Statistics of sequence-structure threading.

Review article

Stephen H Bryant and Stephen F Altschul

Current Opinion in Structural Biology 1995, 5: 236-244.

Outline

- Abstract
- Introduction
- Sequence comparison statistics  
  - Local alignment statistics
  - Global alignment statistics
- Threading statistics  
  - Statistical effects on threading scores
  - Interpretations of threading scores
  - Approximate p-values for threading scores
- Acknowledgements
- References and recommended reading
- Abbreviations
- Copyright

Abstract

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.

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][4][5][6][7][8][9][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][21][22][23], and many 'blind' predictions correct to differing 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
Sequence comparison statistics

Local alignment statistics

Because proteins often contain only isolated regions or domains of similarity, the most widely used algorithms for sequence comparison [24][25][26] employ measures of local similarity. ’Substitution scores’ are assigned to aligned pairs of amino acids, and length-dependent ’gap scores’ to runs of residues inserted or deleted in either sequence; the score for an alignment is simply the sum of these scores. Sequence alignments are considered ’local’ as opposed to ’global’ when only segments of the sequences being compared need be aligned, and these segments may be chosen to optimize the score. If the substitution and gap scores used are too high, the optimal local alignment of two random sequences of roughly equal length will tend to involve virtually the complete sequences [27] as for global alignments, to be discussed 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 $Kmn e^{-S}$ where $K$ and $\lambda$ 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 the $r$ highest segment-pair scores [32]. Similar results are available for the longest run of identical letters in two sequences allowing a specified number of mismatches [33], and weaker ones for a specified proportion of mismatches [34]. Once gaps are permitted, there are no results from which p-values can be calculated for local alignment scores. Nevertheless, assuming that the substitution and gap scores are sufficiently low [27], analogy to the cases just described suggests that the optimal local alignment scores should also follow an extreme value distribution [30]. Computational experiments using real protein sequences or simulated random ones permit the parameters for such a distribution to be estimated [35][36][37][38][39].

A less widely used definition of local sequence similarity involves a ’sliding window’ of fixed length. For any pair of segments with this length, the scores for all aligned residue pairs are summed. Given a random model, the score distribution for a single window location may be calculated explicitly by convolution [40]. Assuming all window positions to be independent introduces only a small error in estimating the maximum score achieved
between two sequences [40].

Global alignment statistics

The scores of global alignments provided the first measure of sequence similarity [41][42][43]. Unfortunately, the statistical distribution of these scores can not as yet be described analytically. The expected score from the alignment of two random sequences of length \( n \) is known in the limit to grow linearly with \( n \), and for certain scoring systems, upper and lower bounds on the constant of proportionality are available [44]. There is no reason to believe that the random score distribution is normal, but for a particular scoring system computational experiments can provide estimates for the mean and standard deviation of random similarity scores [45]. The usual approach to evaluating the significance of a given global 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][49][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][52][53][54][55][56][57][58][59] or with pairs of sites forming a contact [16][17][18][19][49][50][60][61][62][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][16][17][19•][48][52][53][54•][55][56][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 methods reduce the space of alternative alignments to the point that the optimum may be found by enumeration [18•][20] or a branch and bound procedure [65•], or favorable alignments identified heuristically by Monte Carlo sampling ([21]; SH Bryant, unpublished data).

The similarity of the alignment models for threading and sequence comparison also implies similarities in the statistical distribution of matching scores, in particular a common dependence on the number of alternative alignments considered. In threading two randomly 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 to that of sequence alignment with differing gap penalties, 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 significance must clearly take them into account.
Interpretations of threading scores

Threading methods are new, and many investigators have attempted to evaluate their initial results by simply comparing raw scores, perhaps re-scaling them to reflect the range of values obtained in a database search, or the difference in the best and next-best scores [15][16][48][52][53][57][58•][64•]. Fold recognition accuracy in control experiments has been good nonetheless, indicating that threading potentials and associated gap penalties encode much relevant information, although we note that some initial studies considered only matching scores, and not alignment accuracy. In the comparison of raw scores, however, there is no way to allow for statistical effects arising from differences in length, composition, or the effective number of alternative alignments considered. Some investigators have thus proposed corrected scores that take into account one or more of these factors. Johnson and colleagues [54•][55] and Matsuo and Nishikawa [22], for example, divide threading scores by alignment length, to correct in an approximate fashion for the higher scores expected by chance with larger
structures, where more sites may contribute to the sum of residue-environment potentials. Bryant and colleagues \[18•\][20][21] correct for the effect of differences in amino acid composition by determining the distribution of scores expected for random shuffles of the aligned residues, and considering Z-scores relative to this empirical distribution. Sippl and colleagues \[19•\] compare raw scores to the distribution obtained by threading the aligned residues through a 'polyprotein' formed by concatenation of database structures. This measure corrects for the statistical effects of composition differences, and similarly produces a Z-score relative to a set of 'random' models with the same composition. These statistics may be thought of as composition-corrected threading scores, and they appear to improve specificity in a threading search \[18•\][19•].

Some investigators have also attempted to correct for differences in the number of alternative alignments considered in each pairwise comparison of a database search. Bryant and colleagues \[18•\][20] proposed an approximate calculation to correct for differences in the size of the alignment space based on an assumption concerning the effective size of this space, and the form of the expected score distribution. Godzik and colleagues \[17\] and Blundell and colleagues \[54•\] proposed an empirical method to correct for differences in the effective size of the alignment space and for the effect of gap penalties as a function of differences in relative length. They express the raw threading score in standard deviation units relative to the distribution obtained by randomly shuffling the threaded sequence many times, and optimally threading it through the structure. As relative lengths and gap penalties are identical among the shuffled sequences, the threading Z-scores calculated in this way correct for differences in the number of alternative alignments considered and are comparable across the 'hits' obtained in a database search. They are also easy to read. A threading score that is no better than chance has an expected value of zero. A score of 3 is three standard deviation units from the mean value expected for random sequences, and relatively unlikely to arise by chance in a pairwise comparison. Bryant and colleagues \[21\]; SH Bryant, unpublished data) have applied the same idea in the context of an alignment model using core elements, without gap penalties, and in conjunction with adjustment for aligned-residue composition.

**Approximate p-values for threading scores**

We note that these simulations involving shuffled sequences are analogous to the simple 'shuffle and re-align' tests applied for many years in sequence alignment, and that they would appear to be a means to estimate the statistical significance of threading scores. There is certainly some question as to whether the log-odds tables used to calculate threading scores are best interpreted as potentials of mean force in a thermodynamic \[10\] or 'evolutionary' \[71\] ensemble, and as a consequence whether one should consider in a statistical test the distribution of scores for 'random' sequences or 'random' structures. The success of the same potentials in both 'forward' and 'reverse' fold recognition \[17][18•\] argues, however, that both points of view must be at least partially correct, and that threading scores may be interpreted generally as measures of sequence-structure compatibility or fitness \[58•\]. The ready definition of 'randomness' afforded by shuffling a one-dimensional sequence allows one to control in a straightforward way for the statistical effects of composition, length and degrees of freedom in alignment. We therefore suggest that the statistical significance of threading scores should be evaluated by comparison to the distribution obtained for shuffled, optimally aligned sequences, regardless of whether the experiment at hand is "structure seeks sequence" or "sequence seeks structure" \[19•\].

To illustrate this proposal we present two examples, a 'forward' and 'reverse' threading search involving respectively the sequence and structure of myoglobin, a test case employed by many investigators \[15][16][52][62\]. Details are given in the captions to Fig. 1 and Fig. 2. Suffice to say here that the searches are over structures or sequences in a subset of the Protein Data Bank \[72\] that excludes proteins with greater than 30% residue identity and includes nine true positives – globins below this sequence similarity threshold relative to whale myoglobin and to one another. Alignments are global with respect to the core structure of each protein chain, or the core structure of myoglobin as defined in an equivalent, automatic, manner (T Madej, J-F Gibrat,
SH Byrant, unpublished data), and local with respect to the aligned sequence. The scoring potential was calculated with exclusion of globins [18•]. For each search we present raw scores in nominal units and composition-corrected scores in standard deviation units relative to the raw score distribution for shuffles of the aligned residues. We then re-express the composition-corrected score for each sequence-structure pair in standard deviation units relative 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 reduces the number of false positives. Comparison to the composition-corrected score distribution for shuffled and optimally threaded sequences further reduces false positives and yields scores where the expected value for a random sequence is zero. The right tail of this distribution is enriched in true positives, and the false positives in both searches are proteins which are for the most part structurally related. For the 'forward folding' search in Fig. 1, for example, the first false positive, at a threading Z-score of 4.76, is the B chain of phycocyanin, a structurally related protein [73]. For the 'reverse folding' search in Fig. 2, the first false positive at Z = 3.76 is the heme-binding domain of cytochrome p450, an all-helical protein; the fourth at Z = 2.94 is the C-terminal domain of colicin A, a structurally related protein. We note that the greater number of false positives for the reverse folding search is a consequence of the search strategy employed. The threading alignments considered here are global with respect to the structure, but local with respect to the sequence, an "asymmetry" [58•] intended to detect any globin-like domains in larger proteins. Long sequences may thus be aligned with the myoglobin core in more ways than the myoglobin sequence may be aligned with other core structures, which must contain fewer residues. One thus expects to find more false positives as a consequence of the increased alignment space. Sequences which are 'too long' may of course be excluded from consideration in a reverse folding search, by gap penalties or other means, but we have not done so, in order to illustrate the effect of alignment space. Quantifying this effect is one of the issues to be addressed in estimating the numbers of false positives expected in a threading search of an entire database [2•].

The approximate p-values shown in the last panel of Fig. 1 and Fig. 2 may be expected to be accurate only in the region corresponding to the inverse of the number of times the sequence was shuffled and optimally aligned, here around \( p = 0.02 \). It is satisfying to note that the approximate p-values of true positives are uniformly smaller than this value, indicating that the shuffled sequence 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 distinguishing helical proteins of different topology [58•]. We will know the answer to this question only when the same statistical tests are applied with other threading potentials, and when tests that similarly account for the statistical effects of length, composition, and alignment space are compared to the shuffled-sequence test we consider here.
Acknowledgements

We thank the National Institutes of Health Intramural Research Program for support, and T Madej and J Wilbur for many useful discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest.
- of outstanding interest.


A careful review describing differences in the classification of structural environments in the sequence-structure matching potentials found by different investigators.
Return to reference citation 1

[MEDLINE], Return to reference citation 1, 2

Return to reference citation 1

[ MEDLINE], Return to reference citation 1, 2, 3, 4, 5

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6

Introduces a threading model based on pairwise contacts and alignment of core elements without gap penalties. Describes statistical corrections for aligned-residue composition and the number of alternative alignments considered in a database search.
[ MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Describes application of the threading potentials developed by this group to 'forward' fold recognition, using dynamic programming and a one-dimensional representation of structural environments.
Return to reference citation 1, 2, 3, 4, 5, 6, 7

[MEDLINE], Return to reference citation 1, 2, 3, 4, 5


[MEDLINE], Return to reference citation 1, 2, 3, 4


[MEDLINE], Return to reference citation 1, 2


[MEDLINE], Return to reference citation 1


[MEDLINE], Return to reference citation 1, 2


[MEDLINE], Return to reference citation 1


[MEDLINE], Return to reference citation 1


Return to reference citation 1, 2, 3


[MEDLINE], Return to reference citation 1


Return to reference citation 1


Return to reference citation 1, 2


Return to reference citation 1


[MEDLINE], Return to reference citation 1


Return to reference citation 1


47. Altschul SF, Erickson BW: Significance of nucleotide sequence alignments: a method for random
[MEDLINE], Return to reference citation 1.

[MEDLINE], Return to reference citation 1, 2, 3, 4.

[MEDLINE], Return to reference citation 1, 2.

[MEDLINE], Return to reference citation 1, 2.

[MEDLINE], Return to reference citation 1.

[MEDLINE], Return to reference citation 1, 2, 3, 4.

[MEDLINE], Return to reference citation 1, 2, 3.

A thorough discussion of dynamic programming alignment models for threading. Describes construction of a database of 'templates' representing the cores of known structures, for use in 'forward folding' searches.
[MEDLINE], Return to reference citation 1, 2, 3, 4, 5.

[MEDLINE], Return to reference citation 1, 2, 3.

[MEDLINE], Return to reference citation 1, 2.

[MEDLINE], Return to reference citation 1, 2, 3.

Provides a thoughtful discussion of alignment models and scoring schemes for sequence-structure threading.
[MEDLINE], Return to reference citation 1, 2, 3, 4, 5, 6, 7.
   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1, 2

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1, 2

   Return to reference citation 1, 2, 3

   Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1

   [MEDLINE], Return to reference citation 1
Illustrates the sensitivity of dynamic programming alignments to small variations in gap penalties. [MEDLINE], Return to reference citation 1

[MEDLINE], Return to reference citation 1

Return to reference citation 1, 2

73. Holm L, Sander C: **Searching protein structure databases has come of age.** *Proteins* 1994, **19**: 165-173.
[MEDLINE], Return to reference citation 1

---

**Abbreviations**

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

Return to Title

---

**Author Contacts**

SH Bryant and SF Altschul, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA.

Return to author list

---

**Publication Details**

**Publisher:** Current Biology Ltd
London

**Address:** Middlesex House 34-42 Cleveland Street
London
W1P 6LB