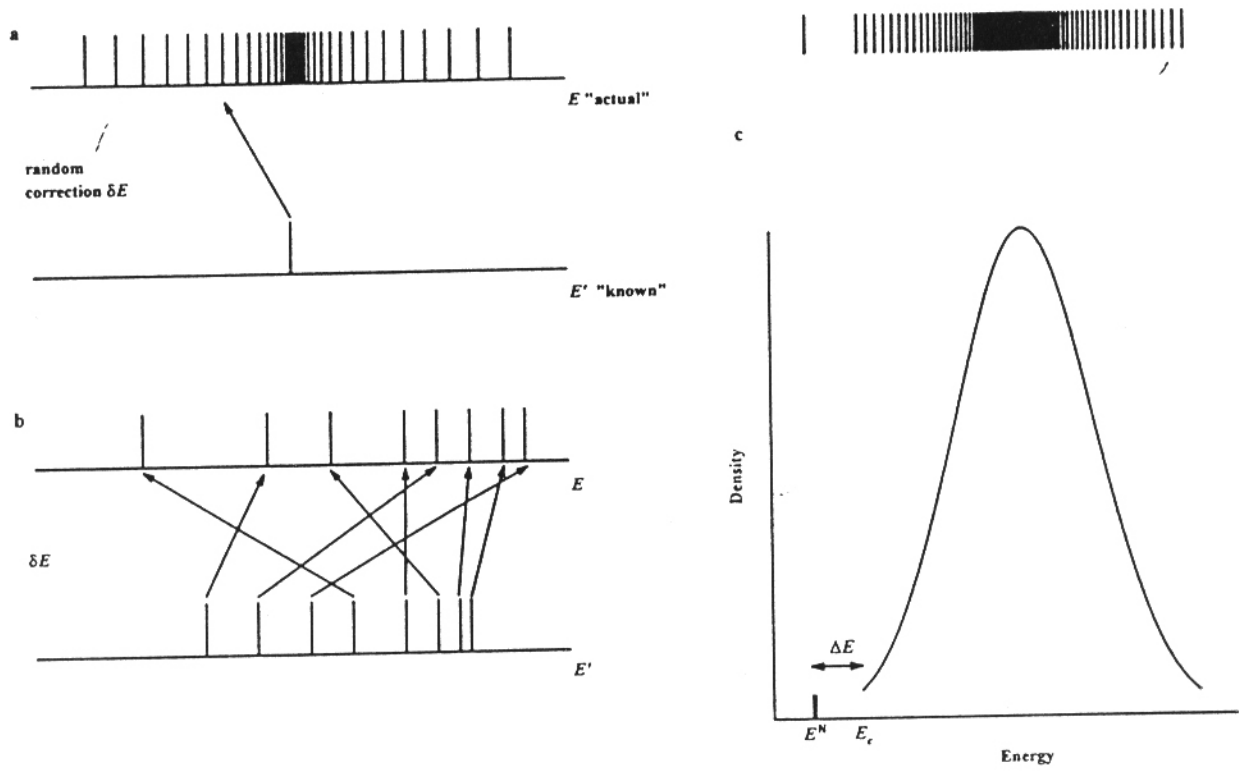

Fig. 1. (a) "Known" (calculated) energy of a fold, E', gives a range of possible values of the "actual" energy of this fold, E. The uncertainty is due to a "random" (as regards the energetic parameters used to calculate the fold energy) value of a correction δE. The arrow shows one of the possible values of this random correction. (b) Spectrum of calculated energies and one of the many possible spectra of actual energies compatible with the calculated one. (c) Energy spectrum (above) and its density (below). E_N is the energy corresponding to the native (the lowest-energy) fold. The Gaussian shows the density of the quasi-continuous part of the spectrum. Below this part there are only a few, if any, energy lines. For the "protein-like" sequences there is at least one such line, and it is separated by a considerable energy gap ΔE from the energy E_c of the beginning of the continuous spectrum.

for simple heteropolymer models (in particular, the proof[15,16] does not take into account secondary structure and flexibility of side chains). However, applied to globular proteins, REM is very helpful in explaining their statistics (from the level of side chain conformations to the level of overall architectures) from a single point of view.[20–22] Thus, it can be assumed that REM accounts for the general properties of energy spectra of globular proteins.

The spectrum of calculated energies is always an approximation of the actual energy spectrum due to inevitable errors in energetic parameters (Fig. 1). The native fold of a chain has apparently the lowest actual energy, at least for small proteins.[23] But what position in the *calculated* energy spectrum does this native fold occupy? And is any prediction possible when the native fold does not have the lowest computed energy? The answers to these questions are important for understanding the possibility and limits of protein structure prediction.

Preliminary Remarks: Actual Energy and Its "Known" Part

The higher the correlation between the computed energies $\hat{E}_1, \ldots, \hat{E}_M$ and their actual values $E_1, \ldots, E_M$, the higher is the possibility of predicting protein structure.

The conventional[24] *coefficient of correlation* between two sets of values has the form

$$C = \frac{<\hat{E}E>}{[<\hat{E}^2><E^2>]^{1/2}} \qquad (1)$$

where $< >$ means the averaging: $<\hat{E}\,E> = \sum_{i=1}^{M} \hat{E}_i\, E_i/M$, *etc.* (*for simplicity, we assume, here and below, that both the actual and the computed energies are counted off their mean values, i.e.*, $<E> = 0$ and $<\hat{E}> = 0$). The correlation is good when $C \approx 1$ and poor when $C \approx 0$ or below.

The best fit of the computed to the actual energies is found by the least-squares method.[24] E_i must be represented as $(\tilde{\lambda}\hat{E}_i + \delta) + \delta E_i$ where δE_i is the "random error" (Fig. 1) and one has to minimize the sum of deviations $\sum_{i=1}^{M}[\delta E_i]^2 = \sum_{i=1}^{M}[E_i - (\tilde{\lambda}\cdot\hat{E}_i + \delta)]^2$ over the values of the *scaling coefficient* $\tilde{\lambda}$ and the shift constant δ. The minimum corresponds to the following values of δ and $\tilde{\lambda}$: $\delta = 0$ and

$$\tilde{\lambda} = \frac{<\hat{E}\,E>}{<\hat{E}^2>} \equiv C\left(\frac{<E^2>}{<\hat{E}^2>}\right)^{1/2} \qquad (2)$$

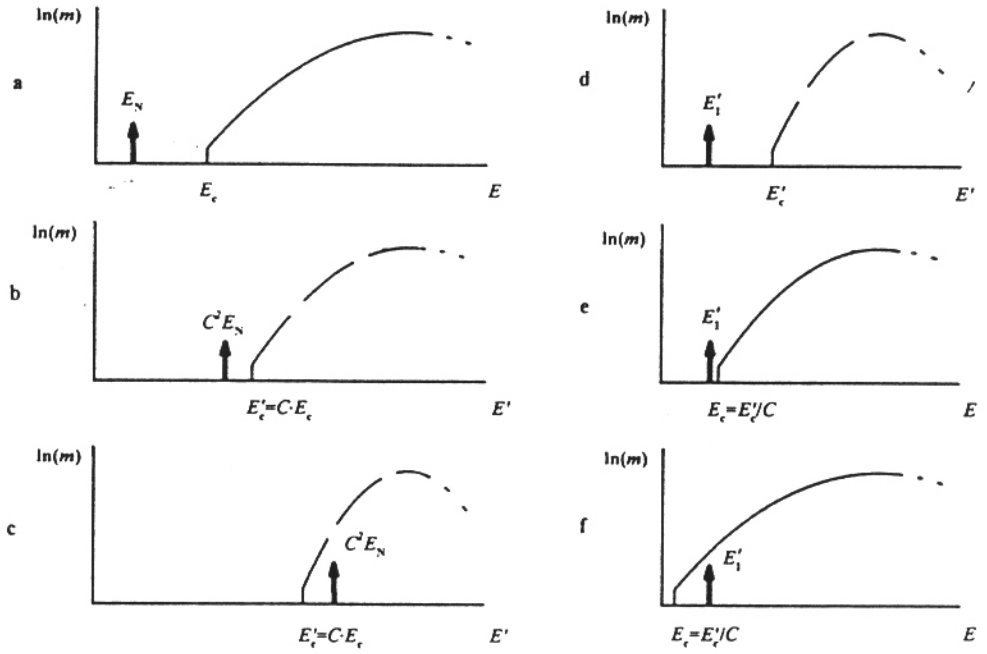

Fig. 2. Schematic representation of densities of the "actual" and "known" energy spectra. Left side: density of the "actual" spectrum (a) and the expected form of densities of the "known" spectra at high (b) and low (c) level of correlation between the "actual" and "known" energies. Right side: density of the "known" spectrum (d) and the expected form of densities of the "actual" spectra at high (e) and low (f) level of correlation. The densities m are given in a logarithmic scale. E is the actual fold energy, E' the "known" one. The parabolas correspond to the continuous parts of the spectra. E_c is the critical energy where the continuous part of the "actual" energy spectrum begins [where the spectrum density $m(E)\sim 1$], E'_c is the critical energy for the beginning of the continuous spectrum of "known" energies. The arrows show the native fold [its "actual" energy is E_N in (a) and the expected "known" energy is C^2E_N in (b)–(c)], and the fold of the lowest "known" energy (its "known" energy E'_1 is equal to the expected "actual" energy) in (d)–(f). C is the coefficient of correlation between the "actual" and "known" energies. For other explanations see the text.

(note that $<E>=0$, $<\hat{E}>=0$). Thus, the normalized values

$$E'_i = \tilde{\lambda}\,\hat{E}_i, \qquad i = 1, \ldots, M \tag{3}$$

give the best possible fit of the computed values $\hat{E}_i$ to the "actual" energies E_i. Therefore, the actual energies can be represented as

$$E_i = E'_i + \delta E_i, \qquad i = 1, \ldots, M \tag{4}$$

where E'_i is the "known" part of the actual energy E_i and δE_i is the correction (Fig. 1). The values of E'_i can be calculated by equations (1), (2), and (3) from the scaling coefficient $\tilde{\lambda}$ and the computed fold energies $\hat{E}_i$. It is readily seen that the dispersions of the "actual" and the "known" energies satisfy the equation

$$<(E')^2> = C^2<E^2> \tag{5}$$

that $<(E')^2> = <E'\,E>$, that the errors δE_i do not correlate with the "known" energies E'_i (i.e., $<\delta E\,E'>=0$), that $<\delta E>=0$, and that the dispersion of errors δE_i is

$$<\delta E^2> = (1 - C^2)<E^2>. \tag{6}$$

Similarly, one can also find the best possible fit of the *actual* energies to the "known" ones by minimization of the error function $\sum_{i=1}^{M}[E'_i - (\lambda\cdot E_i + b)]^2$, and show that the "known" energies can be represented as

$$E'_i = C^2\cdot E_i + \delta E'_i, \qquad i=1, \ldots, M \tag{7}$$

where E_i is the actual energy value, and the error $\delta E'_i$ has a mean value of zero and does not correlate with E_i (i.e., $<\delta E'>=0$, $<\delta E'\,E'>=0$).

It can be said that $\alpha = C^2$ is that part of the actual interactions which is in the set of parameters used for energy computations.

The values of the scaling coefficient $\tilde{\lambda}$ and the correlation coefficient C are inherent to each set of energy parameters, no matter how the set was obtained: from atom–atom potentials in vacuum, from empirical estimates of interactions between residues in water, from protein statistics, etc.

The values $\tilde{\lambda}$ and C can be found by Eqs. (1) and (2) when the experimental and computed stabilities of some molecular structures can be compared. Protein engineering experiments on melting temperature, binding constants, secondary structure stability, etc., are the most suitable objects to test the parameters used for protein structure prediction, as in this case both the experiments and the calculations pertain to polypeptide chains and concern similar conditions (temperature, solvent, pH, etc.).

The $\tilde{\lambda}$ value changes nothing except the scale of computed energies; with this reservation, it can always be assumed that the computed energies are normalized [see Eq. (4)]. On the contrary, the value of C is of crucial importance for the applicability itself of energetic parameters for protein structure prediction.

Energy Spectra of Heteropolymer Globules

The energy spectrum of a chain consists of energy lines corresponding to different folds of the chain. It has been shown[14–18] that the energy spectra of heteropolymer globules are well described by the Random Energy Model.[13]

Here we recall (following refs. 13–22) only those features of the spectra which are essential for this study.

The energy spectra are similar for most of the heteropolymer sequences, though each fold obtains quite a different energy when one sequence is replaced by another. Nearly all the energy lines lie in the continuous part of a spectrum where the intervals between them are negligible, and only a few, if any, lines occur below the continuous spectrum (Fig. 1c).

The density $m(E)$ of the continuous part of energy spectrum of a heteropolymer has a Gaussian form, $m(E) \sim M \exp(-E^2/2\sigma^2)$, as different chain folds acquire their energies virtually independently of one another.[16] Here M is the total number of folds, E is the energy of a fold (the mean fold energy $<E>$ is

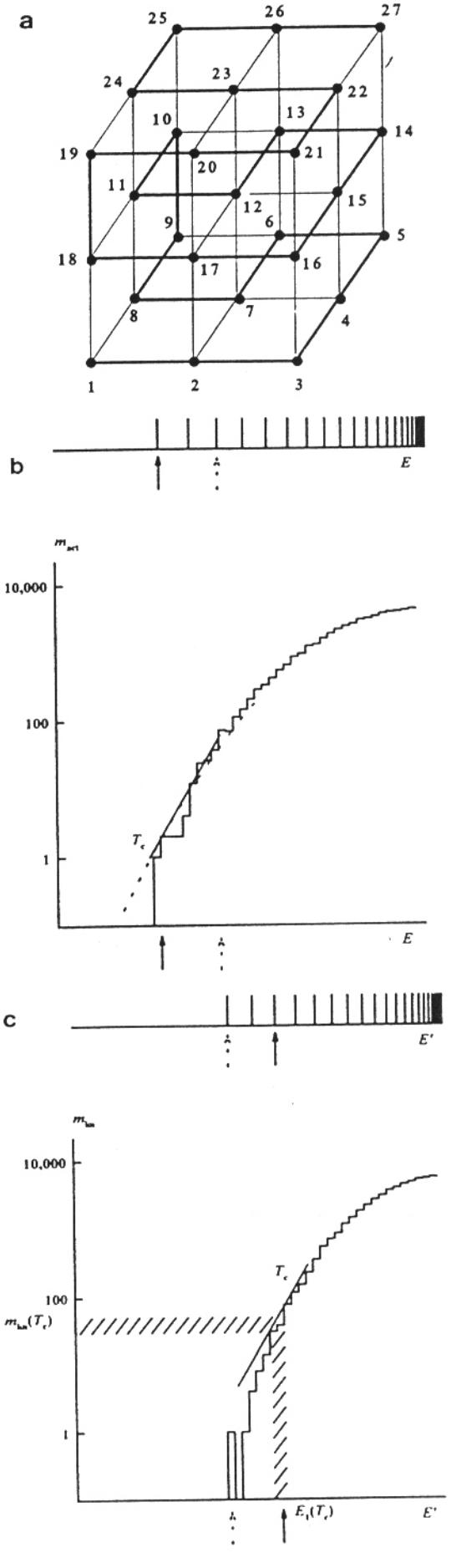

Fig. 3. Simplified model of a protein globule. (**a**) One of 103,346 self-avoiding compact chain folds at a $3\times3\times3$ fragment of a cubic lattice.[17] Each of the compact folds includes 28 link-to-link contacts. The fold energy is equal to the sum of energies of these contacts. (**b**) Low-energy part of the "actual" energy spectrum for some randomly generated sequence ("sequence" here is a synonym of the set of link-to-link contact energies). The spectrum is obtained by an exhaustive enumeration of all 103,346 compact folds shown in (**a**). The "actual" energy is calculated on the grounds of all the interactions operating in the chain. Above: energy lines (only a small part of them are drawn). Below: histogram of the spectrum density given in a logarithmic scale. $m_{act}(E)$ is the number of folds with the energy $E\pm0.25$ (the energy unit is the standard deviation of a link-to-link contact energy from zero). When $m_{act}(E)>>1$, the spectrum can be treated as a continuous one. The dashed curve shows a parabolic extrapolation of its density to the region where $\ln m(E) \sim 1$ and less, which virtually corresponds to $\ln \bar{m}_{act}(E)$; $\bar{m}_{act}(E)$ is the result of the averaging of $\bar{m}_{act}(E)$ over the random sequences. The straight line shows a growth of $\ln \bar{m}_{act}(E)$ at the beginning of the continuous spectrum; its slope gives the characteristic temperature T_c: $d\ln \bar{m}_{act}/dE|_{\bar{m}\to1+0} = 1/RT_c$ ($\bar{m}\to1+0$ means extrapolation to $\bar{m}_{act}=1$ from the side of the continuous spectrum). T_c is the critical temperature of protein melting.[16,17] (**c**) "Known" energy spectrum and its density for the same as in (**b**) random chain calculated on the grounds of $\alpha=54\%$ of all the link-to-link interactions; the other 46% of contacts do not contribute to the "known" energy E'. The straight line shows that the tangent to this histogram with the same slope $1/RT_c$ as the tangent in (**b**). The abscissa of the tangent-to-histogram contact gives $E_1(T_c)$, the most probable value of the "known" energy of the native fold (i.e., of the fold with the lowest "actual" energy). The ordinate of this contact gives $m_{kn}(T_c)$, the number of folds with the "known" energy $E'=E_1(T_c)\pm0.25$. The solid arrow points to the native fold positions in both spectra; the dashed arrow points to the fold which has the lowest "known" energy. Note that the beginning of the "known" spectrum corresponds to the temperature below T_c and does not contain the native fold.

taken as zero), and $\sigma^2 = \langle E^2 \rangle$ is the dispersion of the fold energies.

The spectrum is continuous until $m(E) \gg 1$. According to conventional thermodynamics,[25] $T \equiv (\partial S/\partial E)^{-1}$. Thus, for a Gaussian spectrum,

$$T = (R\partial \ln m/\partial E)^{-1} = -\sigma^2/RE \qquad (8)$$

(where $E<0$) is the temperature corresponding to the spectrum region with the energy E: $R \ln m(E)$ is the entropy of this region (R being the gas constant). Thus, the density of a continuous spectrum obeys the equation

$$\ln m(E) = \ln M - E^2/2\sigma^2 = \ln M - \sigma^2/(2R^2T^2). \quad (9)$$

Equations (8) and (9) are valid until $\ln m > 0$. Thus, the beginning of the continuous spectrum corresponds to the critical region where $\ln m = 0, E \approx E_c = -\sigma\sqrt{2 \ln M}$, and $T \approx T_c = (\sigma/R)/\sqrt{2 \ln M}$.

This temperature T_c corresponds to the freezing point of a heteropolymer[14–18] which is a gradual glass transition for most of the random sequences.[16]

However, some of the sequences are "protein-like." They have stable folds with the energy $E_N = E_c - \Delta E$, where ΔE exceeds a few RT_c. This gap ΔE between the lowest-energy fold and the edge of the continuous energy spectrum (see Fig. 2) is sufficient to convert the freezing into an "all-or-none" transition, as it occurs in proteins.[20] The fraction of the "protein-like" chains among the random sequences can be estimated as $\sim\exp(-\Delta E/RT_c)$. Most of these chains have only one fold with $\Delta E \gg RT_c$ because the probability of having two or more such folds is even smaller: for two folds, it is $\sim\exp(-2\Delta E/RT_c)/2!$, and so on.

"Known" and "Actual' Energy Spectra

The above consideration concerns the actual energies of the folds. However, the same consideration must be valid also for the "known" (computed) fold energies E' which, in essence, summarize some "known" part of the actual interactions operating in the chain.

The only difference is that the dispersion $\langle (E')^2 \rangle$ of the "known" energies is less: $\langle (E')^2 \rangle = C^2\sigma^2$ rather than $\sigma^2 = \langle E^2 \rangle$ [see Eq. (5)]. Thus, one expects (Fig. 2b,c) that the "known" energy spectra also have only a few, if any, lines below the continuous part, and that the density $m_{kn}(E')$ of this continuous part obeys the equation

$$\ln m_{kn}(E') = \ln M - (E')^2/(2\sigma^2C^2) \equiv \ln M - (\sigma^2C^2)/(2R^2T^2). \qquad (10)$$

Here C is the correlation coefficient between the "actual" and "known" energies defined by Eq. (1), and $T = (R\partial \ln m_{kn}/\partial E')^{-1} = \sigma^2C^2/(RE')$ is the temperature corresponding to energy E'.

The continuous part of the "known" spectrum begins when $\ln m_{kn} = 0$ and $E' \approx E'_c$, where the critical energy

$$E'_c = -(\sigma C)\sqrt{2 \ln M} = CE_c. \qquad (11)$$

The critical temperature $T'_c = (C\sigma/R)/\sqrt{2 \ln M} = CT_c$ corresponds to the edge of the continuous spectrum (note that $T'_c \leq T_c$, and $-E'_c \leq -E_c$, as $C \leq 1$).

Now we can consider the possibility of predicting the native fold using a given set of energetic parameters.

This possibility is determined by the expected form of the spectrum of "known" energies. This spectrum, in turn, depends on the correlation coefficient C between the "known" and the "actual" energies and on the size of the gap ΔE which separates the native fold from the continuous part of the actual energy spectrum (see Fig. 2a–c). An unambiguous prediction is possible if such an energy gap between the native and the other folds is retained also in the "known" spectrum.

The expected form of the continuous part of the "known" spectrum is given by Eq. (10).

To estimate the expected "known" energy $\bar{E}'_i$ of fold i having an actual energy E_i the value of E'_i in Eq. (7) must be averaged over the values $\delta E'_i$ which cannot be known and must be treated as random errors

$$\overline{E'_i} = C^2E_i. \qquad (12)$$

This equation is valid for any fold, including the native fold N.

The actual energy of the native fold, $E_N = E_c - \Delta E$, is below the actual energies of all the other folds. The expected "known" energy of this fold, $\overline{E'_n} = C^2E_N$ is also below the expected "known" energy $\overline{E_i}$ of any other fold i. However, there are numerous "other" folds, and some of them must by chance acquire the "known" energies E'_i below the expected $\overline{E'_i}$ values [the standard deviation of E'_i from $\overline{E'_i}$ is $\sigma_e = \sigma(1-C^2)^{1/2}$, see Eq. (6)].

Equation (11) shows that the width of the "known" continuous spectrum is proportional to C rather than to C^2. As a result, the "known" native fold energy, $C^2E_N \pm \sigma_e$, competes with E_cC, the energy of the beginning of the "known" continuous spectrum, rather than with the "known" energy $C^2E_i \pm \sigma_e$ of any particular fold i (including the fold second in the rating list of the "actual" energies). This is especially typical for the "protein-like" sequences where the actual native fold energy is far below the energies of all the other folds.

An unambiguous prediction of the native fold is possible until $\overline{E'_N} = C^2E_N < E'_c = CE_c$, i.e., until the "known" native fold energy is, as a rule, below the continuous spectrum of the "known" energies. This is the case when the correlation coefficient C is greater than

$$C_o = E_c / E_N. \quad (13)$$

When the correlation coefficient C is smaller than C_o, the native fold can be singled out by energy calculations and only the expected rating of the "known" native fold energy in the "known" energy spectrum cannot be estimated. This estimate follows from Eqs. (10), (12), and (13):

$$\ln m_{kn}(\overline{E'_N}) = \ln M - (\overline{E'_N})^2 / (2\sigma^2 C^2) = (1 - C^2/C_o^2) \ln M. \quad (14)$$

When $C \leq C_o$, Eq. (14) gives the number of "candidates" for the native fold role. [More strictly, $m_{kn}(E'_N)$ is the number of folds with $E'_i = \overline{E'_N}$; the folds with $E'_i < \overline{E'_N}$ are also candidates; however, this virtually does not alter the above estimate, since most of the folds with energy below E occur between E and $E-RT$, see Eq. (8).] The number of candidates depends only on three parameters: the total number of folds, the energy gap dividing the native fold from the continuous part of the actual energy spectrum, and the correlation between the computed and actual energies.

The list of candidates for the native fold role is shorter for the "protein-like" sequences where the gap $\Delta E = E_c - E_N$ is large, for those generalized models of protein structures (e.g., "folding patterns") which diminish the variety of conformations, and for accurate sets of energetic parameters.

There is only one candidate when $C > C_o$, the native fold itself.

The above consideration remains virtually the same if we transpose the "actual" and "known" energies (compare the ratios $C^2 E_N / C\ E_c$ and $E'_1 / (E'_c / C)$ in Figure 2a–c and d–f: they are both proportional to C).

Consequently, the fold with the lowest "known" energy must have approximately the same position in the list of "actual" energies, as the native fold has in the list of "computed" energies.

The "Perfect Temperature" T_*

The native fold position in the list of the "known" energies can be found by statistical mechanics. To this end the calculations must be carried out at the "perfect" temperature

$$T_* = C_0\ T_c \equiv (E_c / E_N) T_c \quad (15)$$

independently of value of the correlation between "known" and "actual" energies.

Indeed, when the correlation coefficient $C > C_0$, the expected "known" energy $\overline{E'_N}$ of the native fold is below the edge of the continuous part of the "known" energy spectrum. This edge has a temperature CT_c which is above T_* since $C > C_0$, and the temperatures of the higher-energy regions of the spectrum are higher, see equation (8). Therefore, the native fold alone dominates in the "known" partition function

$$Z_{kn}(T) = \sum_{i=1}^{M} \exp(-E'_i / RT) \quad (16)$$

computed at the temperature $T = T_*$ (as well as at $T < T_*$).

When the correlation coefficient $C = C_0$, the expected "known" native fold energy $\overline{E'_N} = C^2\ E_i$ is at the edge of the "known" continuous spectrum which has just the temperature $T_* = T_c C_0$. This means that the native fold is among the few which give the main contribution to $Z_{kn}(T_*)$.

When the correlation coefficient $C < C_0$, the native fold is expected to be within the continuous part of the "known" energy spectrum. The native fold occurs in that part of the spectrum which has a temperature T_* since

$$T(\overline{E'_N}) = -C^2\sigma^2 / (R\overline{E'_N}) = -C^2\sigma^2/(RE_N C^2) = \frac{E_c}{E_N}[-\sigma^2/RE_c] = C_0 T_c \quad (17)$$

[since $E_N C_0 = E_c$, see Eq. (11)]. Thus, the native fold is again among those which correspond to the temperature T_*, i.e., among the folds which dominate in $Z_{kn}(T_*)$.

The theorem that the "known" energies of the folds and their "actual" energies correspond to the same temperature is proved here under the assumption that both the "known" and the "actual" spectra have a Gaussian form. However, we show in the Appendix that this theorem is valid for any form of the spectra, provided only that the differences between the "known" and "actual" energies can be treated as random ones.

It is noteworthy that the temperature $T_* = C_0 T_c$ is somewhat *below* T_c, while the melting temperature T_m of a "protein-like" chain is somewhat *above*[20,26] T_c (see Discussion).

Consequently, the candidates for the native fold role are singled out by statistical mechanics when the calculations are carried out at the temperature T_* which is somewhat below that of protein melting. The value of this "perfect" temperature is considered in the Discussion.

COMPUTER EXPERIMENTS

To explain the obtained results we used a computer model of protein shown in Figure 3a. This simple model is utilized only as an example to illustrate and test the general analytical theory presented above. In this model, a "sequence" of the chain is given by potential energies of all the link-to-link contacts. To obtain a random sequence, each of the contact energies ε_{ij} (i, j are link numbers) is generated with a Gaussian probability $P(\varepsilon_{ij}) = (2\pi)^{-1/2} \exp(-\varepsilon_{ij}^2 / 2)$. The "native" (lowest-energy) conformation of a chain is found by exhaustive enumeration of all its 103,346 conformations[17] of the cubic folds

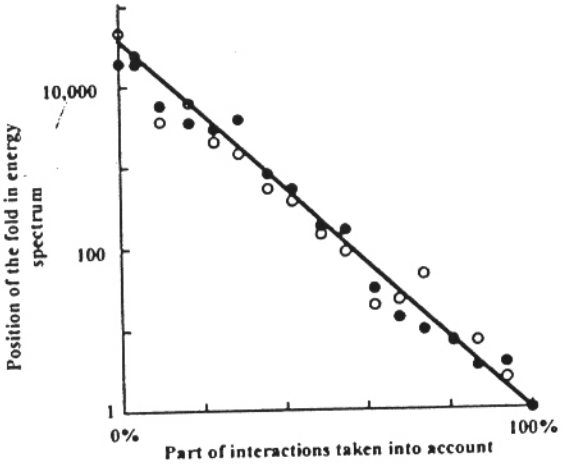

Fig. 4. Position of native folds (filled circles) in the "known" energy spectra calculated on the grounds of some part of all the interactions operating in the chain and position of folds of the lowest "known" energy (open circles) in the spectra of actual energies. Filled and open circles are obtained by computer experiments with the protein globule model shown in Figure 3a; each point is the average over 100 random sequences. The straight line corresponds to the theoretical estimate, see Eq. (14) with $C_0 = 1$. The "known" part of interactions is equal to C^2, where C is the correlation coefficient between "known" and "actual" energies.

(we neglect the less compact folds which means that we assume that the chain is in a very bad solvent).

The simulation of protein structure prediction is done as follows:

We "forget" (replace by zeros) the energies of some contacts chosen at random. The other contact energies (their fraction is α) remain "known" for us, and we calculate the "known" fold energies on this basis. The "forgotten" contact energies (their fraction is $1-\alpha$) contribute to the "random errors" of our estimates of fold energies. Using Eqs. (1) and (6), it can be shown that, in this case, the correlation coefficient C between "known" and "actual" energies is

$$C = \alpha^{1/2} = [1-(\text{fraction of changed interactions})]^{1/2} \quad (18)$$

Then we range the folds according to their "known" energies and find the position of the native fold (that with the lowest "actual" energy) among them.

Figure 4 summarizes the experiments with 100 random sequences. As the sequences are random, one expects that the actual native fold energy E_N is very close to E_c, and thus the expected position of native folds in the spectra of "known" energies must be given by Eq. (14) with $C_0 = 1$. The results shown in Figure 4 confirm this estimate.

We have found also the place that the fold of the lowest "known" energy occupies in the list of the "actual" energies. The corresponding points in Figure 4 also follow the line given by equation (14), as is expected theoretically.

Then we simulated a 3-D structure prediction for the sequences which exhibit a considerable ($\gg RT_c$) energy gap between the lowest-energy fold and the continuous energy spectrum. Thus, now we deal with the "protein-like" chains (see above). This is a more adequate simulation of protein structure prediction.

In essence, the experiment is the same as that described above, only the generation of sequences was different.

A random generation of the sequences which provide a large energy gap is too difficult, as the fraction of these sequences is rather small.[20] Therefore, we used the following procedure: we generate a random sequence, find its native fold, and add the negative energy term $-\Delta E/K$ to each of $K = 28$ link-to-link contact energies observed in this native fold.

The results shown in Figure 5 are in a good agreement with the theoretical estimates (13) and (14), both for the average position of the native folds in the "known" energy spectra, and (in this case, the agreement is somewhat worse) for the average position of the folds of the lowest "known" energy in the "actual" energy spectra.

DISCUSSION

Predictions

The above study shows that energy minimization is not a reliable tool to predict the protein structure from its amino acid sequence. This conclusion does not concern the technique of energy minimization and the troubles caused by a multiminimum problem. These difficulties, at least in some cases, can be overcame by simulated annealing, etc. Our conclusion is that the *result* of the minimization is not reliable.

Even a small uncertainty in energetic parameters can lead to an exponential increase of the number of possible candidates for the native fold role, since the errors increase with the uncertainty of energy, and the number of possible folds grows exponentially with energy. For the given error level, the list of candidates is especially long when the energy gap between the native fold and the continuous spectrum is small.

Under such circumstances one cannot prefer candidates and can rely only on common features of the lowest energy folds. These common features, if they exist, can be singled out by statistical mechanics provided the calculations are carried out in the vicinity of the optimal temperature T_*. These calculations give the best prediction which can be obtained using the given set of energetic parameters. It is noteworthy that these calculations show which structural features are most probable and thus can be predicted reliably, and which are less probable and therefore can be predicted only tentatively.

In other words, one has to compute the equilibrium state of the molecule and pay attention only to its most probable structural features (Fig. 6). This

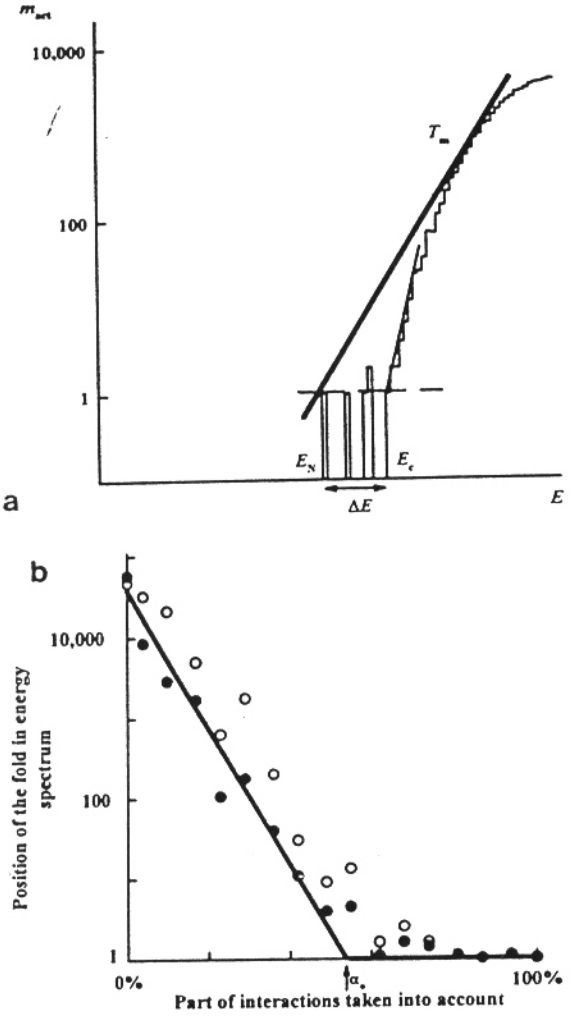

Fig. 5. (a) Density of the "actual" energy spectrum for one of the rare sequences producing a large energy gap ($\Delta E \approx |E_c|/3 \approx 9RT_c$ in this figure) between the native fold energy and the edge of the continuous spectrum. Thin tangent at the beginning of the continuous part has a slope $1/RT_c$, as in Figure 3b. Another tangent (solid line) has a slope $1/RT_m$; T_m is the melting temperature of native protein, i.e., $E_N = E(T_m) - RT_m \ln[m(E(T_m))]$. (b) Position of the native fold (filled circles) in the "known" energy spectra calculated from a part of interactions operating in the chain and the position of folds of the lowest "known" energy (open circles) in the spectra of actual energies. Each point is the average over 100 random sequences with the given energy gap. The solid line corresponds to the theoretical estimate [see Eqs. (13) and (14)]. $\alpha_0 = C_0^2$ is the minimal part of interactions which must be taken into account for unambiguous prediction of the native fold. For $\Delta E/|E_c| = 3$, $C_0 = 0.75$ and $\alpha_0 = 0.562$.

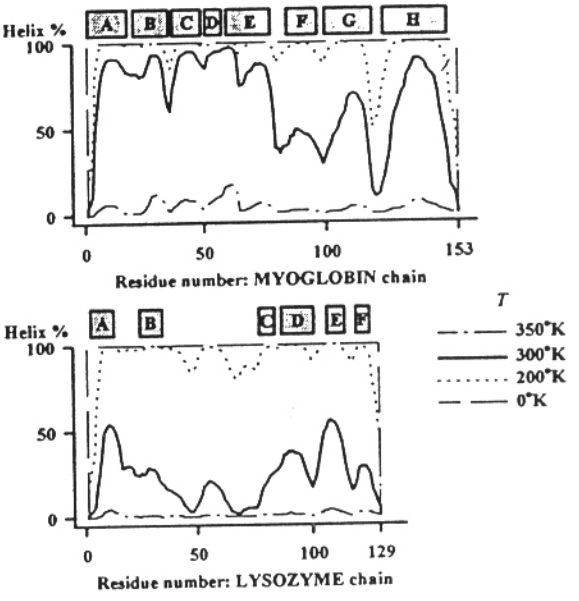

Fig. 6. α-Helical and coil states of the residues of unfolded protein chains. Calculations correspond to different temperatures T; the case 0°K corresponds to the energy minimum. The ordinate shows the average computed helicity for each residue in the chain. The algorithm of calculations has been given[30]; it is based on statistical mechanics and corresponds to a Zimm-Bragg's[27] model of helix–coil transition. Neither the β-structure nor the globular states are taken into account. Energetic parameters of helix–coil transition are the same as those previously given[30]; it has been demonstrated that they allow a rather accurate calculation of the helicities of oligo- and polypeptides in water solutions in a wide range of temperatures. Rectangles show the positions of α-helices in native myoglobin[36] and lysozyme.[37] It is clearly seen that the secondary structure of unfolded protein chains at temperature $T \approx 300$ K reflects the secondary structure of native proteins and that a reasonable prediction of the α-helical and the coil region in globular proteins can be based on statistical mechanics calculations held at $T \approx 300$ K, but it cannot be based on the minimization of secondary structure energy which gives only one continuous α-helix in all cases (minimization corresponds to calculations at $T = 0$ K).

conclusion is a general one, independent of what interactions are taken into account and how precisely they are estimated.

The search for the equilibrium state of a chain and computation of its properties is carried out in the Zimm-Bragg's[27] and related[28–30] models, it is used in protein structure predictions based on the molecular field theory,[11] etc.

Theoretically, the calculations must be carried out at the temperature T_* which is somewhat below that of protein melting (in the next section we show that $T_* \sim 280–300$ K). However, this requires a preliminary "normalization" [see Eq. (3)] of the parameters used for energetic computations: the normalization, in turn, needs a comparison of at least a few computed energies with experiment. When such a preliminary adjustment of the parameters is not possible due to the lack of adequate experimental data, it is worth trying different T_* values and choosing the one given the best results for the proteins whose structures are known.

The quality of prediction is rather sensitive to the accuracy of energetic parameters determined by the value of the correlation coefficient C between actual and computed (with these parameters) energies.

When the quality of these parameters is high enough ["enough" according to Eq. (13), means that the C of the parameter set is above the threshold C_0 determined by the native fold stability], the com-

puted native fold energy is below that of the computed energies of all the other folds. In this (and only in this) case the native fold can be singled out unambiguously—either by a search for the energy minimum, or by a search for the thermodynamic equilibrium at a temperature T_* or below, where the native fold is thermodynamically stable.

When errors are moderate, statistical mechanics singles out a limited set of low-energy folds, the native fold being one of them. The number of these "promising" folds increases exponentially with the increase of errors in energetic parameters (Fig. 4 and 5), and *there is no way* to find the "right" native fold among them using the energy calculations with the same set of parameters. This is one of the main conclusions of this paper. However, the right fold can be chosen, in principle, using appropriate additional experimental information, such as experimentally established residue-to-residue contacts, etc.

It is noteworthy that some native fold features (such as secondary structure location in the sequence[29,30] and space[10]) can be predicted from the common features of the folds with a low "known" energy. If the majority of these folds share some features, they are probably shared also by the native fold (Fig. 6). These common features are revealed using thermodynamic averaging performed by the algorithms of statistical mechanics.[10,28] At the same time, the fold of the minimal "calculated" energy can have little in common with the fold which is actually the lowest-energy one (Fig. 6).

Finally, when the errors are great (or, in other words, when too few of the "actual" interactions are taken into account), the set of "promising" folds is huge, their common features are slurred over, and structure prediction is impossible.

It should be mentioned that the "known" interactions can be too few for any prediction of the overall chain fold, but sufficient for the prediction (again, in a statistical sense) of some elements of this fold, e.g., for secondary structure prediction. Actually, the Boltzmann-like equation of Pohl[31]

$$OCCURRENCE \sim \exp(-ENERGY/RT_*) \quad (19)$$

which relates the internal energy of a structural element with its occurrence in native proteins is an example of the *statistical prediction* done on the grounds of only *one* (among hundreds) of the interactions operating in a protein globule. This prediction, though not perfect, is rather reliable when one has to choose from only a few possibilities and *ENERGY* exceeds a few RT_*.

Some Estimates

From Eqs. (8) and (9), it follows that when $T>T_c$, the molten protein has the energy $E(T)=-\sigma^2/RT > E_c$, the entropy $S(T) = R\,[\ln M-\sigma^2/(2R^2T^2)]$ and, as a result, the free energy $F(T) = -\sigma^2/(2RT)-TR\ln M$. A "protein-like" chain melts at the temperature T_m when the native fold energy is equal to the free energy of the denatured state: $E_N = F(T_m)$ (see Fig. 5a).

The enthalpy of this melting is $\Delta H_m = E(T_m)-E_N$ and the jump of heat capacity is $\Delta C_m\,(T_m) = \sigma^2/(RT^2)$. Now one can find the ratio E_c/E_N as a function of three observable parameters: T_m, ΔH_m, and ΔC_m:

$$E_c/E_N = (1 - [\Delta H_m/(\Delta H_m + T_m\Delta C_m)]^2)^{1/2} \quad (20)$$

Since $E_c/E_N = C_0$ [see Eq. (13)], this is an estimate of the minimal correlation between the "actual" and "computed" energies necessary for the reliable prediction of a unique protein structure.

Taking the values T_m, ΔH_m, and ΔC_m from the experiment,[32] one can see that the value of C_0 is not less than 0.95. The same, in essence, estimate of C_0 follows from the fact that the protein structure is destroyed by 3–5% of random mutations in the chain[34] [see Eq. (18)].

According to a recent estimate of Bryngelson,[33] an unambiguous native fold prediction is possible when $<\delta E^2>/<E^2> \le 1/N$, which, according to Eq. (6), means that C_0 is about $(1-1/N)^{1/2}\approx 0.995$ for a protein of 100 residues. Our estimate of C_0 is less pessimistic because we take into account that the native fold is divided from the continuous spectrum by a considerable gap in the protein chains.

The obtained estimate ($C_0\approx 0.95$) is, of course, rather tentative; however, it gives us the possibility of understanding how far the quality of energetic parameters is from that necessary for the unambiguous prediction of protein structure: even the correlations between different sets of experimental estimates[35] are below 0.85–0.9, to say nothing of the well-known uncertainty as regards the dielectric constant values.

Thus, all the predictions which are possible at present and in the near future can be only probabilistic ones and can hardly give a single unique protein structure.

Folding

All the above considerations concern not only protein structure prediction but protein folding as well.

The early stages of folding are governed by only a part of all the interactions (first, by hydrogen bonding, then, also by hydrophobic forces, and, finally, the specific van der Waals forces come into play[4,5]). Each stage singles out a set of folds which has a sufficiently low energy of interactions operating at this stage of folding.

It is noteworthy that at each stage the interactions do not fix all, but only some degrees of protein chain freedom. For example, at the stage of secondary structure formation, each secondary structure location in the chain corresponds to a multitude of folds with different conformations of the loop regions

which are not fixed by the hydrogen bonds. In the same way, the forces forming the molten globule do not fix side chain rotamers, and so on: in the intermediates, each structure corresponds to many particular conformations.

When the folding intermediate demonstrates some definite structural feature (e.g., some secondary structure type), this means that this feature is inherent in many low-energy folds. Therefore this feature is the most probable (though, not absolutely obligatory) also for the native fold (Fig. 6).

Each new step of folding involves new forces, and the number of dominating folds becomes less and less. When folding takes place at the temperature T_* (which, according to our qualitative estimate, must be "somewhat below that of protein melting"), the folds with the native features (e.g., with the native secondary structure) are always in the set of dominating folds. Thus, T_* can be estimated as the temperature typical for the formation of native-like intermediates in protein folding:[1-8] this temperature is usually ~280–300 K, and namely this temperature T_* should be used in protein structure prediction.

It should be stressed that before using the optimal temperature T_* in calculations, the energetic parameters must be compared with experiment and normalized [see Eq. (3)], and their quality must be estimated by Eq. (1).

In this study we assumed that parameter errors are random, though the existing sets of parameters contain also systematic errors: e.g., some overestimate electrostatics and some underestimate it. It follows that some energetic parameters can be more appropriate for computation of proteins stabilized by hydrophobic forces, and others for those stabilized by electrostatic ones.

The deviation from the simple "random error" model does not change the main conclusion: Statistical mechanics rather than energy minimization can reveal the features of protein structure which can be reliably predicted.

ACKNOWLEDGMENTS

We are grateful to A. G. Raiher for careful reading of the manuscript and helpful remarks. We acknowledge the financial support of the Protein Engineering Scientific Council of the Russian Academy of Sciences (grant 104) and of the Russian Fundamental Research Foundation (grant 93-04-6636).

REFERENCES

1. Kuwajima, K., Yamaya, H., Miwa, S., Sugai, S., Nagamura, T. Rapid formation of secondary structure framework in protein folding studied by stopped flow circular dichroism. FEBS Lett. 221:115–118, 1987.
2. Udgaonkar, J.B., Baldwin, R.L. Early folding intermediate of ribonuclease A. Proc. Natl. Acad. Sci. U.S.A. 82: 8197–8201, 1990.
3. Bycroft, M., Matouschek, A., Kellis, J.T., Jr., Serrano, L., Fersht, A.r. Determination and characterization of a folding intermediate by NMR. Nature (London) 346:488–490, 1990.
4. Ptitsyn, O.B. How does protein synthesis give rise to the 3D-structure?. FEBS Lett. 285:176–181, 1991.
5. Ptitsyn, O.B. The molten globule state. In: "Protein Folding." Creighton, T.E. (ed.) New York: W.H. Freeman, 1992: 243–300.
6. Ptitsyn, O.B., Semistnov, G.V. The mechanism of protein folding. In: "Conformation and Forces in Protein Folding." Nall, B.T., Creighton, T.E. (eds.). Washington DC: AAAS Press, 1991: 155–168.
7. Matouschek, A., Serrano, L., Fersht, A.R. The folding of enzyme. IV. Structure of an intermediate in refolding of barnase studied by a protein engineering procedure. J. Mol. Biol. 224:819–835, 1992.
8. Shakhnovich, E.I., Finkelstein, A.V. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is the first order phase transition. Biopolymers 28:1667–1680, 1989.
9. Covell, D.G., Jernigan, R.L. Conformations of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.
10. Finkelstein, A.V., Reva, B.A. Search for the most stable folds of protein chains. Nature (London) 351:497–499, 1991.
11. Fasman, G.D. "Prediction of Protein Structure and the Principles of Protein Conformation." New York: Plenum Press, 1989.
12. Thornton, J.M., Flores, T.P., Jones, D.T., Swindells, M.B. Prediction of progress at last. Nature (London) 354:105–106, 1991.
13. Derrida, B. Random-energy model: An exactly solvable model of disordered systems. Phys. Rev. B 24:2613–2626, 1981.
14. 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.
15. Bryngelson, J.B., Wolynes, P.G. A simple statistical field-theory of heteropolymer collapse with application to protein folding. Biopolymers 30:177–188, 1990.
16. Shakhnovich, E.I., Gutin, A.M. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys. Chem. 34:187–199, 1989.
17. 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.
18. Shakhnovich, E.I., Gutin, A.M. Implication of thermodynamics of protein folding for evolution of primary sequences. Nature (London) 346:773–775, 1990.
19. 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.
20. 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. 24. Biswas, B.B., Roy, S. (eds.). New York: Plenum Press, 1995:1–26.
21. Finkelstein, A.V., Badretdinov, A.Ya., Gutin, A.M. Why do protein architectures have Boltzmann-like statistics? Proteins 1995, in press.
22. 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.
23. 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.
24. Hudson, D.J. "Statistics." Geneva: CERN, 1964.
25. Landau, L.D., Lifshitz, E.M. "Statistical Physics." London: Pergamon, 1959.
26. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Optimal protein-folding codes from spin-glass theory. Proc. Natl. Acad. Sci. U.S.A. 89:4918–4922, 1992.
27. Zimm, B.H., Bragg, J.R. Theory of the phase transition between helix and random coil in polypeptide chains. J. Chem. Phys. 31:526–535, 1959.
28. Finkelstein, A.V. Theory of protein molecule self-organization. III. A calculating method for the probabilities of the secondary structure formation in an unfolded protein chain. Biopolymers 16:525–529, 1977.

29. Finkelstein, A.V., Ptitsyn, O.B. A theory of protein molecule self-organization. IV. Helical and irregular local structures of unfolded protein chains. J. Mol. Biol. 103:15–24, 1976.
30. Finkelstein, A.V., Badretdinov, A.Ya., Ptitsyn, O.B. Physical reasons for secondary structure stability. Proteins 10: 287–299, 1991.
31. Pohl, F.M. Empirical protein energy maps. Nature New Biol. 234:277–279, 1971.
32. Privalov, P.L., Makhatadze, G.I., Contribution of hydratation and non-covalent interactions to the heat capacity effect on protein folding. J. Mol. Biol. 224:715–723, 1992.
33. Bryngelson, J.D. When is a potential accurate enough for structure prediction? Theory and application to a random heteropolymer model of protein folding. J. Chem. Phys. 100:6038–6045, 1994.
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. Minor, D.L., Jr., Kim, P.S. Context is a major determinant of beta-sheet propensity. Nature (London) 371:264–267, 1994.
36. Watson, H.C. The stereochemistry of the protein myoglobin. Progr. Stereochem. 4:299–333, 1969.
37. Imoto, T., Johnson, L.N., North, A.C.T., Phillips, D.C., Rupley, J.A. Vertebrate lysozymes. Enzymes 7:665–868, 1972.

APPENDIX: EQUIVALENCE OF "KNOWN" AND "ACTUAL" TEMPERATURES

Suppose that the energies of different folds of a protein chain are computed using some set of energetic parameters. The computed energies E_1', . . . , E_M' of the folds "1", . . . ,"M" approximate the actual fold energies E_1, . . . ,E_M which cannot be known precisely.

It can always be assumed that the errors $\delta E_i = E_i - E_i'$ (where $i = 1, \ldots M$) do not correlate with the "known" fold energies E_i' [this corresponds to the best possible fit of the computed energies to the actual ones, see Eq. (4) and below]. As regards the "known" fold energies, this means that the errors δE_i depend on the sequence and folds in some random way. In other words, the values δE can be treated as "random" corrections of the "known" energies E' (Fig. 1a).

Our aim is to elucidate the connection between the "actual" and the "known" (i.e., calculated) energy spectra (Fig. 1b).

To obtain the basic estimate, we assume that the values δE are absolutely random: i.e., the probability of having a given value of the error δE_i does not depend either on the fold "i," or on its "known" energy E_i', or on the values of other errors.

Let $m_{act}(E)$ be the number of folds whose actual energy is E (strictly speaking, the number of folds with energies between E and $E + B$ where the energy unit B is a characteristic difference in interaction energies of different links), and $m_{kn}(E')$ be the number of folds whose "known" energy is E'.

Of course, $m_{act}(E)$ themselves cannot be computed exactly from m_{kn} due to random errors. However, it is possible to estimate the mean values $\bar{m}_{act}(E)$, averaged over the possible distribution of errors:

$$\bar{m}_{act}(E) = \int_{-\infty}^{+\infty} m_{kn}(E')P(E-E')dE'/D. \qquad \text{(A1)}$$

Here $P(\delta E)$ is a function which describes the distribution of errors: $P(\delta E)dE/D$ is the probability of the value of error falling between δE and $\delta E + dE$, and $D = <\delta E^2>^{1/2}$ is a characteristic width of the function P [see Eq. (6)].

It is reasonable to assume that the width D is greater than the distance between the lines of the energy spectrum. Indeed, the range of computation errors is hardly less than 1 kcal/mol, while an average density of a spectrum is very high: for a 100-residue protein, one has something like $\sim 2^{100}$ of lines for the energy range which hardly exceeds thousands of kcal/mol. This density drops at the spectrum edges (Figs. 2 and 3), but even at the edges the distance between the energy lines is usually only $\sim RT_c$, i.e., less than 1 kcal/mol, as T_c is a characteristic temperature of protein melting.[14–18] Thus, one can treat the spectrum density $m_{kn}(E')$ as a continuous function everywhere, with the exception, perhaps, of the very edge.

The integral (A1) can be estimated by the saddle-point method because the error function P has a limited width. This estimate gives

$$\bar{m}_{act}(E) = m_{kn}(E_1)P(E-E_1) \qquad \text{(A2)}$$

when the range of "known" energies is greater than D (this corresponds to the most interesting case of small and moderate errors).

The saddle point $E_1(E)$ is the point where the integrated function $m_{kn}(E')P(E-E')$ has its maximum over E', i.e., where

$$\frac{\partial}{\partial E'}[m_{kn}(E')\,P(E - E')]|_{E' = E_1(E)} = 0. \qquad \text{(A3)}$$

As

$$\frac{\partial}{\partial E'}(m_{kn}P) = m_{kn}P\left(\frac{\partial}{\partial E'}\ln m_{kn} + \frac{\partial}{\partial E'}\ln P\right)$$

and

$$\frac{\partial}{\partial E'}P(E-E') \equiv -\frac{\partial}{\partial E}P(E-E'),$$

Eq. (A3) is equivalent to

$$\frac{\partial}{\partial E'}\ln m_{kn}(E')_{E' = E_1(E)} = \frac{\partial}{\partial E'}\ln P(E-E')|_{E' = E_1(E)}. \qquad \text{(A4)}$$

The point $E_1(E)$ has a simple physical meaning: E_1 is the most probable "known" energy of a fold whose "actual" energy is E.

Finally, we have to exclude the unknown function P. To this end we differentiate Eq. (A2):

$$\frac{d}{dE}\overline{m}_{act}(E) = \frac{dE_1}{dE}\frac{\partial}{\partial E'}[m_{kn}(E')\,P(E-E')]|_{E'=E_1(E)}$$

$$+ m_{kn}(E_1)\frac{\partial}{\partial E}P(E-E')|_{E'=E_1(E)}. \quad \text{(A5)}$$

The first term in the right part of this equation is zero according to Eq. (A3); taking into account also Eqs. (A2) and (A4), we obtain

$$\frac{d}{dE}\ln m_{act}(E) = \frac{d}{dE'}\ln m_{kn}(E')_{E'=E_1(E)}. \quad \text{(A6)}$$

This equation gives a relationship between the energies E and E_1 through the densities of the "actual" and "known" energy spectra.

Thus [since $R \ln m(E)$ is the enthropy $S(E)$, and $dS/dE \equiv 1/T$][25] Eq. (A6) means that the folds constituting the part of the "actual" energy spectrum which corresponds to the temperature T refer to the part of the "known" energy spectrum, which corresponds to the *same* temperature T.