

# Combination of Threading Potentials and Sequence Profiles Improves Fold Recognition

Anna R. Panchenko, Aron Marchler-Bauer and Stephen H. Bryant\*

*Computational Biology Branch  
National Center for  
Biotechnology Information,  
Building 38A, Room 8N805  
National Institutes of Health  
Bethesda, MD 20894, USA*

Using a benchmark set of structurally similar proteins, we conduct a series of threading experiments intended to identify a scoring function with an optimal combination of contact-potential and sequence-profile terms. The benchmark set is selected to include many medium-difficulty fold recognition targets, where sequence similarity is undetectable by BLAST but structural similarity is extensive. The contact potential is based on the log-odds of non-local contacts involving different amino acid pairs, in native as opposed to randomly compacted structures. The sequence profile term is that used in PSI-BLAST. We find that combination of these terms significantly improves the success rate of fold recognition over use of either term alone, with respect to both recognition sensitivity and the accuracy of threading models. Improvement is greatest for targets between 10% and 20% sequence identity and 60% to 80% superimposable residues, where the number of models crossing critical accuracy and significance thresholds more than doubles. We suggest that these improvements account for the successful performance of the combined scoring function at CASP3. We discuss possible explanations as to why sequence-profile and contact-potential terms appear complementary.

\*Corresponding author

**Keywords:** protein evolution; fold recognition; threading; sequence profiles

## Introduction

The process of molecular evolution leads to varying degrees of similarity between proteins related by descent from a common ancestral gene. At short evolutionary distances sequences and structures are recognizably similar and easily aligned. At larger evolutionary distances only certain sequence features or motifs may be conserved (Bork & Koonin, 1996), and structural similarity may extend to only a core substructure (Chothia & Lesk, 1986; Hubbard & Blundell, 1987; Matsuo & Bryant, 1999). At still greater evolutionary distances sequence similarity may be undetectable, and one may be unable to determine whether an observed structural similarity indicates homology, or is instead due to convergence (Sander & Schneider, 1991; Murzin, 1998; Russell *et al.*, 1998).

Sequence comparison methods based on profiles aim to detect significant similarity at as great an evolutionary distance as possible (Gribkov *et al.*,

1987; Henikoff & Henikoff, 1996; Neuwald *et al.*, 1997; Krogh *et al.*, 1994; Tatusov *et al.*, 1994). In the popular PSI-BLAST method, for example, a position-specific score matrix (PSSM) describes sequence conservation among known members of a protein family on a residue-by-residue basis (Altschul *et al.*, 1997; Aravind & Koonin, 1999). Even if a new sequence is similar to others only with respect to a particular region or a highly conserved motif, the profile may detect this similarity and allow one to recognize the new sequence as a member of the family. In the recent CASP3 competition (Moult *et al.*, 1999) a number of fold-recognition targets were correctly identified by sequence-profile methods (Karplus *et al.*, 1999; Murzin, 1999).

Threading methods, in contrast, aim to detect conservation of a core substructure, even in the absence of shared sequence motifs. A sequence is threaded through templates from the structure database and alternative sequence-structure alignments are scored using a conformational energy calculation of some kind (Bowie *et al.*, 1990, 1991; Bryant & Amzel, 1987; Godzik *et al.*, 1992; Goldstein *et al.*, 1992; Hendlich *et al.*, 1990; Johnson *et al.*, 1993; Jones *et al.*, 1992; Mirny *et al.*, 1998). If

Abbreviations used: PSSM, position-specific score matrix.

E-mail address of the corresponding author: [bryant@ncbi.nlm.nih.gov](mailto:bryant@ncbi.nlm.nih.gov)

the (unknown) structure of that sequence is sufficiently similar to a known structure, threading may produce an accurate alignment and structure prediction. This was the case, for example, with a prediction for the Obese gene product, Leptin (Madej *et al.*, 1995a; Zhang *et al.*, 1997). In the CASP3 competition, a number of fold-recognition targets were also correctly identified by threading methods, including some not identified by profile methods (Jones *et al.*, 1999; Koretke *et al.*, 1999; Murzin, 1999; Ota *et al.*, 1999; Panchenko *et al.*, 1999; Domingues *et al.*, 1999).

Sequence-profile and threading methods may be distinguished by their use of similarity metrics based on sequence and structure conservation, respectively. These metrics may be thought of as unrelated, in as much as one involves genotype (sequence) and the other a simple phenotype (three-dimensional structure). And yet both have a common goal, to better detect homology by focusing the comparison on the substructure most highly conserved within a protein family, be it the sequence motif detected by a sequence-profile method, or the conserved core structure detected by a threading method. For this reason one may imagine combining these methods, to provide a similarity metric that simultaneously considers sequence and structure conservation. Such a combined metric might offer improved sensitivity for classification of new sequences.

Other investigators have previously attempted to combine sequence profiles with structure-based similarity scores (Elofsson *et al.*, 1996; Fischer & Eisenberg, 1996; Jaroszewski *et al.*, 1998; Rost *et al.*, 1997; Yi & Lander, 1994). These groups presented evidence that detection rates increased when using a combined score, but they also found that alignment accuracy (using structure-structure alignments as a standard of truth) did not greatly improve. One might expect sequence-profile and threading-potential terms to be sensitive to the presence of the same conserved substructure, and that improvement in fold-recognition specificity and alignment accuracy would thus go hand in hand. Failure to improve alignment accuracy suggests that these metrics may have not been combined in the best way, or that the sequence profile and threading potential somehow encode different and perhaps contradictory features of the template, calling into question the premise of combining them. To address this issue, and to measure as carefully as possible any difference in fold-recognition performance when sequence-profile and contact-potential terms are combined, we return here to an investigation of a combined score function.

The score function we consider is based on well-known sequence-profile and contact-potential terms. We sum a profile score derived from the PSI-BLAST PSSM for the template (Altschul *et al.*, 1997) with a threading score derived from contact potentials used in all CASP competitions to date (Bryant, 1996; Bryant & Lawrence, 1993; Levitt, 1997; Madej *et al.*, 1995b; Panchenko *et al.*, 1999).

We sum terms with a linear weighting factor whose optimal value we seek to determine. Our experimental design is to select a benchmark set containing pairs of structurally similar proteins, with emphasis on medium-difficulty targets which show extensive structural similarity but no pairwise sequence similarity (Marchler-Bauer & Bryant, 1997a, 1999). We then perform a series of threading experiments to identify the value of the weighting factor that gives the overall best performance. Performance is evaluated with respect to both the statistical significance of the combined threading score, an indicator of fold-recognition sensitivity, and the accuracy of the threading alignment, using structure-structure alignments as a standard of truth.

This combined score function has been tested in blind predictions for the 1998 CASP3 competition (Moult *et al.*, 1999; Murzin, 1999; Panchenko *et al.*, 1999) where our method was awarded "first place" in fold recognition. The present control experiments allow us to examine a larger number of test cases, however, and to conduct a more thorough search for optimal weighting factors. We find that combination of contact-potential and sequence-profile terms improves fold recognition sensitivity for "easy" and "medium" targets, in agreement with previous work (Elofsson *et al.*, 1996; Fischer & Eisenberg, 1996; Jaroszewski *et al.*, 1998; Rost *et al.*, 1997; Yi & Lander, 1994). In contrast to earlier results, however, we find that combination of profile and threading scores also clearly improves alignment accuracy across the entire range of target difficulties, and particularly so for medium targets. We conclude that a metric combining sequence and structure similarity improves detection and modeling of conserved substructures, and we discuss why sequence-profile and contact-potential terms appear to be complementary.

## Algorithm

### Scores for sequence and structure similarity

The threading method we employ has been described (Panchenko *et al.*, 1999). Target sequences are threaded through the core of a structural template using the combined score function and the core-element alignment algorithm (Bryant, 1996; Madej *et al.*, 1995b). The combined score function is comprised of a contact-potential term and a sequence-profile term. The contact potential is derived from the frequencies of non-local contacts (separated by five or more peptide bonds in the polypeptide chain) between different residue types, as observed in the 3D-structure database (Bryant & Lawrence, 1993). The sequence-profile term is derived from a PSSM calculated as in PSI-BLAST, by iterative search of the sequence database for sequences similar to that of the template (Altschul *et al.*, 1997).

The contact-potential score  $\Delta G^T$  corresponds to the difference between the sum of residue contact

energies in the threading alignment and the sum expected for random shuffles of the aligned residues (Bryant & Lawrence, 1993). This normalization to a random-sequence reference state corrects for sequence composition effects (Bryant & Altschul, 1995). The sequence-profile score  $\Delta G^S$  is calculated similarly, as the difference between the sum of PSSM elements for the threading alignment and the sum expected for random shuffles of the aligned residues. We find that  $\Delta G^T$  and  $\Delta G^S$  are on approximately the same scale, since expected values for random sequences are set to zero and standard deviations are comparable for the range of sequence similarities considered here (not shown). To modulate the contributions of structure and sequence similarity to the combined score  $\Delta G$  we thus use a linear weighting factor  $w$ :

$$\Delta G = w\Delta G^S + (1 - w)\Delta G^T$$

Here  $w = 0$  corresponds to full weight on the contact potential and  $w = 1$  to full weight on the sequence profile.

### Definition of conserved core structures

We define core structures of template proteins using a knowledge-based procedure. Core elements are defined as those chain-continuous segments that are conserved across multiple structure-structure alignments with homologous neighbors (Matsuo & Bryant, 1999; Panchenko *et al.*, 1999). The allowed maximum lengths of loops (unaligned segments of the target sequence, falling between aligned core elements) are set in proportion to the longest loop seen in the structures of homologous neighbors. Minimum loop lengths are constrained dynamically, to exclude non-physical models with too few loop residues to span the distance between core elements. We note that core definitions are determined automatically based on available structural neighbors, as in CASP3 predictions. For some pairs in the benchmark set, the correct structure-structure alignment may be partially excluded by the core definition if the consensus substructure it describes is more extensive or differs from the substructure shared between the target and template (Madej *et al.*, 1995b).

Target sequences are threaded through structural templates using the core-element alignment algorithm (Bryant, 1996). Chain-continuous segments from the sequence are initially aligned randomly with core elements of the template. Alternative alignments of each core element are then sampled iteratively in the field defined by the score function and the current alignment of other core elements. The alignment search algorithm allows shifts of the target sequence relative to a core element as well as recruitment of additional residue sites by chain-continuous recruitment to the core-element N or C terminus. We note that the core element alignment algorithm does not use the "frozen approximation" (Godzik *et al.*, 1992) or gap penalties, since the

number of alignment variables is small enough to allow Gibbs sampling of alternative alignments with direct evaluation of the contact potential. There is no guarantee this sampling process will converge, but we believe this to be the case, since run times are adjusted to achieve recurrence of top-scoring alignments.

### Significance threshold for fold recognition

To determine whether the threading score obtained for a given target-template pair indicates successful fold recognition we examine its threading  $p$ -value, i.e. the probability that a randomly shuffled target sequence would obtain a score as high. Threading  $p$ -values calculated in this way formally give the expected specificity of reverse fold recognition (a "structure seeks sequence" search). However, a significant  $p$ -value may be understood as a necessary condition for successful fold recognition (a "sequence seeks structure" search), since one can have no confidence in a threading model where a shuffled target sequence scores equally well. Control experiments (Bryant & Altschul, 1995) and blind prediction results for CASP3 (Panchenko *et al.*, 1999) have furthermore shown that significant  $p$ -values very rarely occur for incorrect choices of structural template, for the threading method we consider here. Below, in comparing results for alternative score functions, we therefore interpret occurrence of a significant  $p$ -value as evidence that target-template similarity would be successfully detected in a fold-recognition search.

We estimate threading  $p$ -values by a conventional "shuffle and re-align" procedure. We calculate composition-corrected scores as the ratio of  $\Delta G$  and the standard deviation of scores  $\Delta G^0$  obtained by randomly shuffling the aligned residues 10,000 times (Bryant & Lawrence, 1993). The expected composition-corrected score distribution for the target-template pair is then sampled by shuffling and optimally realigning the full-length target sequence 50 times. Assuming this distribution to be normal, we estimate the probability that a random sequence would score as high or higher than the target sequence as the integrated standard normal density for the that score. The  $p$ -values calculated in this way are measures of statistical significance or that take into account differences in the number of alternative alignments allowed by sequence length, template length and the constraints imposed by the structural core definition for the template (Bryant & Altschul, 1995).

### A benchmark for comparison of threading score functions

We select pairs of structurally similar proteins from the structure-neighbor database distributed with Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>; Wang *et al.*, 2000), using data available as of February, 1998. Structure-structure alignments

in this set are computed by the VAST algorithm (Gibrat *et al.*, 1996), based on complete protein chains and/or domains identified by a compactness algorithm (Madej *et al.*, 1995b) similar to that of Holm & Sander (1994). Detailed structure-structure alignments may be viewed by using Entrez to select the specific target-template pairs in the benchmark set, as listed in Table 1. We note that structural similarities as detected by VAST have previously been compared to SCOP classifications (Matsuo & Bryant, 1999; Murzin *et al.*, 1995; Przytycka *et al.*, 1999). Overall quality of VAST alignments may also be judged by examining search results for targets in the CASP3 competition (<http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/casp3/>).

To select a subset of Entrez structure neighbors suitable as a benchmark for comparison of threading score functions we apply a number of filter criteria. To omit “very easy” targets from

the benchmark set we first of all omit pairs with recognizably similar sequences. We select only structurally similar pairs involving members of a non-redundant subset of chain sequences, constructed by single-linkage clustering with a BLAST *e*-value (Altschul *et al.*, 1997) threshold of  $10^{-7}$  (<http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpd.html>; Matsuo & Bryant, 1999). We furthermore exclude pairs with more than 25% sequence identity in structure-structure alignment. The latter criterion assures that none of the pairs in the test set crosses the threshold for significant sequence similarity suggested by Sander & Schneider (1991).

We also apply filter criteria to select pairs where structural similarity involves a single protein domain. We omit pairs where domain length is greater than 250 residues or where the target domain is formed from more than one chain-continuous segment. We further accept only those

**Table 1.** Structurally similar proteins in the test set

|                |                |               |               |
|----------------|----------------|---------------|---------------|
| 1abe 2-2dri 2  | 1abe 2-2lbp 2  | 1abrB1-1bff   | 1abrB1-1hcd   |
| 1abrB1-2i1b    | 1abrB2-1bff    | 1abrB2-1hcd   | 1abrB2-2i1b   |
| 1acf -1fil     | 1acx -1ctn 1   | 1agjA -1havA  | 1agjA -1hpgA  |
| 1agjA -5ptp    | 1aizA -1aac    | 1aizA -1cur   | 1amm 1-1prs 1 |
| 1ao7D -1cd8    | 1ao7D -1cfb 1  | 1ao7D -1ctn 1 | 1ao7D -1fnf 1 |
| 1ao7D -1tit    | 1ao7D -1zxq 2  | 1ao7D -2ncm   | 1ao7E1-1boy 1 |
| 1ao7E1-1cd8    | 1ao7E1-1ctn 1  | 1ao7E1-1hnf 1 | 1ao7E1-1hnf 2 |
| 1ao7E1-1tit    | 1ao7E1-1zxq 2  | 1ao7E1-2ncm   | 1argB3-1tplA3 |
| 1arhB3-1tplA3  | 1asu -2rn2     | 1avdB -1smpI  | 1bbpA -1epaB  |
| 1bbpA -1fem    | 1btt3 -1bmv21  | 1btt3 -1bmv22 | 1bebA -1epaB  |
| 1bebA -1fem    | 1bff -1abrB2   | 1bff -1hcd    | 1bff -2i1b    |
| 1bhgA2-1ctn 1  | 1bmdA1-1scuA1  | 1bmtA2-1dih 1 | 1bmtA2-1scuB3 |
| 1bmtA2-1yasA1  | 1bmtA2-2dri 1  | 1bmtA2-2lbp 1 | 1bmtA2-2lbp 2 |
| 1bovA -1aszA1  | 1boy 1-1boy 2  | 1boy 1-1cfb 1 | 1boy 1-1ebpA2 |
| 1boy 1-1fnf 4  | 1boy 1-1nfkA2  | 1boy 2-1boy 1 | 1boy 2-1cd8   |
| 1cd8 -1zxq 1   | 1cdg 4-1nfkA2  | 1cfb 1-1boy 1 | 1cfb 1-1hnf 2 |
| 1cgv 4-1nfkA2  | 1chrA2-1pii 2  | 1ciy 3-1tulo  | 1ctn 1-1fieB1 |
| 1cwpB -1stmA   | 1cwpB -2stv    | 1cydA -1dhr   | 1cydA -1dih 1 |
| 1cyw -1cur     | 1din -1oilA1   | 1din -1yasA1  | 1dsuA -1agjA  |
| 1dsuA -1havA   | 1dsuA -1hpgA   | 1dvfD -1neu   | 1dvfD -1vcaA1 |
| 1ebdA2-1tde 1  | 1ebdA2-1tde 2  | 1edhA2-1ksr   | 1edhA2-1vcaA1 |
| 1eft 1-1bmtA2  | 1eft 1-1tndA1  | 1eft 1-5p21   | 1efuA1-1tndA1 |
| 1epaB -1bbpA   | 1epaB -1bebA   | 1fdr 1-1que 1 | 1fdr 2-1ndh 3 |
| 1fem -1bbpA    | 1fem -1bebA    | 1fft -1up1 1  | 1finA2-2csn 2 |
| 1fnb 1-1fdr 1  | 1fnb 1-1ndh 1  | 1fnf 1-1vcaA2 | 1fnf 2-1tit   |
| 1fod2 -1bmv21  | 1gca 2-1abe 2  | 1gca 2-1pea 2 | 1gof 3-1dim 2 |
| 1gof 3-1tbgA1  | 1hcd -1abrB2   | 1hlcA -1lcl   | 1hlpA2-1bmdA2 |
| 1hpgA -1agjA   | 1hurB -1eft 1  | 1hurB -1rcf   | 1hurB -5p21   |
| 1igs 1-1pii 2  | 1igs 1-1sftA2  | 1irp -1bff    | 1irp -1hcd    |
| 1livd 2-1dim 2 | 1livd 2-1tbgA1 | 1kit 1-2ayh   | 1kit 4-1lcl   |
| 1kit 4-1sacA   | 1kit 4-1slcA   | 1krs -1bovA   | 1ldg 2-1bmdA2 |
| 1led -2ayh     | 1mucB2-1nal11  | 1nal11-1pii 2 | 1nal11-1sftA2 |
| 1nif 1-1aozA3  | 1ntr -1e2b     | 1ntr -5p21    | 1pnf 1-1stmA  |
| 1pnrA3-1abe 2  | 1pnrA3-1pea 2  | 1poxA1-1pvdA1 | 1poxA1-1pvdA3 |
| 1poxA3-1pvdA1  | 1poxA3-1pvdA3  | 1prs 1-1amm 1 | 1qrdA1-1rcf   |
| 1rcf -1qrdA1   | 1reqA2-1bmtA2  | 1reqA2-1dih 1 | 1reqA2-2dri 1 |
| 1reqB3-1bmtA2  | 1rtm11-1lit    | 1rvv1 -1abe 1 | 1scuA2-1scuB3 |
| 1stmA -1cwpB   | 1tlk -1ten     | 1vid -2admA1  | 1wba -1bff    |
| 1wba -2i1b     | 1wbc -1bff     | 1wit -1ten    | 1zxq 2-1cfb 1 |
| 2admA1-1vid    | 2cut -1din     | 2cut -1yasA1  | 2fcr -1qrdA1  |
| 2i1b -1bff     | 2i1b -1hcd     | 2rn2 -1asu    |               |

Protein target/template pairs used in the analysis are listed by their four-letter PDB code (lower case), chain identifier (if applicable, upper case) and domain identifiers (numeric, starting with 1 for each chain). Superpositions may be examined using: <http://www.ncbi.nlm.nih.gov/Entrez/Structure.html>. Enter the PDB code, “structure summary”, and then select “structure neighbors” of the target domain. Select “All of MMDb” to show the complete list of neighbors, and use the pattern-search feature of the browser to find the listed template.

pairs where Entrez domain definitions agree with those in SCOP to a threshold of 80% mutual overlap (Matsuo & Bryant, 1999). By including only single-domain pairs in the benchmark we focus on the simplest possible case for fold recognition, detection of a conserved domain substructure. We note, however, that this does not make the benchmark “easy”: Successful recognition of single-domain similarities is by no means guaranteed in blind prediction (Marchler-Bauer & Bryant, 1997a; Murzin, 1999). We also note that use of single-domain pairs allows us to match unambiguously SCOP domain classifications (Murzin *et al.*, 1995) with VAST alignments, to identify those pairs in the benchmark where the authors of SCOP find evidence for homology and descent from a common ancestral gene.

We apply further filter criteria to omit pairs where structural similarity does not involve a compact domain substructure. We require that the fraction of non-local contacts conserved in the structure-structure alignment of each protein with the other be 30% or more and that overall contact density be above 4.5 contacts per residue (counting contacts in distance intervals up to 8 Å). By including only compact domains in the benchmark we again focus on the simplest possible case for fold recognition, where conserved intra-domain interactions are extensive and inter-domain or docking interactions may be ignored. The specific thresholds we employ derive from an analysis of CASP2 and CASP3 predictions, where predictions by our (and other’s) threading methods were found to be uniformly inaccurate for targets that do not meet these criteria (Marchler-Bauer & Bryant, 1999).

Lastly, we select pairs where homologous structure neighbors well define the conserved core structure of the template. We require that the template structure possess at least five such neighbors, each sequence-dissimilar to the template and to one another to a BLAST *e*-value of  $10^{-7}$  (Matsuo & Bryant, 1999; Panchenko *et al.*, 1999). This last criterion is particularly restrictive, since there are relatively few protein families where this many sequence-dissimilar, homologous structures are known. In total these filter criteria, applied automatically, reduce the test set to 163 structurally similar pairs as listed in Table 1. We note that the benchmark set nonetheless remains diverse, since its members are selected from a sequence-dissimilar subset of known structures. According to SCOP classifications (Murzin *et al.*, 1995) the structural templates in the benchmark are drawn from four different structural classes, 34 different folds, and 51 different superfamilies.

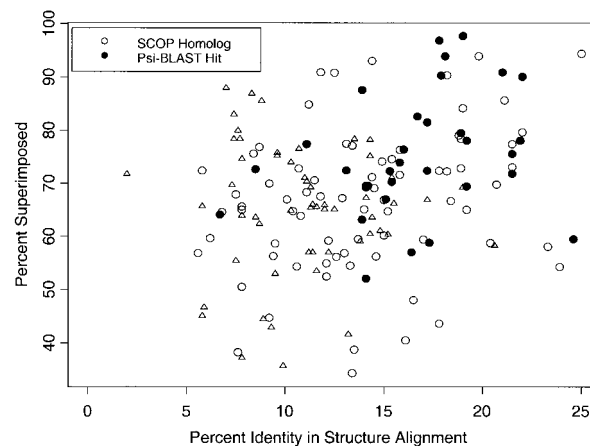
## Results

### Difficulty of targets in the benchmark

In the experiments below we apply a threading algorithm to the structurally similar pairs in the

benchmark set, treating one protein as the structural template and the other as the target sequence. To interpret these results it is important to understand the difficulty of the fold recognition problems presented by this benchmark, and in particular to ask whether the targets are sufficiently difficult to provide a sensitive test of alternative score functions. Target difficulty may be characterized by a “phase diagram” plotting the degree of sequence similarity *versus* the extent of structural similarity (Marchler-Bauer & Bryant, 1997b, 1999). In Figure 1 we apply this analysis to the benchmark set, plotting the percentage of identical residues in structure-structure alignment against the percentage of target residues that may be superimposed on the template. Plotting symbols indicate other aspects of similarity, whether the target sequence may be recognized as similar to the template by PSI-BLAST (Altschul *et al.*, 1997) and whether the target and template have been classified as homologs, belonging to the same SCOP superfamily (Murzin *et al.*, 1995).

One may see from Figure 1 that the benchmark set presents a range of target difficulties. “Hard” fold recognition targets were previously defined as those where structural similarity extends to less than 60% of target residues, and where there are no characteristic sequence motifs (here, as recognized by PSI-BLAST). By this definition the test set contains 40 hard targets (25%), 25 of which are homologous to the template according to SCOP.



**Figure 1.** The phase diagram of target difficulty for target/template pairs in the benchmark set. The percentage of identical residues in structure-structure alignments of the target and template is plotted on the horizontal axis. The percentage of target residues that are superimposed on the template, in the same structure-structure alignment, is plotted on the vertical axis. Structure-structure alignments are computed by the VAST algorithm and may be examined using the Entrez retrieval system (see footnote to Table 1). Homologous target/template pairs as identified by SCOP classifications are shown as circles. Pairs where sequence similarity can be detected by PSI-BLAST (based on the sequence profile for the template, with *e*-value  $\leq 0.01$ ) are shown by filled symbols.

There are 42 medium targets (26%), where the extent of structural similarity rises above 60%, but motifs remain absent and sequence identity remains below 12%. Of these, 17 are homologous to the template, according to SCOP. Lastly, there are 81 easy targets (50%), where motifs are recognized by PSI-BLAST, or structural similarity is above 60% and sequence identity above 12%. Of the easy targets, 65 are classified as homologous to the template by SCOP. We emphasize, of course, that even these easy fold-recognition targets have been selected to have sequence identities of 25% or less, in structural alignment, and BLAST  $e$ -values of  $10^{-7}$  or less, in pair-wise sequence alignment.

Judging from the performance of different methods at CASP3 (Marchler-Bauer & Bryant, 1999; Murzin, 1999), we expect threading to succeed for relatively few of the hard targets, some of the medium targets, and most of the easy targets. Medium targets were rare among the small sets available for CASP workshops, but it appears that the filter criteria described above have particularly enriched the benchmark set for targets in the medium-difficulty range, near the center of the "phase diagram" in Figure 1. Fold-recognition experiments using this set may thus be sensitive to even small differences in performance of alternative scoring functions, since it is with medium targets that one expects to see the greatest change in success rate. We note that the benchmark set may be made larger by relaxing one or another of the filter criteria. Results below suggest, however, that the test set is large enough to allow comparison of score functions with different relative weights on contact-potential and sequence-profile terms.

### Improvement in threading model accuracy

To determine the optimal weighting factor for combination of contact-potential and sequence-profile terms we conduct a series of threading experiments, aligning each target sequence in the benchmark set with the corresponding template, using a series of alternative weights. We judge the success of each fold-recognition experiment by examining the accuracy of the threading model and the statistical significance of the threading score, expressed as a  $p$ -value. The accuracy of threading models is synonymous with threading alignment accuracy, which may be measured in a number of related ways (Marchler-Bauer & Bryant, 1997b). Here we rely on contact specificity, the fraction of the contacts predicted by the threading model that

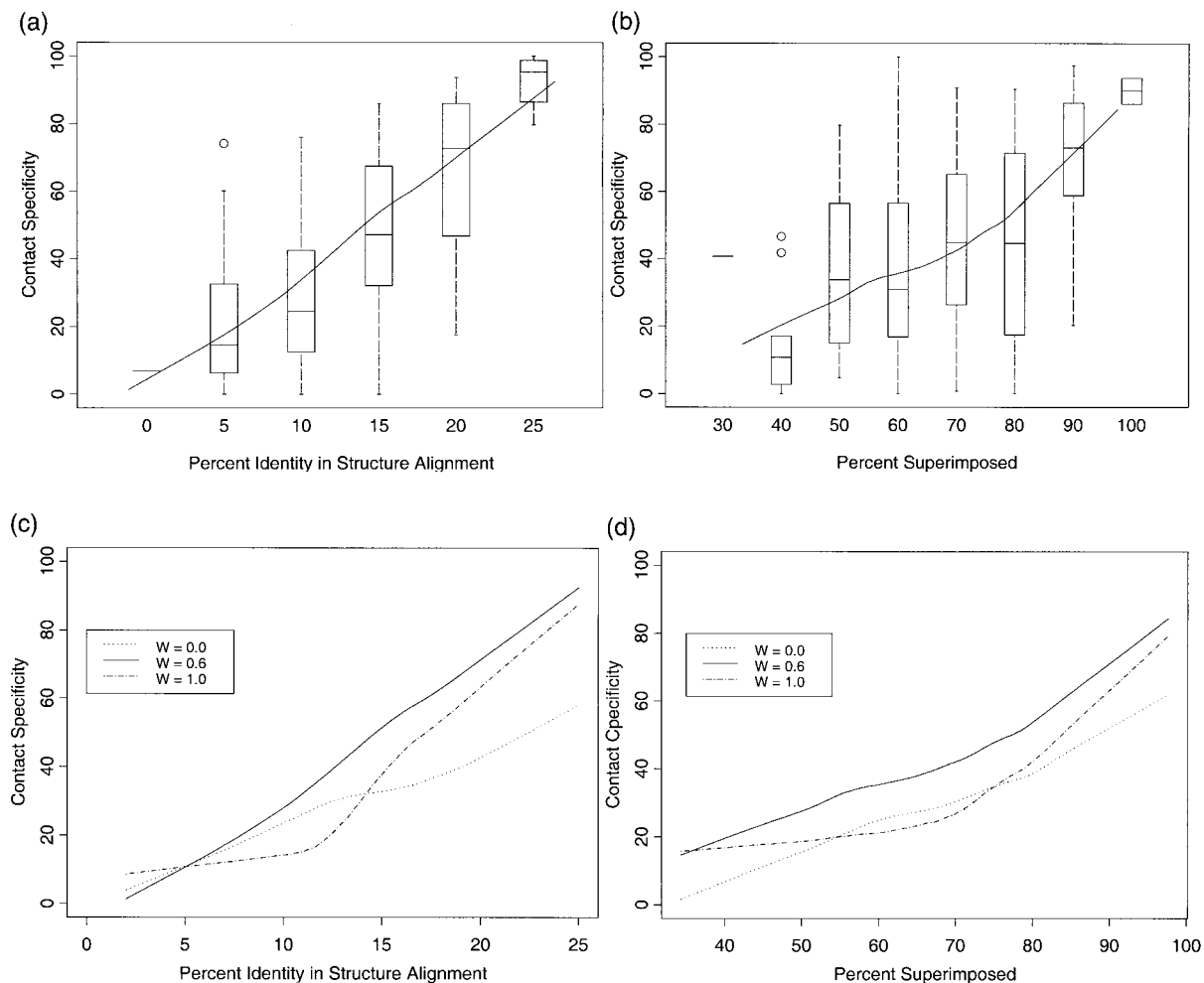
are present in the true three-dimensional structure of the target. Contact specificity is an extensive measure indicating, roughly, the fraction of the threading model that may be considered correct.

Figure 2 shows the "titration" of contact specificity versus target difficulty, for various values of the weighting factor that governs the relative contributions of the sequence-profile and contact-potential terms. As in the phase diagram in Figure 1, the indicators of target difficulty we employ are the percentage of identical residues in structure-structure alignment and the percentage of target residues superimposed on the template in that alignment. While it is apparent from the plots that model accuracy varies widely among individual targets, it is also clear that average model accuracy increases systematically as targets become less difficult, a trend noted previously (Bryant, 1996; Marchler-Bauer & Bryant, 1997a, 1999). Furthermore, for any given range of target difficulty, it is clear that there are significant differences in model accuracy as a function of the weighting factor used to combine the sequence-profile and contact-potential terms. In particular, one may see that threading alignment accuracy is improved, across the entire range of target difficulties, when sequence-profile and contact-potential terms are combined with roughly equal weight ( $w = 0.6$ ).

Tables 2 and 3 present an alternative analysis of these results, listing the number of threading models that exceed a specific accuracy threshold (contact specificity  $\geq 50\%$ ) for targets in different difficulty ranges, as defined by intervals in the percentage of identical residues or superimposable residues. One may see that the improvement in threading alignment accuracy when sequence-profile and contact-potential terms are combined can be large. The greatest difference is observed in the region of identity from 10% to 20% and percentage of superimposable residues between 60% and 80%. For the interval 10% to 15% identity, for example, the number of models surpassing the accuracy threshold with the combined potential ( $w = 0.6$ , 34% of models) is twice the number for the contact potential term alone ( $w = 0$ , 17%), or for the sequence-profile term alone ( $w = 1$ , 17%). One may also see that the number of accurate models reaches a maximum at  $w = 0.4$  or  $w = 0.6$  for the entire difficulty range of 10% to 20% identity, and 50% to 80% conserved contacts. We note that the increase in contact specificity for the combined potential is accompanied by a corresponding decrease in the RMS superposition residuals of the

**Table 2.** Percentage of models with contact specificity  $\geq 50\%$ , by target/template sequence identity and  $w$ -value

|           | $w = 0$ | $w = 0.2$ | $w = 0.4$ | $w = 0.6$ | $w = 0.8$ | $w = 1$ |
|-----------|---------|-----------|-----------|-----------|-----------|---------|
| <10 %     | 12      | 15        | 15        | 8         | 10        | 5       |
| 10 %-15 % | 17      | 31        | 34        | 34        | 23        | 17      |
| 15 %-20 % | 22      | 44        | 66        | 63        | 58        | 52      |
| 20 %-25 % | 50      | 56        | 69        | 75        | 75        | 81      |



**Figure 2.** Dependence of the contact specificity of threading models on target difficulty and the weighting factor used to combine sequence-profile and contact-potential terms. (a) and (b) Details of the experiment with roughly equal weight on the contact-potential and sequence-profile terms ( $w = 0.6$ ). Results are shown as a boxplot (Chambers, 1998) displaying the range of contact-specificity values observed for target/template pairs within a specific difficulty interval, as defined by (a) percentage sequence identity or (b) percentage superimposable residues. The central line in each box shows the median contact specificity observed for that difficulty interval. The upper and lower boundaries of the box show the upper and lower quartiles, and the vertical lines extend to a value 1.5 times the inter-quartile range. Outlier values beyond these ranges are shown as individual points. To provide a further summary of these distributions we plot a smoothed curve through the data points, using the Lowess interpolation method (Chambers, 1998). (c) and (d) Results for other values of the weighting factor combining sequence-profile and contact-potential terms. To allow comparison we show only the smoothed curves computed as in (a) and (b), for full weight on the contact potential ( $w = 0$ ), full weight on the sequence profile ( $w = 1$ ), and the combined score function with roughly equal weight on the two terms ( $w = 0.6$ ). The smoothed curves found for other values of the weighting factor ( $w = 0.2$ ,  $w = 0.4$ , and  $w = 0.8$ ) fall below the curve for  $w = 0.6$  and are omitted for clarity.

threading model *versus* the true structure of the target (not shown).

It seems logical that the combined score function should register the greatest increase in performance in the range of 10% to 20% identity and 60% to 80% superimposable residues. In this range a signal, albeit weak, might be expected from both the sequence-profile and contact-potential terms, since there is sufficient sequence and/or structure similarity. In this range, the data in Tables 2 and 3 clearly suggest that it is useful to combine these

terms. The combined score function appears less important for harder targets, with percentage identical residues below 10% or superimposed residues below 60%. Relatively few models cross the accuracy threshold, no matter how terms are combined, and it seems that neither is likely to detect any signal when sequence and structural similarity is this low. For easier targets, with percentage identical residues above 20% the combined score function also appears less important. From Figure 2 it is apparent that model accuracy does improve for

**Table 3.** Percentage of models with contact specificity  $>= 50\%$ , by percentage of target residues superimposed on template and  $w$ -value

|            | $w = 0$ | $w = 0.2$ | $w = 0.4$ | $w = 0.6$ | $w = 0.8$ | $w = 1$ |
|------------|---------|-----------|-----------|-----------|-----------|---------|
| 50%-60 %   | 10      | 17        | 23        | 23        | 21        | 17      |
| 60 %-70 %  | 12      | 23        | 31        | 29        | 17        | 17      |
| 70 %-80 %  | 24      | 41        | 51        | 47        | 41        | 35      |
| 80 %-90 %  | 30      | 50        | 40        | 60        | 70        | 50      |
| 90 %-100 % | 67      | 92        | 100       | 100       | 92        | 83      |

easy targets, but the number of models surpassing the accuracy threshold is already high for either term alone.

### Improvement in fold recognition sensitivity

Tables 4 and 5 show the number of models where the threading score exceeds a specific threshold ( $p$ -value  $\leq 0.01$ ) for targets in different difficulty ranges, and for different values of the weighting factor used to combine sequence-profile and contact-potential terms. One may see that the effect of combining these terms on threading  $p$ -values is nearly as dramatic as its effect on model accuracy. For structurally similar proteins in the range of 10% to 15% sequence identity, for example, scores cross the significance threshold for almost twice as many models when  $w = 0.6$  (49% of models), as compared to either  $w = 0$  (22%) or  $w = 1$  (26%). In this range of target difficulty it is clearly advantageous to combine sequence-profile and threading scores. Further examining Tables 4 and 5, one may see that fold recognition sensitivity reaches a maximum for the combined score with  $w = 0.6$  for all difficulty ranges below 20% sequence identity, or 90% superimposable residues. The effect is greatest, however, for targets with 10% to 15% sequence identity, and 50% to 70% superimposable residues. For both harder and easier targets combining terms has less effect on the number of models with statistically significant threading scores, presumably for the same reasons that it has less effect on model accuracy, as discussed above.

Another trend is also apparent in Tables 4 and 5: when sequence identity rises above 15%, for full weight on the sequence-profile term ( $w = 1$ ), the number of models with significant  $p$ -values rises abruptly, as it does for the combined score with  $w = 0.6$ . The sequence-profile term reliably detects easy targets with greater than 15% sequence identity, which presumably share recognizable

sequence motifs with the protein family of the template. The number of models with significant  $p$ -values for  $w = 0$  (full weight on the contact potential) increases more slowly, reaching 75% of models only for the range 90% to 100% superimposable residues. Clearly, for easy targets, the sequence-profile term greatly improves fold recognition sensitivity.

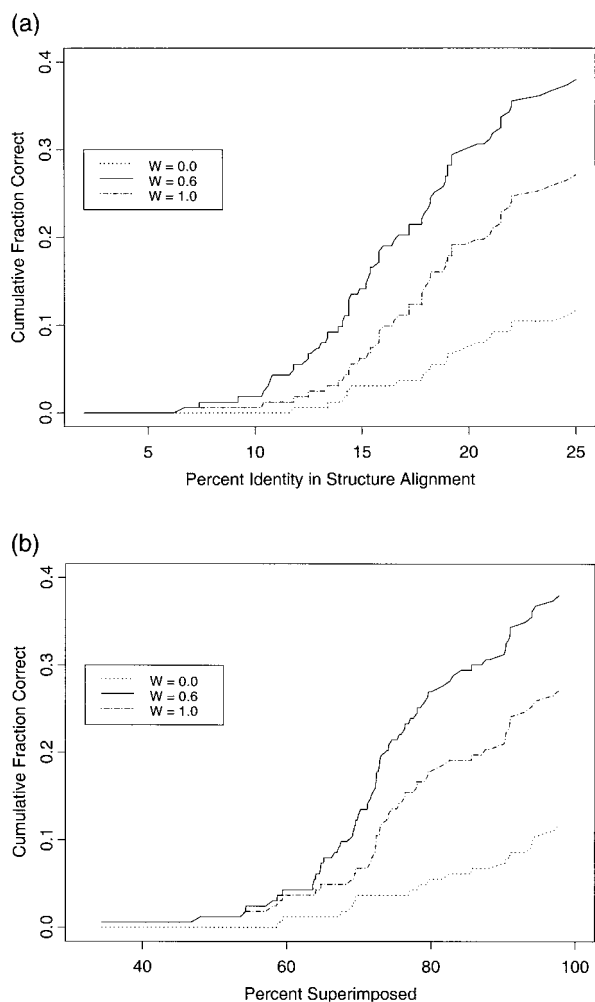
### Improvement in threading success rate

To be considered successful a threading model must identify the correct structural template and produce an accurate sequence-structure alignment. The particular thresholds one chooses are arbitrary, however, as evidenced by the alternative evaluation proposals suggested for CASP2 (Levitt, 1997; Marchler-Bauer *et al.*, 1997) and CASP3 (Marchler-Bauer & Bryant, 1999; Murzin, 1999; Lackner *et al.*, 1999). For the present control experiments we apply the thresholds mentioned above: A successful fold-recognition experiment is one where the threading  $p$ -value is 0.01 or less and the contact specificity of the threading model is 50% or more. This definition of successful recognition is similar to that suggested for CASP2 (Marchler-Bauer & Bryant, 1997a,b). We note, however, that the model accuracy threshold is somewhat more stringent, in that contact specificity of 50% or greater generally implies that more than half of target residues are aligned correctly. A  $p$ -value of 0.01 corresponds formally to 99% confidence in finding the correct template among the top five, when searching a database of 500 alternatives (Bryant & Altschul, 1995), and is roughly equivalent to the 20% confidence threshold suggested for CASP2.

Figure 3 shows the instances of successful recognition as one "titrates" by target difficulty, for scoring functions with alternative weights on the sequence-profile and contact potential terms. One may see that instances of correct recognition

**Table 4.** Percentage of target/template pairs with  $p <= 0.01$ , by target/template sequence identity and  $w$ -value

|           | $w = 0$ | $w = 0.2$ | $w = 0.4$ | $w = 0.6$ | $w = 0.8$ | $w = 1$ |
|-----------|---------|-----------|-----------|-----------|-----------|---------|
| <10 %     | 5       | 12        | 10        | 20        | 15        | 18      |
| 10 %-15 % | 22      | 29        | 46        | 49        | 48        | 26      |
| 15 %-20 % | 29      | 68        | 85        | 90        | 82        | 80      |
| 20 %-25 % | 62      | 81        | 94        | 94        | 94        | 100     |



**Figure 3.** Cumulative distribution of the fraction of target/template pairs with successful fold recognition, as a function of target difficulty, for different values of the weighting factor combining sequence-profile and contact potential terms. Fold recognition is considered successful when the threading score exceeds a threshold value ( $p$ -value  $\leq 0.01$ ) and threading model accuracy exceeds a threshold value (contact specificity  $\geq 50\%$ ). Counts of target/template pairs with successful recognition are shown as a fraction of the total (163 target/template pairs). Values are plotted against (a) percentage sequence identity in structure-structure alignment and (b) percentage superimposable residues in structure-structure alignment. The value shown at each difficulty value is that for all targets of equal or greater difficulty. For 15% sequence identity, for example, the values shown give the fraction of target/template pairs with successful recognition, among all pairs with 15% or less identity. Values at the extreme right, for 25% sequence identity or 100% superimposed residues, show the fraction of successful recognized pairs for the complete benchmark set. Traces are shown for full weight on the contact potential ( $w=0$ ), full weight on the sequence profile ( $w=1$ ), and roughly equal weight on the two terms ( $w=0.6$ ). Traces for other values of the weighting factor ( $w=0.2$ ,  $w=0.4$ , and  $w=0.8$ ) fall below the trace for  $w=0.6$ , with lesser overall performance, and are omitted for clarity.

increase as the targets become easier, as one would expect from the analyses of model accuracy and fold recognition sensitivity shown above. One may also see that with respect to instances of successful recognition, the combined scoring function ( $w=0.6$ ) clearly outperforms either the contact potential alone ( $w=0$ ) or the sequence-profile term alone ( $w=1$ ). For targets with 15% or less identical residues, for example, the number of correct models increases from about 5% (of the total of 163 models) for the sequence-profile only, to about 15% for the combined scoring function ( $w=0.6$ ), a threefold increase. On the whole, for all of the targets in the benchmark test set, 38% are correctly modeled (contact specificity  $\geq 50\%$  and  $p$ -value  $\leq 0.01$ ) with the combined scoring function ( $w=0.6$ ), as compared to 27% for the sequence-profile term alone ( $w=1$ ), and 12% for the contact potential alone ( $w=0$ ).

### Effect of evolutionary distance

None of the target/template pairs in the benchmark set shows significant sequence similarity as judged by BLAST  $e$ -value (Altschul *et al.*, 1997), and it is therefore not simple to measure relative evolutionary distances. By reference to the SCOP classification (Murzin *et al.*, 1995), however, we can identify those structurally similar pairs where there is evidence for descent from a common ancestral gene, such as functional similarity. We may then ask whether threading is more successful for these pairs than for the remaining pairs. Of the 163 target/template pairs in the benchmark set, SCOP classifies 107 as homologs, i.e. belonging to the same superfamily. The remaining 56 pairs have been termed "analogs", where descent from a common ancestral gene is uncertain, and structural similarity may indicate convergent evolution.

We find that threading with the combined score function ( $w=0.6$ ) is successful (contact specificity  $\geq 50\%$  and  $p$ -value  $\leq 0.01$ ) for 59 (55%) of the homologous pairs, but only three (5%) of the analogous pairs. By this classification, fold recognition is much more successful for homologs than analogs, consistent with previous observations (Russell *et al.*, 1998). The distinction is not absolute, however, in that recognition for a few analogous pairs crosses the relatively stringent specificity and accuracy thresholds employed here. The greater success rate for homologs may also simply reflect the titration of success rate *versus* target difficulty shown in Figure 3. As one may see by reference to Figure 1, the homologous pairs identified by SCOP tend to have a greater fraction of superimposable residues and/or greater sequence identity, and to be among the easier targets in the benchmark set. We note that while the success rate for SCOP homologous pairs was greatest with the combined score function ( $w=0.6$ ), two of three "analogous" pairs were recognized with the contact potential only ( $w=0$ ), and none with the sequence-profile only ( $w=1$ ). As noted for CASP3 results, the

**Table 5.** Percentage of target/template pairs with  $p \leq 0.01$ , by percentage of target residues superimposed on template and  $w$ -value

|            | $w = 0$ | $w = 0.2$ | $w = 0.4$ | $w = 0.6$ | $w = 0.8$ | $w = 1$ |
|------------|---------|-----------|-----------|-----------|-----------|---------|
| 50 %-60 %  | 17      | 20        | 40        | 47        | 48        | 34      |
| 60 %-70 %  | 17      | 25        | 44        | 48        | 40        | 27      |
| 70 %-80 %  | 24      | 53        | 59        | 61        | 61        | 57      |
| 80 %-90 %  | 30      | 60        | 60        | 70        | 60        | 50      |
| 90 %-100 % | 75      | 100       | 100       | 100       | 100       | 100     |

sequence-profile of the template structure may be irrelevant for fold recognition of analogs sharing no common sequence motifs (Panchenko *et al.*, 1999).

### Effect of conserved core definitions

The threading method we consider here uses a sequence-profile term computed as in the PSI-BLAST method, but it uses the core-element alignment algorithm (Bryant, 1996) rather than the BLAST alignment algorithm (Altschul *et al.*, 1997). An input to the core-element alignment algorithm is a definition of the core structure of each template, as inferred from structure-structure alignments with homologous neighbors. It is interesting to ask whether this additional information concerning evolutionary constraints on the protein family of the template contributes to specific fold recognition. To do so we may compare the numbers of targets recognized by PSI-BLAST with the number recognized by core-element threading with full weight on the sequence profile term ( $w = 1$ ), i.e. with the same score function. This comparison may also reflect other differences in the alignment algorithms: BLAST finds an optimal local alignment using affine gap penalties, for example, while threading enforces ungapped alignment with all core elements, constraining maximum loop lengths, but employing no gap penalties.

PSI-BLAST (Altschul *et al.*, 1997) identifies 33 of the target/template pairs in the benchmark set (20 %) as significantly similar ( $e$ -value  $\leq 0.01$ ; four additional pairs are identified with  $e$ -value  $\leq 1$ ). All 33 pairs are homologous, as classified by SCOP. Threading with full weight on the sequence profile term, the same score function used by PSI-BLAST, recognizes and correctly aligns 44 pairs (27 %, contact specificity  $\geq 50$  %,  $p$ -value  $\leq 0.01$ ), all of them homologous as classified by SCOP. On the whole one may conclude that the core element alignment algorithm leads to an increase in fold-recognition sensitivity, presumably due to its explicit representation of the core structure of the template. An analysis of cases where PSI-BLAST did not yield a significant score confirms that the correct structure-structure alignment involves many gaps between core elements, which are penalized by the BLAST alignment algorithm (not shown). We also note, however, that eight pairs recognized by PSI-BLAST are not recognized by threading. Threading with the combined score

function ( $w = 0.6$ ) identifies a total of 62 pairs (38 %), and clearly performs better than PSI-BLAST on this benchmark set. This analysis shows that the increased sensitivity is due partly to the use of explicit core definitions, and partly to the use of contact-potential terms in the combined score function.

### Discussion

The benchmark set we consider here is difficult by most standards for evaluation of fold-recognition or sequence-comparison methods. None of the structurally similar pairs it contains are sequence-similar, as judged by BLAST  $e$ -value or sequence identity in structural alignment. Only 20 % are recognized as similar by the PSI-BLAST sequence-profile method (Altschul *et al.*, 1997), and only 38 % are recognized as similar by the most sensitive threading method we consider. Precisely because the benchmark contains many target/template pairs at the limit of detection, however, it allows us to measure differences in the performance of alternative threading score functions. With this benchmark we can show that the optimal threading score function is one combining sequence-profile and contact-potential terms. The improvement in sensitivity is not uniform but greatest for medium-difficulty targets, with 50 % to 70 % of residues superimposable on the template and sequence identities of 10 % to 15 %. For these targets neither the sequence-profile nor contact-potential terms alone may be sufficient for sensitive recognition and accurate alignment, but their combination often is. This combined score function was recently tested at the CASP3 competition (Moult *et al.*, 1999; Murzin, 1999; Panchenko *et al.*, 1999), and the improved performance we document here seems the most likely explanation for this method's relative success in blind prediction.

The improvement we see in fold-recognition success rate derives both from improvement in fold-recognition sensitivity and improvement in alignment accuracy. Examination of the results above, in Tables 4 and 5, shows that the sequence-profile term contributes most to improvement in recognition sensitivity, particularly for easier targets. For targets with 15 % or greater sequence identity,  $p$ -values derived from the score function using contact potentials alone often fail to cross the significance threshold, but those from the sequence-profile term almost always do so. The explanation

for the improved signal from the sequence-profile term is likely to be the effectively larger "alphabet" of sequence profiles. Even a rather short sequence motif may have a relatively low probability of occurring by chance, since residues in a random sequence have roughly a one in 20 chance of matching the motif. For a motif of length  $N$ , for example, chance occurrence probabilities are of the order  $(1/20)^N$ . Contact potentials, on the other hand, are largely sensitive to the alignment of hydrophobic *versus* hydrophilic residues (Bryant & Lawrence, 1993; Park *et al.*, 1997; Yue & Dill, 1995). To the extent that this is the case, they may be understood as having the information content of an "alphabet" with only two letters. When scored by contact potentials, a hydrophobic-hydrophilic motif of length  $N$  thus has a chance occurrence probability of  $(1/2)^N$ , much higher than a sequence motif of the same length. Threading scores using contact potentials alone can be significant, but for this to occur the threading alignment must be long, containing most target and template residues, such that chance occurrence probabilities drop below the "noise" level.

It is also apparent from the above experiments that the contact-potential term contributes to the observed improvement in fold recognition success rate. Examination of Figure 2 and Tables 2 and 3 shows that the score function combining contact-potential and sequence-profile terms produces more accurate threading alignments than does either term alone. The effect is greatest for medium targets, where neither term alone may be sufficient for accurate alignment, beyond the threshold we consider (contact specificity  $\geq 50\%$ ). But the combined score function also improves alignment accuracy for easy targets, where alignment accuracy using the sequence-profile term alone may already exceed this threshold. The most likely explanation for improved alignment accuracy is the ability of the contact potential to consider non-local interactions. To score favorably with respect to the contact-potential term the threading model must, for example, bring hydrophobic residues from one region of the target sequence into contact with those from another region (Bryant & Lawrence, 1993). This effectively imposes a requirement for a globally compact model with many hydrophobic contacts (Bryant & Amzel, 1987), a requirement that cannot be well encoded by a sequence profile, and it would appear that this requirement leads to more accurate alignments. We note that alignments produced by the combined score function are also more accurate than those obtained from the contact potential alone. Threading with contact potentials may allow "register shifts" preserving the periodic hydrophobicity pattern of secondary structure elements (Bryant, 1996), but this effect appears to be counteracted by combination with sequence profiles.

Lastly, it is interesting to consider why the greatest improvement in threading success rate comes when the contact potential and sequence profile

terms are combined with roughly equal weight ( $w = 0.6$ ). The contact potential considered here was derived with considerable effort to avoid "memory" of the template sequence (Bryant & Lawrence, 1993), and one is tempted to suggest that the contact-potential and sequence-profile terms are simply independent measures of target/template complementarity, best combined with equal weight. It is obvious that a sequence profile may encode some of the information represented by the template contact matrix and contact potential, however, and *vice versa*. Either may represent a preference for a hydrophobic residue at a certain template position, for example. The explanation for the improved performance with equal weight on each term seems more likely to be the nature of the benchmark we consider here. We see the greatest improvement for medium-difficulty targets, where both terms contribute a weak signal. One may imagine that the sequence-profile term locks the alignment at any conserved motifs, and that the contact potential term enforces a requirement for a physically plausible and globally compact model. For target/template pairs with low levels of sequence and structure similarity these signals may appear independent, and it is effective to simply add them. From the experiments above, however, one can also see that greater weight on the sequence profile term is acceptable for easier targets, and greater weight on the contact potential for harder targets such as the SCOP analogs. In future it may prove possible to treat the weighting factor as a variable in the threading alignment algorithm, and to optimize its value to produce the most favorable  $p$ -value. Threading is most useful for classification of medium targets, however, and for these targets simply adding sequence-profile and contact-potential terms may continue to prove effective.

---



---

## Acknowledgments

We thank T. Madej for assistance with VAST structure alignment data and A. Schaffer for assistance with PSI-BLAST checkpoint version. We thank the NIH Intramural Research Program for support.

©2000 US Government

## References

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
- Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.* **287**, 1023-1040.
- Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366-376.
- Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching

- hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257-264.
- Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.
- Bryant, S. H. (1996). Evaluation of threading specificity and accuracy. *Proteins: Struct. Funct. Genet.* **26**, 172-185.
- Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236-244.
- Bryant, S. H. & Amzel, L. M. (1987). Correctly folded proteins make twice as many hydrophobic contacts. *Int. J. Pept. Protein Res.* **29**, 46-52.
- Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92-112.
- Chambers, J. M. (1998). *Programming with Data. A Guide to the S Language*, Springer-Verlag, New York.
- Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.
- Domingues, F. S., Koppensteiner, W. A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Floeckner, H., Lackner, P. & Sippl, M. J. (1999). Sustained performance of knowledge-based potentials in fold recognition. *Proteins: Struct. Funct. Genet. Suppl 3*, 112-120.
- Elofsson, A., Fischer, D., Rice, D. W., Le, Grand S. M. & Eisenberg, D. (1996). A study of combined structure/sequence profiles. *Fold. Des.* **1**, 451-461.
- Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.
- Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377-385.
- Godzik, A., Kolinski, A. & Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227-238.
- Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl Acad. Sci. USA*, **89**, 9029-9033.
- Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355-4358.
- Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167-180.
- Henikoff, J. G. & Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.* **12**, 135-143.
- Holm, L. & Sander, C. (1994). Parser for protein folding units. *Proteins: Struct. Funct. Genet.* **19**, 256-268.
- Hubbard, T. J. & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* **1**, 159-171.
- Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci.* **7**, 1431-1440.
- Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* **231**, 735-752.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-89.
- Jones, D. T., Tress, M., Bryson, K. & Hadley, C. (1999). Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins: Struct. Funct. Genet. Suppl 3*, 104-111.
- Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. & Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins: Struct. Funct. Genet. Suppl 3*, 121-125.
- Koretke, K. K., Russell, R. B., Copley, R. R. & Lupas, A. N. (1999). Fold recognition using sequence and secondary structure information. *Proteins: Struct. Funct. Genet. Suppl 3*, 141-148.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
- Lackner, P., Koppensteiner, W. A., Domingues, F. S. & Sippl, M. J. (1999). Automated large scale evaluation of protein structure predictions. *Proteins: Struct. Funct. Genet. Suppl. 3*, 7-14.
- Levitt, M. (1997). Competitive assessment of protein fold recognition and alignment accuracy. *Proteins: Struct. Funct. Genet. Suppl 1*, 92-104.
- Madej, T., Boguski, M. S. & Bryant, S. H. (1995a). Threading analysis suggests that the obese gene product may be a helical cytokine. *FEBS Letters*, **373**, 13-18.
- Madej, T., Gibrat, J.-F. & Bryant, S. H. (1995b). Threading a database of protein cores. *Proteins: Struct. Funct. Genet.* **23**, 356-369.
- Marchler-Bauer, A. & Bryant, S. H. (1997a). A measure of success in fold recognition. *Trends Biochem. Sci.* **22**, 236-240.
- Marchler-Bauer, A. & Bryant, S. H. (1997b). Measures of threading specificity and accuracy. *Proteins: Struct. Funct. Genet. Suppl 1*, 74-82.
- Marchler-Bauer, A. & Bryant, S. H. (1999). A measure of progress in fold recognition? *Proteins: Struct. Funct. Genet. Suppl 3*, 218-225.
- Marchler-Bauer, A., Levitt, M. & Bryant, S. H. (1997). A retrospective analysis of CASP2 threading predictions. *Proteins: Struct. Funct. Genet. Suppl. 1*, 83-91.
- Matsuo, Y. & Bryant, S. H. (1999). Identification of homologous core structures. *Proteins: Struct. Funct. Genet.* **35**, 70-79.
- Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976-4981.
- Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins: Struct. Funct. Genet. Suppl 3*, 2-6.
- Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.
- Murzin, A. G. (1999). Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Struct. Funct. Genet. Suppl 3*, 88-103.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins

- database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
- Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* **25**, 1665-1677.
- Ota, M., Kawabata, T., Kinjo, A. R. & Nishikawa, K. (1999). Cooperative approach for the protein fold recognition. *Proteins: Struct. Funct. Genet. Suppl 3*, 126-132.
- Panchenko, A., Marchler-Bauer, A. & Bryant, S. H. (1999). Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins: Struct. Funct. Genet. Suppl 3*, 133-140.
- Park, B. H., Huang, E. S. & Levitt, M. (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831-846.
- Przytycka, T., Aurora, R. & Rose, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Struct. Biol.* **6**, 672-682.
- Rost, B., Schneider, R. & Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471-480.
- Russell, R. B., Saqi, M. A., Bates, P. A., Sayle, R. A. & Sternberg, M. J. (1998). Recognition of analogous and homologous protein folds - assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng.* **11**, 1-9.
- Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.
- Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091-12095.
- Wang, Y., Address, K. J., Geer, L., Madej, T., Marchler-Bauer, A., Zimmerman, D. & Bryant, S. H. (2000). MMDB: 3D structure data in Entrez. *Nucl. Acids Res.* **28**, 243-245.
- Yi, T. M. & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.* **3**, 1315-1328.
- Yue, K. & Dill, K. A. (1995). Forces of tertiary structural organization in globular proteins. *Proc. Natl Acad. Sci. USA*, **92**, 146-150.
- Zhang, F., Basinski, M. B., Beals, J. M., Briggs, S. L., Churgay, L. M., Clawson, D. P., Zhang, X. Y., Wery, J. P. & Schevitz, R. W. (1997). Crystal structure of the obese protein leptin-E100. *Nature*, **387**, 206-209.

*Edited by B. Honig*

*(Received 27 September 1999; received in revised form 12 January 2000; accepted 18 January 2000)*