Averaging Interaction Energies Over Homologs Improves Protein Fold Recognition in Gapless Threading

Boris A. Reva,[1]* Jeffrey Skolnick,[1] and Alexei V. Finkelstein[2]
[1]*Department of Molecular Biology, The Scripps Research Institute, La Jolla, California*
[2]*Institute of Protein Research, Russian Academy of Sciences, Moscow Region, Russian Federation*

ABSTRACT **Protein structure prediction is limited by the inaccuracy of the simplified energy functions necessary for efficient sorting over many conformations. It was recently suggested (Finkelstein, Phys Rev Lett 1998;80:4823–4825) that these errors can be reduced by energy averaging over a set of homologous sequences. This conclusion is confirmed in this study by testing protein structure recognition in gapless threading. The accuracy of recognition was estimated by the Z-score values obtained in gapless threading tests. For threading, we used 20 target proteins, each having from 20 to 70 homologs taken from the HSSP sequence base. The energy of the native structures was compared with the energy from 34 to 75 thousand of alternative structures generated by threading. The energy calculations were done with our recently developed C_α atom-based phenomenological potentials. We show that averaging of protein energies over homologs reduces the Z-score from ~ -6.1 (average Z-score for individual chains) to ~ -8.1. This means that a correct fold can be found among $3 * 10^9$ random folds in the first case and among $3 * 10^{15}$ in the second. Such increase in selectivity is important for recognition of protein folds. Proteins 1999;35:353–359.** **© 1999 Wiley-Liss, Inc.**

Key words: **homologous sequences; interaction energies; energy averaging; threading; protein structure recognition**

INTRODUCTION

The prediction of protein structure is limited on one side by the enormous amount of different conformations accessible to a chain molecule. On the other side, the predictions are limited by the inaccuracy of energy calculations. The need for the fast sorting and rejection of unfavorable conformations resulted in the development of simplified energy functions where each amino acid residue is usually approximated by a single interaction center. Such energy functions are always approximate. They are usually accurate enough to distinguish the native structure in simple tests, but they are not sufficient for protein folding and the selection of the native fold as the minimal energy structure among a multitude of potential folds. In this work, we study the possibility of improving the accuracy of the native fold selection by using remote homologs of protein chains, i.e., protein molecules with similar folds and diverse sequences.

It has been known empirically that one can improve structure predictions by using many homologous protein (or RNA) sequences,[1–10] but no general algorithm on how to use them has been suggested so far. Recently, one of us suggested a simple analytical theory[11] applicable to simple test cases that showed that the energy errors can be significantly reduced by using a set of proteins with the same fold but different sequences. This theory is applied and tested here for recognition of protein structure in gapless threading.

MATERIALS AND METHODS

Accuracy of protein structure recognition is characterized commonly by the value of the Z-score[12]:

$$Z = \frac{E_{Nf} - \langle E \rangle}{D}, \tag{1}$$

where E_{Nf} is the energy of the native fold,

$$\langle E \rangle = \frac{1}{M} \sum_{i=1}^{M} E_i$$

is the average energy,

$$D = \left[\frac{1}{M} \sum_{i=1}^{M} (E_i - \langle E \rangle)^2 \right]^{1/2}$$

is the standard deviation of energies; E_i is the energy of the i-th sequence of M alternative folds ($i = 1, \ldots, M$). [We should note that the term "energy" is used in this paper for simplicity only. Strictly speaking, we mean the free energy of the sequence in a fold, because the entropically driven

Grant sponsor: National Institutes of Health; Grant number: GM48835; Grant sponsor: National Institutes of Health (Fogarty Research Collaboration Grant); Grant number: TW00546; Grant sponsor: Russian Foundation for Basic Research; Grant sponsor: Howard Hughes Medical Institute (International Research Scholar's Award); Grant number: 75195–544702.

Boris A. Reva is on leave from the Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation.

*Correspondence to: Boris A. Reva, Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037.

Received 21 July 1998; Accepted 11 January 1999

solvent-mediated (hydrophobic and electrostatic) forces also contribute to the stability of the fold.]

The Z-score gives the average number of standard deviations between the native and the random fold energy. It allows one to estimate the expected number of folds, N_Z, among which the native structure can still be selected as the one with the lowest energy. Assuming a normal distribution for energies[13] of competing folds,

$$N_Z = \frac{\sqrt{2\pi}}{\int_{-\infty}^{Z} \exp\left(-\frac{x^2}{2}\right) dx}. \tag{2}$$

When $Z \ll -1$ which corresponds to a reasonable accuracy of predicting methods:

$$N_Z \simeq \sqrt{2\pi} Z \exp\left(\frac{Z^2}{2}\right). \tag{3}$$

The larger N_Z (and hence $-$Z), the more accurate the protein structure recognition. Because of errors in the energy estimates, the energy spectrum of the misfolded structures swells. As a result, the value of $-$Z drops and the calculated native energy fold can be hidden among the calculated energies of misfolded structures.[14]

Homolog-Averaged Energies and Z-Scores

The homolog-averaged fold energy has the general form:[11]

$$U_i = \sum_{\lambda=1}^{\Gamma} A_\lambda E_i^{(\lambda)}, \tag{4}$$

where Γ is the number of homologs, $E_i^{(\lambda)}$ is the computed energy of chain λ ($\lambda = 1, 2 \ldots, \Gamma$) in fold i ($i = 1, 2, \ldots M$), and ($\lambda = 1, 2 \ldots, \Gamma$) are weights, the optimal values of which will be considered below.

The same averaging of the fold energies should improve the Z- scores. The Z-score for the homolog-averaged energies U_i is:

$$Z_h = \frac{\sum_{\lambda=1}^{\Gamma} A_\lambda (E_{Nf}^{(\lambda)} - \langle E^{(\lambda)} \rangle)}{\sqrt{\sum_{\lambda=1}^{\Gamma} \sum_{\mu=1}^{\Gamma} A_\lambda A_\mu \langle (E^{(\lambda)} - \langle E^{(\lambda)} \rangle)(E^{(\mu)} - \langle E^{(\mu)} \rangle) \rangle}}$$

$$= \frac{\sum_{\lambda=1}^{\Gamma} A_\lambda D_\lambda Z_\lambda}{\sqrt{\sum_{\lambda=1}^{\Gamma} \sum_{\mu=1}^{\Gamma} A_\lambda D_\lambda C_{\lambda\mu} A_\mu D_\mu}}. \tag{5}$$

Here, $\langle\ \rangle$, as above, means averaging over all folds, $\langle E^{(\lambda)} \rangle$ is the average energy for homolog λ,

$$D_\lambda = \left[\frac{1}{M} \sum_{i=1}^{M} (E_i^{(\lambda)} - \langle E^{(\lambda)} \rangle)^2\right]^{1/2} \tag{6}$$

is the standard deviation for energies of the corresponding homolog, and $C_{\lambda\mu}$ is the energy correlation for homologs λ and μ:

$$C_{\lambda\mu} = \frac{\frac{1}{M} \sum_{i=1}^{M} (E_i^{(\lambda)} - \langle E^{(\lambda)} \rangle)(E_i^{(\mu)} - \langle E^{(\mu)} \rangle)}{D_\lambda D_\mu}. \tag{7}$$

Note that $C_{\lambda\mu} < 1$ if $\lambda \neq \mu$, while all $C_{\lambda\lambda} = 1$.

The Z-score achieves its extreme value[11] when

$$\frac{\delta Z_h}{\delta A_\lambda} = 0$$

for all $\lambda = 1, \ldots, \Gamma$; this gives

$$A_\lambda^* = C D_\lambda^{-1} \sum_{\mu=1}^{\Gamma} C_{\lambda\mu}^{-1} Z_\mu, \tag{8}$$

where the constant C does not influence the Z_h^* value (cf. Eq.(5)). Then the minimum of the score is:

$$Z_h^* = -\sqrt{\sum_{\lambda=1}^{\Gamma} \sum_{\mu=1}^{\Gamma} Z_\lambda C_{\lambda\mu}^{-1} Z_\mu}, \tag{9}$$

where $C_{\lambda\mu}^{-1}$ is the $\lambda\mu$-th element of the matrices inverse to the correlation matrices defined by Eq. (7).

Unfortunately, one cannot use Equations (8) and (9) in practical protein structure prediction because one does not know the Z-values until the native fold has been found. However, these equations can help estimate the lower limit of Z_h that can be reached with the given potentials and given set of homologs. (The estimate of Eq. (9) will be used to this end below.)

It is easy to see that in the extreme case when pairwise correlations in energies between homologs are close to zero (i.e., when $C_{\lambda\mu} \approx 0$, if $\lambda \neq \mu$), Eq. (8) gives $A_\lambda^* \sim Z_\lambda / D_\lambda$. A similar result, $A_\lambda \sim 1/D_\lambda$, is obtained when all the Z_λ are equal and all the $C_{\lambda\mu}$ (when $\lambda \neq \mu$) are the same. Hence, when Z_λ values are unknown (and one can only suppose that they are more or less the same) it seems reasonable to use a simple approximation:

$$A_\lambda = \frac{1}{D_\lambda}. \tag{10}$$

This gives,

$$Z_h = \frac{\langle Z \rangle_h}{\sqrt{\frac{1}{\Gamma} + \langle C \rangle_h \left(1 - \frac{1}{\Gamma}\right)}}, \tag{11}$$

where

$$\langle Z \rangle_h = \frac{1}{\Gamma} \sum_{\lambda=1}^{\Gamma} Z_\lambda \tag{12}$$

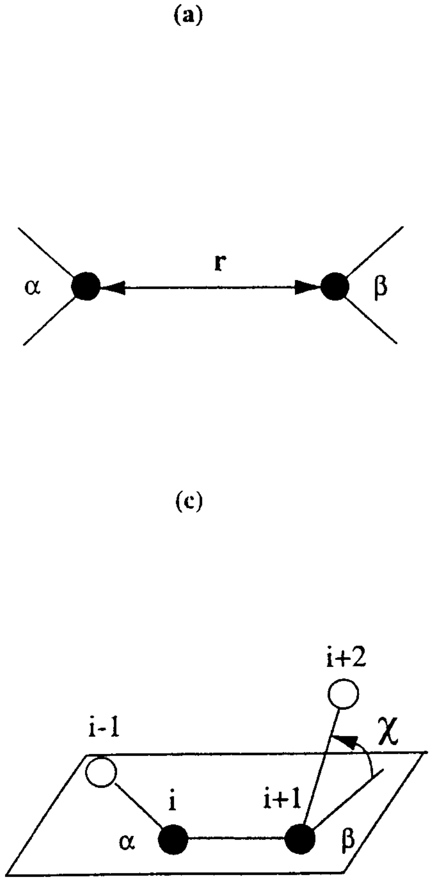

Fig. 1. Scheme of interactions taken into account; residues to which potential are applied are shown by filled circles. **a:** Interactions depending on the distance **r** between residues α and β ("long-range" potentials[15–16]). **b:** Chain bending potential:[15–16] bending at the intervening residue α affects the distance **r** between terminal residues i−1 and i+1. **c:** Chiral potential[16] depending on the dihedral angle χ between two planes of C_α atoms, (i−1,i,i+1) and (i,i+1,i+2), and the residues α and β. Long-range and bending potentials are derived at resolution of 1 Å; angular resolution for chiral potential is 40 degrees. This crude resolution of energy functions makes insignificant the statistical errors connected with inclusion (or deletion) of each individual protein to the database used in derivation of potentials.

is the Z-score averaged over homologs; and

$$\langle C\rangle_h = \frac{1}{\Gamma(\Gamma - 1)} \sum_{\lambda=1}^{\Gamma} \sum_{\substack{\mu=1,\\ \mu\neq\lambda}}^{\Gamma} C_{\lambda\mu} \qquad (13)$$

is the averaged pairwise energy correlation for different chains[11].

One can see from Eqs. (12) and (13) that $Z_h \sim \langle Z\rangle_h\sqrt{\Gamma}$ when the correlations in energies between homologs are close to zero. Thus, the accuracy of predictions can be drastically improved by using a few homologs with low sequence similarity. A practical problem is that the current database must contain enough diverse homologs.

In what follows, all energy calculations were done using our recently developed potentials;[15–16] they are briefly described in Figure1.

According to Eq.(4), the averaging of pairwise interaction energies is done as follows:

$$U_{ij}(r) = \sum_{\lambda=1}^{\Gamma} A_\lambda \epsilon(\alpha_i^{(\lambda)}, \alpha_j^{(\lambda)}, r). \qquad (14)$$

Note that the homolog-averaged energy $U_{ij}(r)$ gives the average interaction between positions i and j of the alignment rather than the interaction between "average residues" occupying these positions.

In Eq.(14), the sum is taken over those homologs belonging to the sequence family being considered; $A_\lambda$is the weight coefficient for the homolog λ; $\epsilon(\alpha, \beta, r)$; is the interaction energy for residues α and β at a distance **r**; $\alpha_i^{(\lambda)}$ is the residue of sequence λ occupying the i-th position of the alignment. When $\alpha_i^{(\lambda)}$ and/or $\alpha_j^{(\lambda)}$ in Eq.(14) correspond to a gap in the alignment of the sequence λ then $\epsilon = 0$.

The same averaging over homologs is applied also to the bending and chirality terms shown in Figure1.

When a separate sequence is threaded onto a target fold, we take the gaps in the sequence to be fixed and the same as in the HSSP alignment, and the energy values corresponding to the gaps are taken to be zero, as above.

Gapless Threading of Protein Families

To see how the averaging of energies over homologs can facilitate protein structure predictions (i.e., how it reduces the Z-scores), we use the gapless threading test.[17] In this test, the energy of the native structure is compared with those of alternative structures obtained by threading the query sequence onto all possible 3D structures provided by the backbones of a set of host proteins. No internal gaps or insertions are allowed; thus, a chain of $\boldsymbol{N}$ residues length can be threaded through a host protein molecule $\boldsymbol{M}$ residues length in $\boldsymbol{M} - \boldsymbol{N} + \mathbf{1}$ different ways. Only a tiny fraction of the possible conformations of the query sequence can be checked in this test; however, it is enough for a crude estimation of the Z-score values.

In threading tests, we used 20 different protein families. Each includes 20 to 70 homologs for the HSSP sequence base.[18] In HSSP, all the sequences are aligned to the first (the "root") sequence of the family; the 3D structure of the root sequence is known. The root sequences are 64 to 200 residues in length for the 20 considered protein families. For each family, we used the native 3D structure of the root sequence as the target fold for all sequences of this family. We did not try to improve the alignments given by HSSP.

To choose remote homologs for energy averaging and to estimate the corresponding weight coefficients A_λ, we have to know the correlation coefficients $C_{\lambda\mu}$ and the standard deviations D_λ. For each family, we computed them in the preliminary threading tests using only 300 alternative structures. These 300 structures were taken from backbone fragments of our database that are of appropriate length (see below). None belongs to the tested proteins. The backbone fragments from the same protein were allowed with a shift of 10 residues or more along a chain to avoid structural similarity between subsequent fragments of the same structure.[19]

All the protein structures used in threading were taken from the PDB[19] according to Sander's 25% similarity list[21] of October 1997. Actually, we selected only 364 structures from this list (those with no chain breaks, with a resolution better than 2.5 Å and R factor less than 0.2 and with no structural homologs[19]). They formed our final list. The native structures of the 20 root sequences used in our threading tests were also taken from this final list.

Each of the root sequences (and all its aligned homologs) has been threaded onto the folds of all the larger proteins. On average, this set of 364 folds provided us with ~50,000 alternative conformations for a given query sequence (from 74,676 for the shortest sequence of 64 residues to 33,792 alternatives for the longest query sequence of 200 residues).

RESULTS AND DISCUSSION

We investigated the Z-score values for the 20 protein families in the extensive threading test, using all the alternative structures provided for these families by the 364-protein database. The aim was to compare the Z-scores for individual sequences with those calculated with homolog-averaged energies.

To maximize the effect of the energy averaging (see Eq. (14) and the subsequent paragraphs), we selected homologs having no significant gaps in the alignment. We did not consider homologs where gaps comprise more than 20% of the aligned residues (as compared to the root sequence of the family). The optimal threshold for the gap content, 20%, has been determined by preliminary experiments (detailed results are not shown here).

Since sequences that are too similar are useless for the improvement of Z-scores (see Eq. (11)), we select only remote homologs as follows. We take the family sequences one by one (in their order in the HSSP). For each new sequence λ, we calculated the correlation $C_{\lambda\mu}$ (Eq.(7)) of its fold energies with those of the already selected sequences μ (by using a small set of 300 alternative folds). We excluded any new sequence when its energy correlation $C_{\lambda\mu}$ with any of already selected sequence exceeds a chosen threshold. After some preliminary experiments (see Table II below), we saw that the best correlation threshold is equal to 0.7 (i.e., we selected only the sequences with $C_{\lambda\mu} \leq 0.7$).

The results of the extensive threading tests with the homologs selected in this way are given in Table I.

The main result is that all the resulting Z_h-scores obtained with the homolog-averaged interactions are significantly lower than the $\langle Z \rangle_h$ values, i.e., than the mean Z-scores computed for the same, but separately taken, homologs. This result directly confirms the theoretical conclusion summarized by Eq.(11).

Averaged over 20 tested protein families, the Z-scores decrease from -6.14 (the average Z-score for threading of individual chains) to -8.08 (for threading with the homolog-averaged interactions). According to Eq. (3), this means that a correct fold can be found among $\sim 3 * 10^9$ random folds in the first case (i.e., when a threading is done for one randomly chosen homolog) and among $3 * 10^{15}$ in the second, when the fold energies are averaged over several aligned homologs.

This shows that protein fold recognition should be significantly more successful when performed for a protein family with the homolog-averaged energies, rather than for one separate member of this family.

The Z_h-score obtained for a family not only surpasses the Z- score of a randomly chosen homolog: it also surpasses as a rule (exceptions: 1thx, 2tgi, 1rie, 2cpl) the Z-score of the "root" sequence of the family, Z_0. The latter is impressive because the Z_0 is computed with the exact native 3D structure of the root sequence, while for other homologs, this 3D structure is not their exact native one.

Although $Z_h < Z_0$ as a rule, a great and apparently random scattering of Z-scores over homologs is observed. In 13 out of 20 cases, the minimal Z-score was obtained for a homolog different from the root sequence.

Nevertheless, one can see that the homolog-averaged Z_h scores are lower than the minimal Z-scores, Z_{min}, found within the families of homologs; this is observed for most of the protein families (exceptions: 1thx, 2tgi, 1rie, 2cpl, 1xnb). However, the average difference is only ~ -0.5.

This small difference suggests another (but for averaging of energies) way how to use homologs to increase the sensitivity of protein fold recognition; namely, one can

TABLE I. The Z-Score Values Obtained in Threading Tests With Averaging Energies Over Homologs†

Protein family	L	Number of folds	Γ	S	$\langle C\rangle_h$	$\langle Z\rangle_h$	Z_h	Z_0	Z_{min}	Z^*_h
1ptx	64	74676	7	0.55	0.55	**−5.96**	**−7.62**	−7.26	−7.26	−8.08
1fkj	107	59577	14	0.52	0.54	**−5.99**	**−7.92**	−5.53	−7.49	−8.88
1thx	108	59237	11	0.45	0.51	**−6.20**	**−8.30**	−8.42	−8.42	−9.77
1ccr	111	58230	9	0.68	0.59	**−3.97**	**−4.98**	−3.49	−4.84	−5.30
2tgi	112	57897	8	0.40	0.56	**−5.62**	**−7.10**	−7.25	−7.25	−8.36
1whi	122	54611	5	0.42	0.40	**−6.78**	**−9.40**	−7.76	−7.76	−9.24
1bp2	123	54285	12	0.69	0.51	**−5.47**	**−7.35**	−5.83	−7.00	−7.85
4fgf	124	53962	6	0.45	0.49	**−5.83**	**−7.72**	−6.09	−7.08	−8.02
7rsa	124	53962	9	0.42	0.41	**−5.96**	**−8.66**	−6.81	−7.18	−9.12
1rie	127	53010	4	0.42	0.52	**−4.02**	**−4.96**	−6.04	−6.04	−7.05
3chy	128	52696	15	0.39	0.55	**−6.83**	**−8.99**	−6.68	−7.82	−9.21
1vsd	146	47359	11	0.39	0.54	**−5.28**	**−6.93**	−6.19	−6.19	−7.60
1jcv	153	45377	10	0.48	0.53	**−6.53**	**−8.57**	−7.04	−8.09	−9.02
1gpr	158	44000	12	0.51	0.52	**−6.91**	**−9.27**	−7.54	−7.76	−9.81
2cpl	164	42411	14	0.47	0.52	**−5.87**	**−7.81**	−7.99	−8.34	−9.71
5p21	166	41895	7	0.83	0.57	**−6.99**	**−8.81**	−7.57	−8.16	−9.12
1amm	174	39865	11	0.52	0.45	**−7.29**	**−10.36**	−9.46	−10.27	−11.76
1xnb	185	37199	7	0.47	0.58	**−6.62**	**−8.20**	−7.67	−8.75	−9.20
1gen	200	33792	8	0.45	0.58	**−8.07**	**−10.12**	−8.31	−9.25	−10.40
1iae	200	33792	12	0.45	0.56	**−6.58**	**−8.53**	−8.15	−8.15	−9.78
Average	140	49892	10	0.50	0.52	**−6.14**	**−8.08**	−7.05	−7.65	−8.86

†L is the length of root sequence; Γ is the number of selected homologs; S is the mean HSSP similarity of the root sequence to the rest $\Gamma - 1$ homologs; $\langle C\rangle_h$ is the mean correlation in energies of the homologs (Eqs. (7, 13)); $\langle Z\rangle_h$ is the mean Z-score for Γ individual homologs, Eq. (12); Z_h is the Z-score achieved with the homolog-averaged energies, the energy averaging is done with A_λ coefficients defined by Eq. (10); Z_0 is the Z-score for the root sequence; Z_{min} is the minimal Z-score found among Γ homologs; and Z^*_h is a theoretical lower limit for Z_h values defined by Eq. (9).

calculate the Z-scores of different folds for a group of homologs and base the fold recognition on the minimally found Z score.

For comparison, Table I presented also Z^*_h, the theoretical lower limit for Z_h values. A comparison of $\langle Z\rangle_h$, Z_h, and Z^*_h values shows that Z_h is rather close to its theoretical limit Z^*_h.

The found Z_h values exceed the $\langle Z\rangle_h$ values by approximately 30%. This is very close to that which can be expected (see Eq.(11)) for $\Gamma = 10$ (on the average) homologs whose energies correlate at a level of $\langle C\rangle_h = 0.52$. Thus, one can estimate the expected improvement of the Z-scores, having a given number of homologs and a given $\langle C\rangle_h$ value. The latter can be precisely calculated from a rather small set of alternative folds, as one can see from comparison of the corresponding mean $\langle C\rangle_h$ values in Tables I and II (although the former refers to threading over ~50,000 folds and the latter over only 300). A rough estimate of the $\langle C\rangle_h$ value can be obtained even without threadings since, as Tables I and II show, $\langle C\rangle_h$ is usually rather close to the mean HSSP similarity S of the root sequence to the remaining $\Gamma - 1$ homologs.

TABLE II. Average Z_h-Score Values Obtained in Threading Tests With Averaging of Energies Over Homologs at Different Thresholds of Energy Correlation†

Threshold	Γ	$\langle C\rangle_h$	S	$\langle Z\rangle_h$	Z_h
0.99	38	0.63	0.58	−6.18	−7.73
0.90	26	0.61	0.56	−6.14	−7.79
0.80	17	0.57	0.52	−6.04	−7.84
0.70	10	0.53	0.50	−6.02	−7.90
0.60	5	0.49	0.47	−6.07	−7.83

†Averaging is done over 20 protein families (see Table I) using a small set of 300 alternative folds. Γ, $\langle C\rangle_h$, S, $\langle Z\rangle_h$ and Z_h are described in the footnote to Table I; here they are averaged over 20 protein families.

The improvement in the accuracy of protein structure recognition due to energy averaging depends on the "quality" of homologs used in averaging. The selection of homologs is a controversial procedure because from the very beginning, homologs are selected for the database by their similarity. On the other hand, homologs selected for averaging should be as dissimilar as possible to give the least correlation in energy as required by Eq.(11). Furthermore, the same equation shows that the more homologs we have the better, and their number drops only when very remote homologs are taken.

To optimize the selection of homologs, we did some preliminary tests; the results are given in Table II.

Table II shows that most of the homologs in the original HSSP sets of alignments have too much similarity. They are optimally excluded at a 0.7 energy correlation threshold; here, the number of homologs is not too small yet. At smaller values of this threshold, the number of selected homologs drops drastically and this limits the advantage of the homolog-averaging approach.

Besides calculations with the basic estimate of A_λ coefficients given by Eq.(10), we investigated a few more

TABLE III. Number of "False Positives" Found in Threading Energies of Individual Chains and in the Homolog-Averaged Threading Energy

Individual chain[a]	Number of false positives
cyc_orysa[b]	2
cy2_rhogl	1
c550_thino	1
cy2_rhoph	2
cyc_tetpy	8
Energies averaged over the above 5 homologs of 1ccr family	0

[a] All these chains belong to the cytochrome *c* family of homologs (PDB code 1ccr); chain names below are given by EMBL/SWISSPROT identifiers.

[b] Corresponds to a cytochrome *c* molecule (PDB code 1ccr).

approximations of these coefficients. For some of them, we observed some reduction in *Z*-scores; however, neither of them performed better than the basic one. In addition, we explored the possibility of deriving the weights in a way that does not depend on preliminary threading-based estimates of deviations D_λ. To this end, we used the simplest weights $A_\lambda = 1/\Gamma$ for all the homologs $\lambda = 1, 2, \ldots, \Gamma$. The obtained (with $A_\lambda = 1/\Gamma$) Z_h scores were essentially the same as those obtained with $A_\lambda = 1/D_\lambda$. Thus, for immediate applications, one can use the simplest estimate of the averaging coefficients, $A_\lambda = 1/\Gamma$.

The averaging can be done only over the residues that are aligned with no-gap characters in homologs, and this also does not change the increase of the Z_h scores: when all the aligned residues are taken into the averaging, the Z_h scores are 32% better than the $\langle Z \rangle_h$ ones; when the averaging is applied only to the positions that are aligned with no-gap characters in homologs, the resulting scores are 27% better than $\langle Z \rangle_h$ scores.

One additional experiment has been done with the chains demonstrating "false positives," i.e., the chains for which some wrong folds have an energy below that of the native fold. All these chains belong to the family of cytochrome *c* (PDB code 1ccr); in other families, no false positives were observed in our tests. The results are presented in Table III.

As predicted, we obtained an impressive improvement: after averaging over the chains, *each* of which demonstrates some false positives, the averaged chain energy demonstrates no such error.

CONCLUSIONS

In this work, we examined how much one can increase the accuracy of protein structure recognition by energy averaging over homologs. The accuracy of recognition was estimated by the Z-score values obtained in threading tests. These tests showed that energy averaging over a family of homologous sequences leads to a perceptible improvement of the resulting Z-score value compared to the Z-scores of any individual homologs of the family.

The highest improvement in accuracy was found when the number of deletions in a sequence does not exceed 20% as compared with the root sequence, and a pairwise energy correlation between homologs does not exceed 0.7.

Although the obtained improvement of Z-scores is perceptible (on average, by 32%: from −6.1 for individual chains to −8.1 for the family), it is not as large as one could desire since the current databases contain a relatively small number of really diverse sequences (with sequence similarity below 0.50). Nevertheless, the obtained improvement of Z scores means (cf. Eq.(3)) that a correct fold can be found (on the average) among $3 * 10^{15}$ random folds when the energy is averaged over homologs, while for individual chains, it can be found within $3 * 10^{9}$ random folds only.

The results of this work suggest that the problem of errors in energy functions can be partly but not completely overcome by using a set of homologous sequences for prediction of their common native fold.

ACKNOWLEDGMENTS

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 the Russian Foundation for Basic Research and by an International Research Scholar's Award No. 75195–544702 from the Howard Hughes Medical Institute.

REFERENCES

1. Levitt M. Detailed molecular model for transfer ribonucleic acid. Nature 1969;224:759–763.
2. Maxfield FR, Scheraga HA. Improvements in the prediction of protein backbone topography by reduction of statistical errors. Biochemistry 1979;18:697–704.
3. Schulz GE, Schirmer RH. Principles of protein structure. New York: Springer-Verlag; 1979. 314 p.
4. Gutell RR, Weiser B, Woese CR, NollerHF. Comparative anatomy of 16-S-like RNA. Prog Nucleic Acid Res 1985;32:115–216.
5. Benner SA, Gerloff D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: the catalytic domain of protein kinases. Adv Enzyme Regul 1990;31: 121–181.
6. Rost B. PHD: predicting one-dimensional protein structure by profile-based neutral networks. Meth Enzymol 1996;226:525–539.
7. Keazar C, Elber R, Skolnick J. Simultaneous and coupled energy optimization of homologous proteins: a new tool for structure prediction. Fold Des 1997;2:247–259.
8. Dunbrack RL Jr, Gerloff DL, Bower M, Chen X, Lichtarge O, Cohen FE. Meeting review: the second meeting on the critical assessment of techniques for protein structure prediction (CASP2), Asilomar, California, December 13–16, 1996. Fold Des 1997;2: R27–42.
9. Finkelstein AV. Protein structure: what is possible to predict now. Curr Opin Struct Biol 1997;7:60–71.
10. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
11. Finkelstein AV. 3D Protein Folds: homologs against errors. A simple estimate based on the random energy model. Phys Rev Lett 1998;80:4823–4825.
12. Sippl MJ, Jaritz M. Predictive power of mean force pair potentials. In: Bohr H, Brunak S, editors. Protein structure by distance analysis. Amsterdam: IOS Press; 1994. p 113–134.
13. Reva BA, Finkelstein AV, Sanner MF, Olson AJ, Skolnick J. Recognition of protein structure on coarse lattices with residue-residue energy functions. Protein Eng 1997;10:1123–1130.
14. Finkelstein AV, Gutin AM, Badretdinov AY. Perfect temperature for protein structure prediction and folding. Proteins 1995;23:151–162.

15. Reva BA, Finkelstein AV, Sanner MF, Olson AJ. Residue-residue mean force potentials for protein structure recognition. Protein Eng 1997;10:865–876.
16. Reva BA, Finkelstein AV, Skolnick J. Derivation and testing residue-residue mean-force potentials for use in protein structure recognition. In: Protein structure prediction methods and protocols. Totowa, NJ: Humana Press, Inc.; 1999. In press.
17. Hendich M, Lackner P, Weitckus S et al. Identification of native folds amongst a large number of incorrect models. J Mol Biol 1990;216:167–180.
18. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
19. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an rmsd of 6Å? Fold Des 1998;3:141–147.
20. Bernstein FC, Koetzle TF, Williams GJB et al. The protein bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80:319–324.
21. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci. 1992; 1: 409–417.