Statistics of sequence–structure threading

Stephen H Bryant and Stephen F Altschul

National Institutes of Health, Bethesda, USA

The past two years have seen the rapid development of new recognition methods for protein structure prediction. These algorithms 'thread' the sequence of one protein through the known structure of another, looking for an alignment that corresponds to an energetically favorable model structure. Because they are based on energy calculation, rather than evolutionary distance, these methods extend the possibility of structure prediction by comparative modeling to a larger class of new sequences, where similarity to known structures is recognizable by no other means. The strength of the evidence they offer should be judged by objective statistical tests, however, so as to rule out the possibility that favorable scores arise from chance factors such as similarity of length, composition, or the consideration of a large number of alternative alignments. Calculation of objective p-values by analytical means is not yet possible, but it would appear that approximate values may be obtained by simulation, as they are in gapped, global sequence alignment. We propose that the results of threading experiments should include Z-scores relative to the composition-corrected score distribution obtained for shuffled and optimally aligned sequences.

Current Opinion in Structural Biology 1995, 5:236–244

Introduction

Today, in the age of genome projects, it would be hard to find a biologist unaware of the importance of methods for automatic sequence comparison. Searches of sequence databases routinely identify molecules homologous to a newly discovered protein, and often allow reliable inference concerning its biological function. Researchers engaged in this work are also well aware of the 'twilight zone' phenomenon: that there exists a range of similarity scores where statistical significance must be examined very carefully. Calculating reliable significance estimates has been a difficult problem in the past, and biologists have often relied on 'rules of thumb', based on experience, to decide if a given score is significant and indicative of evolutionary relationship. This situation has changed dramatically in the past few years, however. For some alignment models accurate p-values may be calculated analytically, and are available as a search is performed. For other alignment models the distribution of scores expected by chance remains less tractable, but in this day of fast computers approximate p-values may be had rapidly by simulations that employ random sequences similar in length, composition, and other variables that affect the score distribution (for reviews, see [1,2•]). In either case one may answer the question, "Are these sequences significantly similar?" with an answer of the form, "The probability that the observed score would be obtained by chance is x or less."

In the past three years a new class of molecular comparison algorithms have appeared based on the idea of 'threading' a sequence through a known three-dimensional structure (for reviews, see [3–10,11•]). These methods offer a means of recognizing similarity in cases where evolutionary relationship is distant, and where the protein 'fold' has been conserved to a greater extent than its sequence [12]. It is also widely believed that natural proteins will fall into a relatively small number of discrete folds [13,14], and that the general problem of predicting protein three-dimensional structure may approach that of fold recognition within the database of known structures. Though new, threading methods already offer some hints of their ultimate success. The structural similarity of actin and heat-shock protein 70 can be recognized, even though sequence similarity is well within the 'twilight zone' [15], and accurate threading alignments have also been reported in cases of low sequence similarity such as globins and phycocyanin, or immunoglobulin domains [16,17,18•,19•]. Several predictions have appeared recently in the literature, which will be tested as the corresponding experiments are done [20–23], and many 'blind' predictions correct to differ-

Abbreviations

p-value—probability value; **Z-score**—number of standard deviations from the mean.

© Current Biology Ltd ISSN 0959-440X

ing degrees were reported at a recent workshop devoted to critical assessment of these new techniques (Meeting on The Critical Assessment of Techniques for Protein Structure Prediction, Asilomar, California, December 1994) [24].

The statistical interpretation of threading scores has, to date, largely followed rules of thumb developed with reference to scores for known true positives (proteins that are structurally similar to each other). It is well known to investigators in this field, however, that other variables such as length, composition and the number of alternative alignments affect the distribution of threading scores one may expect by chance, and that statistical significance must be evaluated critically. It has even been suggested, for example, that the favorable score of the heat-shock protein sequence, when threaded through the actin structure, is not due to recognition of a common fold by the sequence–structure scoring potential, but instead to the chance fact that their sequences

later. Little is known about the score distributions of such alignments. For sufficiently low scores, however, the optimal local alignment of two random sequences will tend to involve only a short segment from each [27]. It is about this scoring regime that much can be said.

The case for which the asymptotic score distribution is fully understood is that of local alignments with gaps disallowed. Briefly, one assumes a probability distribution over a set of letters, two random sequences of lengths m and n of independently sampled letters, and a set of substitution scores with negative expected value. Then the number of distinct segment pairs with a score of at least S is approximately Poisson distributed, with parameter $Kmn\,e^{-\lambda S}$ where K and λ are calculable parameters [28,29]. This implies that the highest score follows an 'extreme value distribution' [30]. The theory has been extended to sequences of Markov dependent letters [31], and to the distribution of the sum of

[redacted]

alignment is thus to align and score a large number of shuffled versions of the original sequences [46,47]. One difficulty with this procedure is that, unless one may assume that the shuffled scores follow a particular known distribution, the smallest p-value that can be rigorously claimed is the reciprocal of the number of shuffled alignments performed. In the case of multiple tests, one may require a very small nominal p-value in order to claim significance, and practical limitations due to available computer time may arise. The simulation method is quite practical in most cases of pairwise comparison, however, where one asks, "What are the odds that the similarity score I see for sequences A and B would arise by chance?" If the alignment score is greater than that for any of a 1000 pairs of shuffled sequences, then the p-value may be estimated as 0.001 or less. P-values calculated in this way may obviously be used to eliminate 'false positives' encountered in a database search, for example those due to unusual amino acid composition. Alignment scores may also be expressed in standard deviation units relative to the distribution for shuffled and optimally aligned sequences, as Z-scores, and used in this way to rank the 'hits' obtained in a database search.

Threading statistics

Statistical effects on threading scores

What distinguishes threading methods from sequence alignment is the matching scores they employ. Rather than the cost of a residue substitution, threading methods consider the energetic cost of placing an amino acid of a given type at a particular site in the structure, with a characteristic structural environment. In place of a table of log-odds scores for residue-residue substitution, threading methods use tables giving the log-odds of a residue type occurring in a given environment, as observed in the database of known three-dimensional structures, or perhaps as estimated by other means [48–50]. The detailed manner in which structural environments are classified differs greatly among current methods. They may be grouped loosely as methods which associate an environment category with individual residue sites [15,48,51–53,54•,55–57,58•,59] or with pairs of sites forming a contact [16,17,18•,19•,49,50,60–63,64•,65•], but there are other differences as well, which we will not attempt to describe here. We note only that the primary component of the threading score is in all cases a sum taken over residue-environment energies, similar in form to the sum of substitution costs used in sequence alignment.

As a result of this similarity in the form of score calculation one may expect the statistical distribution of 'random' scores in threading and sequence comparison to have some similarities. When the expected score for a residue, site pair is positive, as when the alignment space is large, then the expected effect of increasing alignment length is to increase the score. Thus, in the optimal alignment, against two different structures, of a long, randomly shuffled sequence, one may expect the longer alignment to obtain the better score, in rough proportion to the number of residue sites it contains. One may also expect composition effects, in the sense that the mean and variance of the score distribution obtained for random shuffles of an aligned sequence need not be the same between sequences that differ in their amino acid content. This effect is a consequence of using a scoring table derived from a particular database, with a certain composition. The scoring tables are not intended to measure composition preferences, but sequences which differ from the implicit composition model used in their derivation will nonetheless have different expected scores. The effects of local composition bias on sequence comparison scores are well known [2•,66,67]. Threading scores are perhaps more sensitive, as they are strongly affected by overall hydrophobicity of the aligned residues, and sometimes employ potentials where 'composition' must be interpreted to include the interval separation of residue types, and may be quite different among candidate alignments [68].

Threading methods also bear some resemblance to sequence comparison algorithms in the way in which they constrain alignments. Threading is intended to detect remote relationships, where protein evolution is expected to conserve a 'core' substructure consisting of helices and β-strands dispersed throughout the sequence [12]. Threading methods thus consider alignments that are global with respect to the known structure, so that they include most of its core, but gaps are allowed, so that the expected variation in the length and conformation of loop regions will not prevent recognition of the common fold. The techniques by which such alignments are determined differ among current methods. Many employ variations of the dynamic programming algorithms used for sequence alignment, with gap penalties that effectively exclude alignments that do not contain most of the core substructure, or that imply large variation in loop lengths [15–17,19•,48,52,53,54•,55–57,58•,59]. Some methods in this group also penalize gaps at the ends of the aligned sequence [54•]. For methods using gap penalties, the exact choice of penalty is quite important [19•,58•], as in sequence comparison [69,70•], and an additional complication arises for the subset that defines structural environment in terms of pairs of residue sites, where alignment scores are non-local, and heuristic application of dynamic programming may find favorable but not necessarily optimal alignments [16,17]. Another group of threading alignment methods avoids gap penalties altogether. They instead define 'core elements' which correspond to the β-strands and helices of a structure, and consider only alignments that contain no gaps internal to a core element [18•,65•]. By making explicit the assumption that core elements are conserved, these

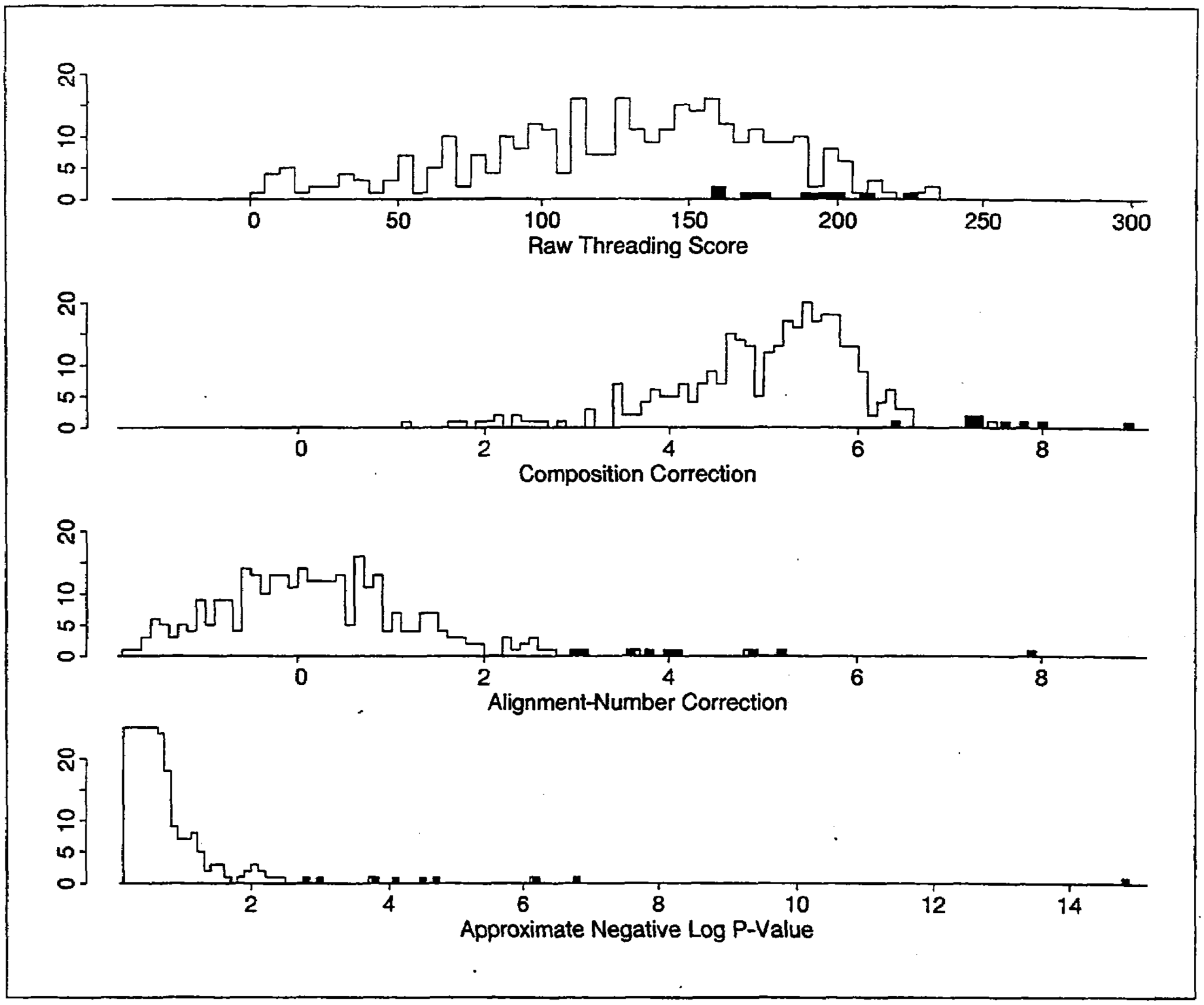


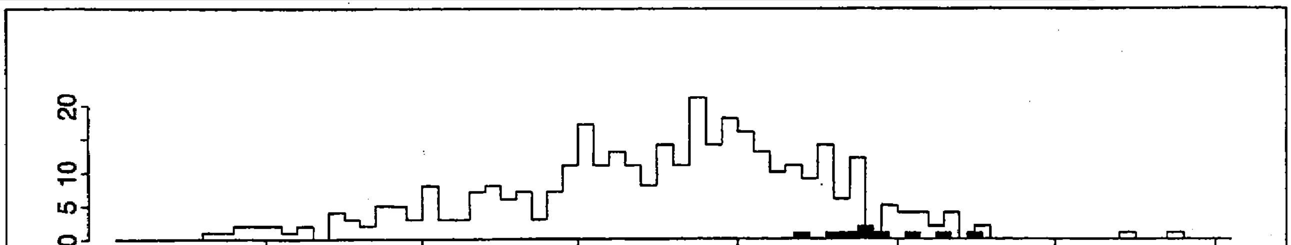
Fig. 1. The sequence of whale myoglobin threaded through the cores of known structures (an example of a 'forward folding' search). Open bars in the histograms indicate non-globin scores, and solid bars the scores for true globins. Bars for globins are staggered by half of the bar width so that no false positives are obscured. The uppermost panel gives the raw threading scores for the most favorable alignment identified for each of 321 cores structures small enough to be threaded by the myoglobin sequence, the quantity defined as $-\Delta G(r|m)$ in [18•]. The

[redacted]

shuffled sequences through a structure, for example, one may expect that the longer sequence will obtain a better score: when more alternatives are considered, one can expect to find a better alignment by chance. A similar effect occurs when comparing threading scores for structures that may be of the same length, but where one structure allows more gaps than another, because of differences in position-dependent gap penalties or the presence of a larger number of core elements. In this case one may expect to find a better score for the structure with more gaps allowed, because the effective number of alternative alignments is greater. This effect is similar

ties, where an alignment with lower gap penalties will always get a better score. For threading methods that employ gap penalties the dependence of the score distribution on relative lengths is similar to that for sequence comparison. In threading a sequence of length N through structures of lengths N and $2N$, for example, one may expect that scores for the latter will be lower, because all alignments must contain gaps of greater aggregate length. These statistical factors may clearly affect the raw threading scores obtained in a database search for structures compatible with a sequence, or sequences compatible with a structure, and calculations of statistical

[redacted]



tive to the distribution of such scores obtained by randomly shuffling the complete sequence involved in each comparison, and optimally aligning it to corresponding structure. We convert these scores to p-values, assuming arbitrarily, and for purposes of illustration, that the distribution of optimal threading scores across randomly shuffled sequences is normal. This procedure illustrates our suggestion as to how score distributions relative to shuffled sequences may be used to control for the statistical effects we mention, and to derive an approximate p-value indicating the odds that the threading score for a pairwise comparison would arise by chance.

The plots show the effect on the rank ordering of the true- and false-positive 'hits' of the successive corrections for composition and numbers of alternative alignments. As noted before [18•,20], correction for the statistical effect of aligned-residue composition dramatically

test offers clear evidence of non-random complementarity of sequence and structure in either the forward or reverse folding experiments. One may conclude that the threading potential is sufficiently sensitive to recognize this complementarity, even among the billions of alternative alignments allowed for each of the shuffled and optimally threaded sequences. It would be desirable, of course, if there were no false positives below some objectively defined level of structural similarity, and it is unclear whether this can be achieved. One may well imagine that some false positives are due to the strict nature of the shuffled-sequence test proposed, in the sense that many all-helical proteins might be expected to fit a globin core better than would a purely random sequence. It is impossible to tell, however, whether false positives are a consequence of the statistical test or of the threading potential, which, after all, examines only local contacts, and might be expected to have some difficulty in dis-

8. Jones D, Thornton J: **Protein fold recognition.** *J Comput Aided Mol Des* 1993, 7:439–456.

28. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, 87:2264–2268.

[redacted]

sets with solvent accessibility patterns of known structures. *Proteins* 1990, 7:275–264.

52. Goldstein R, Luthey-Shculten ZA, Wolynes PG: **Protein tertiary structure recognition using optimized Hamiltonians with local interactions.** *Proc Natl Acad Sci USA* 1993, 89:9029–9033.

53. Overington J, Donnelly D, Johnson MS, Sali A, Blundell T: **Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds.** *Protein Sci* 1992, 1:216–226.

54. Johnson MS, Overington JP, Blundell TL: **Alignment and search-**
• **ing for common protein folds using a data bank of structural templates.** *J Mol Biol* 1993, 231:735–752.
A thorough discussion of dynamic programming alignment models for threading. Describes construction of a database of 'templates' represent-

[redacted]

64. Kocher JPA, Rooman MJ, Wodak SJ: **Factors influencing**
• **the ability of knowledge-based potentials to identify native sequence-structure matches.** *J Mol Biol* 1994, 235:1598–1613.
Describes the contribution to fold-recognition specificity deriving from the different energetic components of threading potentials.

65. Lathrop RH, Smith TF: **A branch and bound algorithm for opti-**
• **mal protein threading with pairwise (contact potential) amino acid interactions.** In *Proceedings of the 27th Hawaii international conference on system sciences*, vol 5. Edited by Hunter L. Los Alamitos: IEEE Computer Society Press; 1994:365–376.
Introduces a novel algorithm for alignment of core elements. Illustrates how the reduced alignment space of this model allows optimal alignments to be found without use of dynamic programming or gap penalties.