Evaluation of Threading Specificity and Accuracy

Stephen H. Bryant
Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894

ABSTRACT **Threading experiments with proteins from the globin family provide an indication of the nature of the structural similarity required for successful fold recognition and accurate sequence-structure alignment. Threading scores are found to rise above the noise of false positives whenever roughly 60% of residues from a sequence can be aligned with analogous sites in the structure of a remote homolog. Fold recognition specificity thus appears to be limited by the extent of structural similarity, regardless of the degree of sequence similarity. Threading alignment accuracy is found to depend more critically on the degree of structural similarity. Alignments are accurate, placing the majority of residues exactly as in structural alignment, only when superposition residuals are less than 2.5 Å. These criteria for successful recognition and sequence-structure alignment appear to be consistent with the successes and failures of threading methods in blind structure prediction. They also suggest a direct assay for improved threading methods: Potentials and alignment models should be tested for their ability to detect less extensive structural similarities, and to produce accurate alignments when superposition residuals for this conserved "core" fall in the range characteristic of remote homologs.** © 1996 Wiley-Liss, Inc.*

Key words: **contact potentials, fold recognition, protein threading, Gibbs sampling**

INTRODUCTION

Sequence-structure threading methods offer a means to recognize similarity to a protein of known structure in the absence of detectable sequence similarity.[1–6] A recent test of these methods, in blind structure prediction, provided some clear indications of success, in that several nontrivial similarities were detected.[7,8] Predictions were by no means uniformly successful, however, and the results in some ways raised as many questions as they answered. One would like to know, in particular, why the correct folds were recognized for some sequences but not others, and why threading alignments sometimes agreed with structural superpositions, but were other times grossly wrong. Do these results reflect crucial differences among threading methods, and the chance success or failure of the heuristic alignment algorithms they employ[9–13]? Or do they instead reflect differences among the proteins in the prediction sample, with sequence-structure compatibility being more easily recognized in some cases than others?

In this paper I address one of these questions, the dependence of successful fold recognition on the nature and extent of structural similarity. Using one of the threading methods employed at Asilomar,[12] I conducted an all-against-all threading comparison of sequences and structures from the globin family, proteins that span a wide range of sequence and structural similarity. This test, it seems, reproduces the phenomenon seen in blind predictions. The common fold of these proteins is recognized with statistically significant threading scores in some cases, but not all, and threading alignments are quite accurate in some sequence-structure comparisons, but not all. Since the number of threading trials is large, however, and the true structures known, I may examine systematically the relationship of threading scores and alignment accuracy to measures of structural and/or evolutionary similarity. In particular, I may search for any threshold of sequence or structural similarity beyond which threading cannot reliably recognize a related fold or produce an accurate model.

These experiments suggest that successful fold recognition depends very critically on the extent of structural similarity. Threading scores for globin comparisons appear to be independent of the degree of sequence similarity per se, but are statistically significant, on average, only when 60% of residues occupy structurally analogous sites. Alignment accuracy also depends quite critically on the degree of structural similarity, with the majority of residue pairs aligned correctly, in exact agreement with structural alignments, only when root mean square (RMS) superposition residuals are under 2.5 Å. One must be cautious in the interpretation of a small number of blind predictions, but it appears that the success and failures of this threading method at

Received December 15, 1995; accepted April 1, 1996.
Address reprint requests to Dr. Stephen H. Bryant, Computational Biology Branch, National Center for Biotechnology Information, NIH, 8600 Rockville Pike, Bethesda, MD 20894.

Asilomar are consistent with these expectations. Statistical significance of threading scores indeed depended on the extent of structural similarity, the number of superposable residues, and alignment accuracy on the superposition residual in structural alignment.

Proteins with similar folds, as commonly defined,[12,14–17] show a range of structural similarities, often involving fewer than 60% of residues. In the future one would like to detect and accurately model by threading a greater proportion of such cases. If fold recognition is generally limited by the extent of structural similarity, as these experiments suggest, then it seems one should evaluate possible improvements in threading methods with respect to precisely this criterion. Specifically, as an assay system for threading methods, one should examine the dependence of scores and alignment accuracy on the extent and degree of structural similarity, in a manner similar to that considered here. A potential or alignment model is likely to perform better in blind prediction if it can detect above the noise of false positives[2] a less extensive structural similarity, encompassing, for example, 50% of residues forming a protein domain. Model accuracy is likely to improve if such a method can produce alignments in agreement with those of structural superposition when residuals fall, for example, in the 3-Å range.

METHODS

Structural Data

A total of 24 globins, spanning a range of sequence and structural similarity, were selected from the Protein Data Bank (PDB).[18] The sample includes vertebrate myoglobins, vertebrate hemoglobins, and other globins from invertebrates and plants. Two structurally related C-phycocyanins[19] and colicin A[20] are also included. PDB identification codes for these proteins are: 1mbd, 1mbs, 1pmb-a, 2mml, 1yma, 1myt, 2mhb-b, 2hhb-b, 1fdh-g, 1hds-b, 1pbx-b, 2hhb-a, 2mhb-a, 1hds-a, 1pbx-a, 1mba, 1lh4, 3sdh-a, 2lhb, 1ecd, 1ith-a, 1cpc-a, 1cpc-b, 1col-a.

Structures were superposed manually by molecular graphics. From this multiple alignment a common core substructure of 70 residues was identified, consisting of the central 3 or 4 turns of helices A, B, E, F, G, and H.[21] In sperm whale myoglobin, 1mbd, these 6 core elements are formed by residue sites 5–17, 24–32, 64–76, 85–93, 102–114, and 129–141. In this superposition the fraction of identical residues in the pairwise comparisons ranges from 4% to 100%, with half showing fractional identities of 20% or less. Pairwise RMS superposition residuals for C_α atoms range from 0 to 3.9 Å, with 82% of comparisons showing residuals less than 3 Å. RMS values quoted below similarly refer to superposition residuals of aligned C_α atoms.

Threading Alignment Model

Sequences were threaded through core structures by using the contact potential and core-element alignment model described previously.[2,12,22] Contacts are defined on the basis of virtual C_β coordinates, and contain no implicit memory of the side chains in native sequences. The contact potential was derived statistically from nonlocal contacts in a set of proteins that excluded all globins. Minimum-size globin core elements were defined from the conserved substructure identified by multiple structure comparison. All threading models were thus required to contain at least 70 residue sites, corresponding to an alignment of residues from a sequence with the central sites of the 6 helices in a globin structure. Constraints on the maximum lengths of the 5 intervening loops were set to 7, 31, 16, 10, and 20 residues, respectively, values that exclude loops of much greater length than seen in members of the globin family.[21] Minimum loop lengths were determined dynamically, based on the number of residues required to span the distance between the endpoints of sequentially adjacent core elements.

This use of core definitions based on known structural similarity is intended to rule out the possibility of poor threading scores or alignments due to incorrect specification of minimum-size core elements or loop length constraints. Model quality as measured here is thus limited by the inherent accuracy of the potential and convergence properties of the alignment algorithm, and is intended to represent a most favorable case for fold recognition, when an accurate description of the conserved core of a protein family is known. I note, however, that threading scores and alignments for globins are generally similar when core elements and loop-length constraints are determined automatically, based on geometrical criteria intrinsic to each structure.[2] The exceptions are threading models based on the structures of colicin A and C-phycocyanin, which contain large helices not present in other globins. With the current experimental design I may include these additional models in the analysis without the confounding effects of core-definition error.[12]

Alignment Optimization

Favorable sequence-structure alignments were identified by the Gibbs sampling algorithm shown schematically in Figure 1. In this procedure the alignment of subsequence blocks with core elements is sampled iteratively in the field defined by the pairwise contact potential and the alignment of other core elements. The locations in the structure of core element endpoints are also sampled iteratively, and core elements thus allowed to extend beyond their pre-defined minimum sizes by chain-continuous addition of new residue sites. Recruitment of new sites into the threading model is similarly governed by the contact potential and endpoint locations and alignments of other core elements. Threading alignment with this algorithm does not require a "frozen approximation" to construct profilelike terms from the pairwise contact potential[1,10,23] and does not involve gap penalties.

Threading alignments for all sequence-structure pairs were optimized using an annealing schedule found in test cases to yield reproducible threading scores and alignments (not shown). It is possible to verify convergence to the global minimum contact energy only in test cases with fixed core element endpoints, however, where exhaustive enumeration[22] remains feasible (not shown). This annealing schedule calls for 50 random-alignment starts, each with 40 iterations of core element alignment and endpoint sampling. Each iteration calls in turn for 10 cycles of alignment and 10 cycles of endpoint refinement, with each cycle involving a potential move of each core element's alignment or endpoints. Nominal temperature for Boltzmann sampling of alternative alignments was reduced from 10 to 8 and from 8 to 5 kT units, at 20 and 30 iterations, respectively, and maintained at 5 kT units for endpoint sampling. For p-value calculations, only 25 random starts and the first 10 iterations were performed, to reduce computer time. This change has a minor effect on threading scores, and, in test cases, no systematic effect on p values (not shown).

Programs implementing the Gibbs alignment algorithm were written in the S and C languages, and make use of the PKB database and program library.[24] Source code for threading and analysis is available via internet at http://www.ncbi.nlm.nih.gov/Structure/. Computer time requirements with the complete annealing schedule above were approximately 16 minutes per alignment on a Silicon Graphics R4400 processor.

Threading Score Evaluation

Threading scores reported below correspond to the quantity $Z(r|m)$ described previously.[22] They refer the sum of contact potentials to the distribution obtained upon randomly shuffling the aligned residues 10,000 times, and may be understood as a composition-corrected raw score expressed in standard deviation units.[2] Sequence-sequence comparison scores were calculated in the same manner, with the PAM 250 matrix[25] and the sequence of a database structure taking the place of the threading contact potential and contact list of that structure. The sequence-dissimilar subset referred to below was defined by a sequence-sequence comparison Z score of 7 or less, which corresponds in this sample to 22% or less residue identity in the conserved globin structural core.

Threading p values are estimated empirically by referring the composition-corrected threading score to the distribution obtained for 100 randomly shuffled copies of the threaded sequence, each optimally aligned with the structure in question. The threading score of the unshuffled sequence is expressed in standard deviation units relative to this distribution, which is assumed to be normal, and the p value calculated as the integral of the standard normal for scores greater than or equal to this value. Threading p values give the odds that a randomly chosen sequence would fit the structure as well, and are a measure of the probability that false positives would obscure globin sequence-structure compatibility in a database search.[2]

Model Accuracy Evaluation

Alignment accuracy is measured in two ways: as a shift error in residues, and as a percentage of residues aligned in strictly correct agreement with structural superposition. Shift error is the number of positions to the left or right that the threaded sequence must be moved to place it in agreement with structural alignment. In the core-element alignment model residues from a sequence are aligned as an ungapped block with residue sites from a structure, and the shift error of all residues aligned with a core element is necessarily the same. The mean shift error of an alignment is thus calculated from the shift error of each core element, weighted by the relative numbers of residues each contains. Alignment accuracy as shown below is very similar if these weights are set to 1 or if measured by the shift errors of individual core elements rather than as the mean across the 6 core elements of sequence-structure alignments (not shown).

The optimization algorithm employed here produces an ensemble of alternative sequence-structure alignments, each with an associated conformational potential ΔG_i, where i indicates a particular threading model. ΔG_i corresponds to the quantity $\Delta G(r|m)$ defined previously,[22] the sum of pairwise contact energies less their expected value for a randomly assigned sequence of that length and composition. The statistical weight of each alternative alignment was calculated by a conventional Boltzmann factor,

$$e^{-\Delta G_i/kT} / \sum_i e^{-\Delta G_i/kT}$$

with the final annealing "temperature" of 5 kT units. Alignment accuracy was evaluated as the weighted average across this ensemble. To conserve computer storage this ensemble was represented by the 4000 lowest-energy alignments and endpoint locations identified by the Gibbs sampling algorithm, a sample found in test cases to include all alignments with significant weight (not shown). In practice relatively few low-energy alignments are identified, and alignment accuracy as shown below is

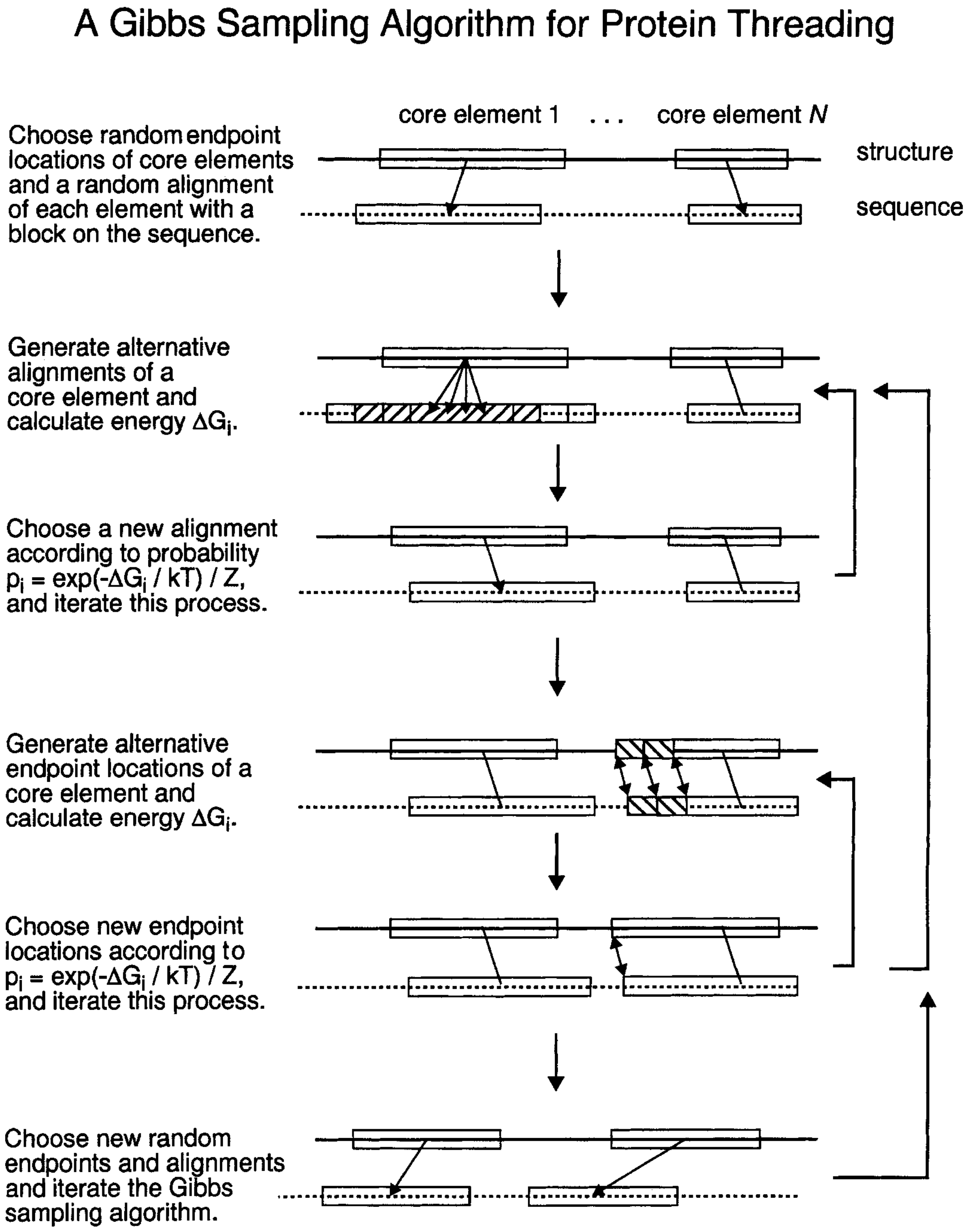

A Gibbs Sampling Algorithm for Protein Threading
core element 1 ... core element N
structure
sequence
Choose random endpoint locations of core elements and a random alignment of each element with a block on the sequence.
Generate alternative alignments of a core element and calculate energy ΔG_i.
Choose a new alignment according to probability $p_i = \exp(-\Delta G_i / kT) / Z$, and iterate this process.
Generate alternative endpoint locations of a core element and calculate energy ΔG_i.
Choose new endpoint locations according to $p_i = \exp(-\Delta G_i / kT) / Z$, and iterate this process.
Choose new random endpoints and alignments and iterate the Gibbs sampling algorithm.

Fig. 1. Schematic representation of a Gibbs sampling algorithm for protein threading. The boxes indicate subsequence blocks which are aligned with equal-length, chain-continuous segments from a known structure. Alternative alignments of these blocks and extensions of their endpoints by chain-continuous recruitment are sampled in tandem, as shown. The partition function for sampling is defined as $Z = \Sigma_i \exp(-\Delta G_i/kT)$, where the sum is taken in each cycle over all alignments or endpoint positions allowed by the lengths of the sequence and structure, the alignment of other blocks, and the distances between residue sites in the structure (see text).

similar if only the top-scoring alignment for each sequence-structure pair is considered (not shown).

Core element endpoint definition error is examined below as the relationship between the frequency with which a residue site is included in threading models in the Gibbs sampling ensemble and the superposition residual at that site in structural alignment. These residuals are calculated by extending without gaps the structural alignment of core elements, and may reach large values when adjacent loop regions differ greatly in length or conformation. Core element alignment recruits new sites to the threading model in a similar manner, by chain-continuous addition, and if the common substructure it identifies is accurate one expects to see that sites with large residuals are rejected. The analysis below considers only residue sites beyond the minimum-sized core elements included in all 576 threading models, 46,488 from a total of 86,808 sites. Site inclusion frequencies are calculated as a weighted average across the Gibbs sampling ensemble using the Boltzmann weights described. For calculation of alignment accuracy, which involves weights based on the relative lengths of core elements, endpoints are taken as the N- or C-terminalmost residue site with recruitment frequency above 0.8. The precise value chosen affects only slightly the mean shift error or percentage of residues correctly aligned, and has little effect on results below (not shown).

RESULTS

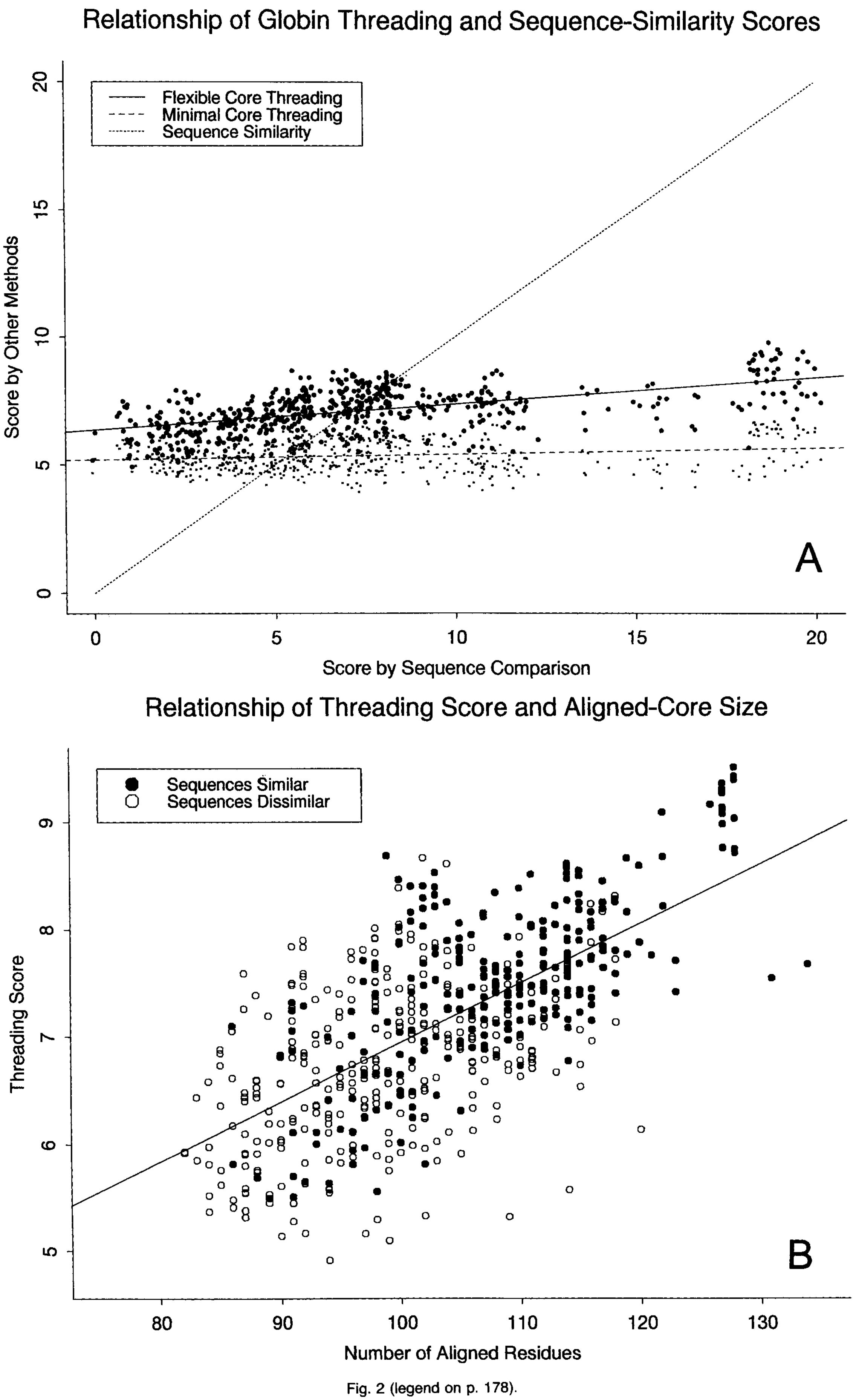
Fold Recognition Specificity

One of the most obvious questions, in seeking to understand the nature of the structural similarity detectable by threading methods, is to ask whether they in fact measure anything other than evolutionary distance, as traditionally measured by sequence comparison with substitution probability matrices.[26] Data relevant to this question is plotted in Figure 2A, which shows the relationship of threading and sequence-similarity scores, calculated in an equivalent fashion, for $24 \times 24 = 576$ comparisons from the globin family. It may be seen that threading scores vary in a manner quite unlike sequence similarity scores, and that the slope of the fitted regression line is near 0, not near 1. The contact potential which forms the basis of the threading score is sensitive only to the coordinates of atoms in the polypeptide backbone, and it apparently does not "remember" well the sequence of the known structure,[4,22] or function as a measure of evolutionary distance comparable to sequence similarity. The absence of gap penalties in the threading score also contributes to it's apparent independence from sequence similarity, since differences in the lengths of loops increase with evolutionary distance.[21,27,28]

The strongest association observed between threading scores and many structural parameters examined is shown in Figure 2B. Scores increase regularly with the length of the sequence-structure alignment, which is equivalent to the number of residues in the threading model, with the quantitative relationship indicated by the fitted regression line. The figure is based on the size of the threading model as identified by the Gibbs sampling algorithm, via recruitment of additional residue sites to the minimum-size structural core of the globin family. The same relationship is apparent, however, if one substitutes the number of residues superposable to a 2-, 3-, or 4Å threshold (not shown), since recruitment typically includes only such sites (see below). It appears that threading scores are primarily determined by the extent of structural similarity, which one may define as the proportion of residues which occupy structurally analogous sites.

The size of the common substructure among globins decreases systematically with evolutionary distance, precisely because of the accumulation of insertions and deletions in loop regions.[21,27,28] With respect to threading scores, these variables are thus confounded, and dependence on one cannot be formally distinguished from dependence on the other. Several observations argue that the size of the common substructure is best interpreted as the causal factor determining threading scores, however. One is the change in threading scores when recruitment of additional residues sites is enabled, as shown in Figure 2A. For individual pairwise comparisons, where evolutionary distance is constant, threading scores increase when the size of the identified common substructure is allowed to increase. Another relevant observation is provided by the stratification of threading models in Figure 2B into low- and high-sequence similarity groups. A roughly comparable dependence of threading score on the size of the aligned substructure is apparent, regardless of evolutionary distance. For theoretical reasons one might also expect threading scores to increase with the number of residues correctly aligned with analogous sites in a structure. Since residues in native structures tend to have comparable, favorable, interactions with their environment,[29,30] threading scores, as a sum over these interactions, should increase in proportion to the numbers of such sites.

The extent of structural similarity necessary for specific fold recognition is suggested by the data shown in Figure 2C, which plots the dependence of threading p values on the number of residues in the threading model structure. Threading p values give the probability that sequence-structure complementarity would be masked by false positives, the probability that a randomly chosen sequence would score as well as the particular sequence threaded through each structure.[2] The value $p = .05$ might be considered significant in pairwise comparison, and is indicated on the plot. One may see that the fitted regression line crosses $-\log(.05) = 1.3$ at 87 residues, which for globin sequences with a median length of 146 corresponds to 60% of residues occupying analogous sites. This may be interpreted as a threshold of structural similarity above which one can expect, on average, to recognize structural similarity. In searching a large database of protein core substructures a lower p value is required for inference of statistical significance, since the number of false positives crossing a given threshold will increase in proportion to database size.[2,26] For a database of 500 structures a value of $.05/500 = .0001$ might be considered significant.[12] As can be seen in Figure 2,



Relationship of Globin Threading and Sequence-Similarity Scores

Relationship of Threading Score and Aligned-Core Size

Fig. 2 (legend on p. 178).

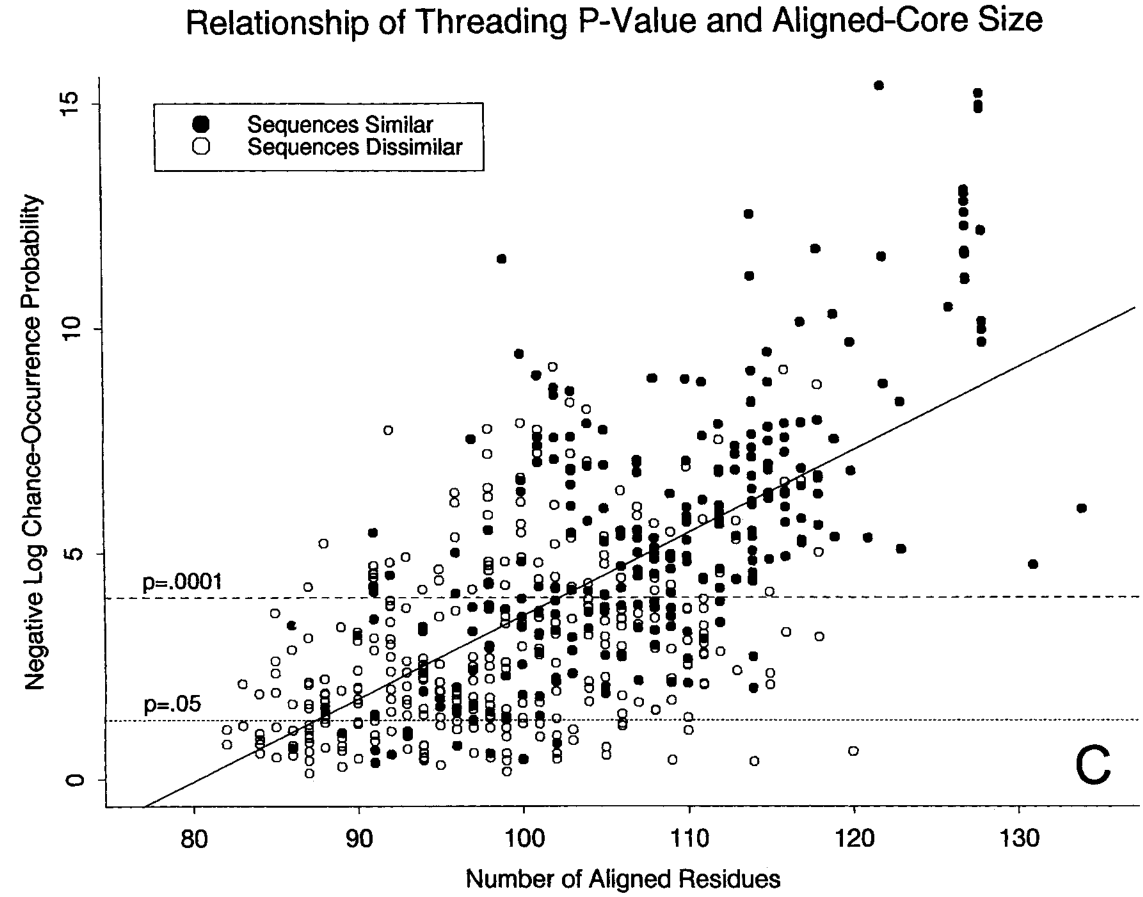


Fig. 2. Threading scores for globin family comparisons. **A:** Comparison between threading and sequence similarity scores for threading models with and without recruitment of additional residue sites beyond their common minimum-size core structure. Fitted regression lines are shown, and a line with slope 1 plotted to facilitate comparison of threading and sequence similarity scores. **B:** Relationship of threading scores and the number of residue sites in each threading model, and a fitted linear regression line. **C:** Relationship of the negative logs to the base 10 of threading p values and the number of residue sites in the threading model. A fitted regression line is shown, together with significance thresholds of $p = .05$ and $p = .0001$.

this level of significance is reached, on average, when 102 residues are aligned, 70% of the median globin sequence length. The strong dependence of threading score statistical significance on the extent of structural similarity is apparent.

Model Accuracy

A threading model is as accurate as possible when residues from a sequence are correctly aligned with the sites from a structure that are found to be analogous and superposable by structure-structure comparison. The most direct quantitative measures of model accuracy thus involve comparison of the sequence-structure and structure alignments. Figure 3 presents an example from the globin threading alignments considered here and it's comparison to the results of structural superposition. The figure shows the threading model for the C-phycocyanin from the cyanobacterium *Fremyella diplosiphon* (1cpc-a),[31] based on the structure of hemoglobin from the antarctic fish *Pagothenia bernacchii* (1pbx-a).[32] These proteins show 10% residue identity in their structural core, and superpose to a RMS C_α residual of 2.9 Å. Because of the high superposition residual this is one of the less accurate threading alignments in the sample.

Figure 3A displays the frequency with which the six hemoglobin helices are aligned with various residues of the phycocyanin sequence in the alignment ensemble produced by the Gibbs sampling algorithm. A perfect alignment would center each helix on the residues indicated by arrows, corresponding to exact agreement with structural alignment. It may be seen that the central sites of helices 2, 5, and 6 are most frequently aligned with the correct residues, and those of helices 3 and 4 with residues 3 and 1 positions displaced from the correct ones, respectively. All of the helices show a high frequency of "register shifts" of 1, 3, or 4 residues, however, a shift which corresponds to zero or one helical turns

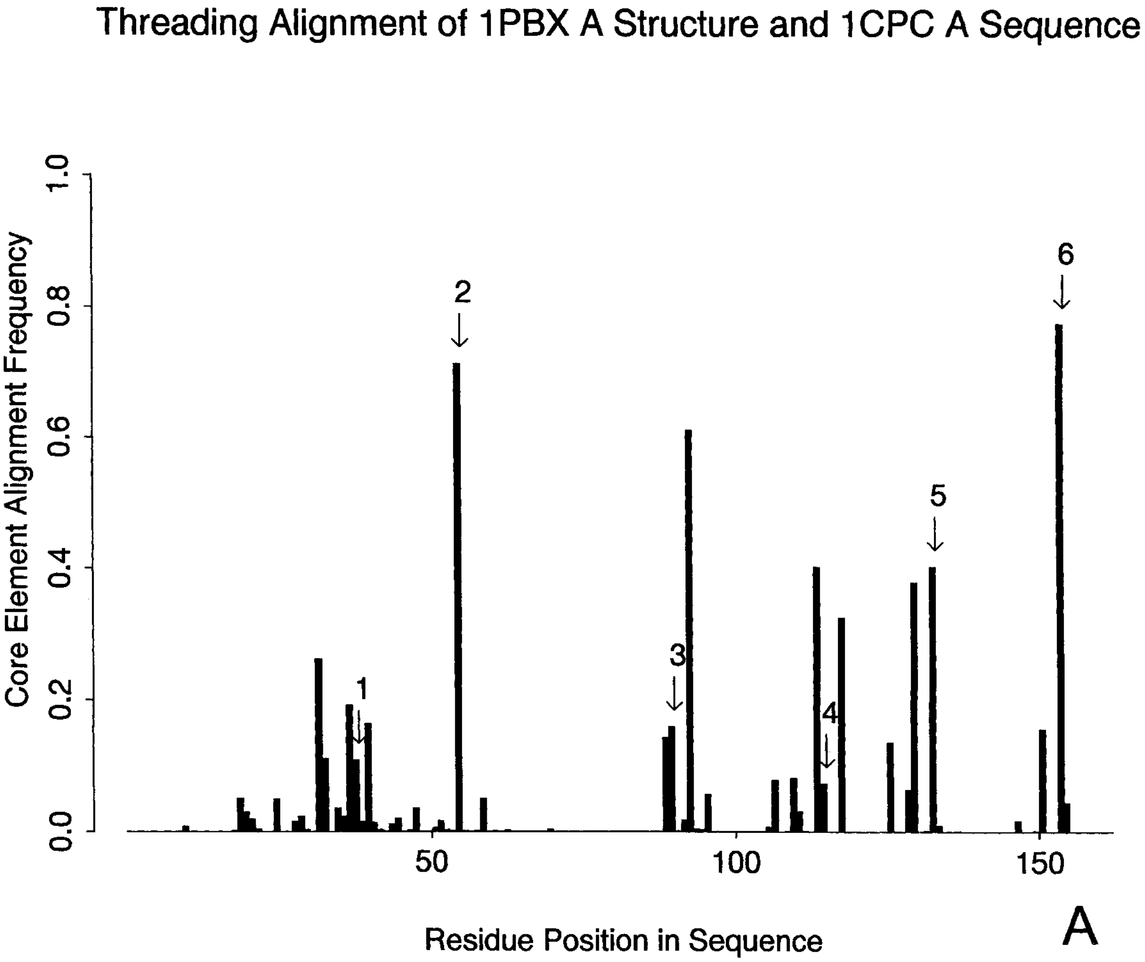


Fig. 3 (legend on p. 180).

and which preserves the amphipathic orientation of the helix. Helix 1, which superposes poorly between the two structures, shows a larger displacement of 2 turns, and with low frequency even larger displacements toward the N terminus, a region that forms a different, nonequivalent helix in the structure of phycocyanin. The threading alignment ensemble nonetheless localizes the core globin helices in the correct regions of the phycocyanin sequence, a characteristic pattern even when the true structures differ as much as they do in this example.

The results of recruitment of additional sites into the phycocyanin-from-hemoglobin threading model is summarized in Figure 3B. The frequencies with which different sites are added to the minimum-size globin core is plotted together with the root mean square residuals of each site in the phycocyanin-hemoglobin structural alignment. For these proteins structural similarity does not extend much beyond the minimum-size core, and one may see in the figure that recruitment tends to add few sites to the threading models in the ensemble. About one turn is frequently added to the C terminus of helix 2, however, up to the point where the superposition residual rises above about 5 Å, an effect which is more pronounced in comparisons of other proteins where helical extensions and loop conformations are more similar (see below). This example nonetheless shows that sites tend to be recruited to the threading model only when the true structures are roughly superposable in that region.

To characterize the overall accuracy of the 576 threading models one must employ summary measures of alignment accuracy, such as the percentage of correctly aligned residues and the mean shift error. Of many structural parameters examined, these summary measures of accuracy alignment seem to be most associated with the RMS superposition residual of the proteins compared. One may see in Figures 4A and 4B that the percentage of residue pairs correctly aligned decreases steadily with increasing RMS residual, on average, and mean shift error

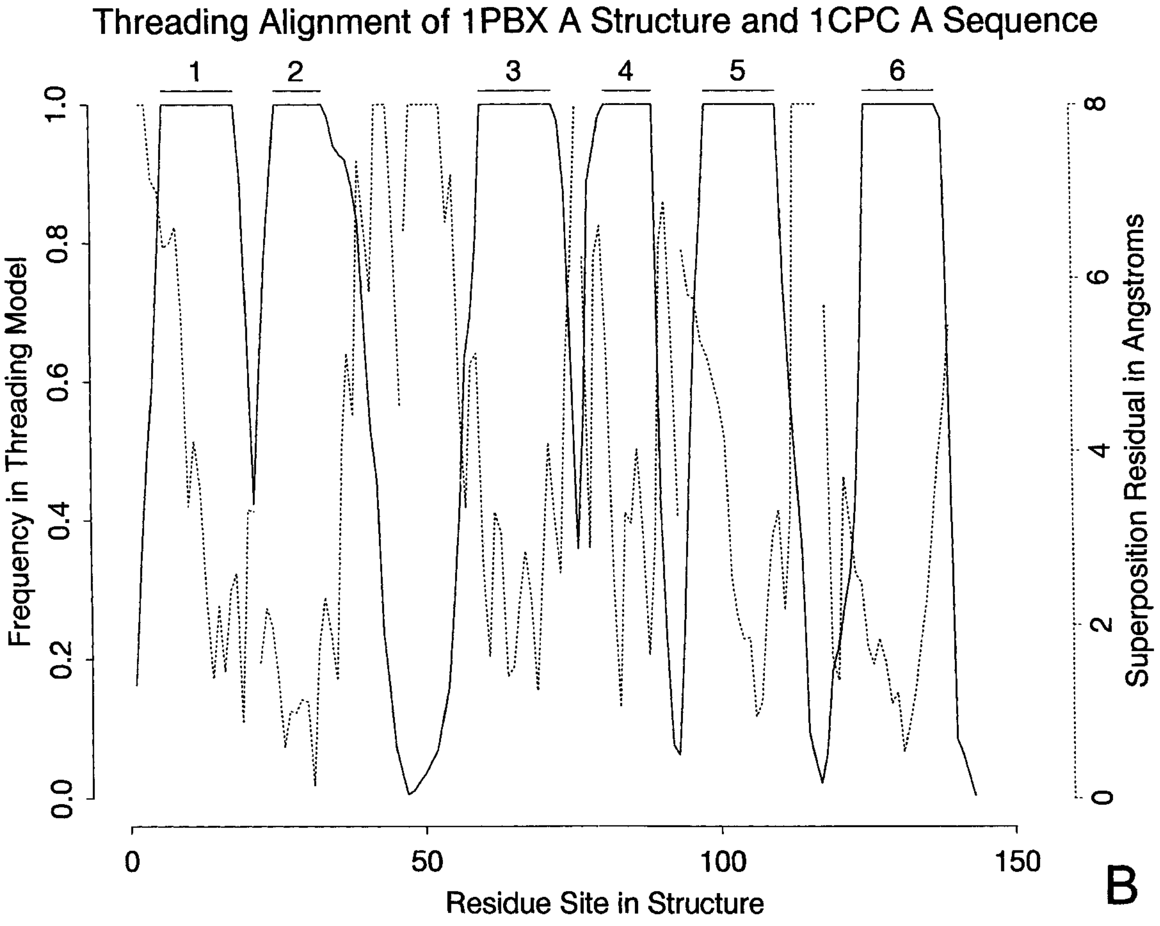


Fig. 3. Threading alignment of the phycocyanin (1cpc-a) sequence and hemoglobin (1pbx-a) core structure. **A**: The marginal frequencies with which the 6 hemoglobin helices are aligned at various positions on the phycocyanin sequence in the alignment ensemble produced by Gibbs sampling. For clarity, only the alignment of the central residue of each helical core element is plotted, corresponding to 1pbx-a residues 11, 28, 65, 84, 103, and 130 in PDB numbering. The correct alignment according to structural superposition is indicated by arrows. **B**: The marginal frequencies of core element endpoint locations in the 1pbx-a structure, the frequencies with which residues to the N or C terminus of each globin helix were recruited into the threading model. Residue sites forming the minimum-size globin core and necessarily present in all models are indicated by horizontal bars. The residue-by-residue superposition residual in the correct structural alignment is plotted as a dotted line, with the scale in Å units to the right (see text).

rises in proportion. Below 2 Å RMS the models are quite accurate. The "self" alignments in the sample, for example, are all completely correct (not shown). A critical value is reached at around 2.5 Å RMS superposition residual, a point where only about half of the residues are aligned with the corresponding structurally equivalent sites. Mean shift error is about 2 residues at this point, on average, suggesting that the other half of the residues are displaced from correct alignment by 4 residues, or one helical turn.

It is perhaps not surprising that the alignment ensemble begins to include "register shifts" at 2.5 Å, since one may imagine how an axial movement of a helix by this distance will produce contacts that are intermediate between the correct and one-turn-shifted alignment. It has been noted previously that side chain packing interactions generally begin to differ at 2.5 Å RMS.[33,34] The ensemble of threading alignments may necessarily broaden at this point, since the side-chain contacts of the database structure correspond less exactly to those of the protein one is attempting to model.

It should be noted that alignment accuracy also depends critically on the extent of structural similarity, the fraction of residues aligned with analogous sites. Alignment accuracy decreases for all models when recruitment of additional sites to the minimum-size globin core was not allowed (not shown), reminiscent of the decrease in threading scores shown in Figure 2A. Mean shift error increases by roughly 2 residues for all ranges of RMS residual, in fact, and the improvement in model accuracy brought about by the flexible core definition is clear. Alignment accuracy still deteriorates markedly above 2.5 Å RMS, however, and it seems reasonable to conclude that this degree of structural

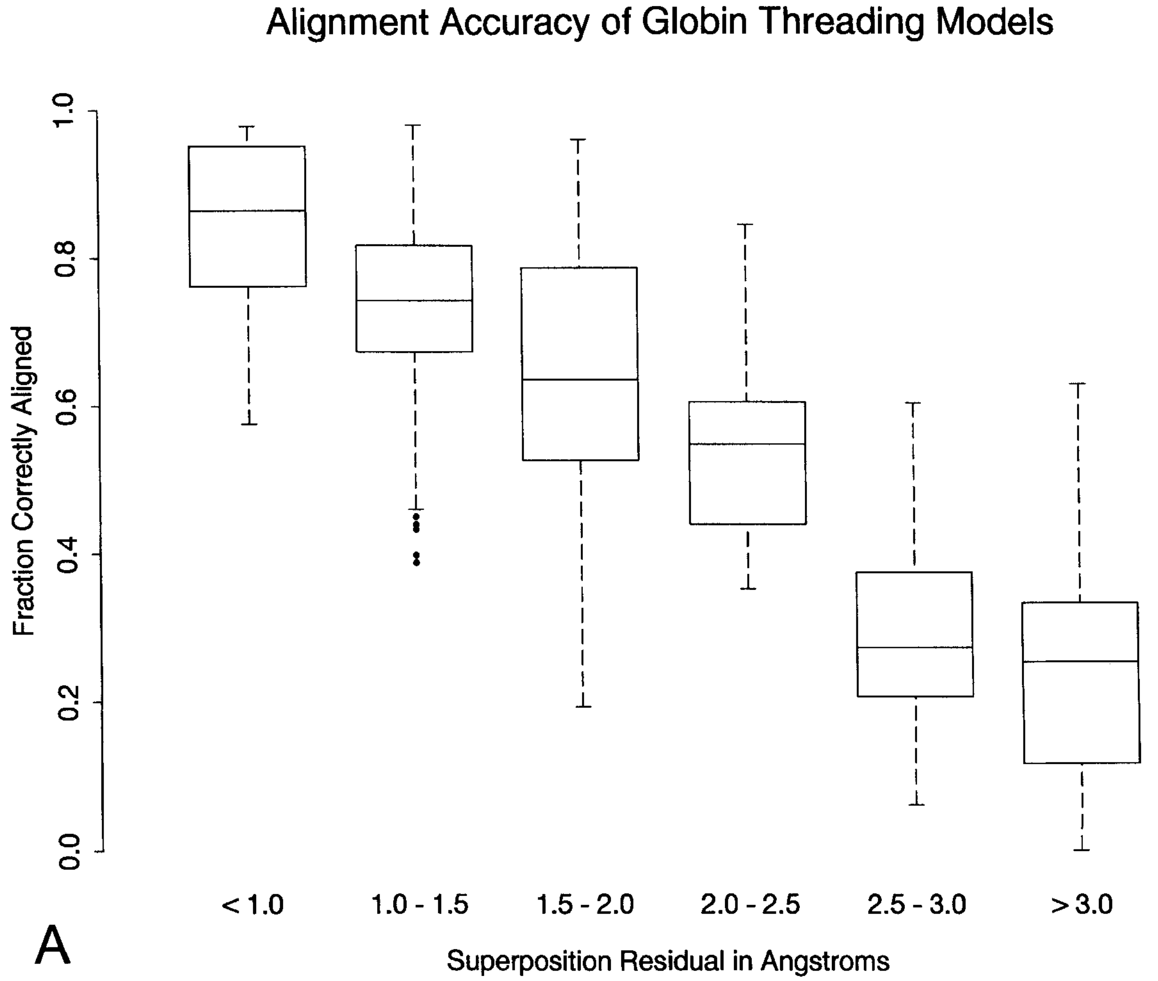


Fig. 4A (legend on p. 183).

similarity is a necessary if not sufficient condition for highly accurate threading alignments.

Accuracy in threading definition of the boundaries of the common substructure shared between globins is summarized in Figure 4C. The figure shows the frequencies with which adjacent helical-extension and loop sites were added to the minimal globin core, as a function of the RMS residual of those sites in structural alignment. It may be seen that for sites with superposition residuals under about 2.5 Å recruitment frequency is above 0.8, and that for larger superposition residuals values recruitment frequency decreases steadily. There is no point at which recruitment frequency drops completely to zero, perhaps because the alignments in the threading ensembles are not completely accurate as discussed, and perhaps because of chance sequence-structure compatibilities that do not correspond to similar loop conformations as judged by superposition residuals. Changes in conformational potential of the threading model are also necessarily small as individual sites are recruited, because they add only a few contacts, and the Gibbs sampling algorithm at the relatively high "temperature" employed here can be expected to produce only a stochastic definition of core element endpoints. It is clear, nonetheless, that sites are added to the threading models with high frequency only when they are structurally superimposable. It is the addition of this largely correct pairwise contact information that presumably accounts for the improved alignment accuracy of threading models based on flexible as opposed to fixed core definitions.

DISCUSSION

These experiments allow two observations relevant to the nature of the structural similarity detected by sequence-structure threading and the accuracy of the three-dimensional models it produces. For the globins compared, threading scores and their statistical significance depend above all on the extent of structural similarity, the size of the common substructure they share. Furthermore, the models are accurate, placing residues from the se-

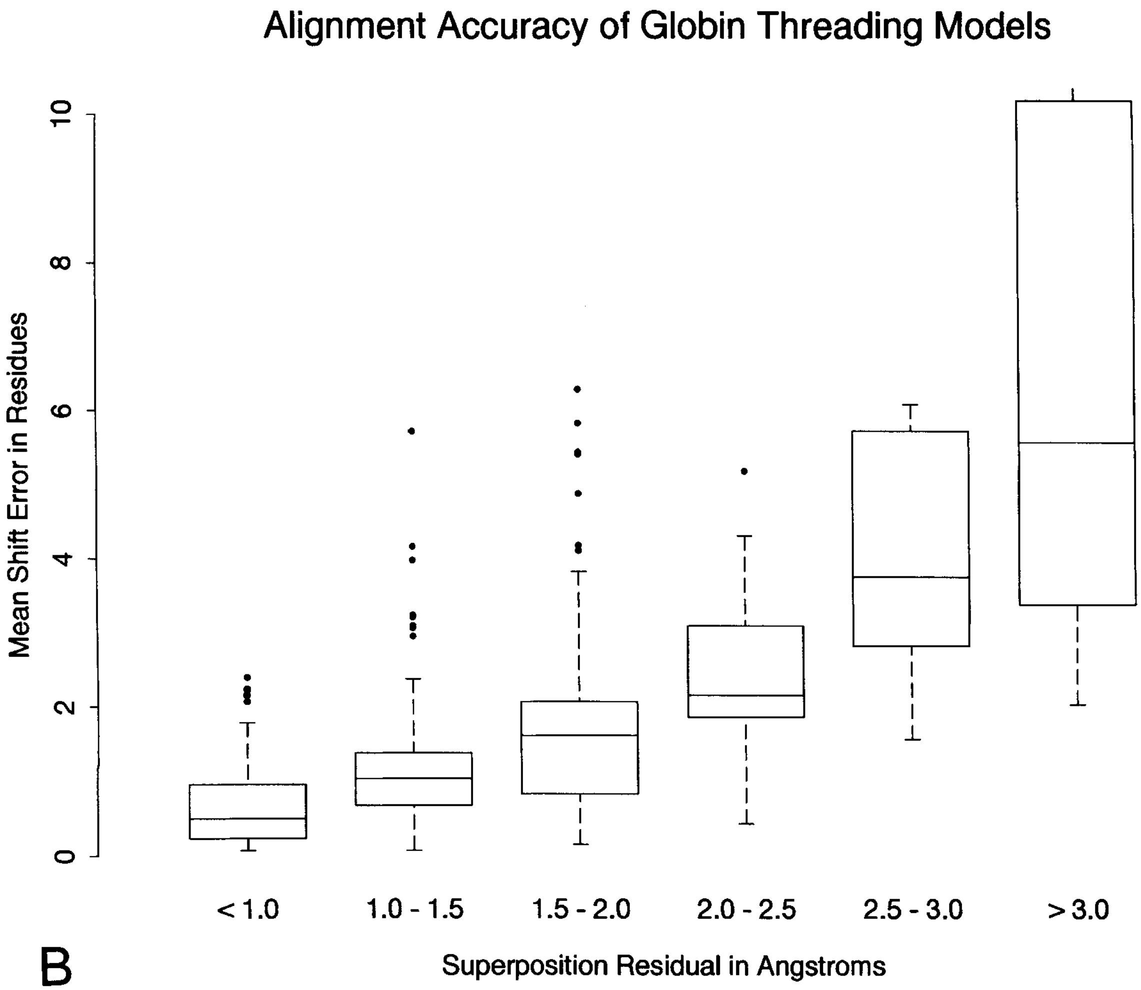


Fig. 4B (legend on p. 183).

quence in the correct, structurally analogous sites, only when true structures are similar to a degree that implies similar side-chain environments and pairwise contacts. From the experiments one can derive quantitative models for these relationships. Threading scores reach values that are statistically significant for pairwise comparison, on average, when the fraction of residues from the sequence that can be aligned with analogous sites in the structure approaches 60%. The majority of residue pairs in the threading alignment agree exactly with those of structural alignment only when superposition residuals of the proteins compared are, on average, 2.5 Å or less.

One cannot be sure that these suggested criteria for successful fold recognition and accurate modeling apply to other proteins. They appear to be consistent, however, with results of blind structure predictions undertaken with the same method for the recent Asilomar workshop.[8,12] In two cases true structures were later found to show significant similarity to a core structure defined in the search database.[12] In one, "prosub",[35] the threading model included 75% (58 of 77) of residues above a recruitment probability of 0.8 and yielded a top-ranked threading p value of 0.001. In the other, "rtp,"[36] 51% (62 of 122) of residues were included in the threading model, and the p value was only 0.1. Not all of the aligned residues in "prosub" corresponded to strictly superimposable sites,[12] and its p value does not fall exactly on the regression line of Figure 2C, but threading scores nonetheless reflect the extent of structural similarity. Alignment accuracies for "prosub" and "rtp" were 4.0 and 2.7 residues mean shift error, for RMS superposition residuals with respect to the database structures of 2.4 and 1.9 Å, respectively. Minor core definition errors in both cases forced slight misalignment,[12] and these values are somewhat greater than those expected for best-case, error-free core definitions as shown in Figure 4A, but they nonetheless fall within the range of variation seen there.

It is more difficult to tell whether the suggested criteria for specific fold recognition and model accu-

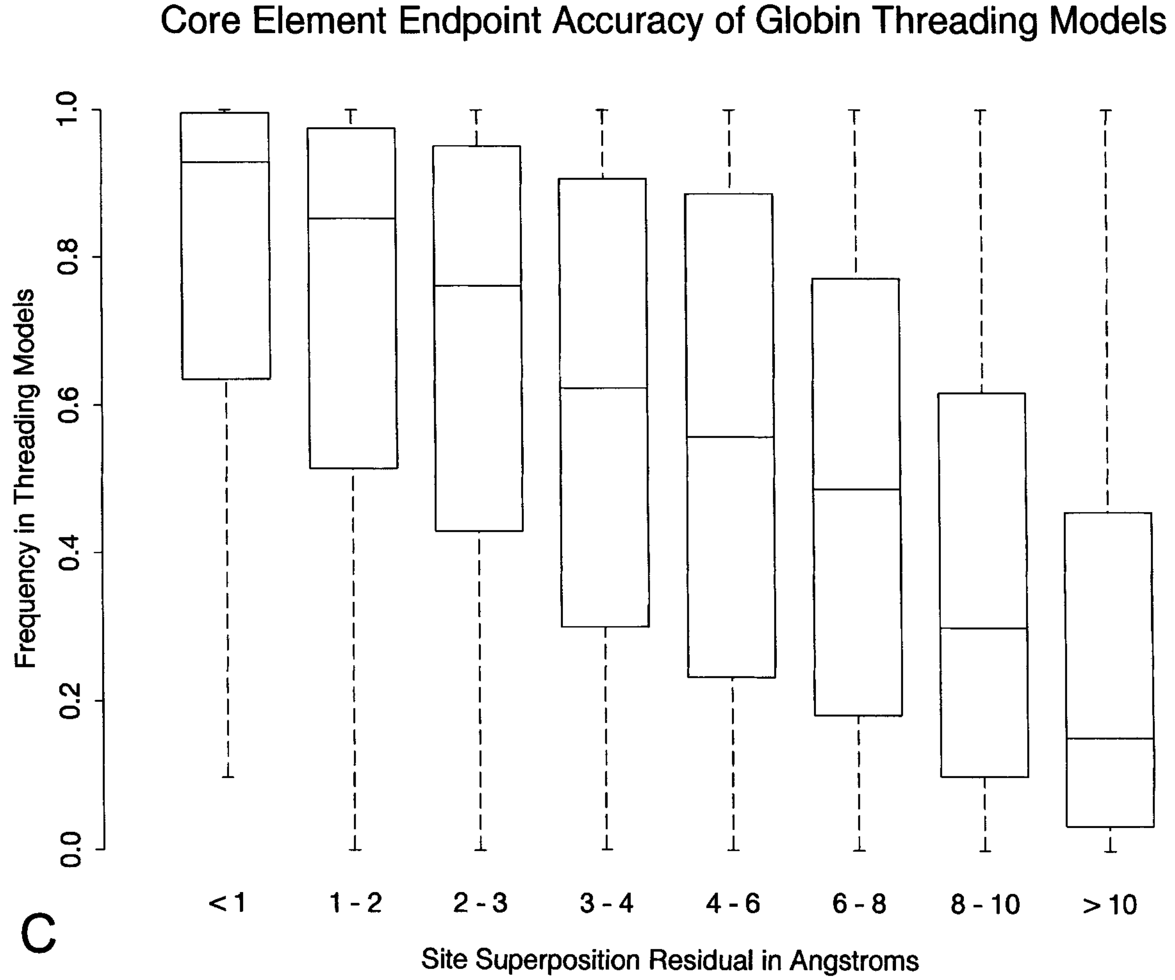


Fig. 4. Threading alignment accuracy for globin family models. **A**: The fraction of aligned residue pairs in exact agreement with structural alignment, as a function of the RMS superposition residual of the two structures. The boxplot shows the median value across the 576 sequence-structure alignments considered, the interquartile range, and "whiskers" drawn at 1.5 times the interquartile range. Alignment errors for individual models which are outliers are plotted explicitly. **B**: The mean shift error in residues for each sequence-structure alignment, again as a boxplot versus RMS superposition residual. **C**: Recruitment frequencies for residue sites not forming part of the minimum-size globin core, as a function of the superposition residual at those sites (see text).

racy apply to prediction results for other proteins and other threading methods tested at Asilomar.[8,12] Threading alignment constraints may lead to false negatives by either incorrect core definitions or incorrect gap-penalty assignments, and threading models may fail to reach the expected scores or accuracy for this reason. Two predictions with the present method suffered from this problem, "pbdg" and "ppdk-4".[12] Extents of similarity to database structures are less than 45% in both cases, however, and threading p values remain insignificant even when the correct alignments are allowed, as one would predict from the suggested criteria (not shown). The effect of alignment constraints is not easily ascertained for other methods, and their threading scores were furthermore not expressed in the same manner as chance occurrence p values.[9,10,11,13] For these reasons one can as yet draw no firm conclusion as to consistency with the suggested criteria. I note, however, that the only cases where predictors claim both good fold recognition and accurate alignment are proteins where a large fraction of residues within a chain-continuous domain were shared between the target and database proteins, and where RMS similarity was below 2.5 Å.[9,11,12] A recent control study of alignment accuracy of threading models also supports the conclusion that accuracy depends on the degree of structural similarity, though for this method the quoted accuracies are generally lower, and a critical similarity value of 2.0 Å RMS is suggested.[37]

If fold recognition specificity is indeed limited by the extent of structural similarity, as these experiments suggest, then it seems one should gauge improvements in these methods precisely by their ability to detect folds with less extensive similarity. It seems quite possible that an improved empirical potential, more sensitive to details of packing interactions, might detect the signal of a "core" substructure similarity encompassing somewhat fewer than 60% of residues. Similarly, with an improved representation of the substructure one expects to be conserved in protein evolution, the number of alternative sequence-structure alignments can be reduced and specificity improved by reducing the noise of false positives.[2] The best assay, in either case, is likely to be the ability of a method to detect less extensive similarities. The dependence of threading specificity on the extent of structural similarity is dramatic, since differences of 10% in the fraction of residues mapped to specific sites in a database structure can lead to threading *p* values which differ by two orders of magnitude as was shown in Figure 2C. Even a small improvement in potentials or alignment models may thus have a large effect on the proportion of structural similarities one might recognize by threading, and it seems worthwhile to pursue research along these lines.

These experiments also raise further questions concerning the nature of the structural similarity "seen" by sequence-structure threading. They show, in particular, a large range of variation for both threading scores and model accuracies, even with proteins from the same family, and the criteria they suggest must clearly be seen only as the expected behavior over many trials. These variations may be inherent in application of a statistically derived potential to individual model structures, which realize very few side-chain interactions as compared to the database from which potentials were derived. They may also reflect the relative simplicity of the potential used here, which includes only nonlocal pairwise contacts, and largely reflects the tendency of hydrophobic residues to make many contacts.[22] It is also possible, however, that the conventional coordinate superposition residual used as a measure of structural similarity is not the one most closely associated with threading scores. Side-chain interactions are a short-range phenomenon, and the contact maps of two proteins may be similar even though they show some large-scale deformation such as twist of a β sheet, and a high RMS residual.[38,39] Measures of structural similarity that are tuned to short-range interactions might thus yield a tighter association with threading scores and allow one to characterize the variation inherent in the potential employed, and perhaps in this way differentiate among threading potentials. It seems likely, however, that the general dependence of threading scores on the extent of structural similarity will be observed again, perhaps with changes in detail, as other similarity measures and other threading methods are examined.

ACKNOWLEDGMENTS

I thank my colleagues S. Altschul, J.-F. Gibrat, C. Hogue, T. Madej, A. Neuwald, H. Ohkawa and J. Wilbur for discussions and comments on the manuscript.

Support for this work was provided by the intramural research program of the National Institutes of Health.

REFERENCES

1. Bowie, J.U., Eisenberg, D. Inverted protein structure prediction. Curr. Opin. Struct. Biol. 3:437–444, 1993.
2. Bryant, S.H., Altschul, S.F. Statistics of sequence-structure threading. Curr. Opin. Struct. Biol. 5:236–244, 1995.
3. Johnson, M.S., Overington, J.P., Blundell, T.L. Alignment and searching for common protein folds using a data bank of structural templates. J. Mol. Biol. 231:735–752, 1993.
4. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from three-dimensional structure. J. Mol. Biol. 232:805–825, 1993.
5. Sippl, M.J. Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. 5:229–235, 1995.
6. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3:247–259, 1993.
7. Moult, J., Pedersen, J.T., Judson, R., Fidelis, K. A large scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv, 1995.
8. Lemer, C.M.-R., Rooman, M.J., Wodak, S.J. Protein prediction by threading methods: Evaluation of current techniques. Proteins 23:337–355, 1995.
9. Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M.J. Progress in fold recognition. Proteins 23:376–386, 1995.
10. Godzik, A., Kolinski, A., Skolnick, J. Topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227: 227–238, 1992.
11. Jones, D.T., Miller, R.T., Thornton, J.M. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 23:387–397, 1995.
12. Madej, T., Gibrat, J.-F., Bryant, S.H. Threading a database of protein cores. Proteins 23:356–369, 1995.
13. Matsuo, Y., Nishikawa, K. Assessment of a protein fold recognition method that takes into account four physicochemical properties: Side-chain packing, solvation, hydrogen bonding, and local conformation. Proteins 23:370–375, 1995.
14. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein families and domain superfolds. Nature 372:631–634, 1994.
15. Holm, L., Sander, C. Searching protein structure databases has come of age. Proteins 19:165–173, 1994.
16. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
17. Russell, R.B., Barton, G.J. Structural features can be unconserved in proteins with similar folds: An analysis of side-chain to side-chain contacts, secondary structure and accessibility. J. Mol. Biol. 244:332–350, 1994.
18. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. Protein data bank. In: "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn: International Union of Crystallography, 1987:107–132.
19. Pastore, A., Lesk, A. Comparison of the structures of globins and phycocyanins: Evidence for evolutionary relationship. Proteins 8:133–155, 1990.
20. Holm, L., Sander, C. Structural alignment of globins, phycocyanins and colicin A. FEBS Lett. 315:301–306, 1993.
21. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold: Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
22. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through folding motif. Proteins 16:92–112, 1993.
23. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature 358:86–89, 1992.
24. Bryant, S.H. PKB: A program system and data base for analysis of protein structure. Proteins 5:233–247, 1989.

25. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure." Vol. 5, Suppl. 3. Silver Spring, MD: National Biomedical Research Foundation, 1978:345–352.
26. Altschul, S.F., Boguski, M.S., Gish, W., Wootton, J.C. Issues in searching molecular sequence databases. Nature Genet. 6:119–129, 1994.
27. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
28. Benner, S.A., Cohen, M.A., Gonnet, G.H. Empirical and structural models for insertions and deletions in divergent evolution of proteins. J. Mol. Biol. 229:1065–1082, 1993.
29. Go, N. Theoretical studies of protein folding. Annu. Rev. Biophys. Bioeng. 12:183–210, 1983.
30. Bryant, S.H., Lawrence, C.E. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: A statistical model for nonbonded interactions. Proteins 9:108–119, 1991.
31. Duerring, M., Schmidt, G.B., Huber, R. Isolation, crystallization, crystal structure analysis and refinement of constitutive C-phycocyanin from the chromatically adapting cyanobacterium Fremyella diplosiphon at 1.66 Å resolution. J. Mol. Biol. 217:577–592, 1991.
32. Camardella, L., Caruso, C., D'Avino, R., Di Prisco, G., Rutigliano, B., Tamburrini, M., Fermi, G., Perutz, M.F. Haemoglobin of the antarctic fish Pagothenia bernacchii: Amino acid sequence, oxygen equilibria and crystal structure of its carbonmonoxy derivative. J. Mol. Biol. 224:449–460, 1992.
33. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.
34. Chung, S.Y., Subbiah, S. The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. Protein Sci. 4:2300–2309, 1995.
35. Gallagher, T., Gilliland, G., Wang, L., Bryan, P. The prosegment-subtilisin BPN′ complex: Crystal structure of a specific 'foldase.' Structure 3:907–914, 1995.
36. Bussiere, D.E., Bastia, D., White, S.W. Crystal structure of the replication terminator protein from B. subtilis at 2.6 Å. Cell 80:651–660, 1995.
37. Wilmanns, M., Eisenberg, D. Inverse protein folding by the residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences. Protein Eng. 7:627–639, 1995.
38. Godzik, A., Sander, C. Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng. 2:589–596, 1989.
39. Godzik, A., Skolnick, J., Kolinski, A. Regularities in interaction patterns of globular proteins. Protein Eng. 6:801–810, 1993.