Defrosting the Frozen Approximation: PROSPECTOR—A New Approach to Threading

Jeffrey Skolnick* **and Daisuke Kihara**
Laboratory of Computational Genomics, Danforth Plant Science Center, Creve Coeur, Missouri

ABSTRACT **PROSPECTOR (*PRO*tein *S*tructure *P*redictor *E*mploying *C*ombined *T*hreading to *O*ptimize *R*esults) is a new threading approach that uses sequence profiles to generate an initial probe-template alignment and then uses this "partly thawed" alignment in the evaluation of pair interactions. Two types of sequence profiles are used: the close set, composed of sequences in which sequence identity lies between 35% and 90%; and the distant set, composed of sequences with a FASTA E-score less than 10. Thus, a total of four scoring functions are used in a hierarchical method: the close (distant) sequence profiles screen a structural database to provide an initial alignment of the probe sequence in each of the templates. The same database is then screened with a scoring function composed of sequence plus secondary structure plus pair interaction profiles. This combined hierarchical threading method is called PROSPECTOR1. For the original Fischer database, 59 of 68 pairs are correctly identified in the top position. Next, the set of the top 20 scoring sequences (four scoring functions times the top five structures) is used to construct a protein-specific pair potential based on consensus side-chain contacts occurring in 25% of the structures. In subsequent threading iterations, this protein-specific pair potential, when combined in a composite manner, is found to be more sensitive in identifying the correct pairs than when the original statistical potential is used, and it increases the number of recognized structures for the combined scoring functions, termed PROSPECTOR2, to a total of 61 Fischer pairs identified in the top position. Application to a second, smaller Fischer database of 27 probe-template pairs places 18 (17) structures in the top position for PROSPECTOR1 (PROSPECTOR2). Overall, these studies show that the use of pair interactions as assessed by the improved Z-score enhances the specificity of probe-template matches. Thus, when the hierarchy of scoring functions is combined, the ability to identify correct probe-template pairs is significantly enhanced. Finally, a web server has been established for use by the academic community (http://bioinformatics.danforthcenter.org/services/threading.html). Proteins 2001;42:319–331.** © 2000 Wiley-Liss, Inc.

Key words: threading; protein structure prediction; PROSPECTOR; sequence profiles; Fischer database; tertiary contact prediction; protein specific potentials

INTRODUCTION

Sequence-based approaches to functional annotation typically make functional assignments for 40–60% of the ORFS in a given genome. An essential issue in the post-genomic era is to develop methods that can assign the function of the remainder of the ORFS, termed ORFans, about which nothing is known. Because of their widespread use and success, sequence alignment methods such as PSI-BLAST[1,2] and sequence motif (that is, local sequence descriptors) methods such as Prosite,[3] Blocks,[4] Prints,[5,6] and Emotif[7] set the bar for structure-based methods. However, sequence-based approaches increasingly fail as the protein families become more diverse.[8] Thus, an extension of these approaches, which combines one-dimensional information about sequence and structure, has been developed, with some success reported.[9] An alternative structure-based approach to function prediction that uses the sequence-structure-function paradigm has been introduced recently.[8,10–15] Here, models predicted by threading are screened for matches to known active sites; if a match is found, then a functional assignment is made. One key to the success of this approach is to use the best threading algorithm possible so that more distant cases can be recognized. In this regard, the recent CASP3 results have shown both the strengths and weaknesses of contemporary threading algorithms. Based on insights obtained from CASP3, our goal is to develop an improved algorithm that makes better use of pair interactions in particular and to demonstrate its efficacy on a number of standard benchmarks, especially the Fischer databases.[16]

At this juncture, it is important to review the current status of various threading approaches so that, in the development of a new threading algorithm, we can exploit their advantages while avoiding their pitfalls. All threading algorithms are essentially defined by three choices. First, the nature of the interactions and the functional form of the "energy" must be selected. For scoring functions composed of a variety of terms, their relative weight must be established. The type of energy terms considered in the past has included the local burial status of residues, secondary structure propensities or predicted secondary structure as well as additional energy penalty terms[16,17] (for example, terms that compensate for different protein

Grant sponsor: National Institutes of Health; Grant number: GM-48835.

*Correspondence to: Jeffrey Skolnick, Laboratory of Computational Genomics, Danforth Plant Science Center, 893 North Warson Rd., Creve Coeur, MO 63141. E-mail: skolnick@danforthcenter.org

Received 23 March 2000; Accepted 14 September 2000

lengths), and the inclusion of pair or higher-order interactions. Contemporary algorithms often include an essential term that is related to the sequence identity between the template and the probe sequence.[18] This evolutionary component is designed to improve both the template protein recognition ability and the quality of the predicted structural alignment.[16,19–22] Second, if pair interactions are included, then the type of interaction centers must be selected. Commonly used choices are the Cαs,[23,24] the Cβs,[25,26] the side-chain centers of mass, specially defined interaction centers,[27,28] or any side-chain atom.[29] The functional form of the pair energy ranges from contact potentials[29,30] to continuous distance-dependent potentials[25,31] to interaction environments.[32] Third, given an energy function, a search procedure that finds the optimal alignment between the probe sequence and each structural template must be used. When all of the interactions are local in nature (for example, a fitness score defined by mutation matrices and secondary structure propensities), then dynamic programming[33] is the best choice. If a nonlocal scoring function is used (pair interactions, for example), then the key question is how the interactions are updated in the template structure to reflect the probe sequence. Some approaches use dynamic programming with the "frozen" approximation (where the interaction partners or a set of local environmental preferences are taken from the template protein in the first threading pass).[29,34] This might be followed by iterative updating.[29,32,35] Still other workers use double dynamic programming, which updates some interactions recognized as being the most important in the first pass of the dynamic programming algorithm.[25] Other variants evaluate the nonlocal scoring function directly and search for the optimal probe-template alignment by Monte Carlo[27] or branch-and-bound search strategies.[28]

It should be recognized that almost no search protocols allow the actual template structure to adjust in order to reflect the actual structural modifications in the probe structure relative to that of the template. Algorithms such as Monte Carlo and branch-and-bound strategies permit the partner from the probe sequence found in the current alignment to be used, but they do not allow the template's backbone structure to dynamically readjust to reflect the probe sequence. Such readjustments might be quite important when the probe and template structure differ substantially, for example, when a template protein's GLY is replaced by the probe's TRP. Unfortunately, this is precisely the realm in which threading would be expected to be the most valuable as compared with pure sequence-based methods.

In principle, the advantage of threading over pure sequence-based approaches is that it uses structural rather than evolutionary information. However, as evidenced by CASP3, many of the successful fold-recognition approaches are pseudo one-dimensional in nature and use evolutionary information (typically implemented in the form of sequence profiles) plus predicted secondary structure. Furthermore, the evolutionary component contributes a significant fraction of the selectivity.[36] Of the top performing groups in fold recognition in CASP3, this type of approach was typified by Jones et al.[37] and Koretke et al.[22] Here, structure (in this case secondary structure) played an ancillary role. Ota et al.[38] also used a hierarchy of local scoring functions to describe side-chain packing, hydration, secondary structure, and hydrogen bonding.

Moving to approaches in which structure played a more prominent role in CASP3, Domingues et al.[39] used a burial energy and the frozen approximation to evaluate pair interactions. However, they used a single sequence rather than sequence profiles; this represents a more structure-based approach to threading, but all interactions are still implemented at the pseudo one-dimensional level to enable the use of dynamic programming. Panchenko et al.[40] was unique among the predictors in CASP3 in that they explicitly treated interactions in a structural core identified on the basis of evolutionary conservation of the structure across a protein family. In some sense, this approach is closest to the original idea of threading; yet they too use a PSI-BLAST sequence-profile component and conclude that the combination of both sequence profiles and contact potentials improves the success rate over that when either term is used alone. Because they use a nonlocal scoring function, dynamic programming cannot be used to search for the best match of a sequence to a given structure. Rather, a Monte Carlo search procedure is used to search for the best sequence-structure fitness. Such calculations take a considerable amount of computer time; therefore, application of the method on a genomic scale would require considerable computer resources. Further, for the identification of the core, a number of structures in the protein family must be solved. Overall, the general consensus is that progress was made in CASP3, with improvement in alignment quality since CASP2.[36,41,42] But, as Murzin[36] observed, threading "performs better on distant homology recognition targets than on 'pure' folding recognition targets. This bias probably resulted from the implementation of 'distant homology' filters."

Thus, techniques that extend the ability of threading techniques to address "pure" fold recognition situations are still required. But, as indicated in the work of Panchenko et al.,[18] the best results seem to occur when a sequence-profile term is combined with threading potentials. We proceed in this spirit by presenting the PROSPECTOR method (*PRO*tein *S*tructure *P*redictor *E*mploying *C*ombined *T*hreading to *O*ptimize *R*esults), which shows that pair interactions can improve the sequence-structure specificity over that of sequence-profile terms used alone. However, when multiple scoring functions are combined, the resulting recognition ability is even larger. The organization of the presentation of this methodology in this article is as follows. In the Materials and Methods section, we describe the approach, the scoring functions, the way pair interactions are updated, a methodology for using consensus contacts in threaded structures to construct a protein-specific pair potential that is used in a subsequent threading iteration, and a new means of assessing the quality of the predicted structures based on the significance of the predicted contacts. Then, in the Results section,

we present results for the 68 probe-template pairs of the Fischer database, and a second database of 27 probe-template pairs compiled by Fischer (http://www.doe-mbi.ucla.edu/peopleFischer/BENCH/tablepairs2.html). Finally, in the Discussion section, we summarize the results of the present work and highlight future research directions.

MATERIALS AND METHODS

Background

During the course of developing PROSPECTOR, we noticed that the sequence profiles generated from the BLOSUM 62 matrix[43] often provided quite reasonable alignments between the probe and template, even when the alignment score itself was insignificant (see also Table IIIA). This suggested that the first stage of a hierarchical threading approach should use a sequence profile[44–46] (using a sequence profile plus a three-state secondary structure prediction scheme gave worse results) to generate the initial alignment between the probe sequence and the template structure. We call this the "partly thawed" approximation because the resulting alignment of the probe sequence in the template structure is used to calculate the partners for the evaluation of the pair interactions. That is, in all cases, the probe sequence itself is used to evaluate the pair interactions.

Previously, in the first iteration of the frozen approximation,[29] the partners were taken from the template structure. In practice, this worked well when the probe and template structures had similar environments, but more often than not the environments were quite different. For example, the probe sequence might be entirely devoid of any TRP, but in the frozen approximation, a given residue might be forced to interact with a TRP from the template. On successive iterations, in the so-called defrosted approximation where the partners were taken from the previous alignment,[29] there were times when the resulting alignments never converged. This resulted from the poor environment provided by the initial frozen approximation that selected the partners from the template.

A schematic overview of the entire threading approach is shown in Figure 1. All alignments are generated using dynamic programming. In the upper half of Figure 1, we present PROSPECTOR1. It is a hierarchical approach consisting of close (distant) sequence profiles that, for each structure, generate the probe-template alignment to be used in the evaluation of the pair interactions in the second pass. A total of 20 structures (the four scoring functions times five structures for each scoring function) are reported, as are the consensus predictions. In PROSPECTOR2, we pool these structures and select consensus contacts in the set of these 20 best structures. Using a recently developed formalism,[47] we then convert these consensus contacts into a protein-specific pair potential. Again, using the sequence-based profile to generate alignments, we use these to evaluate the pair interactions in the second cascade of the threading algorithm. In what follows, we describe how each of these terms is derived.

Generation of Sequence Profiles

A sequence database combining Swissprot (http://www.expasy.ch/sprot/) and the genome sequence database (ftp://kegg.genome.ad.jp/genomes/genes)[48] is used for selecting sequences. First, we use FASTA[49,50] to select those sequences whose sequence identity lies between 35% and 90% of the probe sequence. Then, multiple sequence alignments are generated by using CLUSTALW.[51] We term this the "close" set of alignments. The sequence profile for the ith position in the probe sequence for amino acid type γ is

$$P^{\text{close}}(\gamma, i) = \frac{\sum_{l=1}^{N_{\text{close}}} B(\gamma, a_{il})}{N_{\text{close}}}. \quad (1a)$$

Here, N_{close} is the number of sequences that are aligned in the "close" alignment, $B(\gamma, \eta)$ is the BLOSUM 62[52] mutation matrix for residues type γ and η, and a_{il} is the amino acid at position i in the lth sequence.

To this set, we add additional sequences whose E-value in FASTA is less than 10, and we generate a profile[53] for these distantly related sequences; these are termed the "distant" set of alignments.

$$P^{\text{dist}}(\gamma, i) = \frac{\sum_{l=1}^{N_{\text{dist}}} B(\gamma, a_{il})}{N_{\text{dist}}}. \quad (1b)$$

Here, N_{dist} represents the "distant" sequences that are aligned. The goal is to have two sequence profiles: one that is more sensitive to more closely related sequences and another that can sometimes detect more distantly related sequences. Note that gaps are assigned a value of $B = 0$, but are counted in the averaging process. If a region has a large number of gaps, then its contribution to the alignment is diminished relative to a gap-free region, where $B > 0$, i.e., favorable mutations have occurred.

First-Pass Sequence-Profile Score Matrix

The score matrix for the first pass through the structural database associated with aligning residue i with the Jth residue in the Kth structure is

$$\Xi_K^{\varphi,1}(i, J) = P^{\varphi}(a_{JK}, i) \quad (2a)$$

where a_{JK} is the residue at position J in the Kth structure. Here, we use the shorthand notation

$$\varphi = \left\{ \begin{array}{c} \text{close} \\ \text{dist} \end{array} \right\} \quad (2b)$$

which refers to the close or distant set of multiple sequence alignments.

Secondary Structure Propensities and Pair Interactions

In the next stage of the alignment process, we consider secondary structure propensities and pair interaction terms. For the secondary structure propensities, we con-

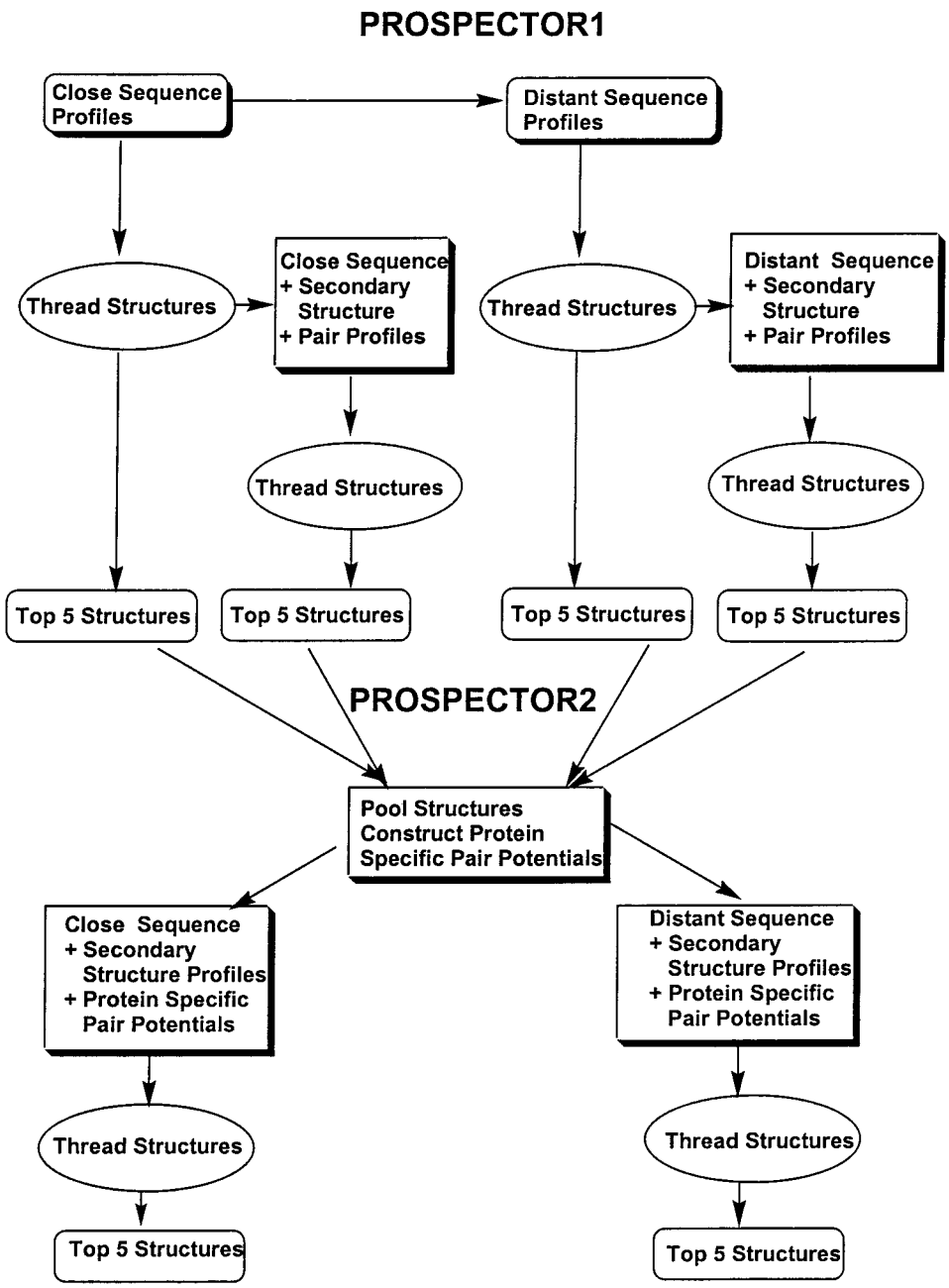

Fig. 1. Overview of PROSPECTOR1 and PROSPECTOR2 threading routes. The protocol for PROSPECTOR1 is the following. First, close and distant sequence profiles are generated. Then, each of these sequence profiles is used to scan a structural database. The probe-template alignments provided by the sequence profile scoring function are used to identify the partners in the probe sequence for use in the next threading iteration that uses sequence plus secondary structure plus pair interaction profiles. The five top-scoring structures for each scoring scheme are collected and composite results are reported. For PROSPECTOR2, the consensus contacts (occurring in structures with Z-scores greater than 1.2 and are found at least three times) in this set of the 20 top-scoring structures provided by PROSPECTOR1 are then used to construct a protein-specific pair potential that is used in a subsequent iteration of threading, again based on close and distant sequence profiles.

sider the homology averaged secondary structure profile energy defined by

$$S^{\varphi}(\Theta_1, \Theta_2, i) = \frac{-\sum_{l=1}^{N_{\varphi}} \varepsilon(\Theta_1, \Theta_2, a_{il}, a_{i+1,l})}{N_{\varphi}} \quad (3)$$

where $\varepsilon(\Theta_1, \Theta_2, a_{il}, a_{i+1,l})$ is the energy of a consecutive pair of amino acids a_{il}, $a_{i+1,l}$ in consecutive secondary structure environments Θ_1 and Θ_2, respectively, which can be helix, beta, or turn. This term is newly derived, but it is related to a six-state conformational descriptor that was developed previously.[54] Here we consider three, rather than six, conformational states (which is a finer-grained

description) and use multiple-sequence averaging rather than just considering a single sequence.

The next step is to use the alignment provided by the sequence-only scoring profile to generate the partners in the evaluation of the pair potentials. First, we consider the homology averaged pair interaction matrix defined by

$$E^{\varphi}(i,j) = \frac{-\sum_{l=1}^{N_{\varphi}} \varepsilon(a_{il}, a_{jl})}{N_{\varphi}} \quad (4)$$

where $\varepsilon(\gamma, \eta)$ is the arithmetic average of quasi-chemical pair potentials which describes interactions between side chains of amino acid types γ, η that are in contact (that is, have one pair of heavy atoms within 4.5 Å of each other). This protein-specific pair potential was derived previously by using weak local sequence fragment similarity[47] and φ is defined as in Eq. (2b). The minus sign arises because we want to maximize the score, that is, gap penalties are negative.

The results described below are insensitive to the choice of the contact definition cutoff, provided that atoms in the first solvation shell are considered. The effect of the use of alternative, distance-dependent potentials has not yet been explored, but will be examined in future work.

Second-Pass Scoring Matrix Using the Partially Thawed Approximation

Let $M_{1K}^{\phi}(J)$ be the alignment between the Jth residue in the Kth structure and the probe sequence generated by the φth sequence profile after the first iteration. There is one of two possible values for $M_{1K}^{\phi}(J)$: either there is a gap in the probe sequence that aligns to the Jth position in the Kth structure or the probe sequence position aligns the Jth position.

We can now construct the score matrix, $\Xi_K^{\varphi,2}(i, J)$ associated with aligning the ith probe position with the Jth position in the Kth structure:

$$\Xi_K^{\varphi,2}(i,J) = \lambda_1 \Xi_K^{\varphi,1}(i,J) + \lambda_2 S^{\varphi}(s_j, s_{j+1}, i) + \lambda_3 \sum_{m=1}^{nc_K(J)} E^{\varphi}\{i, M_{1K}^{\phi}[C_{JK}(m)]\}. \quad (5)$$

Here, $\Xi_K^{\varphi,1}(i, J)$ is given by Eq. (2a) or (2b) depending on which sequence profile is used, the $\{\lambda_\omega\}$ are the weight factors of the various scoring functions (taken on optimization to be 1, 5, and 5 respectively, as this set of parameters gave the best results on the 68 pairs of proteins). Here, s_J and s_{J+1} are the conformations of residues J and $J + 1$ in the Kth structure. $nc_K(J)$ is the number of contacts the Jth residue makes in structure K, $C_{JK}(m)$ is the identity of the mth contact partner that residue J makes in structure K, $M_{1K}^{\phi}[C_{JK}(m)]$ is the alignment to the corresponding position in the probe sequence associated with residue $C_{JK}(m)$ that was generated using the first pass, and the sequence-profile score matrix is given by either Eq. (2a) or (2b) depending on the sequence profile that is used. As before, dynamic programming is used to generate the alignments in the second pass.

TABLE I. Compilation of Gap Penalties for the Four Scoring Functions Used in PROSPECTOR1 and PROSPECTOR2†

Method	Gap opening penalty	Gap propagation penalty
"Close" sequence-profile	−6.0	−0.8
"Close" sequence-profile plus secondary structure plus pair profile	−12.0 (−10.0)	−1.0 (−0.8)
"Distant" sequence-profile	−4.0	−1.2
"Distant" sequence-profile plus secondary structure plus pair profile	−8.0 (−6.0)	−1.2 (−1.0)

†The numbers in parentheses refer to those cases in PROSPECTOR2 that differ from PROSPECTOR1.

We allow for the possibility of different gap opening and gap propagation penalties. Table I shows a summary of the set of values of the gap penalties optimized to select the maximum number of correct structures as compared with the Fischer database for each of the four scoring functions. In this optimization procedure, gap insertion penalties were allowed to assume all even integer values from −2 to −12, and gap propagation penalties were scanned from −0.1 to −1.2. The set of gaps used in PROSPECTOR2 may also be found in Table I. Interestingly, when secondary structure propensities and pair interactions are considered, the gap-opening penalties are larger for the close profile cases than the distant profile cases. This reflects the fact that when a distant sequence profile is used, the gap penalties may have to be smaller to allow the favorable alignment regions to be found.

For each of the four scoring functions, we report the top five scoring structures, for a total of 20 structures (four scoring functions times the best five structures for each scoring function). Alignments for each of these probe-template assignments are also generated.

Generation of Protein-Specific Pair Potentials from Threading

In the twilight zone of score significance where a probe cannot be assigned to a specific template, there may be fragments of template structures that display some relationship to the native structure of the probe sequence. If this were true, then at least some of the selected structures should have consistent substructure features such as side-chain contacts. By trial and error on a set of protein structures previously used in an earlier ab initio folding study,[55] the following criteria were found to work best. As described in detail elsewhere, on average about 35–40% of the predicted contacts are correct if they belong to pairs of residues that are at least four residues apart and occur in at least 25% of the top scoring structures having a Z-score greater than 1.3.

The question then remains as to how to incorporate such information into the subsequent iteration of the threading

algorithm. Because the predicted contacts are inexact, rigorously demanding that they all be satisfied would lead to spurious results. Rather, such contacts could be converted into a pseudo potential that reflects a bias toward such contacts, and we could then repeat the second pass of the threading procedure using the newly derived, now protein-specific, pair potential. Because PROSPECTOR2 uses consensus information, which includes a significant number of correct or near-correct contacts, it should be more specific than when such information is absent, as in PROSPECTOR1. As in PROSPECTOR1, the alignment that assigns the partners is provided by either the close or distant sequence profile. Then, we use dynamic programming to evaluate the probe-template fitness with an energy function that now includes the modified pair potentials. We call the threading method that uses these terms PROSPECTOR2. A schematic overview of PROSPECTOR2 is given in the lower half of Figure 1.

The protein-specific pair potential is constructed as follows. If there are more than three contacts predicted between residues i and j, whose total number is q_{ij}, then we calculate the pair potential for these positions as

$$V(i,j) = -\ln\left(\frac{q_{ij}}{q_{ij}^0}\right) \tag{6a}$$

where the expected number of contacts, q_{ij}^0 is given by

$$q_{ij}^0 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}}{n^2} \tag{6b}$$

and n is the number of residues in the probe sequence. For those pairs of positions in which no consensus contacts are found, we simply use the profile-based pair potential of Eq. (4). We then use the arithmetic average of this potential and our profile-based pair potential given by Eq. (4) in the second threading pass for the close and distant cases of protein-specific pair potentials.

Analysis of the Structural Predictions

We have argued that the use of pair potentials improves the fold specificity. There are a few ways to test this hypothesis. One is to examine the mean Z-score of the correctly identified template structure as a function of the various potentials used, where the Z-score for the Kth structure having energy E_K is given by

$$Z_K = \frac{(E_K - \langle E \rangle)}{\sigma} \tag{7}$$

with $\langle E \rangle$ and σ being the mean and standard deviation values of the energy of the probe in all templates of the structural database. This is one measure of the utility of a particular scoring function. Because we do not randomize the sequence in the evaluation of Eq. (7), our reported Z-scores will be lower than when this is done. However, sequence randomization is a computationally expensive process, and it would be a significant advantage to be able to avoid it, especially when threading is done on a genomic scale.

Another area to be investigated is the accuracy of the predicted structure. One means of assessing accuracy is to examine the predicted side chain contact maps. Among the quantities that we report are f_c, the fraction, and N_c, the number of correctly predicted contacts. Again, we need some measure of the significance of these quantities. One such measure arises by generating random alignments of the probe sequence in the correct template structure. However, this does not necessarily indicate how significant the contact map prediction is. Consider the case in which one has a library of homologous structures and predicts 95% of the contact map correctly. By randomizing the contact map, one would conclude that this is a highly significant prediction. However, one could just as easily have selected the structure at random. Here, the specificity of the prediction is in fact close to zero. In general, if one focuses on a single structure, relative to the library of all structures, one has no idea of the significance of the value of N_c. To address this issue, we suggest the following metric: for the entire structural template library, let us calculate the average number, N^0, of correctly predicted contacts for the best probe-template alignments of the probe sequence in all template structures as well as the standard deviation of this quantity, σ^0. Then, we report the Z-score for the number of correctly predicted contacts for the correct probe-template pair as

$$Z_{\text{con}} = \frac{(N_c - N^0)}{\sigma^0}. \tag{8}$$

This quantity more appropriately measures the significance of a given number of predicted contacts.

RESULTS

Application to the Original Fischer Benchmark

We focused on the Fischer database.[56] Composed of 301 template structures and 68 probe sequences, the Fischer database represents a standard benchmark in the threading field. We tried a variety of approaches on this database before deciding on the aforementioned combination of parameters described in Materials and Methods. We summarize the results of these earlier studies below.

For a given scoring function, the Needleman-Wunsch global alignment algorithm[33] recognized more correct probe-template pairs on average than did the Smith-Waterman local alignment algorithm.[57] We also tried using the secondary structure profiles alone as the initial step in generating the probe-template alignment for pair evaluation. Secondary structure profiles alone only recognize 18 cases in the first position, whereas secondary structure profiles plus pair profiles recognize 29 cases. This clear improvement shows the utility of pair potentials in this approach. Nevertheless, 29 correctly recognized pairs represent rather poor performance. The major improvement in fold recognition is achieved, as others have observed, when sequence profiles are used.[18] If we use the sequence-profile-based alignment, but ignore the sequence-profile term in the calculation of the energy [i.e., using Eqs.

TABLE II. Summary of Threading Results Using Different Scoring Functions for the Fischer Database[†]

Method	Number of Fischer pairs in the first position	Number of Fischer pairs in the top 5 (4) positions	Number of Fischer pairs in the top 10 (8) positions	Mean Z-score of correctly predicted pairs
PROSPECTOR1				
"Close" sequence-profile	44	46 (46)	49 (47)	2.65
"Close" sequence-profile plus secondary structure plus pair profile	45	55 (53)	56 (55)	3.32
"Distant" sequence-profile	46	53 (51)	53 (53)	2.5
"Distant" sequence-profile plus secondary structure plus pair profile	52	56 (56)	59 (57)	3.06
Hierarchy of four scoring methods	59	63 (62)	65 (63)	
Hierarchy of three scoring functions (as above but without the "distant" sequence-profiles)	58	62	64	
PROSPECTOR2				
"Close" PROSPECTOR2 sequence-profile plus protein-specific pair and secondary structure potentials profile	48	51 (51)	58 (58)	3.95
"Distant" sequence-profile plus protein-specific pair and secondary structure potentials	51	59 (59)	59 (59)	3.84
Hierarchy of four scoring methods	61	64 (64)	65 (65)	
Hierarchy of three scoring functions (as above but without the "distant" sequence-profiles)	60	64	65	
Other methods				
Simple Blast[1]	27	—	—	
PSI-BLAST restricted to the Fischer database[46,59]	24	37 (36)	40 (39)	
PSI-BLAST using extensive sequence database and PSSM constructured using IMPALA[60]	41	46 (46)	47 (46)	
Original GKS threading program[29]	22	30	34	
Hybrid threading[58]	52	57	60	
Best UCLA benchmark results as of 2/4/00 which is prediction of secondary structure plus multi-gonnet[16]	52	(56)	(58)	

[†]Results are reported in both the top 5 (4) and top 10 (8) positions,[58] with the number in parenthesis given by the UCLA benchmark website (http://www.doe-mbi.ucla.edu/people/fischer/BENCH/table1.html).

(3) and (4)], then 34 probe-sequence-template structure pairs are matched as the top score. This is to be contrasted to 52 cases (see Table II) that are correctly assigned when the entire sequence plus secondary structure plus pair profiles are used. In all cases, when combined in a hierarchical manner, we have found that inclusion of pair interactions improves the yield of correct probe-template matches.

In the top part of Table II, we summarize our results by using PROSPECTOR1 and its hierarchy of four scoring functions. Note that the "distant" sequence profile recognizes a somewhat greater number of correct pairs (46 pairs or 67%) than does the "close" profile (44 pairs or 65%). This is very interesting in that it shows that these distant profiles contain additional information that can be profitably used to increase the recognition abilities of these threading algorithms. However, the best single scoring function is the combined distant sequence profile plus secondary structure plus pair interaction scoring function that recognizes 52 (72%) cases in the top position. In itself, this single scoring function is a competitive threading algorithm (see below). This is an improvement of six correctly matched structures relative to the best (distant) sequence profile case. Further, it recognizes the most proteins in the top four, five, eight, and 10 positions. The performance of the close sequence-profile plus secondary structure plus pair interaction scoring function is also quite good. It recognizes more top-scoring proteins than the close sequence-profile case alone (45 versus 44), and also recognizes considerably more proteins in the top four, five, eight, and 10 positions, for example, 55 versus 46 proteins in the top five positions. Clearly, the best performance is when all four scoring functions are combined. Then 59, 63, and 65 proteins are recognized in the top, top five, and top 10 positions, respectively.

Another means of assessing the utility of a given scoring function is to measure the mean Z-score of the correctly

identified proteins. The close sequence profile plus secondary structure plus pair interaction scoring function has a mean Z-score of 3.32 that is the best of all scoring functions and is significantly better than the close sequence profile, which has a mean Z-score of 2.65. Note that the distant profile recognizes 46 proteins in the top position and has a marginally poorer mean Z-score of 2.50 as compared with the close profile value of 2.65. For both close and distant cases, the use of pair interactions plus secondary structure propensities increases the sequence-structure specificity relative to the use of a sequence profile alone, with the mean Z-score of the distant case of 3.06. In other words, the use of structural information confers an advantage over the cases in which pure evolutionary information is used, both in terms of the number of proteins placed in the top position as well as in the sequence-structure specificity as assessed by the Z-score.

One of the best alternative methods reported on the UCLA website as of August 1, 2000 (http://www.doembi.ucla.edu/people/fischer/BENCH/table1.html) is that of Gonnet (which is a pairwise sequence-alignment method that also uses predicted secondary structure); it recognizes 52 proteins in the top position. This is the same number that the combined distant sequence profile plus secondary structure plus pair interaction scoring function recognizes. We also recognize the same number of proteins in the top four positions (56) and one less protein in the top eight (57 versus 58).

If any method, in particular a hierarchical method such as PROSPECTOR1, is considered, then ours is clearly the best, as 59 proteins are recognized in the top position, with a total of 65 pairs recognized in the top 10 positions. It is clearly superior to all of our early efforts in threading as well as to the hybrid method,[58] BLAST,[1] and PSI-BLAST.[46,59] In particular, for PSI-BLAST, we report two sets of results. The first is when only sequences from the Fischer database are used to generate the profiles, and the second is when an extended version of the same sequence database that we use to generate the sequence profiles is used. For the FISCHER-only database, only 24 probe-template pairs are correctly identified in the first position. Next, we use a larger sequence database (consisting of all the sequences in Swiss Prot, the genome sequence database from KEGG, and the trEMBL database (http://expasy.proteome.org.au/sprot/) to generate position-specific score matrices, PSSM, using the IMPALA package with default settings.[60] Now, 41 cases are assigned to the top position. Note that this performance is worse than when either the close or distant sequence profiles are used alone. With respect to the top five positions, the close (distant) profile places 46 (53) and 49 (53) in the top five and 10 positions. In contrast, PSI-BLAST places 46 and 47 proteins in the top five and 10 positions, respectively.

It might be argued that, because we use four scoring functions and the hybrid threading method[58] only uses three, this is not a fair comparison. If we eliminate those results obtained from the "distant" sequence profiles, then we obtain 58, 62, and 64 cases in the top one, five, and 10 positions, respectively. Thus, with respect to this test, PROSPECTOR1 is certainly a very competitive algorithm.

In Table IIIA,[5] we further analyze the distant sequence-profile scoring function. Here we show that the Z-score for the number of correctly predicted side-chain contacts, Z_{con} (see column 6) is, in general, significantly better that one would expect from random; indeed, it has a mean value of 7.57. At first glance, it might be argued that this is simply an artifact in that the sequence profile generates a good probe-template match based on the score significance, and because the two structures are similar, this result is trivial. A number of the correctly ranked structures have a rather poor energy Z-score, yet their contact prediction is highly significant, e.g., (1bbt1 in 2plv1 has a Z-score of 1.64 and Z_{con} is 13.6). Furthermore, some of the probe-template pairs that do not lie near the top scores can also have a significant Z_{con}. For example, the score of 1ten_ in 3hhrB is at position 127, yet Z_{con} is 4.6. Note that 4 of 11 of the poorly ranked structures ($1 < \text{rank} < 16$) have a Z_{con} greater than three, which is much better than one might guess based on the rank of the correct template structure. Of course, there are some cases that are much worse than random as well. This substantiates our earlier observation that a sequence profile can often generate a reasonable set of correct contacts (on average 25% correct) even when the score of the alignment is not significant. Of course, because there is a substantial fraction of incorrect contacts as well, the pair-potential contribution cannot be made too large because these incorrect contributions could dominate the score.

In Table IIIB, we present the results for the distant sequence profile plus secondary structure plus pair profile scoring function. As would be expected, compared with the distant profile case, the mean Z-score (now over all, not just correctly predicted pairs) has increased from 1.95 to 2.59. Now, 29% of the contacts are, on average, correct, and the mean Z-score of correctly predicted contacts has increased from 7.57 to 8.39. The ranking of 16 probe-template pairs improve and six cases get worse. Of the six cases that have a worse ranking, all have Z-scores less than 1.6, a range in which 16 of 32 cases are correctly assigned. Furthermore, the misassignment of 1omf_ (a membrane protein), by a pair potential derived for water-soluble proteins is understandable. In another two cases (3hlaB and 1sacA), the rank moves from first to second. For 3hlaB, the best scoring fold, 3cd4_, has its best structure alignment with a root-mean-square deviation of 2.52 Å on the Cαs over a slightly smaller part of the structure as compared with the best structural alignment of 2rhe_ of 2.56 Å over a slightly longer piece of structure. For 1sacA, the correct fold is 2ayh_ and the misassignment, 8fabA, are all β barrels, with the latter having a significant structural superposition over roughly half of the 1sacA native structure. For 3rubL the rank of 6xia_ moves from third to 21st, with 1ipd_ rated as the best match. The best structural superposition of 1ipd_ and 3rubL is 2.78 Å whereas that of 6xia and 3rubL is 2.52 Å, over about two-thirds of the structure. Finally, for 1tahA, the first- to second-pass ranks move from 88th to 20th.

Turning now to PROSPECTOR2, and using the formalism of Eqs. (6a) and (6b), we derive a set of protein-specific

TABLE IIIA. Compilation of Results on the Fischer Benchmark for the Distant Sequence-Profile Scoring Function

Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}	Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}
2mnr_	4enl_	1	5.95	118	0.09	6.48	22.50	1hip_	2hipA	1	2.03	104	0.57	21.26	7.94
1tahA	1tca_	88	0.52	80	0.14	4.34	23.53	1arb_	4ptp_	1	2.85	22	0.04	0.96	13.90
1ltsD	1bovA	40	0.82	6	0.04	−1.07	12.98	1atnA	1atr_	1	1.69	178	0.21	9.35	26.14
1mdc_	1ifc_	1	2.78	0	—	−1.91	12.92	2sarA	9rnt_	4	1.01	26	0.18	2.98	10.01
3chy_	4fxn_	60	0.75	46	0.16	2.79	18.80	1sacA	2ayh_	1	1.32	32	0.12	1.37	19.05
2sga_	4ptp_	1	1.34	72	0.19	13.06	6.80	1hom_	1lfb_	1	1.37	70	0.55	8.27	11.86
1fc1A	2fb4H	1	1.52	134	0.24	11.39	17.76	2snv_	4ptp_	11	1.06	50	0.19	5.02	14.06
1onc_	7rsa_	1	4.30	176	0.61	30.37	11.09	1cewI	1molA	15	0.95	86	0.34	13.18	12.23
1fxiA	1ubq_	19	0.90	20	0.10	2.22	9.44	1cid_	2rhe_	15	0.79	30	0.12	1.70	16.52
3hlaB	2rhe_	1	1.17	54	0.25	8.91	8.52	2hhmA	1fbpA	1	1.27	128	0.20	9.39	20.15
3rubL	6xia_	3	1.53	46	0.07	1.20	25.56	1tie_	4fgf_	1	1.14	36	0.14	3.76	1.82
1chrA	2mnr_	1	5.33	556	0.44	33.87	24.21	1rcb_	1gmfA	1	1.01	38	0.15	2.08	17.37
2pia_	1fnr_	175	0.31	42	0.10	1.33	23.78	1tlk_	2rhe_	1	1.14	82	0.39	11.74	10.76
1aep_	256bA	1	1.07	34	0.14	1.60	18.23	1stfI	1molA	3	0.95	22	0.10	1.74	11.64
2ak3A	1gky_	1	2.06	114	0.27	6.13	26.59	2omf_	2por_	1	2.11	38	0.08	1.80	18.78
3cd4_	2rhe_	1	2.24	66	0.29	8.37	11.55	4sbvA	2tbvA	1	2.49	110	0.21	11.38	15.36
1cauB	1cauA	1	1.94	198	0.47	23.00	16.06	1dxtB	1hbg_	1	1.80	246	0.64	23.88	19.62
1c2rA	1ycc_	1	4.05	164	0.54	21.19	15.07	2cmd_	6ldh_	1	8.68	430	0.40	23.70	27.20
1aaj_	1paz_	1	1.70	102	0.40	14.93	11.55	2fbjL	8fabB	2	1.02	24	0.09	0.49	19.47
1gky_	3adk_	1	1.37	182	0.34	12.67	22.61	2sas_	2scpA	1	2.04	176	0.33	10.43	24.37
1mioC	3minB	1	7.18	542	0.31	23.18	32.50	2pna_	1shaA	1	1.18	6	0.06	−0.70	9.66
1eaf_	4cla_	1	1.98	174	0.32	13.20	21.89	1osa_	4cpv_	1	1.86	142	0.46	8.53	26.71
1pfc_	3hlaB	1	1.74	62	0.22	12.30	6.27	2hpdA	2cpp_	1	3.64	284	0.25	14.48	27.08
5fdl_	2fxb_	5	1.20	26	0.41	2.72	10.93	1lgaA	2cyp_	1	3.69	442	0.46	28.53	23.80
2afnA	1aozA	1	2.66	54	0.06	3.02	18.62	1bbhA	2ccyA	1	2.36	118	0.38	12.22	15.12
1hrhA	1rnh_	1	1.62	82	0.28	8.32	15.30	1isuA	2hipA	1	1.56	42	0.30	8.98	6.33
1npx_	3grs_	1	7.72	492	0.36	26.44	26.53	2mtaC	1ycc_	1	1.13	26	0.08	1.13	16.74
1bbt1	2plv1	1	1.64	106	0.34	13.55	13.63	1dsbA	2trxA	5	0.93	66	0.24	3.62	22.83
1mup_	1rbp_	1	1.56	108	0.29	10.74	17.31	2sim_	1nsbA	256	−0.49	28	0.06	0.45	22.41
1aba_	1ego_	1	2.08	78	0.32	11.94	10.43	2gbp_	2liv_	16	0.82	58	0.10	1.68	28.46
1crl_	1ede_	36	1.10	46	0.07	1.19	25.39	1gplA	2trxA	84	0.55	28	0.17	0.50	22.54
1cpcL	1colA	3	1.06	34	0.10	1.51	18.81	8ilb_	4fgf_	1	1.17	60	0.19	6.10	15.18
2azaA	1paz_	24	0.82	46	0.21	4.71	13.64	1gal_	3cox_	1	2.63	322	0.28	17.87	25.42
1bgeB	1gmfA	11	0.88	62	0.27	4.28	18.63	Average			1.95		0.25	7.57	
1ten_	3hhrB	127	0.52	28	0.12	4.56	7.52								

[a]Z-score for the score significance is given by Eq. (7).
[b]Number of correctly predicted contacts for the correct probe-template pair.
[c]Fraction of correctly predicted contacts for the correct probe-template pair.
[d]Z-score of correctly predicted contacts given by Eq. (8) for the correct probe-template pair.
[e]Number of correctly predicted contacts averaged over the entire structural template library.

potentials, generated by consensus contacts in the top threaded structures as provided by PROSPECTOR1. We use the arithmetic average of this potential given by Eqs. (6a) and (6b) and the original profile-based pair potential given by Eq. (4) in the next threading iteration. This case is termed the "close" and "distant" protein-specific pair potentials. The results of this calculation as well as the entire composite result of all four scoring functions ("close" sequence profiles, "close" sequence profiles plus secondary structure plus protein-specific pair potentials, "distant" sequence profiles, "distant" sequence profiles plus secondary structure plus protein-specific pair potentials) are reported in Table II. It is shown that the "distant" case alone recognizes a total of 51 proteins. This is somewhat worse than in PROSPECTOR1, where 52 proteins are recognized. However, the mean Z-score of the correctly predicted proteins increases from 3.06 to 3.84. However, the number of proteins in the top five positions increases from 56 to 59. The close sequence profiles plus secondary structure plus protein-specific pair potentials recognizes 48 proteins in the top position as compared with 46 in PROSPECTOR1, with an increase in the mean Z-score of correct cases (3.32 to 3.95) and the recognition of one new protein in the top position. For the composite prediction of PROSPECTOR2, 61, 64, and 65 proteins are recognized in the top, top five, and top 10 positions. Interestingly, as shown in Table IV, the average fraction of side-chain contacts that are selected in the probe-template structure increases slightly from 0.29 (see Table IIIB) for the pair profiles of PROSPECTOR1 to 0.30 with a slight increase in the average Z_{con} from 8.39 to 8.66. Finally, the mean threading Z-score for all structures increases from 2.59 to 3.23.

Application to the Second Fischer Benchmark

Fischer has prepared another benchmark composed of 29 probe-template pairs scanned against the original Fischer structural database plus an additional 19 tem-

TABLE IIIB. Compilation of Results on the Fischer Benchmark for the Distant Sequence Profile Plus Secondary Structure Plus Pair Interactions Scoring Function in PROSPECTOR1

Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}	Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}
2mnr_	4enl_	1	3.60	150	0.12	5.42	33.17	1hip_	2hipA	1	2.74	96	0.54	15.43	11.60
1tahA	1tca_	20	0.85	142	0.15	4.67	37.55	1arb_	4ptp_	1	2.23	36	0.05	1.66	18.89
1ltsD	1bovA	9	1.01	14	0.06	−0.50	17.67	1atnA	1atr_	1	2.27	156	0.18	5.02	38.40
1mdc_	1ifc_	1	3.32	0	—	−2.21	18.57	2sarA	9rnt_	1	1.38	70	0.29	7.78	15.10
3chy_	4fxn_	1	1.33	76	0.20	3.87	26.65	1sacA	2ayh_	6	1.28	58	0.13	2.28	29.08
2sga_	4ptp_	1	1.47	86	0.18	11.33	11.00	1hom_	1lfb_	1	2.05	88	0.64	7.77	18.03
1fc1A	2fb4H	1	2.50	242	0.40	15.27	27.69	2snv_	4ptp_	14	1.15	62	0.23	4.67	20.36
1onc_	7rsa_	1	4.98	176	0.61	22.15	16.41	1cewI	1molA	1	1.48	112	0.38	11.67	18.76
1fxiA	1ubq_	2	1.39	30	0.15	2.61	13.80	1cid_	2rhe_	9	1.01	24	0.10	0.16	22.29
3hlaB	2rhe_	2	1.55	66	0.29	5.95	14.59	2hhmA	1fbpA	1	2.83	236	0.24	11.05	36.72
3rubL	6xia_	21	1.34	78	0.09	1.69	36.12	1tie_	4fgf_	1	1.51	66	0.18	4.87	19.52
1chrA	2mnr_	1	8.28	638	0.46	20.19	45.57	1rcb_	1gmfA	1	1.35	90	0.30	4.76	28.30
2pia_	1fnr_	50	0.63	148	0.18	4.99	40.00	1tlk_	2rhe_	1	1.46	140	0.56	16.58	15.80
1aep_	256bA	1	1.48	18	0.06	−0.36	22.19	1stfI	1molA	1	1.36	48	0.22	4.29	15.92
2ak3A	1gky_	1	2.66	132	0.24	4.62	41.02	2omf_	2por_	12	1.17	132	0.20	6.02	29.10
3cd4_	2rhe_	1	3.02	96	0.38	9.97	15.00	4sbvA	2tbvA	1	2.52	116	0.21	8.57	23.28
1cauB	1cauA	1	3.28	268	0.57	24.58	20.96	1dxtB	1hbg_	1	3.40	260	0.68	18.63	27.79
1c2rA	1ycc_	1	4.75	170	0.56	18.93	19.06	2cmd_	6ldh_	1	9.54	462	0.42	15.53	43.85
1aaj_	1paz_	1	2.80	146	0.50	15.80	16.69	2fbjL	8fabB	1	2.25	140	0.20	9.84	24.64
1gky_	3adk_	1	2.33	172	0.32	7.68	37.14	2sas_	2scpA	1	2.52	176	0.33	7.27	38.76
1mioC	3minB	1	8.01	642	0.37	17.40	46.84	2pna_	1shaA	1	1.42	8	0.07	−0.77	12.18
1eaf_	4cla_	1	3.56	244	0.38	13.59	33.14	1osa_	4cpv_	1	3.17	152	0.45	6.31	39.87
1pfc_	3hlaB	1	2.72	66	0.22	9.11	9.55	2hpdA	2cpp_	1	8.86	652	0.42	20.45	46.73
5fd1_	2fxb_	1	1.40	34	0.22	2.28	17.34	1lgaA	2cyp_	1	6.48	502	0.50	21.52	38.11
2afnA	1aozA	1	2.50	72	0.07	2.46	30.94	1bbhA	2ccyA	1	2.98	176	0.53	14.89	21.18
1hrhA	1rnh_	1	2.38	96	0.34	7.76	22.13	1isuA	2hipA	1	1.81	40	0.30	7.68	7.50
1npx_	3grs_	1	7.91	536	0.37	16.48	45.61	2mtaC	1ycc_	1	1.25	44	0.14	1.77	24.58
1bbt1	2plv1	43	0.77	132	0.41	13.07	17.10	1dsbA	2trxA	1	1.69	88	0.31	3.31	34.67
1mup_	1rbp_	1	2.31	208	0.42	14.79	30.07	2sim_	1nsbA	112	0.39	78	0.08	1.91	37.13
1aba_	1ego_	1	2.47	78	0.32	8.41	14.82	2gbp_	2liv_	1	1.57	164	0.21	4.21	46.57
1crl_	1ede_	1	1.94	138	0.13	3.80	39.09	1gplA	2trxA	25	0.97	54	0.18	1.56	32.03
1cpcL	1colA	1	1.34	36	0.08	0.88	24.96	8ilb_	4fgf_	1	1.80	66	0.19	4.47	22.40
2azaA	1paz_	2	1.47	72	0.29	5.67	19.61	1gal_	3cox_	1	4.68	472	0.29	17.49	38.49
1bgeB	1gmfA	3	1.39	76	0.21	3.90	25.35								
1ten_	3hhrB	37	0.96	40	0.15	5.35	10.11	Average			2.59		0.29	8.39	

[a]Z-score for the score significance is given by Eq. (7).
[b]Number of correctly predicted contacts for the correct probe-template pair.
[c]Fraction of correctly predicted contacts for the correct probe-template pair.
[d]Z-score of correctly predicted contacts given by Eq. (8) for the correct probe-template pair.
[e]Number of correctly predicted contacts averaged over the entire structural template library.

plate structures (http://www.doe-mbi.ucla.edu/people/fischer/BENCH/tablepairs2.html). We have only been able to find 27 of the probe sequences, and report our results accordingly. We do not know whether the lack of the two additional sequences would change our results, and thus we report on those probes in which sequences are available. PROSPECTOR1 recognizes 17 pairs in the top position, as compared with the best-reported results of 17 correctly identified pairs as well as 21 and 22 in the top four and eight positions, respectively. However, in our case, one probe, "stel," which is supposed to be matched to 2azaA, selects 2pcy in the top position. Then, we have 18, 19 (19), and 20 (20) correct matches in the top position and top five (four) and 10 (eight) positions, respectively. Thus, we have somewhat better results for the first position than what has been reported previously. If we consider PROSPECTOR2, then a total of 17, 20, and 20 proteins are recognized in the top, top five, and top 10 positions, respectively.

Web-Based Access

We have set up a web server, available to the academic user community, for threading calculations that may be found at http://bioinformatics.danforthcenter.org/services/threading.html. A single sequence is all that is required, and the results will be e-mailed back to the user along with the probe-template alignments of the top 20 scoring structures and the multiple sequence alignment files.

DISCUSSION

One of the problems with earlier, pseudo one-dimensional treatments of threading is the problem of correctly selecting the partners for evaluation of the pair interactions. Originally, to address the problem, we introduced

TABLE IV. Compilation of Results on the Fischer Benchmark for the Distant Sequence Plus Secondary Structure Plus Protein-Specific Pair Profiles Scoring in PROSPECTOR1

Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}	Probe	Template	Rank	Z-score	N_c[b]	f_c[c]	Z_{con}[d]	N^{0e}
2mnr_	4enl_	1	5.27	150	0.12	5.28	33.79	1hip_	2hipA	1	3.79	114	0.62	18.79	11.40
1tahA	1tca_	1	1.96	148	0.17	5.42	37.51	1arb_	4ptp_	1	3.24	38	0.06	2.05	18.18
1ltsD	1bovA	81	0.71	44	0.22	3.15	19.04	1atnA	1atr_	1	2.09	220	0.23	7.48	40.09
1mdc_	1ifc_	1	2.80	0	0.00	−2.25	19.87	2sarA	9rnt_	1	1.67	54	0.27	5.54	15.50
3chy_	4fxn_	1	1.73	64	0.16	2.64	27.89	1sacA	2ayh_	14	1.39	66	0.15	2.55	31.69
2sga_	4ptp_	1	1.82	104	0.25	17.72	7.36	1hom_	1lfb_	1	2.56	88	0.64	8.40	17.85
1fc1A	2fb4H	1	3.78	226	0.38	13.82	28.49	2snv_	4ptp_	1	2.13	88	0.28	7.38	21.85
1onc_	7rsa_	1	6.55	176	0.60	21.34	17.53	1cewI	1molA	1	1.95	90	0.31	8.50	19.68
1fxiA	1ubq_	3	1.61	52	0.25	6.09	14.25	1cid_	2rhe_	12	1.26	32	0.13	0.68	24.41
3hlaB	2rhe_	2	2.06	72	0.29	6.41	16.17	2hhmA	1fbpA	1	3.62	258	0.29	11.44	38.07
3rubL	6xia_	48	1.15	82	0.10	1.83	36.19	1tie_	4fgf_	1	2.13	72	0.20	5.10	21.93
1chrA	2mnr_	1	9.62	646	0.46	19.66	46.23	1rcb_	1gmfA	1	1.81	100	0.30	5.87	27.59
2pia_	1fnr_	11	1.27	178	0.22	6.07	41.38	1tlk_	2rhe_	1	1.70	154	0.61	15.72	16.85
1aep_	256bA	4	1.56	20	0.08	−0.05	20.50	1stfI	1molA	1	1.85	42	0.18	3.43	17.03
2ak3A	1gky_	1	2.86	178	0.35	7.17	41.55	2omf_	2por_	37	1.06	74	0.15	2.97	27.89
3cd4_	2rhe_	1	4.26	100	0.41	8.97	16.94	4sbvA	2tbvA	1	4.11	116	0.21	9.08	23.42
1cauB	1cauA	1	4.72	268	0.59	23.67	23.47	1dxtB	1hbg_	1	4.91	260	0.68	20.60	29.20
1c2rA	1ycc_	1	6.86	170	0.56	16.85	19.50	2cmd_	6ldh_	1	9.12	464	0.42	15.79	44.05
1aaj_	1paz_	1	4.10	146	0.53	14.18	18.26	2fbjL	8fabB	1	3.13	142	0.21	10.20	25.14
1gky_	3adk_	1	3.77	194	0.36	9.09	38.53	2sas_	2scpA	1	4.00	188	0.33	7.92	39.11
1mioC	3minB	1	6.96	608	0.34	16.13	47.75	2pna_	1shaA	1	1.53	8	0.07	−0.66	11.68
1eaf_	4cla_	1	4.97	234	0.38	13.04	34.14	1osa_	4cpv_	1	4.54	152	0.44	6.68	39.96
1pfc_	3hlaB	1	3.90	70	0.23	9.56	9.97	2hpdA	2cpp_	1	6.48	604	0.40	17.87	50.09
5fdl_	2fxb_	2	1.53	34	0.21	2.32	17.93	1lgaA	2cyp_	1	6.82	500	0.50	21.94	38.54
2afnA	1aozA	1	4.47	78	0.07	2.51	32.04	1bbhA	2ccyA	1	4.16	178	0.53	16.94	19.69
1hrhA	1rnh_	1	3.32	100	0.34	8.97	21.57	1isuA	2hipA	1	2.30	40	0.32	7.56	7.63
1npx_	3grs_	1	8.48	542	0.36	15.78	47.45	2mtaC	1ycc_	2	1.29	32	0.11	0.81	23.67
1bbt1	2plv1	29	1.07	120	0.38	10.88	17.89	1dsbA	2trxA	1	2.06	76	0.28	2.53	36.07
1mup_	1rbp_	1	3.72	212	0.43	14.14	32.73	2sim_	1nsbA	3	1.50	122	0.12	3.89	37.88
1aba_	1ego_	1	3.18	80	0.33	7.79	16.17	2gbp_	2liv_	1	2.53	222	0.26	6.11	46.93
1crl_	1ede_	13	1.53	130	0.12	3.78	38.34	1gp1A	2trxA	23	1.28	52	0.17	1.39	32.72
1cpcL	1colA	1	2.25	54	0.13	2.34	25.98	8ilb_	4fgf_	1	2.88	62	0.18	3.95	24.79
2azaA	1paz_	1	2.39	84	0.33	6.39	20.88	1gal_	3cox_	1	4.89	568	0.37	21.42	39.20
1bgeB	1gmfA	3	2.00	64	0.21	2.96	26.03	Average			3.23		0.30	8.66	
1ten_	3hhrB	2	1.53	40	0.16	5.17	9.97								

[a]Z-score for the score significance is given by Eq. (7).
[b]Number of correctly predicted contacts for the correct probe-template pair.
[c]Fraction of correctly predicted contacts for the correct probe-template pair.
[d]Z-score of correctly predicted contacts given by Eq. (8) for the correct probe-template pair.
[e]Number of correctly predicted contacts averaged over the entire structural template library.

the frozen approximation in which the partners from the template structure are used in the evaluation of the pair potentials.[29] If the environments were similar, the approximation worked well. Otherwise, it performed poorly. However, for practical reasons, it would be desirable to retain the advantages of a local scoring function that enables dynamic programming to be used as the search scheme. Here, we have suggested an iterative approach in which a sequence profile is used to generate the initial alignment of the probe sequence in the template structure; in subsequent iterations, this alignment is used to evaluate the partners. We term this the "partly thawed" approximation. We have demonstrated that this approximation works quite well, not only in the selection of the template, but also in the construction of a protein-specific pair potential whose recognition capabilities are enhanced as assessed by the Z-score of the correctly predicted structures. When the entire hierarchical approach of four scoring functions is used in PROSPECTOR2, this method correctly recognizes 61 proteins in the top position. In addition, the use of pair potentials enhances the number of correctly identified side-chain contacts when the correct probe-template pair is considered. In future work, we will explore this issue further and report on the use of suboptimal scoring structures in threading for the prediction of tertiary contacts and secondary structure.

The question as to why PROSPECTOR does comparatively better than alternative approaches developed previously and described on the UCLA website (http://www.doe-mbi.ucla.edu/people/fischer/BENCH/tablepairs2.html) is not a simple one to answer. Certainly, the use of a hierarchical approach can provide additional information over the case when just a single scoring function is used. It might be argued that the present approach is superior because it handles pair interactions better than the methods described on the UCLA website, which either do not consider pair

terms or treat them using the original frozen approximation. Also, the pair potentials we use have been highly optimized to give the best available results in gapless threading.[47] A potential weakness of the current approach is that we do not use state-of-the-art sequence profiles (e.g., as provided by PSI-BLAST). But our goal was to use a straightforward sequence-profile implementation that would provide a baseline for future work. However, surprisingly, our naive profile implementation works better than PSI-BLAST.

One of the more surprising results in this series of calculations is that the particular secondary structure implementation we use only imparts a marginal improvement relative to its absence in standard benchmarks. Alternative secondary structure prediction schemes (e.g., using standard secondary structure prediction schemes such as PHD[61] first, and then implementing these predicted secondary structures as a bias) need to be explored. Similarly, the choice of how pair interactions are implemented has not been fully explored, and alternatives such as Cα-based, and side-chain orientation-dependent potentials[62] have to be examined to see where additional improvements in sequence-structure specificity can be made.

Given that active site descriptors can correctly select threading structures well into the twilight zone of sequence-structure specificity and make both structural and functional assignments with a low false-positive rate,[8,10–15] the demands on a threading algorithm that uses such information are much less stringent than if structure prediction alone is to be done. That is, what one really requires is an algorithm that can get the correct fold near the top with a score of at least moderate significance with a reasonably good alignment, and then an active site filter can assist in fold as well as biochemical function identification. This is the origin of our hierarchical method of multiple scoring functions that, in combination, recognizes 59 of the 68 Fischer probe pairs in the top position in PROSPECTOR1 and 61 in the top position in PROSPECTOR2. Nevertheless, it is clearly important to have an excellent threading algorithm to ensure that the correct structure is within this threshold in order to be certain that all proteins in a genome having the particular fold and function are identified.

Finally, we observe that very distant sequence profiles possess significant information and can profitably assist in fold recognition. Indeed, quite often it is this set of sequence plus secondary structure pair interactions that has the best Z-score for the correct probe-template pair on threading. Others have noticed the utility of using distant sequences as well, including Simons et al.,[63] Koretke (personal communication), and our group in an earlier derivation of local fragment-based protein-specific pair potentials.[47] Clearly, better ways remain to be developed to more fully extract the information latent in the set of weakly related sequences.

In summary, a new threading algorithm, PROSPECTOR, has been developed, which is at the state-of-the-art of contemporary threading algorithms, as assessed by its performance on standard benchmarks. In future work, we will apply this methodology both to structure prediction on a genomic scale as well as to the problem of tertiary contact and secondary structure prediction. Moreover, in the way the algorithm is constructed, known experimental restraints (e.g., disulfide bonds or nuclear magnetic resonance restraints) can be readily integrated into this threading algorithm. This can be done both by biasing the pair potential toward known contacts and by eliminating structures that do not satisfy the constraints in a post-threading selection step. In preliminary results, this is found to work quite well for some simple cases. Although sequence profiles still play an important role in structure prediction in the present threading algorithm, pair interactions are seen to play a comparable role. They increase the number of correctly identified Fischer pairs and increase the Z-score of the correct sequence-structure matches. Although this is not yet a purely structure-based threading algorithm, it represents a significant step in that direction.

ACKNOWLEDGMENTS

Stimulating discussions with Drs. J. Fetrow, A. Kolinski, and A. R. Ortiz are gratefully acknowledged. We are indebted to Drs. K. Koretke and A. Lupas for describing their work on the use of distant sequences in PSI-BLAST, which motivated the development of the "distant" sequence profiles used in this work. We also thank Dr. Alejandro Schaffer for his technical help in setting up the IMPALA package. Finally, we acknowledge K. White for her assistance in the preparation of this article.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
2. Pearson WR. Effective protein sequence comparison. Methods Enzymol 1996;266:227–258.
3. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1995. Nucleic Acids Res 1995;24:189–196.
4. Henikoff S, Henikoff JG. Protein family classification based on searching a database of blocks. Genomics 1994;19:97–107.
5. Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ. PRINTS: a database of protein motif fingerprints. Nucleic Acids Res 1994;22: 3590–3596.
6. Attwood TK, Beck ME, Bleasby AJ, Degtyarenki K, Michie AD, Parry-Smith DJ. Novel developments with the PRINTS protein fingerprint database. Nucleic Acids Res 1997;25:212–216.
7. Nevill-Manning CG, Wu TD, Brutlag DL. Highly specific protein sequence motifs for genome analysis. Proc Natl Acad Sci USA 1998;95:5865–5871.
8. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 1998;281:949–968.
9. Yu L, White JV, Smith TF. A homology identification method that combines protein sequence and structure information. Protein Sci 1998;7:2499–2510.
10. Fetrow JS, Godzik A, Skolnick J. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 1998;282:703–711.
11. Zhang L, Godzik A, Skolnick J, Fetrow JS. Functional analysis of *E. coli* proteins for members of the α/β hydrolase family. Fold Des 1998;3:535–548.
12. Fetrow JS, Siew N, Skolnick J. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. FASEB J 1999; 13:1866–1874.

13. Siew N, Skolnick J, Fetrow J. Prediction of disulfide oxidoreductase function in nine genomes. In preparation.
14. Zhang B, Rychlewski L, Pawlowski K, Fetrow JS, Skolnick J, Godzik A. From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. Protein Sci 1999;8:1104–1115.
15. Skolnick J, Fetrow J. From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol 2000;18:34–39.
16. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput 1996;300–318.
17. Wilmanns M, Eisenberg D. Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. Proc Natl Acad Sci USA 1993;90:1379–1383.
18. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
19. Yi T-M, Lander ES. Recognition of related proteins by iterative template refinement (ITR). Protein Sci 1994;3:1315–1328.
20. Matsuo Y, Nishikawa K. Protein structural similarities predicted by a sequence-structure compatibility method. Protein Sci 1994;3: 2055–2063.
21. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
22. Koretke KK, Russell RB, Copley RR, Lupas AN. Fold recognition using sequence and secondary structure information. Proteins 1999;Suppl 3:141–148.
23. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;277:876–888.
24. Tropsha A, Singh RK, Vaisman II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. Pac Symp Biocomput 1996;614–623.
25. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
26. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. Protein Sci 1996;5:1043–1059.
27. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through folding motif. Proteins 1993; 16:92–112.
28. Lathrop R, Smith TF. Global optimum protein threading with gapped alignment and empirical pair scoring function. J Mol Biol 1996;255:641–665.
29. Godzik A, Skolnick J, Kolinski A. A topology fingerprint approach to the inverse protein folding problem. J Mol Biol 1992;227:227–238.
30. Selbig J. Contact pattern induced pair potentials for protein fold recognition. Protein Eng 1995;8:339–351.
31. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Proteins 1992;13:258–271.
32. Wilmanns M, Eisenberg D. Inverse protein folding by the residue pair preference profile method. Protein Eng 1995;8:626–639.
33. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
34. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
35. Thiele R, Zimmer R, Lengauer T. Recursive dynamic programming for adaptive sequence and structure alignment. Ismb 1995;3: 384–392.
36. Murzin AG. Structure classification-based assessment of CASP3 predictions for the fold recognition targets. Proteins 1999;37:88–103.
37. Jones DT, Tress M, Bryson K, Hadley C. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins 1999;Suppl 3:104–111.
38. Ota M, Kawabata T, Kinjo AR, Nishikawa K. Cooperative approach for the protein fold recognition. Proteins 1999;Suppl 3:126–132.
39. Domingues FS, Koppensteiner WA, Jaritz M, et al. Sustained performance of knowledge-based potentials in fold recognition. Proteins 1999;Suppl 3:112–120.
40. Panchenko A, Marchler-Bauer A, Bryant SH. Threading with explicit models for evolutionary conservation of structure and sequence. Proteins 1999;Suppl 3:133–140.
41. Moult J, Hubbard T, Fidelis K, Pedersen JT. Critical assessment of methods of protein structure prediction (CASP): round III. Proteins 1999;Suppl 3:2–6.
42. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
43. Henikoff JG, Henikoff S. Blocks database and its applications. Methods Enzymol 1996;266:88–105.
44. Ogiwara A, Uchiyama I, Takagi T, Kanehisa M. Construction and analysis of a profile library characterizing groups of structurally known proteins. Protein Sci 1996;5:1991–1999.
45. 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 3D structures. J Mol Biol 1993;232:805–825.
46. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST: a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
47. Skolnick J, Kolinski A, Ortiz AR. Derivation and testing of protein specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
48. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.
49. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 1994;24:307–331.
50. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276:71–84.
51. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
52. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
53. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
54. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 1994;18:338–352.
55. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
56. Fischer Database. http://www.doe-mbi.ucla.edu/people/fischer/BENCH/benchmark1.html. UCLA; 1996.
57. Waterman MS, Eggert M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. J Mol Biol 1987;197:723–728.
58. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 1998;7:1431–1440.
59. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
60. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
61. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
62. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. A method for improvement of threading based models. Proteins 1999;37:592–610.
63. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;Suppl 3:171–176.