3D Protein Folds: Homologs Against Errors—a Simple Estimate Based on the Random Energy Model

Alexei V. Finkelstein

Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

(Received 1 December 1997)

One cannot predict the 3D structure of a protein directly from its sequence because of errors in the energy estimates. However, using a set of homologs (proteins with nearly identical 3D structures despite numerous amino acid mutations in their chains) it is possible to average the fold energies over the homologs (this diminishes energy errors) and to predict the common 3D fold of these proteins from the set of their amino acid sequences. [S0031-9007(98)06152-3]

PACS numbers: 87.15.By, 36.20.Ey, 75.10.Nr

The determination of the three-dimensional (3D) protein structure from the amino acid sequence of its chain is a long-standing challenge to molecular biophysics. This problem is known to be soluble in principle, since a protein chain itself resolves it by spontaneous folding *in vitro* as well as *in vivo* [1]. However, attempts of *a priori* computations of protein structures from the sequences have met with little success, unless the protein in question is similar to a protein with an already known 3D fold [2]. Protein fold recognition is more successful, but the number of "recognizable" folds is restricted by the set of already known 3D structures [2,3]. The principal obstacle to an *a priori* protein structure prediction is connected with unavoidable random errors in the energetic parameters used to estimate the stability of competing 3D structures of a protein chain [3]. Actually, one can state that a protein structure is unpredictable from its sequence alone.

A well known observation is that homologous proteins have nearly identical 3D folds despite very numerous amino acid mutations [4]. It has been observed that one can improve protein structure predictions by using predictions made for many homologous chains [2–6]. In secondary structure prediction, the statistical errors are reduced by averaging the predictions at each position of the aligned sequences [5]; however, a direct generalization of this approach to prediction of 3D protein structures causes a problem.

The goal of this paper is to show how the homologs should be used in the course of 3D structure prediction by energy calculations and to prove that, using a set of homologs, it is possible to make an *a priori* prediction of their common 3D fold from the set of their amino acid sequences—despite the roughness of the energy estimates.

Experiments [1] and theory [7,8] point out in a very suggestive way that a choice of the native (i.e., biologically functional) protein structure is determined by its high stability, i.e., a fold serves as the native fold of a given protein chain only if its energy is at the very bottom of the chain's energy spectrum. [It should be noted that the term "energy" is used here only for simplicity: actually, one has to speak about free energy of a fold, since the protein is in water and the solvent-mediated (hydrophobic and electrostatic) forces contribute to its stability.]

The protein is a dense heteropolymer globule. Its chain has a fixed sequence which looks very much like a random sequence of amino acid residues, but this sequence has one predominant ("native") fold [4].

It is widely recognized that the energy spectra of the dense globules formed by random heteropolymers are sufficiently (for thermodynamics) [9] described by the spectra characteristic for the random energy model (REM) [10]. This model has been borrowed from the statistical mechanics of spin glasses.

The REM assumes that each of the M possible structures (of the M chain folds, as far as polymers are concerned) acquires its energy E_i with a Gaussian probability $p_1(E_i) \propto \exp(-E_i^2/2\sigma^2)$, $i = 1, \ldots, M$, independently of the energies of the other structures $j \neq i$. As a result, the density of the REM energy spectrum has a Gaussian form with the mean $\langle E \rangle = 0$ and the dispersion $\langle E^2 \rangle = \sigma^2$; here $\langle \cdots \rangle$ means averaging over all the M structures: $\langle E \rangle = \frac{1}{M} \sum_{i=1}^{M} E_i$, etc. The number of folds, M, is very large: it exponentially depends on L, the number of chain links; σ^2 is proportional to the number of intrachain interactions, i.e., to L.

For a vast majority of heteropolymers (with random sequences of links) the low-energy end of the spectrum corresponds to the critical energy $E_C = -(2 \ln M)^{1/2} \sigma$ [here $M \times p_1(E_C) \sim 1$], the energy spectrum being dense and quasicontinuous above E_C [8,9].

A protein chain differs from the majority of random sequences by one unusually stable native fold. The energy E_N of this fold N is sufficiently below E_C, while the whole quasicontinuous energy spectrum (above E_C) remains the same, and, as a result, the averaged characteristics of the spectrum ($\langle E \rangle$, $\langle E^2 \rangle$, E_C) remain essentially the same as for random heteropolymers [9].

To predict protein structure, one has to identify the fold N with the minimal energy.

However, an attempt to identify the lowest-energy fold by the energy calculations encounters unavoidable errors in the energetic parameters [11]. Now we put apart all

the difficulties connected with sorting of the folds—in principle they can be solved *when* the native fold energy is well below the energies of its competitors [7]. We stress that due to the errors in the energy estimates the calculated energy of the fold N is above the calculated energies of some other folds.

Indeed, instead of the true energy E_i of any structure i ($i = 1, \ldots, M$) one knows only its "calculated" energy E_i^* which can be represented as $E_i^* = E_i + \Delta E_i$, where ΔE_i is a random energy estimate error which does not correlate with E_i [actually [11], E_i^* should be represented as $\mu \times (E_i + \Delta E_i)$; however, the multiplier μ plays no role in the comparison of structure energies, and one can omit it; neither does the mean value of error play any role in the comparison, so one can assume that $\langle \Delta E \rangle = 0$].

Let us assume that not only the actual chain fold energies follow the random energy model, but also that the calculated (with errors) energies are randomly distributed around their true values (i.e., we consider the "best" case when the systematic energy errors are excluded).

Then the probability of error ΔE_i has a Gaussian form: $p_{\text{err}}(\Delta E_i) \propto \exp(-\Delta E_i^2/2\delta^2)$—like the energy E_i, the ΔE_i value summarizes many independent terms (i.e., the errors in many independent interactions); and the dispersion of errors $\delta^2 = \langle \Delta E^2 \rangle$ is (like σ^2) proportional to L, the number of chain links. The absence of the correlation between the true energies and the errors means that $\langle E \Delta E \rangle = 0$. Hence, the dispersion of the calculated energies is $\langle (E^*)^2 \rangle = \sigma^2 + \delta^2$ and the coefficient of correlation between them and the true energies is $C_{\text{comp}} = \langle E^* E \rangle / [\langle (E^*)^2 \rangle \langle E^2 \rangle]^{1/2} = (1 + \delta^2/\sigma^2)^{-1/2}$. The calculated energies E_i^* of the non-native structures have a Gaussian distribution $p_1^*(E_i^*) \propto \exp[-(E_i^*)^2/2(\sigma^2 + \delta^2)]$ [this is a result of integration of the two-dimensional correlated Gaussian distribution $p_2(E_i, E_i^*) \propto \exp[-E_i^2/2\sigma^2 + (E_i^* - E_i)^2/2\delta^2]$ over E_i] and are independent from one another.

One can estimate the low-energy edge of the computed energy spectrum: $E_C^* = -(2 \ln M)^{1/2}(\sigma^2 + \delta^2)^{1/2} = E_C/C_{\text{comp}}$. Since C_{comp} is always below 1, the computed E_C^* is below the true E_C. At the same time, the expected calculated value of the native fold energy E_N^* is about $E_N \pm \delta$ and (since one can neglect $\delta \sim L^{1/2}$ because $|E_N| > |E_C| \sim L$) E_N^* is close to E_N. Thus, due to the errors in the energies, the native structure is *not* the structure of the lowest *calculated* energy and it *cannot* be distinguished by *any* energy calculation when the accuracy of the energy estimates is insufficient, i.e., when $C_{\text{comp}} < C_0 = E_C/E_N$ [11].

Up to now there is no reliable estimate of C_0 (while very approximate estimates show that $C_{\text{comp}} \approx 0.5$–$0.8$ for different employed sets of energy parameters; see [3], and references therein), but it definitely seems [3,11] that current accuracy of the energy estimates is much below the limit necessary for protein structure prediction, and therefore the protein structure is now unpredictable from *its* sequence *alone*.

However, one can use additional information to predict protein structures.

It is known that homologous proteins have nearly identical 3D structures despite the very numerous amino acid mutations in their sequences [4]. This opens a possibility to predict the common fold of homologous proteins from the *set* of their amino acid sequences despite the errors in energy calculations. Such a prediction can be possible because all the homologs observed among natural proteins (i.e., hemoglobins of horse, carp, sea worm, etc.) are not randomly mutated initial sequence: they are naturally selected to stabilize the same fold (all the hemoglobins stabilize the globin fold, all the immunoglobulins stabilize the immunoglobulin fold, etc.).

It is noteworthy that homology of natural proteins can be reliably established when they have as small as 35%–40% of identical amino acid residues in identical positions [4], while protein engineering can create completely different folds with as much as 60% of identical residues [12]; this stresses that protein fold is preserved by natural selection rather than by the sequence identity itself.

Let us consider Γ homologous chains, and average the computed energy of each structure over the homologs. To reveal a sense of the averaging, let us start with its simplest form:

$$\epsilon_i^* = \frac{1}{\Gamma} \sum_{\lambda=1}^{\Gamma} E_i^{(\lambda)*}. \tag{1}$$

Here $E_i^{(\lambda)*}$ is the computed energy of chain λ ($\lambda = 1, \ldots, \Gamma$) in structure i ($i = 1, \ldots, M$). As mentioned above, the native fold's calculated energy $E_N^{(\lambda)*} \approx E_N$ for each chain λ, so that $\epsilon_C^* \approx E_N$, and the calculated energies of the other folds have Gaussian distributions with the average characteristics $\langle E^{(\lambda)*} \rangle = 0$ and $\langle (E_i^{(\lambda)*})^2 \rangle = (\sigma/C_{\text{comp}})^2$.

Thus, the mean averaged computed energy $\langle \epsilon^* \rangle = \langle \frac{1}{\Gamma} \sum_{\lambda=1}^{\Gamma} E^{(\lambda)*} \rangle = 0$, and the dispersion of the homolog-averaged computed energies ϵ^* is

$$\begin{aligned}
\langle (\epsilon^*)^2 \rangle &= \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{\Gamma} \sum_{\lambda=1}^{\Gamma} E_i^{(\lambda)*} \right) \left(\frac{1}{\Gamma} \sum_{\mu=1}^{\Gamma} E_i^{(\mu)*} \right) \\
&= \frac{1}{\Gamma^2} \sum_{\lambda=1}^{\Gamma} \langle (E^{(\lambda)*})^2 \rangle + \frac{1}{\Gamma^2} \sum_{\lambda=1}^{\Gamma} \sum_{\mu=1, \mu \neq \lambda}^{\Gamma} \langle E^{(\lambda)*} E_i^{(\mu)*} \rangle \\
&= (\sigma/C_{\text{comp}})^2 \left(\frac{1}{\Gamma} + \frac{1}{\Gamma^2} \sum_{\lambda=1}^{\Gamma} \sum_{\mu=1, \mu \neq \lambda}^{\Gamma} C_h^{(\lambda\mu)} \right).
\end{aligned} \tag{2}$$

Here $C_h^{(\lambda\mu)} = \langle E^{(\lambda)*} E^{(\mu)**} \rangle / [\langle (E^{(\lambda)*})^2 \rangle \langle E^{(\mu)*})^2 \rangle]^{1/2}$ is the coefficient of correlation between the computed fold energies for chains λ and μ. Using the averaged correlation coefficient

$$\bar{C}_h = \left(\sum_{\lambda=1}^{\Gamma} \sum_{\mu=1, \mu \neq \lambda}^{\Gamma} C_h^{(\lambda\mu)} \right) \Big/ \Gamma(\Gamma - 1) \tag{3}$$

one can present Eq. (2) in a simple form

$$\langle(\epsilon^*)^2\rangle = (\sigma/C_{\rm comp})^2[\bar{C}_h + (1 - \bar{C}_h)/\Gamma]. \quad (4)$$

The dispersion $\langle(\epsilon^*)^2\rangle$ of the homolog-averaged computed energies is *smaller* than the dispersion $\langle(E^*)^2\rangle = (\sigma/C_{\rm comp})^2$ of the energies, separately computed for each of the homologs.

The computed energies $E_i^{(1)*}, \ldots, E_i^{(\Gamma)*}$ of fold i in the homologs $1, \ldots, \Gamma$ correlate between themselves, but they are independent of the energies of the other folds $j \neq i$. In the spirit of the REM, one can treat $E_i^{(1)*}, \ldots, E_i^{(\Gamma)*}$ as a result of random sampling from the Γ-dimensional (and now correlated) Gaussian distribution. The probability of ϵ_i^* is the result of integration of this distribution over the $E_i^{(1)*}, \ldots, E_i^{(\Gamma)*}$ values under the condition $\frac{1}{\Gamma}\sum_{\lambda=1}^{\Gamma} E_i^{(\lambda)*} = \epsilon_i^*$, i.e., the probability of ϵ_i^* has a conventional Gaussian form. Thus, the spectrum of the homolog-averaged computed energies ϵ_i^* $(i = 1, \ldots, M)$ also has a Gaussian form; and the above estimated $\langle(\epsilon^*)^2\rangle$ is its dispersion.

Correspondingly, the low-energy edge of the quasicontinuous part of the homolog-averaged computed energy spectrum is $\epsilon_C^* = -(2\ln M)^{1/2}\langle(\epsilon^*)^2\rangle^{1/2} = (E_C/C_{\rm comp})[\bar{C}_h + (1 - \bar{C}_h)/\Gamma]^{1/2}$. The ϵ_C^* is significantly above the point $E_C^* = -(2\ln M)^{1/2}\langle(E^*)^2\rangle^{1/2} = E_C/C_{\rm comp}$, where the edges of all the individual computed energy spectra were before the averaging over the homologs.

Now one can find the native fold by its computed energy—*when* ϵ_C^* is higher than $\epsilon_N^* \approx E_N$. The above estimates show that this condition is fulfilled when

$$(C_{\rm comp})^2 > (C_0)^2[\bar{C}_h + (1 - \bar{C}_h)/\Gamma]. \quad (5)$$

Protein structure prediction can be successful when $\bar{C}_h$ is small and Γ is large.

In practice Γ can be about 5 or 10 and $\bar{C}_h \approx 0.15$–0.2—the latter means that a pair of homologs has 15%–20% of identical amino acid residue pairs, i.e., 40%–45% of identical residues in identical chain positions. With $\Gamma = 5$–10 and $\bar{C}_h = 0.15$–0.2, a successful prediction can be done when $C_{\rm comp}$ is as low as 0.5–0.6, which does not exceed the current level of correlation between theoretical and experimental estimates of protein structure stability (see [3], and references therein). Thus, a homolog-based fold prediction must be feasible, and, in particular, a folding simulation must bring a correct native structure [7] at least for small proteins.

Actually, Eq. (1) gives only the simplest form of the homolog-averaged computed energy; the general form is $\epsilon_i^* = \sum_{\lambda=1}^{\Gamma} a_\lambda E_i^{(\lambda)*}$ $(i = 1, \ldots, M)$, where $\sum_{\lambda=1}^{\Gamma} a_\lambda = 1$. It is easy to show that the minimal possible dispersion of the ϵ^* values,

$$\langle(\epsilon^*)^2\rangle = \left(\frac{\sigma}{C_{\rm comp}}\right)^2 \Big/ \sum_{\eta=1}^{\Gamma}\sum_{\mu=1}^{\Gamma}[X^{-1}]_{\eta\mu}, \quad (6)$$

and consequently the highest ϵ_C^* value, is achieved when $a_\lambda = \sum_{\mu=1}^{\Gamma}[X^{-1}]_{\lambda\mu}/\sum_{\eta=1}^{\Gamma}\sum_{\mu=1}^{\Gamma}[X^{-1}]_{\eta\mu}$; here $[X^{-1}]_{\lambda\mu}$ is the $\lambda\mu$th element of the inverse of a matrix $[X]$ having elements $[X]_{\lambda\mu} = C_h^{(\lambda\mu)}$ when $\lambda \neq \mu$ and $[X]_{\lambda\lambda} \equiv 1$. When all $C_h^{(\lambda\mu)} = \bar{C}_h$ at $\lambda \neq \mu$, all $a_\lambda = 1/\Gamma$ and Eqs. (5) and (6) coincide.

Certainly, a transition from the simplified REM to real proteins needs further investigations which can modify the numerical estimates given here, but even the qualitative results of this work show that, using a large set of distant homologs, one must be able to predict their common 3D fold from the set of their amino acid sequences.

I am grateful to O. B. Ptitsyn for stimulating discussions. This work has been supported in part by the Russian Foundation for Basic Research and by an International Research Scholar's Award from the Howard Hughes Medical Institute.

[1] C. B. Anfinsen, Science **181**, 223 (1973); A. R. Fersht, Curr. Opin. Struct. Biol. **7**, 3 (1997).
[2] R. L. Dunbrack, Jr., D. L. Gerloff, M. Bower, X. Chen, O. Lichtarge, and F. E. Cohen, Folding & Design **2**, R27 (1997).
[3] A. V. Finkelstein, Curr. Opin. Struct. Biol. **7**, 60 (1997).
[4] G. E. Schulz and R. H. Schirmer, *Principles of Protein Structure* (Springer-Verlag, New York, 1979); W. R. Taylor, J. Theor. Biol. **119**, 205 (1986).
[5] F. R. Maxfield and H. A. Scheraga, Biochemistry **18**, 697 (1979); S. A. Benner and D. Gerloff, Adv. Enzyme Regul. **31**, 121 (1990); B. Rost, Methods Enzymol. **226**, 525 (1996).
[6] C. Keasar, R. Elber, and J. Skolnick, Folding & Design **2**, 247 (1997).
[7] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996); A. V. Finkelstein and A. Ya. Badretdinov, Folding & Design **2**, 115 (1997).
[8] D. Thirumalai, J. Phys. I (France) **5**, 1457 (1995); P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997).
[9] J. B. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987); E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989); J. B. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990); E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990); A. V. Finkelstein, Curr. Opin. Struct. Biol. **4**, 422 (1994); V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).
[10] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[11] A. V. Finkelstein, A. M. Gutin, and A. Ya. Badretdinov, Proteins **23**, 151 (1995).
[12] R. F. Service, Science **277**, 179 (1997).