What is the probability of a chance prediction of a protein structure with an rmsd of 6 Å?

Boris A Reva[1]*, Alexei V Finkelstein[2] and Jeffrey Skolnick[3]

Background: The root mean square deviation (rmsd) between corresponding atoms of two protein chains is a commonly used measure of similarity between two protein structures. The smaller the rmsd is between two structures, the more similar are these two structures. In protein structure prediction, one needs the rmsd between predicted and experimental structures for which a prediction can be considered to be successful. Success is obvious only when the rmsd is as small as that for closely homologous proteins (< 3 Å). To estimate the quality of the prediction in the more general case, one has to compare the native structure not only with the predicted one but also with randomly chosen protein-like folds. One can ask: how many such structures must be considered to find a structure with a given rmsd from the native structure?

Results: We calculated the rmsd values between native structures of 142 proteins and all compact structures obtained in the threading of these protein chains over 364 non-homologous structures. The rmsd distributions have a Gaussian form, with the average rmsd approximately proportional to the radius of gyration.

Conclusions: We estimated the number of protein-like structures required to obtain a structure within an rmsd of 6 Å to be 10^4–10^5 for chains of 60–80 residues and 10^{11}–10^{12} structures for chains of 160–200 residues. The probability of obtaining a 6 Å rmsd by chance is so remote that when such structures are obtained from a prediction algorithm, it should be considered quite successful.

Addresses: [1]The Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation. [2]Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation. [3]Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

*Present address: Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

Correspondence: Boris A Reva
E-mail: breva@scripps.edu

Key words: rmsd distribution, threading, protein-like structure, protein structure prediction

Received: **17 November 1997**
Revisions requested: **19 December 1997**
Revisions received: **27 January 1998**
Accepted: **04 February 1998**

Published: **13 March 1998**
http://biomednet.com/elecref/1359027800300141

Folding & Design 13 March 1998, **3**:141–147

Introduction

Protein structure prediction is the focus of interest of many research groups. Indicative of this interest is the special CASP conference [1] for objective testing of different prediction methods. The comparison of predicted and experimentally resolved structures is usually done by calculating the root mean square deviation (rmsd) between the predicted and the experimental structures (see Equation 8 in the Materials and methods section). The rmsd value gives the average deviation between the corresponding atoms of two proteins: the smaller the rmsd, the more similar the two structures. Efficient algorithms have been developed to find the best orientation of two structures that gives the minimal possible rmsd [2,3].

A common question in protein structure prediction is what rmsd value between predicted and experimentally determined structures can be considered a successful prediction and what value indicates a failure? Success is obvious when the rmsd is small (< 3 Å; a typical rmsd for homologous proteins [4]). When the rmsd is ~6 Å, however, as frequently reported [5–7], there is serious doubt as to whether one can consider such a result a prediction at all. It is also clear [8,9] that when one uses rmsd as a measure of similarity between structures, one needs to take into account the length of the protein chain: an rmsd of 3 Å between two tripeptides indicates that their structures are different, whereas the same rmsd for two 100-residue chains indicates that their structures are similar. To obtain a more objective estimate of the significance of a given rmsd between two structures, one could compare it to the rmsd values typical of random structures of the same size and compactness.

In their early work, Cohen and Sternberg [8] compared the rmsd between 12 native protein structures and the random compact-chain structures. They showed that the rmsd of a protein from a random compact structure is proportional to R, the radius of the protein (but further considered a linear, with respect to the number of residues, approximation of the rmsd values).

Maiorov and Crippen [9] suggested that globular structures are "intrinsically similar if their rmsd is smaller than that when one of them is mirror inverted". They have shown that the minimal rmsd with the mirror-inverted structure is proportional to $N^{1/3}$ (where N is the number of residues in the chain) and that ~1% of pairs of compared equal size protein fragments have an rmsd below this minimal rmsd.

The above works do not, however, answer the general question of how to estimate the significance of a given rmsd between two protein folds of a given size (e.g. between a true protein structure and its computed model).

Here, we studied the rmsd distributions for 142 proteins using threading to generate a representative set of protein-like alternative structures. The average rmsd between pairs of randomly chosen compact folds is close to their mean radii of gyration (R_g) and the average rmsd scales directly as $N^{1/3}$ (cf. [8,9]). At the same time, the rmsd dispersion is found to be virtually independent of chain length. We investigated the shapes of the obtained rmsd distributions and show that they are rather close to Gaussian form. The normal distribution is further applied to estimate the number of randomly chosen protein-like structures that have to be generated to find one within a given rmsd from a given protein fold. The result is that the probability of finding a structure with a 6 Å rmsd from a given fold is 10^{-5} for a 70-residue chain and 10^{-12} for a 180-residue chain, and for 3 Å rmsd the probabilities are 10^{-7} and 10^{-17}, respectively.

Results and discussion

Generation of alternative structures by gapless threading

To obtain a desirable rmsd distribution for a given protein, it is necessary to have a sufficiently large number of alternative structures that preserve the characteristic features of the protein structure, such as layer organization, secondary structure, and compactness. Ideally, such a set of alternative structures should consist of true native structures of globular proteins. Unfortunately, there are not enough known structures of globular proteins to ensure the statistics necessary for the derivation of distributions. One possible way of overcoming this difficulty is to generate artificial random-walk structures, as was done in [8]. In this case, however, one loses such important protein features as the layer organization and the secondary structure. There is no algorithm available today that would generate truly random protein-like structures.

In this study, we generate a set of alternative protein-like structures using an approach resembling the gapless threading method of Hendlich *et al.* [10]. The alternative structures are obtained as continuous, equal-length backbone fragments taken from non-homologous proteins. No gaps or insertions are allowed; thus, a probe chain of N residues can be threaded onto a protein molecule of M residues ($M \geq N$) in $M - N + 1$ different ways. Not all the possible structures generated by such a procedure can be counted into the statistics of 'protein-like structures', however. First, the extracted protein fragments must be approximately as compact as the considered protein. Second, the subsequent structures obtained by threading with the shift of a few residues along the chain will be essentially the same, having a small rmsd between them. In practice, we find (Table 1) that a 10-residue shift between two subsequent threadings is sufficient to produce essentially different structures as assessed by their rmsd.

Table 1

The rmsd for protein fragments depending on their shift along the chain.*

	Protein					
Shift	1ptx N = 64	1kpt chain A N = 105	2phy N = 125	1lba N = 146	1rmi N = 160	1gky N = 186
1	3.64	3.77	3.78	3.77	3.77	3.81
2	5.74	6.12	6.07	6.01	5.52	5.86
3	7.54	7.97	7.89	7.73	6.02	7.22
4	9.16	9.78	9.61	9.33	7.36	8.87
5	10.42	11.46	11.34	10.83	9.33	10.67
6	11.06	12.73	12.83	11.97	10.60	12.06
7	11.04	13.65	13.99	12.69	11.54	13.18
8	10.50	14.44	14.88	13.07	12.92	14.33
9	9.90	14.86	15.42	13.18	14.33	15.29
10	9.55	14.77	15.09	13.18	15.31	16.02
11	9.56	14.34	14.73	13.15	16.26	16.56
12	9.92	13.82	14.54	13.09	17.36	16.97
Minimum†	7.69	9.44	11.11	10.70	12.02	12.57
Average	11.85	14.54	15.41	16.17	18.32	18.40
Maximum	15.00	17.39	18.55	20.07	21.88	23.03

*The rmsd is calculated between fragments (13, N) and (13 – S, N – S), where S is a shift and N is a sequence length.
†Minimum, average and maximum rmsd values for each protein found in the total statistics are given for comparison.

Thus, we assume that the fragments of proteins used in this study as 'protein-like structures' preserve the most important features of true proteins that will result in unbiased rmsd statistics. This assumption will be tested.

Distribution of rmsd

Distributions of rmsd values for 142 protein chains obtained by the threading procedure are presented in Figure 1. For two randomly chosen folds of equal size the rmsd values are plotted versus $N^{1/3}$ because one can expect that the distance between corresponding residues is, on average, proportional to the radius of the globule, that is to $N^{1/3}$ [8,9].

The least square approximation of the mean rmsd values for R is given by the power law:

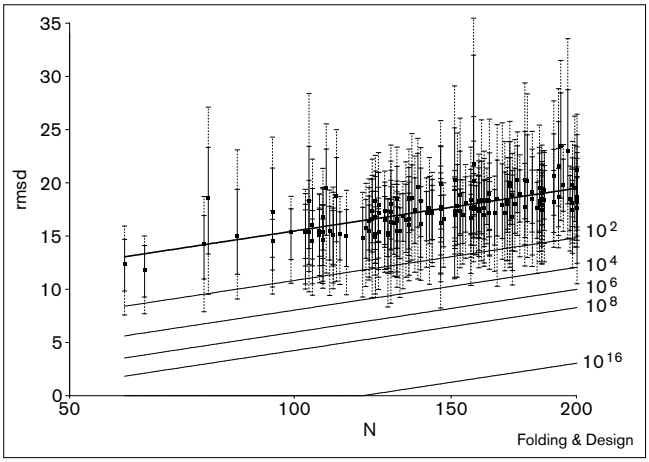
$$\langle R \rangle \cong aN^{p} \quad (1)$$

which gives $a = 3.416$ and $p = 0.3279$. The obtained value of p is indeed very close to 1/3. The best fit of the mean $\langle R \rangle$ data to the $N^{1/3}$ scaling law gives:

$$\langle R \rangle_{fit} \cong 3.333N^{1/3} \quad (2)$$

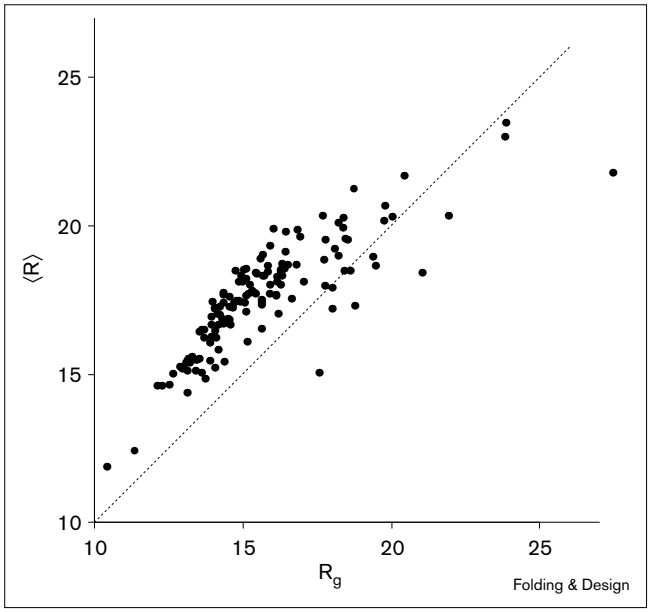
with practically the same accuracy as the best power law (Equation 1).

Figure 1



The rmsd distributions for 142 protein chains. For each chain, the rmsd distribution is obtained from comparison of its native structure with the native structure of ~2500 alternative folds taken from the protein database (see the text for more details). The mean rmsd value ⟨R⟩ for each chain is marked by a filled square. A thin vertical line passing through the square presents the central part of the distribution; the dotted continuations of the thin line embrace 2.5% of the highest and 2.5% of the lowest rmsd values. The distributions are arranged in progression of increasing N, the number of chain residues. The abscissa is scaled as $N^{1/3}$, in accordance with the expected dependence of ⟨R⟩ on N. The thick line presents the theoretical estimate $\langle R\rangle = 3.333N^{1/3}$, the proportionality coefficient 3.333 being obtained from the least square fit. The dotted curve (almost completely overlapped with the thick inclined line) presents the best power law approximation $\langle R\rangle = 3.416N^{0.3279}$, where both the coefficient 3.416 and the index 0.3279 are obtained from the least square fit. The correlation coefficient is 0.72 in both cases. The thin lines show how many protein-like structures one must attempt to have a given rmsd from a protein structure of a given size estimated according to Equations 3 and 4; the lines correspond to 10^2, 10^4, 10^6, 10^8 and 10^{16} such structures.

Figure 2



The mean rmsd values, ⟨R⟩, versus the radii of gyration, R_g, for 142 tested proteins. The correlation coefficient is 0.84 and the thin dotted line corresponds to $\langle R\rangle = R_g$.

Figure 2 demonstrates that the average rmsd values are close to R_g values for the corresponding structures: the proportionality coefficient is close to 1, and the correlation coefficient is 0.84. The standard deviations (sd) of the rmsd distributions are given in Figure 3. One can see a significant dispersion of the sd values for proteins of equal size and practically no dependence on the residue number N (see legend to Figure 3). To see if there is a difference between rmsd distributions obtained with protein fragments and with true whole protein structures (because one could suspect that compact fragments of larger proteins have a construction different from that of small proteins), we especially considered the rmsd distributions obtained by the comparison of equal size proteins. The values of $\langle R\rangle \approx 3.31N^{1/3}$ and $sd \approx 1.5 \pm 0.7$ found in this test are in good agreement with those presented in Figures 1 and 2. These results show that $\langle R\rangle \propto N^{1/3}$, whereas sd depends on the structure of a protein molecule rather than on its size. Approximately, the sd can be estimated to be 1.5 ± 0.4 Å. Taking into account both the deviations of the ⟨R⟩ values from the mean $\langle R_{fit}\rangle$ values and their dispersions sd, one can estimate that most of the rmsd values for an N-residue protein fall in the interval:

$$3.333N^{1/3} - 2.0 \le R \le 3.333N^{1/3} + 2.0 \qquad (3)$$

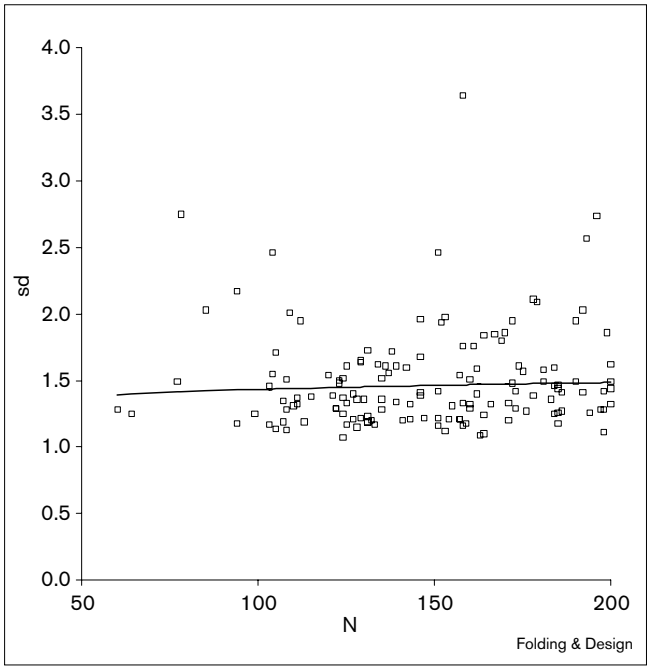
It is noteworthy that $\langle R\rangle \approx 3.333N^{1/3}$ and $sd \approx 2.0$ scale differently with protein size and this prevents a universal scaling of rmsd distributions (cf. [11]).

Typical examples of the rmsd distributions are given in Figure 4. The distribution of Figure 4a is very well approximated by the normal law according to the χ^2 analysis ($\chi^2 = 0.77$). Figure 4b is an example of one of the greatest observed deviations of the rmsd distribution from the normal law, according to the χ^2 criterion ($\chi^2 = 6.93 >> 1$). Even in this case, however, a close similarity of the observed rmsd distribution and of its normal approximation is evident.

The most interesting region of the rmsd distribution is the one in which the rmsd is small. We therefore tried to find the best approximation to the 'left side' of the distribution (i.e. the region where the rmsd is less than the average value for a given protein). We tested four plausible expected statistical laws described in the Materials and methods section (Equations 11–14) and varied their parameters to achieve the minimal χ^2 value for each of the proteins. Figure 5 shows typical examples of such best fitted approximations.

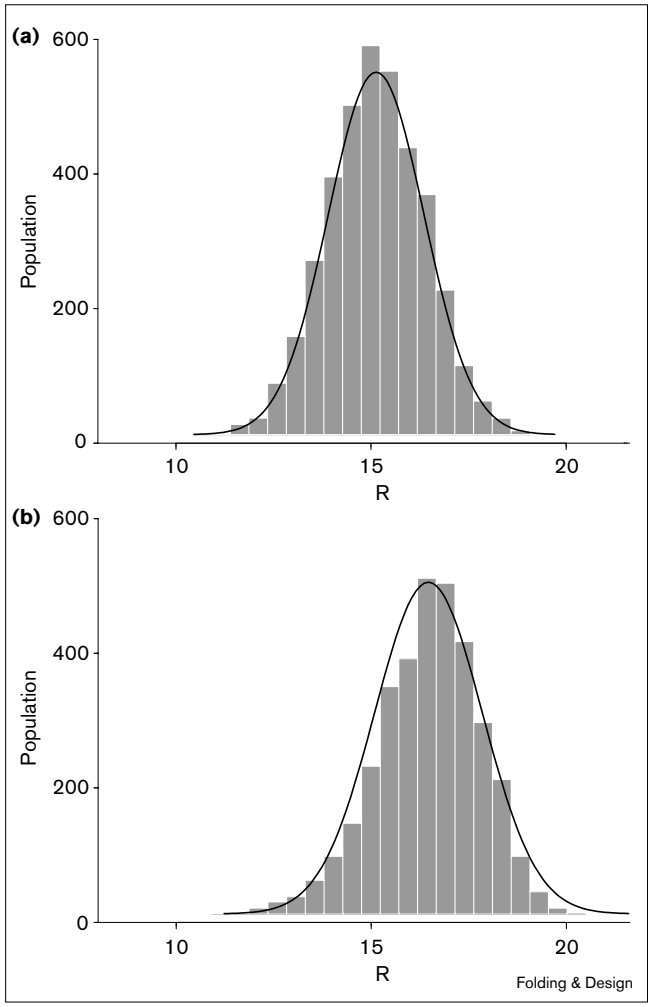
In a conventional χ^2 test, all the observed values $n^{(0)}$ (see Equation 9) are treated as independent. In this study, we

Figure 3



Standard deviation (sd) values for all the 142 rmsd distributions shown in Figure 1. The least square fit gives $sd = 1.1278N^{0.05201}$, a nearly straight line with a correlation coefficient of only 0.046.

Figure 4



Histograms of rmsd distributions approximated by the normal law. **(a)** A molecule of lectin (PDB code: 1jpc), 108 residues. 3924 rmsd values form the histogram; the average rmsd is 15.4 Å, sd = 1.28 Å, the χ^2 value is 0.77. **(b)** Ribonuclease A (PDB code: 7rsa), 124 residues. 3472 rmsd values form the histogram; the average rmsd is 16.8 Å, sd = 1.37 Å, the χ^2 value is 6.93.

cannot guarantee such an independence because there can still remain some traces of related structures among the proteins used as threading targets. The obtained χ^2 values can therefore be greater than those for absolutely independent threadings. Thus, these χ^2 estimates have a relative sense (they tell which 'expected' distribution fits better and which fits worse to the observed rmsd distribution), but not an absolute one (in this case, $\chi^2 > 1$ values do not prove that the tested theoretical distribution does not fit to the real one).

Table 2 gives averaged results of approximations of the experimental rmsd distributions observed for individual proteins by different statistical distributions.

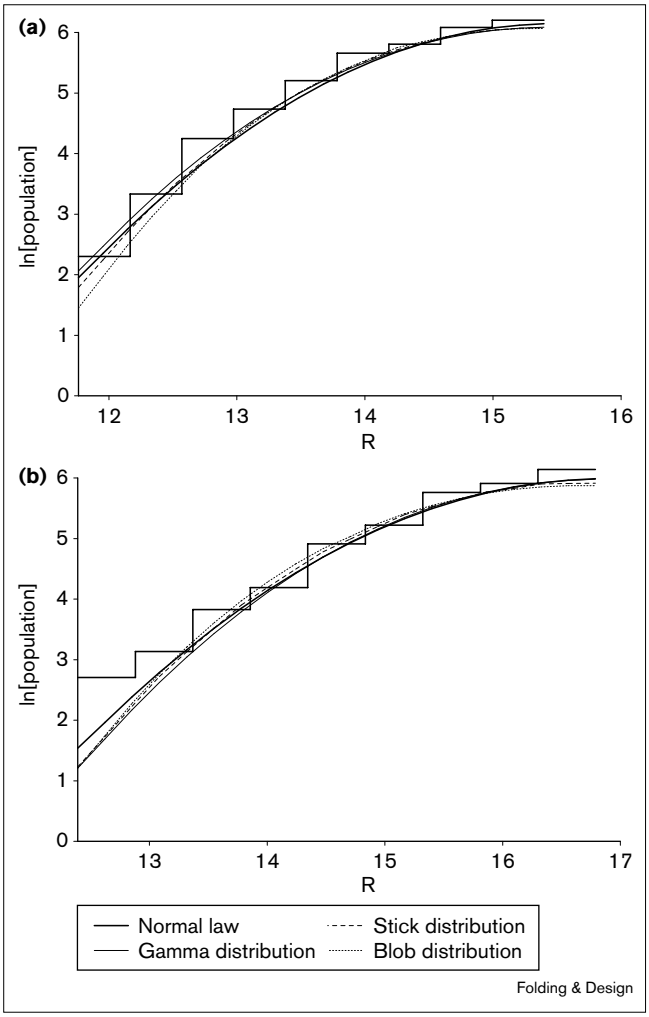
The normal distribution fits the experimental data better than the others (although not ideally: $\chi^2 = 2.78 > 1$). The 'gamma' distribution is somewhat worse, and the 'stick' and 'blob' distributions show a significantly greater deviation from the experimental data. The last result is a surprise for us because the normal distribution does not converge to zero when the rmsd converges to zero, whereas the 'gamma', 'stick' and 'blob' distributions were used because they converge to zero when the rmsd does.

Although the normal law is not accurate at very small rmsd values, it gives a reasonable approximation of the observed low rmsd region (Figure 5), and one can use it to extrapolate the observed rmsd distributions to reach the 'moderately small' rmsd values in the range of 3–4 Å. We therefore used the normal law to estimate the number of protein-like structures required to find one structure within a given rmsd threshold. This number can be estimated easily as [9]:

$$N_R = \frac{\left((sd)\sqrt{2\pi}\right)}{\int_{-\infty}^{R} \exp\left(-\left(x - \langle R \rangle\right)^2 / 2(sd)^2\right) dx} \qquad (4)$$

If Equation 3 is applied to individual proteins, the numbers N_R (the number of protein-like structures) are rather

Figure 5



Histograms (shown by stepping line) of rmsd distributions in regions of small rmsd values for **(a)** lectin and **(b)** ribonuclease A and their best approximations by the normal law; the gamma distribution; the 'stick' distribution; and the 'blob' distribution. The plots are presented on a logarithmic scale to emphasize the sparsely populated region of low rmsd values. The χ^2 values for the corresponding distributions are 0.75, 1.09, 1.51 and 2.23 for lectin and 3.75, 5.22, 7.55 and 10.07 for ribonuclease A.

different, mainly because of a great dispersion in the sd values (Figure 3); N_R is considerably smaller (by a few orders of magnitude) for the proteins with a large sd (especially if they have a small $\langle R \rangle$) than for other proteins.

To get an average estimate of N_R, we used the averaged estimates of $\langle R \rangle$ and sd given in Equation 3: $\langle R \rangle = 3.333N^{1/3}$ and sd = 2.0 Å. The lines corresponding to N_R, equal to 10^2, 10^4, 10^6, 10^8 and 10^{16}, are presented in Figure 1. The line $N_R = 10^2$ is rather close to the 1% similarity level line from [9]. Figure 1 shows that the numbers N_R increase very quickly with increasing the chain length and with decreasing the rmsd threshold. We see in Figure 1 that almost

Table 2

Average χ^2 values for four distribution laws used to approximate observed rmsd statistics for 142 proteins.*

Normal (Equation 11)	Gamma (Equation 12)	Stick (Equation 13)	Blob (Equation 14)
2.78 (2.38)	3.52 (3.57)	4.69 (5.15)	6.32 (7.22)

*Standard deviations are given in parentheses. These values are large because χ^2 values for some proteins are ~10 for normal distribution and as large as 54 for blob distribution.

every protein has rmsd points in the range $N_R = 10^2$–10^4 (as must be so because we have ~2500 points per protein and $10^2 < 2500 < 10^4$); only eight out of 142 proteins have rmsd points at $N_R > 10^4$; and no protein has rmsd points in the range where $N_R > 10^6$.

Although Figure 1 gives a rather approximate average estimate (the numbers for different proteins can differ significantly), one can certainly conclude that protein structure prediction for these chains of 60 or more residues with an rmsd of 5–6 Å is practically impossible by chance. Hence, such a prediction should be considered a successful one. Figure 1 also shows that some equal size proteins (with small $\langle R \rangle$ and large sd) are easier to 'predict by chance' with a lower rmsd than the others. A further analysis is needed to show which features of protein structures are responsible for such relatively low rmsd values.

Conclusions

In this work, we examined the distribution of rmsd values between the native and the protein-like structures of equal compactness that were produced by a threading approach. We found that the observed medians of rmsd distributions satisfy a simple relationship, rmsd ~ $N^{1/3}$, whereas standard deviations do not depend on the chain lengths. The normal distribution gives a reasonable approximation to the observed rmsd distribution. Using the normal distribution, we estimated a probability of protein structure prediction within a given accuracy by chance (see threshold lines in Figure 1) and showed, in particular, that this probability is negligible for an rmsd of 6 Å. The thresholds shown in Figure 1 can also help to estimate *a priori* what protein-structure prediction accuracy one can expect with a given type of (quasi) energetic parameters used in a prediction. To this end, one has to find the Z score for the native-protein fold with these parameters (e.g. using a gapless threading), calculate the number of folds corresponding to this Z score as:

$$N_Z = \frac{\sqrt{2\pi}}{\int_{-\infty}^{Z} \exp(-z^2/2)\,dz} \sim \exp(-Z^2/2) \qquad (5)$$

and find the rmsd point that corresponds to this N_Z (and the given chain length) in Figure 1.

Materials and methods

Preparation of the database

Protein structures used in threading were taken from the 25% similarity list [12]. Any pair of proteins in this list has a similarity of < 25% according to the Smith and Watermann [13] gap-allowing sequence alignment (open gap penalty 3.0, gap elongation penalty 0.05). From the Hobohm *et al.* [12] list of October 1997, our database consists of 377 proteins having no chain breaks, with a resolution better than 2.5 Å and an R factor < 0.2.

In this database, there are 155 proteins of 200 residues or less. Each of them was threaded onto the greater protein structures of the database. When we used a 10-residue shift between two subsequent threadings (see above) and selected only the compact structures (see below), the threading gave from 1500–5000 alternative folds per protein tested.

To select an unbiased set of protein structures, we determined all the cases of low rmsd between structural pairs. Each of these pairs was analyzed with SCOP [14] to determine if the proteins of the pair belonged to the same protein superfamily. In 22 out of 27 of the low rmsd cases, we found that both proteins belonged to the same superfamilies. (In one pair, both proteins belonged to a bacterial pathogens superfamily; for the other 21 pairs, the proteins belonged to the globin-like superfamily.) For three out of the five other low rmsd pairs, the proteins belonged to different superfamilies, and for the final two pairs, SCOP did not classify the proteins. When this analysis was applied to the protein pairs with an rmsd of 8–9 Å, one more pair of homologous proteins (superfamily EF hand) was found. Two more pairs (superfamilies ConA-like and Lipocalins) were found within an rmsd range of 9–10 Å. Thus, 13 protein chains belonging to the abovementioned superfamilies (1tiiD, 1eca, 3sdhA, 1babB, 2fal, 1ash, 2hbg, 1mbd, 2gdm, 1cpcA, 1sltB, 1mup and 2scpA; a capital letter after the PDB code identifies the chain in the protein molecule), were chosen as the shortest among the homologous pairs and were removed from the database of 377 proteins. The resulting database includes 364 proteins, 142 of which are 60–200 residues in length.

Selection of the compact structures

In threading, to maintain the same level of compactness of the alternative structures as that of the original structure, we chose only those structures whose R_g did not exceed 1.2 times the native value. R_g is calculated as:

$$R_g = \left(\frac{1}{N} \sum_{i-1}^{N} (r_i - r_c)^2 \right)^{\frac{1}{2}} \tag{6}$$

where $\{r_i\}$, $i = 1, ..., N$ is a set of coordinate vectors for the Cα atoms of the mainchain of the molecule N residues in length, and:

$$r_c = \frac{1}{N} \sum_{i-1}^{N} r_i \tag{7}$$

are three-dimensional coordinates of the center of mass of the molecule.

Comparison of the native and alternative structures

We consider two folds of protein chains with the same number of residues, N. The rmsd value between the structures is defined as

$$\text{rmsd} = \left(\frac{1}{N} \sum_{i-1}^{N} (r_i - r_i')^2 \right)^{\frac{1}{2}} \tag{8}$$

where we trace the structures through the corresponding sets $\{r_i\}$ and $\{r_i'\}$ of the three-dimensional coordinates for the Cα atoms of the two molecules. The value of rmsd as defined by Equation 8 depends on the mutual position and orientation of two proteins. We used our FITT program [15], based on the algorithm of Lesk [3], which guarantees that the minimal possible rmsd is very quickly and precisely found.

Estimating the quality of approximation of rmsd distribution

The χ^2 value is the usual measure of deviation between an observed statistic and the expected one [16]. To compute this quantity, we divide an rmsd distribution into bins and calculate the observed and expected bin populations. The χ^2 value for protein p is computed as:

$$\chi_p^2 = \frac{1}{M} \sum_{k=1}^{K} \frac{(n_{p,k}^{(o)} - n_{p,k}^{(e)})^2}{n_{p,k}^{(e)}} \tag{9}$$

where $n_{p,k}^{(o)}$ and $n_{p,k}^{(e)}$ are the observed and expected populations, respectively, of rmsd values in bin k for protein p; K is the number of bins taken into account and M is the number of degrees of freedom. In this study, $M = K - 3$ because the total population, mean rmsd value and dispersion of the expected statistics are adjusted to the observed values.

The averaged (over all the 142 examined proteins) χ^2 value is:

$$\langle \chi^2 \rangle = \frac{1}{142} \sum_{p=1}^{142} \chi_p^2 \tag{10}$$

When $\chi^2 \leq 1$, the experimental data can be treated as confirming the expected statistics [16]. When $\chi^2 > 1$, the experimental statistics deviate from the expected one, and this deviation grows with χ^2.

Because the small rmsd region is of major interest, we took the interval $(0, \langle R_p \rangle)$, where $\langle R_p \rangle$ is the average rmsd between protein p and the alternative folds and divided it into 10 bins: $(0, r_p)$, $(r_p, r_p + \delta)$, $(r + \delta, r_p + 2\delta)$, ..., $(r_p + 8\delta, \langle R_p \rangle)$. Here r_p is the tenth lowest rmsd value and $\delta = (\langle R_p \rangle - r_p)/9$ is the bin width. Taking the entire rmsd region into account (not just the left-hand side), the δ-wide bins $(\langle R_p \rangle, \langle R_p \rangle + \delta)$, etc., are extended to the right-hand side of the rmsd distribution until the first δ-wide bin containing < 10 observed rmsd values is found; this bin is then extended to infinity to comprise all remaining rmsd values.

Statistical laws for fitting of rmsd distribution

We explore four probable statistical distributions to choose the one that best fits the observed data. First, we consider the normal or Gaussian distribution:

$$P_n(R) = \frac{1}{(sd)\sqrt{2\pi}} \exp\left(-\frac{(R - \langle R \rangle)^2}{2(sd)^2} \right) \tag{11}$$

where $\langle R \rangle$ and sd are the mean and the standard deviation, respectively, for the distribution. The rmsd distribution must converge to zero when R turns to zero, however, and the normal distribution does not. We therefore tried a few other plausible distributions that converge to zero with R.

The first distribution:

$$P_g(R) \sim R^{(m-1)} \exp(-BR^2) \tag{12}$$

where $m \geq 1$, results from a product of m independent normal distributions centered in $r = 0$ over the coordinates r_1, ..., r_m; $P_g(R)$ and describes the probability of having a given $R = (r_1^2 + ... + r_m^2)^{1/2}$ value; it is related to the gamma function [17].

The distributions:

$$P_s(R) \sim \exp\left(-\frac{A}{R} - BR^2 \right) \tag{13}$$

$$P_b(R) \sim \exp\left(-\frac{A}{R^2} - BR^2 \right) \tag{14}$$

come from consideration of a polymer, each link of which is within a distance R from the corresponding link of another randomly folded chain. The deviations are penalized by the term $\exp(-BR^2)$. The entropy difference between the confined polymer and its random coil is $S_{conf} - S_c = \alpha / n_R$, where n_R is the number of residues per 'blob' (chain region with a characteristic radius R) and α is a conformationally

independent coefficient proportional to the number of residues in the chain [18,19]. Usually, the blobs are treated as Gaussian coils, giving $n_R \sim R^2$. Because the probability is the exponential of the entropy, Equation 13 describes a probability of rmsd ~ R for a Gaussian 'blob' model. When R is comparable to the distance between adjacent chain residues, however, the Gaussian model of a 'blob' must be very rough and a 'stick' model where $n_R \sim R$ must be better; Equation 13 describes the probability of rmsd ~ R for this 'stick' model.

Acknowledgements

This work was supported by NIH Grant GM48835 (to J.S.) and by NIH Fogarty Research Collaboration Grant No. TW00546 (to J.S. and A.V.F.). A.V.F. acknowledges support by an International Research Scholar's Award No. 75195-544702 from the Howard Hughes Medical Institute. The authors are grateful to A.R. Ortiz for help in analysis of the structural database.

References

1. Dunbrack, R.L., Jr., *et al.*, & Cohen, F.E. (1997). Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2). *Fold. Des.* **2**, R27-R42.
2. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **34**, 827-828.
3. Lesk, A.M. (1986). A toolkit for computational molecular biology. II. On the optimal superposition of two sets of coordinates. *Acta Crystallogr. A* **42**, 110-113.
4. Chothia C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.
5. Dandekar, T. & Argos, P. (1994). Folding the mainchain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844-861.
6. Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* **266**, 645-660.
7. Skolnick, J., Kolinski, A. & Ortiz, A. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.
8. Cohen, F. & Sternberg, M.J.E. (1980). On the prediction of protein structure: the significance of the root-mean-square deviation *J. Mol. Biol.* **138**, 321-333.
9. Maiorov, V.N. & Crippen, G.M. (1994) Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* **235**, 625-634.
10. Hendlich, M., *et al.*, & Sippl, M. (1990). Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* **216**, 167-180.
11. Maiorov, V.N. & Crippen, G.M. (1995). Size-independent comparison of protein three-dimensional structures. *Proteins* **22**, 273-283.
12. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* **1**, 409-417.
13. Smith, T. & Watermann, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
14. Murzin, A., Brenner, S., Hubbrad, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
15. Finkelstein, A.V. (1987). Program FITT for rmsd calculation. Institute of Protein Research, Russian Academy of Sciences.
16. Mathews, J. & Walker, B.L. (1964). *Mathematical Methods of Physics.* W.A. Benjamin, Inc. New York.
17. Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1994). Numeric Recipes in FORTRAN. Cambridge University Press.
18. Shakhnovich, E.I. & Gutin, A.M. (1989). Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. *Biophys. Chem.* **34**, 187-199.
19. Wall, F., Seitz, A.W., Chin, J.C. & de Gennes, P.G. (1978). Statistics of self-avoiding walks confined to strips and capillaries. *Proc. Natl Acad. Sci. USA* **75**, 2069-2070.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.