Development and Large Scale Benchmark Testing of the PROSPECTOR_3 Threading Algorithm

Jeffrey Skolnick,* Daisuke Kihara, and Yang Zhang
Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Suite 300, Buffalo, New York

ABSTRACT **This article describes the PROSPECTOR_3 threading algorithm, which combines various scoring functions designed to match structurally related target/template pairs. Each variant described was found to have a Z-score above which most identified templates have good structural (threading) alignments, Z_{struct} (Z_{good}). 'Easy' targets with accurate threading alignments are identified as single templates with $Z > Z_{good}$ or two templates, each with $Z > Z_{struct}$, having a good consensus structure in mutually aligned regions. 'Medium' targets have a pair of templates lacking a consensus structure, or a single template for which $Z_{struct} < Z < Z_{good}$. PROSPECTOR_3 was applied to a comprehensive Protein Data Bank (PDB) benchmark composed of 1491 single domain proteins, 41–200 residues long and no more than 30% identical to any threading template. Of the proteins, 878 were found to be easy targets, with 761 having a root mean square deviation (RMSD) from native of less than 6.5 Å. The average contact prediction accuracy was 46%, and on average 17.6 residue continuous fragments were predicted with RMSD values of 2.0 Å. There were 606 medium targets identified, 87% (31%) of which had good structural (threading) alignments. On average, 9.1 residue, continuous fragments with RMSD of 2.5 Å were predicted. Combining easy and medium sets, 63% (91%) of the targets had good threading (structural) alignments compared to native; the average target/template sequence identity was 22%. Only nine targets lacked matched templates. Moreover, PROSPECTOR_3 consistently outperforms PSI-BLAST. Similar results were predicted for open reading frames (ORFS) ≤200 residues in the *M. genitalium*, *E. coli* and *S. cerevisiae* genomes. Thus, progress has been made in identification of weakly homologous/analogous proteins, with very high alignment coverage, both in a comprehensive PDB benchmark as well as in genomes. Proteins 2004;56:502–518.**

© 2004 Wiley-Liss, Inc.

Key words: protein structure prediction; fold recognition; structural alignment; weakly homologous/analogous proteins; *M. genitalium*; *E. coli*; *S. cerevisiae*; genomes

INTRODUCTION

A key goal in this postgenomic era is to assign the function of genes and gene products in all genomes.[1–5] Here, the corresponding three-dimensional native structures of the associated proteins can play an important role via the sequence-to-structure-to-function paradigm.[2,6–11] This idea, along with the goal of delineating all possible protein folds, has provided the impetus for structural genomics.[2,9–12] To make this process as efficient as possible, it is important to establish whether the target protein adopts an already known fold or if it is likely to adopt a novel fold. In this respect, the development of approaches that not only recognize homologous proteins (*viz.* comparative modeling) but also analogous proteins (proteins that adopt the same topology but at the very least are evolutionarily distant) are very important.[13] One might expect that such analogous predicted structures would be of substantially lower resolution than those obtained from comparative modeling.[14–16] It is important to ask whether such low-resolution structures have any predictive value. Over the last few years, we have demonstrated that, even if the resulting predicted structures are of low resolution with a backbone root mean square deviation (RMSD) from native of 4–6 Å, they can often be used to predict the biochemical function of the protein of interest if the protein is an enzyme.[6,13,17,18] Furthermore, if there is a known ligand, these low-resolution structures can be used to identify the ligand-binding site in about two-thirds of the cases.[19] Thus, it is important to develop methods that extend threading to treat proteins of low sequence identity to the template. Motivated by this goal, in this paper we describe our threading algorithm, PROSPECTOR_3, which extends the applicability and reliability of these methods and provides templates (where appropriate), as well as predicting contacts for use in subsequent fold assembly/refinement.

At present, there are two basic approaches to assess whether a new protein sequence, the target, matches a known fold, the template. Sequence-based approaches are designed to establish an evolutionary relationship between the target and template proteins.[4,14,20–25] Because protein structure is better conserved than protein function, if an evolutionary relationship exists between two proteins, then it is possible can exploit this fact to assign the structure of the target sequence.[26,27] Over the last few

*Correspondence to: Jeffrey Skolnick, Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Suite 300, Buffalo, NY 14203. E-mail: skolnick@buffalo.edu.

Received 11 August 2003; Revised 7 November 2003; Accepted 6 January 2004

Published online 28 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20106

© 2004 WILEY-LISS, INC.

years, sequence-based methods have used not just single sequences but multiple sequence alignments to define sequence families.[28] These approaches construct a sequence profile by pooling sequences identified on successive iterations, as in PSIBLAST.[20] More recent approaches employ sequence profile–profile comparisons that compare the sequence conservation patterns of the target and template sequences.[29] Another powerful class of algorithms is the Hidden Markov Models (HMMs).[24,25,28,30,31] For example, Pfam HMMs are optimized to recognize family members with a small false positive rate, but they my miss more distantly related sequences.[23,24,32] SAM T-99 represents another powerful class of HMMs.[25]

Threading algorithms also attempt to identify related template structures, but unlike sequence-only approaches, they can also include structure-based information, such as secondary structure, burial patterns, and/or side chain pair (or higher order) interactions.[33–37] The goal of threading is to identify proteins that adopt similar structures, whether they are evolutionarily related or not; in its purest form, no evolutionary information is used in the comparison.[38−41] Threading should be able recognize not only homologous proteins but also analogous folds.[13] CASP5 found a number of convincing cases in which threading has begun to demonstrate its ability to recognize such analogous folds.[13] Finally, after over a decade of development, there are a number of threading approaches that significantly outperform sequence-only approaches. These include the PROSPECT algorithm of Xu and coworkers,[35] the GENTHREADER algorithm of Jones and coworkers[36,42] and PROSPECTOR.[13] CASP5 also demonstrated the power of metaservers that combine consensus information from a variety of threading and sequence based servers to make more accurate structural predictions.[43,44] While metaservers perform almost as well as the best human predictors, their success rests on the success of the individual input servers. Thus, there is a significant impetus to improve the individual fold recognition algorithms. In this spirit, we describe the benchmarking and application of a more recent version of our threading algorithm, PROSPECTOR_3, an earlier version of which was a key factor in our performance in CASP5.[13]

A key issue for the success of any threading algorithm is the completeness of the library of solved structures in the protein data bank (PDB).[45] If a related structure of a target sequence is not already solved, then fold recognition algorithms will not work. Recently, we demonstrated that at the level of low to moderate resolution protein structures, the PDB is essentially complete for single domain proteins.[46] For example, low to moderate resolution proteins of 100 residues or fewer have significant coverage, even by proteins from a different secondary structure class (by far the worst case scenario), with an average backbone Cα RMSD from native of 3.8 Å that covers 86% of the target protein. Furthermore, protein structure space is very dense. For larger proteins, non-related proteins cover a significant portion of their structure, with different top hit proteins aligned to different regions; the top ten hit proteins can give 90% coverage for proteins up to 320 residues in length. Thus, in principle, a perfect threading algorithm should be able to assign most, if not all, single domain proteins to a template. At worst, it might identify the correct fold, but with alignment errors; at best, it should be able to provide significant models for use in subsequent refinement.

An outline of this article is as follows. In the Material and Methods section, we present the improved PROSPECTOR_3 threading algorithm, which includes a cascade of various scoring functions. Then, in the Results section, we describe our algorithm's ability to successfully identify analogous/weakly homologous templates in a representative set of PDB structures composed of single domain proteins 200 residues or smaller that are no more than 30% identical to proteins in the threading structural template library; on average, the sequence identity of the target sequence to the assigned templates was 22%. We present an analysis of fold assignments, the percent of residues assigned to structures and the ability to identify reliable alignments. We then summarize a comparison of PROSPECTOR_3 to PSIBLAST.[20−22] Then, the application of PROSPECTOR_3 to the set of ORFS 200 residues or smaller in the *M. genitalium*,[47] *E. coli*[48] and *S. cerevisiae*[1] genomes is reported. In the Conclusions section, we highlight the important results and suggest directions for future research.

MATERIALS AND METHODS

As described above, we recently demonstrated that the PDB is complete at the level of low resolution of single domain protein structures (even if only nonhomologous proteins are considered).[46] Thus, a good threading algorithm should be able to detect these analogous proteins – in other words, given a set of unrelated sequences that fold to a single domain protein, all (or more realistically almost all) should be assigned to templates. In practice, because of deficiencies in threading algorithms, some of these template structures are not identified, but one might imagine that, with a combination of different scoring functions, we should be successful in assigning a significant number of them; this idea provides the motivation for our current approach to threading.

Overview of Methodology

A schematic overview of PROSPECTOR 3.0 is shown in Figure 1. All alignments were generated using a Needleman–Wunsch type of global alignment algorithm.[49] PROSPECTOR 3.0 is an iterative threading approach consisting of close (distant) sequence profiles that, for each structure, generate the probe-template alignment (first pass, $m = 1$) to be used in the evaluation of the pair interactions in the second through fourth ($m = 2–4$) passes. A total of 20 structures (the top five structures for each of the two $m = 1$, close and distant sequence profile scoring functions, plus the top five structures from the two mth pass scoring functions) were used to generate predicted contacts and subsequent protein specific potentials[50] to be used for the $m + 1^{st}$ pass. At the end of the fourth pass, predicted contacts, continuous local fragments, and, if Z exceeded

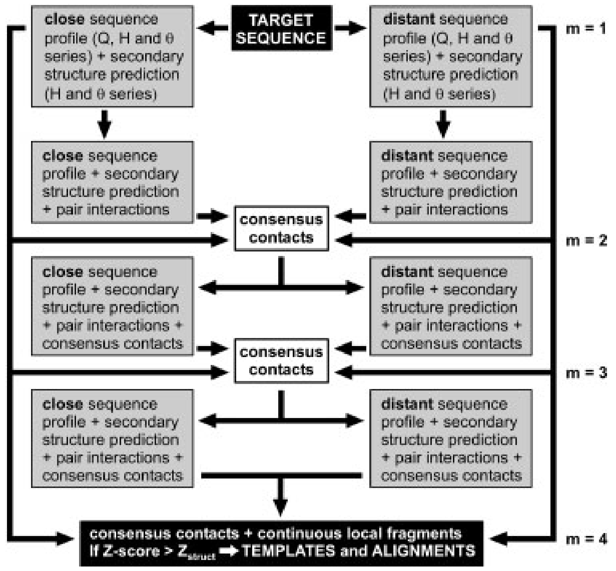
close sequence profile (Q, H and θ series) + secondary structure prediction (H and θ series)
TARGET SEQUENCE
distant sequence profile (Q, H and θ series) + secondary structure prediction (H and θ series)
m = 1
close sequence profile + secondary structure prediction + pair interactions
consensus contacts
distant sequence profile + secondary structure prediction + pair interactions
m = 2
close sequence profile + secondary structure prediction + pair interactions + consensus contacts
consensus contacts
distant sequence profile + secondary structure prediction + pair interactions + consensus contacts
m = 3
close sequence profile + secondary structure prediction + pair interactions + consensus contacts
distant sequence profile + secondary structure prediction + pair interactions + consensus contacts
consensus contacts + continuous local fragments
If Z-score > Z_struct ➡ TEMPLATES and ALIGNMENTS
m = 4

Fig. 1. Flow chart of the PROSPECTOR_3ϑ series of threading algorithms. Close and distant sequence profiles were independently employed (along with predicted secondary structure in the H and θ variants) to generate target sequence alignments to a template to identify the partners to be used in subsequent evaluation of pair interactions. Top scoring alignments from the close, distant profiles ($m = 1$, first pass), and a corresponding set of pair interactions plus secondary structure predictions ($m = 2$) were used to generate predicted consensus contacts to be used in the third ($m = 3$) and fourth ($m = 4$) iterations of threading. The resulting algorithm provides: (1) consensus predicted side chain contacts, (2) predicted continuous local fragments and (3) a predicted template, if the Z-score of the target template match is greater than Z_{struct}. In more than 90% of the cases, the predicted template has a good structural alignment.

Z_{struct}, predicted templates and corresponding alignments were developed.

In order to assign a good structural template to as many target sequences as possible, a variety of pair potentials and scoring approaches are used,[50] the idea being that different scoring functions might be able to assign different target sequences to templates. This is somewhat different from the idea of metaservers that attempt to identify consensus regions assigned by multiple algorithms.[43,44] In practice, there are three classes of pair potentials employed – the Quasichemical based pair potential, indicated by Q;[51] the protein specific, local sequence fragment based, orientation independent pair potential, indicated by H;[50] and the protein specific, local sequence fragment based, orientation dependent pair potential indicated by θ.[50]

There are two ways to evaluate the score significance of a target in a template. In the first, we employ the ideas of the original PROSPECTOR algorithm[37] and evaluate the score significance in terms of the Z-score of the sequence mounted in a given structure defined, in general, by

$$Z = (E - \langle E \rangle)/\mathrm{sqrt}(\langle E^2 \rangle - \langle E \rangle^2) \qquad (1)$$

the quantity in ⟨⟩ denotes the average of the best alignment given by dynamic programming over the template library, and E is the score or energy. This generates 3 versions of PROSPECTOR_3 – PROSPECTOR_3Q, PROSPECTOR_3H and PROSPECTOR_3θ. We also follow the idea of Karplus,[30] and for each target template alignment and corresponding scoring function, the score is evaluated as the energy difference between the best score of the target sequence aligned to the template and the reversed sequence aligned to the template. This is much faster than the more traditional method of evaluating the score relative to that of the randomized sequence. As shown below, this improves the sensitivity of the threading algorithm. For those cases where the score is relative to that of the reversed sequence, we identify each set of scoring functions by 'rev,' thereby, generating the three versions of PROSPECTOR_3 – PROSPECTOR_3Q_{rev}, PROSPECTOR_3H_{rev} and PROSPECTOR_3θ_{rev}. Thus, in total, there are six different versions of PROSPECTOR_3.

In what follows, we establish the Z-score cutoff, Z_{struct}, above which templates with good structural alignments are confidently identified (above 95% of the target/template pairs have a maximum RMSD to native of 6.5 Å on the basis of the best structural alignment, whose average coverage is 72%[46]). This turns out to be a Z-score of 7 (5) for the rev (non rev) series. We also identify the Z-score threshold, Z_{good}, above which the alignment has a RMSD from native below 6.5 Å for the coordinate alignment obtained from threading, (hereafter called 'the threading-based alignment') in more than 81% of the cases. (On the basis of the PDB benchmark described in the results section, $Z_{\mathrm{good}} = 15$ for the rev series, and $Z_{\mathrm{good}} = 10$, 10 and 8.2 for the Q, H and θ series respectively).

The Template Library

Because PROSPECTOR_3 is an iterative process, in which consensus side chain contacts are assumed to have predictive value and are used in subsequent threading iterations, it is important that the template library be sufficiently diverse. Otherwise, the results will be spurious. Our template library was constructed as follows: the entire PDB was clustered into representative families, no pair of which had less than 35% global sequence identity to other members of the family, nor greater than 35% identity to members outside the family. A randomly chosen member of this family was selected as a representative. This protein was added to the template library if one of these two conditions held: (1) it had no more than 35% sequence identity over the aligned regions to any template or (2) it had no more than 35% global sequence identity (ratio of the number of identical residues to the number of residues in the target sequence). As of February 2003, there were 3575 templates in our template library. The list may be found at http://www.bioinformatics.buffalo.edu/threading/LIST.templates.

Generation of Sequence Profiles

Our sequence database was a combination of the SWISSPROT (http://www.expasy.ch/sprot/)[52] and the KEGG genome sequence databases (ftp://kegg.genome.ad.jp/genomes/genes).[53] FASTA[54,55] was employed to select sequences with a sequence identity to the target sequence between 35% and 90%, as well as all sequences with an E-value to the target sequence of less than 10. The former comprise the 'close' and the latter the 'distant' set of sequences whose alignment to the target sequence is generated from CLUSTALW.[56] The goal here was to have sets of sequences that spanned the regime from closely related proteins to distantly related sequences. (We note that we also tried PSIBLAST[20] to generate the distant sequence set; unfortunately, unlike with FASTA, sometimes highly spurious results are generated when PSI-BLAST converges to a completely unrelated family, while other times the distant profiles using PSIBLAST are more sensitive than those from FASTA. Given the instability of the PSIBLST results, we opted to use FASTA.)

The sequence profile for the ith position in the probe sequence for amino acid type γ is very simple and is given by

$$P^{\varphi}(\gamma, i) = 5 \sum_{\ell=1}^{N_{\phi}} B(\gamma, a_{i\ell})/N_{\varphi} \tag{2a}$$

where we use the shorthand notation

$$\varphi = \begin{Bmatrix} \text{close} \\ \text{distant} \end{Bmatrix} \tag{2b}$$

Here N_{φ} is the number of sequences in the 'close' set or 'distant' set. $B(\gamma,\eta)$ is the BLOSUM 62[57,58] mutation matrix for amino acid residues of type γ and η (Other mutation matrices were tried, but BLOSUM62[57,58] worked best) and $a_{i\ell}$ is the amino acid at position i in the lth sequence of the set.

Secondary Structure Propensities

Next, we considered a term to evaluate the consistency of the secondary structure prediction (helix, coil, beta) of the target obtained from PSIPRED[59] with that observed in the template structure,

$$S_i^{\varphi}(\Theta_i, \Theta_{tem,J}) = 3\delta_{\Theta_i,\Theta_{tem,J}} \tag{3}$$

with $\delta_{\kappa\lambda}$ the representing Kronecker delta (=1 if $\kappa = \lambda$ and 0 otherwise). Θ_i and $\Theta_{\text{tem},J}$ are the predicted target sequence secondary structure at residue i and the actual secondary structure of the Jth residue in the template, respectively.

First Pass Scoring Function

The score matrix associated with aligning residue i with the Jth residue in the Kth structure for PROSPECTOR_3Q and PROSPECTOR_3Q_{rev} is

$$S_{K,Q}^{\varphi,1}(i, j) = P^{\varphi}(a_{JK}, i) \tag{4a}$$

while for PROSPECTOR_3ϖ ($\varpi = H,\theta$) and PROSPECTOR_3ϖ_{rev}, the corresponding scoring function is:

$$S_{K,\varpi}^{\varphi,1}(i, J) = P^{\varphi}(a_{JK}, i) + S_i^{\varphi}(\Theta_i, \Theta_{tem,J}) \tag{4b}$$

Pair Interactions

The next step is to use the alignment provided by the $m = 1$ first pass scoring profile to generate the partners in the evaluation of the pair potentials. For PROSPECTOR_3Q and PROSPECTOR_3Q_{rev}, the homology-averaged, orientation-independent pair potential for the second pass, $m = 2$, is given by

$$E_Q^{\varphi,2}(i, j) = -5 \sum_{\ell=1}^{N_{\varphi}} \varepsilon(a_{i\ell}, a_{j\ell})/N_{\varphi} \tag{5a}$$

where $E_Q{}^{\varphi,2}(i,j)$ is the arithmetic average over the N_{φ} sequences of the quasichemical pair potentials symbol for episol (8.n) which describe interactions between contacting side chains of amino acid types γ,η (that is, they have at least one pair of side chain heavy atoms within 4.5 Å of each other).[51] φ is defined in eq. (2b). If there is a gap in the alignment, then the pair potential is assigned a value of zero. The minus sign arises because we want to maximize the score; that is, gap penalties are negative.

For PROSPECTOR_3H and PROSPECTOR_3H_{rev}, the homology-averaged, orientation-independent pair potential for the second pass, $m = 2$, is given by

$$E_H^{\varphi,2}(i, j) = -2.5 \sum_{\ell=1}^{N_{\varphi}} \varepsilon(a_{i\ell}, a_{j\ell})/N_{\varphi} - 0.5\varepsilon_H(i, j) \tag{5b}$$

Where $\varepsilon_H(i,j)$ is the local sequence fragment based pair potential derived previously.[50] For PROSPECTOR_3θ and PROSPECTOR_3θ_{rev}, we employ the analogous protein specific, side chain orientation dependent pair potential defined by

$$E_{\theta}^{\varphi,2}(i, j, \vartheta) = -0.5 * E_H^{\varphi,2}(i, j) - 0.2 * E_{\vartheta}^{\varphi}(i, j, \theta) \tag{5c}$$

where $E_{\vartheta}{}^{\varphi}(i,j,\theta)$ is the protein specific, side chain orientation dependent pair potential derived previously.[50] ϑ is divided into three bins for antiparallel (>120°), parallel (<60°), and acute orientations (between 60 and 120°) of the pair of vectors, each from the side chain center of mass to the Cα.

For iterations $m = 3$ and 4, we add a potential based on contacts predicted from the previous iteration as follows. For a given iteration, if a contact is present in a minimum of four of the top 20 scoring structures, each with a minimum Z-score greater than 1.3 (5 each from the close/distant sets from the $m = 1$ pass, plus 5 each from the close/distant mth pass that includes pair interactions), this constitutes a subset of the predicted contacts. To further expand their number, we also examined the two adjacent residues to i and j, and included their contacts if they had a favorable BLOSUM 62[57] (positive) mutation matrix value. Let the resulting total number of contacts between residues i and j be $q_{ij}{}^{m}$ associated with iteration $m = 2$, 3 or 4, then the contact-based pair potential for these positions where there is a predicted contact for scoring function type $\varpi = Q, H$, or θ is[37]

$$V_{\varpi}^{m}(i, j) = -\ln(q_{ij}^{m}/q_{ij}^{0m}) \tag{6a}$$

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

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

Here, we used the convention that

$$V_{\varpi}^{1}(i, j) = 0 \tag{6c}$$

Here, n is the number of residues in the target sequence.

For those pairs of positions where no consensus contacts were found, depending on which scoring function was used, we found different terms to be optimal. For $\varpi = Q$, if residues i and j are not in contact, then for $m = 3\text{-}4$,

$$V_{Q}^{m}(i, j) = 0 \tag{6d}$$

The situation is somewhat more complicated if $\varpi = H$ or θ; then, for $m = 2$,

$$V_{\varpi}^{2}(i, j) = E_{\varpi}^{\text{near},2}(i, j) \tag{6e}$$

with $E_H{}^{\text{near},2}(i,j)$ given by eq. (5b) and $E_\theta{}^{\text{near},2}(i,j)$ by eq. (5c) respectively.

While for $m = 3$,

$$V_{H}^{3}(i, j) = -E_{H}^{\text{near},3}(i, j)/5 \tag{6f}$$

and

$$V_{\theta}^{3}(i, j) = -E_{\theta}^{\text{near},3}(i, j) \tag{6g}$$

where $E_H{}^{\text{near},3}(i,j)$ is defined below in eq. (8a).

Third and Fourth Iteration Pair Potentials

For PROSPECTOR_3Q and PROSPECTOR_3Q$_{\text{rev}}$, the homology-averaged, orientation-independent pair potential for the third and fourth pass ($m = 3$ and 4 respectively) is given by

$$E_{Q}^{\varphi,m}(i, j) = 0.5 * (E_{Q}^{\varphi\text{second}}(i, j) - V_{Q}^{m-1}(i, j)) \tag{7}$$

where $V_Q{}^{m-1}(i,j)$ is given by eq. (6a) or (6d) where appropriate.

For PROSPECTOR_3H and PROSPECTOR_3H$_{\text{rev}}$, the homology-averaged, orientation-independent pair potentials for the third ($m = 3$) and fourth ($m = 4$) passes are given by

$$E_{H}^{\varphi,3}(i, j) = 0.5 * E_{Q}^{\varphi,2}(i, j) - 0.25 * (V_{H}^{2}(i, j) + \varepsilon_{H}(i, j)) \tag{8a}$$

where $V_H{}^{2}(i,j)$ is given by eq. (6a) or (6e) where appropriate. Similarly,

$$E_{H}^{\varphi,4}(i, j) = 0.5 * E_{Q}^{\varphi,2}(i, j) - 0.25 * (V_{H}^{2}(i, j) + V_{H}^{3}(i, j)) \tag{8b}$$

where $V_H{}^{2}(i,j)$ and $V_H{}^{3}(i,j)$ are defined by eq. (6a), or (6e) and (6f) respectively, where appropriate.

For PROSPECTOR_3θ and PROSPECTOR_3θ$_{\text{rev}}$, we employ the analogous protein specific, side chain orientation dependent pair potential that is given by (for $m = 3, 4$)

$$E_{\theta}^{\varphi,m}(i, j, \vartheta) = 0.5 * E_{Q}^{\varphi,2}(i, j) - 0.2 * E_{\vartheta}^{\text{second}}(i, j, \theta) - 0.4 * V_{\theta}^{m-1}(i, j) \tag{8c}$$

with $V_\theta{}^{m-1}(i,j)$ given by eqs. (6a), (6e), or (6g) where appropriate.

TABLE I. Summary of Gap Initiation/Propagation Parameters

	Prospector_3Q series		Prospector_3H/θ series	
Scoring	Open[a]	Propagation[a]	Open[a]	Propagation[a]
Close profile	−10	−1.2	−8.0	−1.2
Second-fourth pass	−10	−0.4	−10	−1
Distant profile	−8.0	−1.2	−8.0	−0.6
Second-fourth pass	−10	−0.8	−8.0	−0.6

[a] "Open" and "Propagation" refer to the gap open and propagation parameters, respectively.

Second–Fourth Pass Scoring Matrix

Let $A_{K,\varpi}{}^{\phi}(J)$ be the alignment between the J^{th} residue in the K^{th} structure and the target sequence generated by the φth sequence profile alignment for first pass scoring function of type ϖ [see eqs. (4a,b)]. Then, the matrix, $S_K{}^{\varphi,m}(i,J)$ associated with aligning the ith residue in the target sequence with the Jth position in the K^{th} structure for iteration $m = 2–4$ is

$$S_{K,\varpi}^{\varphi,m}(i, J) = S_{K,\varpi}^{\varphi,1}(i, J) + \sum_{m=1}^{nc_K(J)} E_{\varpi}^{\varphi,m}\{i, A_{1K,\varpi}^{\phi}[C_{JK}(m)]\} \tag{9}$$

The values for $E_\varpi{}^{\varphi,m}$ are given by eq. (5–8c) where appropriate. The term $nc_K(J)$ represents the number of contacts the J^{th} residue makes in structure K, and $C_{JK}(m)$ is the identity of the m^{th} contact partner that residue J makes in structure K.

Gap Penalties and End Effects

To allow for better domain identification, there were no gap penalties before the beginning and after the end of the aligned regions. The gap penalties for the various scoring functions are summarized in Table I.

The Easy Set

We also established measures independent of knowledge of the native state to identify which target sequences were accurately aligned to their templates; these target sequences were designated the 'easy' set. For a given version of PROSPECTOR_3ϖ, with $Z > Z_{\text{struct}}$, we identified target sequences with structurally similar, aligned regions in two or more templates (local RMSD <5 Å between identical aligned residues in the top two Z-score templates.). These were termed 'consensus regions,' and for the rev (non rev) series generated good alignments in more than 95% of the cases. We added to this set target sequences matched to a single template with $Z > Z_{\text{good}}$. The resulting set of sequences, ϖ-easy, for a given version of PROSPECTOR_3ϖ, provided template/alignment candidates for the easy set. Since different versions of PROSPEC-

TOR_3ϖ have different accuracies (see Table IIA, below), the easy set (along with associated templates, template alignments, contact predictions and fragment predictions) was constructed by taking all members of the θ_{rev}-easy set, then all members of the H_{rev}-easy set not included in the θ_{rev}-easy set, then all remaining members of the θ-easy set, then all remaining members of the *H*-easy set, then members of the Q_{rev}-easy set, and finally all remaining members of the *Q*-easy set. While this order was constructed on the basis of empirical results from the PDB benchmark, it was subsequently applied to all cases.

The Medium Set

Those proteins having single templates with $Z_{\mathrm{struct}} < Z < Z_{\mathrm{good}}$ were designated the medium set because they had, on average, good structural alignments to the template but sometimes poor threading-predicted alignments. Also included in this group were target sequences that had at least two templates with Z-scores $> Z_{\mathrm{struct}}$, but for which the threading-based alignments lacked structural consensus regions. Due to the iterative nature of the algorithm, sometimes there is one good template and one bad template. This can lead to ambiguous contact predictions. The medium set was constructed as follows. First, all members of the easy set were eliminated (e.g. if a target was a member of at least one ϑ-easy set, then it was assigned to the easy set, even if it was a medium target according to other versions of PROSPECTOR_3ϑ). The medium set was constructed by taking all members of the θ_{rev}-medium set, then all members of the H_{rev}-medium set not included in θ_{rev}-medium, then all remaining members of the θ-medium set, then all remaining members of the *H*-medium set, then additional members of the Q_{rev}-medium set, and finally the remaining members of the *Q*-medium set.

The Hard Set

Finally, there were proteins that could not be assigned to any template whatsoever by any of the scoring functions of PROSPECTOR_3ϑ; these are referred to as the hard set in what follows.

RESULTS AND DISCUSSION

Selection of Benchmark Proteins

To assess the ability of PROSPECTOR_3ϑ to identify analogous templates, we constructed a subset of the representative template library composed of proteins between 41 and 200 residues that form compact tertiary structures and are not coiled coils. (In the absence of the former condition, we often correctly identified domains, but their mutual orientation was wrong; there was also significant misassignment of coiled coils.) We removed all proteins from the template library that had 30% or greater sequence identity to the target sequence, as calculated over the set of aligned residues. These restrictions yielded a representative set of 1491 proteins for testing. The list of these may be found on our web site at http://www.bioinformatics.buffalo.edu/services/threading/LIST.benchmark. Of these, 29% are α-proteins, 32% are β-proteins and 35% are

TABLE IIA. Summary of Results for Top Five Templates for Different Versions of PROSPECTOR_3ω[a]

Criterion	Original[c]	*Q*	Q_{rev}	*H*	H_{rev}	θ	θ_{rev}	Secondary[e]
$z > z_{\mathrm{struct}}$	781/5.0/0.88	1472/4.8/0.58	1291/5.6/0.68	920/5.2/0.88	1045/5.6/0.87	962/5.6/0.87	1122/6.2/0.87	607/4.3/0.89
Structural alignments[b]	764/2.6/0.85	1324/3.3/0.70	1179/3.2/0.71	885/2.9/0.81	983/2.9/0.80	917/3.0/0.77	1033/3.1/0.78	589/2.6/0.87
RMSD < 6.5	(98%)	(90%)	(91%)	(96%)	(94%)	(95%)	(92%)	(97%)
RMSD < 2 Å	57/1.6/0.95	164/1.4/0.56	116/1.5/0.73	137/1.6/0.91	145/1.6/0.92	132/1.6/0.91	147/1.6/0.91	90/1.6/0.93
RMSD < 3 Å	203/2.3/0.93	448/2.1/0.64	345/2.2/0.78	366/2.1/0.91	389/2.1/0.90	365/2.1/0.91	402/2.2/0.90	257/2.2/0.92
RMSD < 4 Å	368/2.8/0.92	672/2.6/0.64	543/2.7/0.78	531/2.5/0.89	581/2.6/0.88	537/2.6/0.89	595/2.6/0.88	377/2.6/0.91
RMSD < 5 Å	493/3.2/0.91	894/3.1/0.63	704/3.1/0.75	631/2.8/0.89	699/2.9/0.87	641/2.9/0.88	706/2.9/0.88	460/2.9/0.91
RMSD < 6 Å	597/3.6/0.90	1099/3.5/0.61	873/3.6/0.74	712/3.1/0.86	774/3.1/0.86	711/3.1/0.87	779/3.1/0.87	502/3.2/0.90
RMSD < 6.5 Å[d]	627/3.8/0.90	1205/3.8/0.60	944/3.7/0.72	724/3.6/0.89	789/3.6/0.88	724/3.5/0.89	791/3.2/0.87	513/3.2/0.90
	(80%)	(82%)	(73%)	(79%)	(76%)	(75%)	(70%)	(85%)

[a] Listed in columns 2–8 are the numbers of target proteins whose best of the top five templates has an RMSD below the threshold specified in column 1, the average RMSD and the average coverage.

[b] The top *Z*-scoring template is taken for structural comparison with the target sequence's native structure; the numbers in parentheses indicate the fraction of targets whose templates have a best structural alignment less than 6.5 Å calculated by SAL.[46]

[c] Values from original version of PROSPECTOR;[37] related to PROSPECTOR_3Q but with full strength gap penalties at the beginning and ends of the alignments.

[d] Fraction of targets with at least one template having an RMSD below 6.5 Å indicated in parentheses.

[e] Calculated using eq. (4b) for both close and distant sequence profiles.

TABLE IIB. Summary of Results for Top Five Templates for Easy and Medium Sets[a]

Summary criterion	East set	H_{rev}: distant profile[d]	Medium set	H_{rev}: distant profile[d]
Total	878/4.1/0.85	878/4.9/0.86	606/9.6/0.60	606/12.1/0.65
Structural alignments RMSD < 6.5[b]	846/2.8/0.83	830/2.8/0.83	502/4.3/0.56	490/505/0.58
RMSD < 2 Å	151/1.6/0.89	119/1.6/0.90	47/1.2/0.24	16/1.2/0.59
RMSD < 3 Å	401/2.1/0.89	332/2.2/0.9	77/1.7/0.28	27/1.7/0.62
RMSD < 4 Å	588/2.6/0.87	495/2.6/0.9	102/2.1/0.30	38/2.3/0.61
RMSD < 5 Å	688/2.9/0.87	599/2.9/0.89	134/2.7/0.35	56/3.0/0.65
RMSD < 6 Å	753/3.1/0.86	677/3.2/0.88	167/3.3/0.4	74/3.6/0.66
RMSD < 6.5 Å	760/3.1/0.86	698/3.3/0.87	191/3.7/0.4	89/4.0/0.65
% < 6.5 Å[c]	87%	81%	31%	15%

[a] Listed in columns 2–4 are the numbers of target proteins whose best of top five Z-score templates has an RMSD below the threshold specified in column 1, the average RMSD and the average coverage.
[b] The top Z-scoring template is taken for structural comparison with the target sequence's native structure.
[c] Fraction of targets with assigned templates at least one of which has an RMSD below 6.5 Å.
[d] Obtained from PROSPECTOR_3H$_{\text{rev}}$ distant profile $m = 1$ pass top scoring templates.

α/β-proteins, with the remaining having little if any secondary structure. This set represents a total of 440 different CATH[26] numbers (counting the first 3 of 4 CATH digits). Considering only CATH numbers starting from 1–4 reduces the total to 290. Among these, the α-class has 78 different topologies, the β-class has 65 different topologies, and the α/β-class has 118 different topologies.

Comparison of Results from Various Versions of PROSPECTOR

In Table IIA, we present a summary of the results from the various versions of PROSPECTOR, including the original version.[37] The entries in columns 2–8 show the number of target sequences that satisfy the criterion given by the corresponding entry in column 1, as well as the average RMSD that satisfies this criterion and the average coverage (fraction of residues aligned relative to the target protein's length). Comparing the original version of PROSPECTOR,[37] which mainly matches the easiest cases of proteins to their templates (column 2), to that of PROSPECTOR_3*Q* (column 3), using the same version but with zero gap penalty at the beginning and end of the alignment, it is clear that roughly twice the number of good target/template matches are found with PROSPECTOR_3*Q*, but at the cost of significantly smaller alignments. If the goal is to identify good regions for as many templates as possible, then reducing the gap penalty on the ends is useful. If, however, one wants as many proteins as possible with essentially comparable coverage to the original version of PROSPCTOR, then PROSPECTOR_3*H* (column 5) and PROSPECTOR_3θ (column 7) are clearly superior to the original PROSPECTOR (with PROSPECTOR_3θ giving a slightly better RMSD distribution than PROSPECTOR_*H* for most RMSD values below 6.5 Å), with each assigning an additional 97 targets to templates.

If further improvement is desired, then for the H and θ series of PROSPECTOR_3, the rev series clearly performs better than the set where the score of the sequence in the template alone is used to assess significance (compare columns 5-6, 7-8). Comparing the H_{rev} and θ_{rev} series, the θ_{rev} series performs slightly better over all RMSD ranges and would be the one to use, if one had to choose a single variant of PROSPECTOR_3 with the highest average coverage and accuracy. The Q and Q_{rev} series (columns 3-4) behave differently in the sense that the Q_{rev} set has higher average coverage, but fewer cases with acceptable (low) RMSDs.

Another issue is how much the performance of PROSPECTOR_3 is enhanced by the inclusion of pair potentials relative just using sequence profiles and predicted secondary structure. In column 9 of Table IIA, we present the results for this situation, calculated using close and distant sequence profiles [see eq. (4b)]. Clearly, for all RMSD thresholds, the number of assigned target proteins is significantly smaller than when any of the pair potentials are used. At worst, using PROSPECTOR_3*H*, for a RMSD threshold of 6.5 Å, an additional 211 targets have acceptable templates. For a RMSD threshold of 4 Å, at worst, again using PROSPECTOR_3*H*, an additional 154 targets have acceptable templates. This clearly shows that the improvement of using pair potentials over sequence profiles plus secondary structure predictions is highly significant. Furthermore, comparing these results with the full suite of PROSPECTOR_3ϑ algorithms as well as with other fold recognition algorithms (see Table VI below), use of secondary structure without pair interactions is significantly worse than use of the full easy set of PROSPECTOR_3ϑ, (indeed only 5 proteins would be added to the easy set), with somewhat better performance than PROSPECT,[35] but worse than SAM-T99[28] and SPARK.[60] Thus, we conclude that it is the presence of pair interactions, enhanced by the predicted contacts, that are responsible for the relatively good performance of PROSPECTOR_3ϑ.

Structural Alignment of Test Template

As shown in the last row of Table IIA, the percentage of targets that have good RMSD values is less when the rev type of scoring function is used, but in an absolute sense the rev type correctly assigns more targets to templates. Certainly no less than 70% of the targets in all cases have a good template in the top five selected templates (ranked on the basis of their Z-scores), but we would like to increase this rate of success. Techniques that do this will be described below (namely the construction of the 'easy' set), but first we address issues of how good the structural alignment is between the target and the top Z-scoring template. Here, we employ our recently developed structural alignment algorithm SAL.[46]

The second row of Table IIA shows that, in more than 90% of the cases, there is good structural alignment, with the percent of residues aligned on the basis of SAL tracking the trend of the average coverage from the given variant of threading. Reflecting the fact that the original version of PROSPECTOR[37] (column 2) recognizes the most easy-to-identify target/template pairs, its best structural alignment also has the highest coverage. In all cases, the fraction of target/best template pairs having good structural alignment is more than 90% on average for all variants of PROSPECTOR. Moreover, on average, 15–20% more target/templates have good structural alignments than threading-based alignments (compare row 3 with row 9), but the average structural alignment coverage is about 10% less for the templates from Q_{rev}, H, H_{rev}, θ and θ_{rev} threading. As shown in the next section, this reflects the fact that the alignment errors in the templates identified by PROSPECTOR_3ϑ responsible for the high threading based RMSD to native often arise from the N- and/or C-terminal fragments having native secondary structure but different orientations from that in the native state (similar errors are seen for the mutual orientation of domains). On the basis of secondary structure, they are similar, but better structural alignments can be obtained by truncating these regions. In the next section, this effect is shown in a number of representative examples.

The Q and Q_{rev} series are qualitatively different in that the average coverage of the best of the top five templates is either the same or lower than the best structural alignment. This reflects the fact that for the more difficult cases (unrecognized by the other threading algorithms), these versions can identify good local regions but not global regions (note that these algorithms use an entirely different type of secondary structure prediction scheme with no structural information used in the first pass; thus, sometimes they can pull locally similar regions, probably from evolutionary information alone). In such cases, the structural alignment algorithms can then recognize large, similar regions.

Analysis of Apparent Errors for Best Template with $Z > Z_{good}$

We consider here a representative set of cases taken from the PROSPECTOR_3H_{rev} results, because they are typical. Considering those proteins that are aligned to a single template with Z-score $> Z_{good}$, of which there are 76 with 27% average sequence identity between target and template; 62 of the 76 have acceptable threading-based alignments with RMSD values < 6.5 Å, with an average RMSD value of 3.6 Å and 91% coverage. However, most of the 14 apparently poorly predicted target proteins actually have well predicted regions. Indeed, 73 of the 76 proteins have an average RMSD of 2.0 Å and 74% coverage, with a minimum of 20 residues aligned. The problem is to detect the well-predicted regions within those 11 proteins that have a relatively poor global RMSD, but good RMSD over a significant fraction of the alignment. Ideally, the threading algorithm itself should eliminate the poorly predicted regions, but in practice sometimes it doesn't. What is happening? Are there general trends observed for these unsuccessful cases?

The first of the 14 cases with a relatively high global RMSD is the 1an7A (136 residues in length)/1iu6A, S8 ribosomal protein/electron transport protein target/template pair which has a Z-score of 16.6 and a global RMSD of 7.4 Å over the 121 aligned residues, but a RMSD of 1.9 Å over 89 residues. The native structure has two domains; each domain has a good RMSD to native but their mutual orientation is different from native. This is the source of the relatively high global RMSD. In fact, the N-terminal 68 residues have a RMSD from native of 2.2 Å. Similarly, the RMSD of residues 94–136 is 3.7 Å. The superposition of each of these domains onto its native structure is shown in Figure 2(a), in the upper and lower sets of superimposed structures respectively.

The next target/template pair examined was 1baq_ (139 residues in length)/1eyvA. Both are transcription termination factors. They are matched with a Z-score of 30.5, with the former an obsolete entry, and a global RMSD of 10.8 Å over the 124 aligned residues, 19 residues of which have a RMSD of 3.5 Å. While both are helical proteins, there are significant differences in their topologies that may reflect problems with the 1baq_ entry.

The 1bmqB (88 residues in length)/1cp3A, the interleukin 1beta converting enzyme/tetrapeptide inhibitor, target/template pair is matched with a Z-score of 15.9. It has a global RMSD of 6.9 Å over the 84 aligned residues, with 79 residues aligned with a RMSD of 1.5 Å. The source of this error is a large gap between residues 2 and 3, due to a break in the chain in the PDB file of 1cp3A. If residue 2 is eliminated, then the RMSD is 3.7 Å over the remaining 83 residues. The resulting superposition is shown in Figure 2(b).

The next problematic pair is 1c20A (128 residues)/1bmy_, DNA binding domain from the dead ringer protein/DNA binding protein, which are matched with a Z-score of 19.9, and have a RMSD of 10.7 Å over the 98 aligned residues. Here, the N-terminal helices differ in orientation. Also, the C-terminal helix has a kink in the native structure, whereas the corresponding helix is straight in the template. If the central 57 residues are aligned, then the corresponding RMSD is 3.8 Å. Again, there are subtleties in the orientation that are not fully captured by threading, but the core is well described. The resulting superposition is shown in Figure 2(c).

The next pair is 1f2rI (100 residues in length)/1d4bA, DNA binding protein/ apoptosis protein, which is matched with a Z-score of 19.4, with 97 residues aligned with a RMSD of 8.8 Å, but 44 residues can be aligned with a RMSD of 3.0 Å. This particular pair has very long, poorly aligned beta strands on the N-terminus that do not interact with the core of the protein, with smaller errors in the C-terminal strands. If both are excised, then the

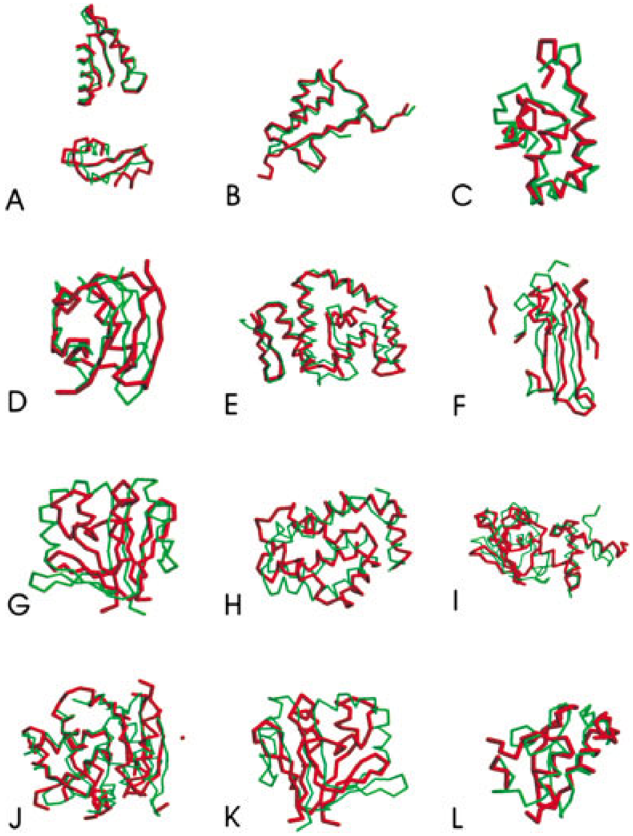

Fig. 2. Selected target/template alignments from PROSPECTOR_3H$_{rev}$ of the 14 target (green)/templates (red) for which $Z > Z_{good}$, but the global RMSD of the alignment was greater than 6.5 Å. (a) 1an7A/1i6uA target/template pair. The N-terminal 68 residues (upper) have a RMSD from native of 2.2 Å. Similarly, the RMSD of residues 94–136 (lower) is 3.7 Å. The superposition of each domain is shown. (b) 1bmqB/1cp3A pair without residue 2. The RMSD is 3.7 Å over 83 residues. The resulting superposition is shown. (c) 1c20A/1bmy_ target/template pair. The central 57 residues are aligned, and then the corresponding RMSD is 3.8 Å. The resulting superposition is shown. Again, there are subtleties in the orientation that are not fully captured by threading, but with the core well described. (d) 1f2rI/1d4bA target/template pair. Eliminating the N/C termini, the RMSD over the remaining 70 atoms is 3.5 Å. The resulting superposition is shown. (e) 1fm2A/1ai4A target/template pair. Eliminating the C-terminal helix, then the resulting RMSD is 6.4 Å over 122 residues. The resulting superposition is shown. (f) 1hpwA/1dzoA target/template pair. Eliminating the C-terminal β-strand as well as the N-terminus, and a hairpin between residues 40–48, the RMSD is 6.6 Å over 83 residues. The resulting superposition is shown. (g) 1hqi_/1ckv_, target/template pair has a RMSD of 7.3 Å over its 89 aligned residues. There is a 36-residue region with a RMSD of 2.9 Å. As shown, the global topologies of the two proteins are absolutely identical with relatively small shifts between corresponding secondary structural elements. (h) 1k3kA/1af3_, target/template pair shown without the N terminal helix. The resulting RMSD is 4.6 Å over 106 residues. (i) 1k5xD/1c05A target/template pair has a RMSD of 8.7 Å over its 136 aligned residues, with a 66 residue fragment having a RMSD of 2.4 Å. The errors in the orientation of the unattached N-terminal helix/strand not associated with interactions with the core. Also, there are some shifts associated with the C-terminal strand. Eliminating these regions gives the 66-residue fragment described above. The full alignment is shown. (j) 1mgtA/1eh6A target/template pair has a RMSD of 7.5 Å for its 151 aligned residues and a RMSD of 1.6 Å for 97 residues. The target and template structures differ somewhat in the N-terminus, with one helix replaced by a beta strand. The full-length alignment is shown. (k) 1qb2A/1hq1A target/template pair has a RMSD of 9.4 Å for its 71 aligned residues, and for 38 of these, the RMSD is 1.1 Å with 40% sequence identity. Here, the two proteins differ by the location of a non-interacting N-terminal helix. If these residues are eliminated, then the resulting RMSD is 1.5 Å over 58 residues. The full length alignment is shown. (l). The 1qbhA/1f3hA target/template pair. After removing the N/C terminal fragments, as shown, the RMSD is 5.2 Å over 65 residues.

RMSD over the remaining 70 atoms is 3.5 Å. The resulting superposition is shown in Figure 2(d).

The 1fm2A (151 residues in length)/1ai4A, cephalosporin acylase/penicillin acylase pair was matched with a Z-score of 23.0 and a RMSD of 12.0 Å over the 143 aligned residues, with 17 residues aligned with a RMSD of 1.8 Å. The major difference is in the location of the C-terminal helix that has a significant gap in the alignment to the template structure. (The secondary structure is correct, but the location is incorrect). If this C-terminal helix (which is not interacting in the native structure with the protein core) is eliminated, then the resulting RMSD is 6.4 Å over 122 residues. The topology is completely correct, with minor shifts in secondary structure. The resulting superposition is shown in Figure 2(e).

The 1hpwA (129 residues in length)/1dzoA/ contractile protein, pilin/truncated pilin pair has a Z-score of 19.9, and a RMSD of 7.4 Å over its 110 aligned residues, with 44 residues aligned with a RMSD of 3.4 Å. Here, the major difference between the two structures is at the C-terminal β-strand as well as in the N-terminal residues at the end of a helix. There is also a difference in geometry of a hairpin between residues 40 and 48, but overall the fold is the same. If these regions are deleted, the RMSD is 6.6 Å over 83 residues. The resulting superposition is shown in Figure 2(f).

The match of 1hqi_(90 residues in length)/1ckv_, phenol-hydroxylase/hydroxylase regulatory protein pair has a Z-score of 18.9, with a RMSD of 7.3 Å over its 89 aligned residues. There is a 36-residue region with a RMSD of 2.9 Å. As shown in Figure 2(g), the global topologies of the two proteins are absolutely identical, with relatively small shifts between corresponding secondary structural elements.

We next consider the target/template pair, 1k3kA (146 residues in length)/1af3_, apoptosis protein, BCL-2 homolog from Kaposi's Sarcoma/ BCL-XL apoptosis inhibitory protein, which is matched with a Z-score of 19.3 and a RMSD of 7.5 Å over the 129 aligned residues. There is an 83-residue fragment that aligns with a RMSD of 3.0 Å. The predicted template alignment has an alignment error associated with the N-terminal helix. If this region is eliminated, as shown in Figure 2(h), the resulting RMSD is 4.6 Å over 106 residues.

Another example is 1k5xD (178 residues in length)/1c05A, 80S Ribosomal protein/the S4 Delta 41 ribosomal protein, target/template pair, which is matched with a Z-score of 18.0 and a RMSD of 8.7 Å over its 136 aligned residues, with a 66 residue fragment having a RMSD of 2.4 Å. Similar to many previous cases described above, the error is in the orientation of the unattached N-terminal helix/strand not associated with interactions with the core. Also, there are some shifts associated with the C-terminal strand. Eliminating these regions gives the 66-residue fragment described above. In Figure 2(i), we show the full alignment in order to illustrate these issues and the fact that the global topology of the target and template are clearly related.

The next example is the 1mgtA (169 residues in length)/1eh6A, methylguanine-DNA methyl transferase/alkylguanine-DNA methyl transferase, target/template pair, which is matched with a Z-score of 26 and a RMSD of 7.5 Å for its 151 aligned residues. It has a RMSD of 1.6 Å over 97 residues, where the pairwise sequence identity between target and template is 42%. The target and template structures differ somewhat at the N-terminus, with one helix replaced by a beta strand. The full-length alignment is shown in Figure 2(j).

Another pair is 1qb2A (106 residues in length)/1hq1A, signal recognition particle/RNA signaling protein, which match with a Z-score of 17.1 and a RMSD of 9.4 Å for 71 aligned residues. For 38 of these, the RMSD is 1.1 Å with 40% sequence identity. Here, the two proteins differ by the location of a non-interacting N-terminal helix. If these residues are eliminated, then the resulting RMSD is 1.5 Å over 58 residues. The full 71-residue alignment is shown in Figure 2(k).

The 1qbhA (101 residues in length)/1f3hA, apoptosis inhibitor/antiapoptotic protein, target/template pair is matched with a Z-score of 22.9 and a RMSD of 8.9 Å for its aligned 100 residues; of these, there is a 24 residue fragment with a RMSD of 3.1 Å. The prediction here repeats the pattern of errors in the N/C terminal secondary structural elements whose secondary structure type is essentially correct, but whose orientations are in error. If we remove the N/C terminal fragments, the RMSD is 5.2 Å over 65 residues, as shown in Figure 2(l).

Finally, the 1ehdA (88 residues in length)/1baa_, isolectin/endochitinase, pair is matched with a Z-score of 55! However, the RMSD between the template and native is 12.1 Å over the 76 template aligned residues. The global folds of both proteins are entirely different. All versions of PROSPECTOR_3ϑ match 1ehdA to 1baa_, and both the close and distant sequence profiles also assign this particular target/template pair. Unfortunately, while a plausible structural alignment can be made, it appears that this prediction is incorrect, but we have been unable to identify its cause.

In summary, of the 14 proteins that do not have good alignments from PROSPECTOR_3H$_{rev}$, all but one of them are in fact highly significant, with many errors resulting either from the misprediction or misorientation of N- or C-terminal secondary structural elements whose secondary structure itself is correctly predicted, the misprediction of the mutual orientation of two domains, or, in rare cases, insertions in loops [see Fig. 2(f)]. Algorithms that readjust the orientation of such fragments could in principle extend the length of the accurate regions as well as refine the alignment. This forms the basis of the TASSER algorithm that is described elsewhere.[13]

Properties of The Easy Set

Based on the above considerations and the fact that the PROSPECTOR_3ϑ algorithm usually provides templates with reasonable accuracy when $Z > Z_{\text{struct}}$, we the proceeded to construct the easy set. On the basis of the performance of the PROSPECTOR_3ϑ variants, we took the templates in the following order as described in the Materials and Methods section: θ_{rev}, H_{rev}, θ, H, Q_{rev} and $Q$the K^{th} template structure. . The results are compiled in column 2 of Table IIB. There are a total of 878 easy targets with an average global pairwise sequence of identity between the target and template proteins of 22%. Of these, 761 have an acceptable RMSD in at most the top five templates. Furthermore, relative to the last row of Table IIA, 87% of the templates in the easy set have RMSD from native below 6.5 Å, which is a significant improvement in the percentage of good templates selected and average coverage over that if any one of the PROSPECTOR_3ϑ methods alone is used (see Table IIA, row 10). Thus, the easy target selection procedure generates an enriched set of good target/template pairs.

TABLE III. Summary of Side Chain Contact Prediction Results for the Easy and Medium Sets

Criterion	Easy set	Medium set
$\delta = 0^{a}$	0.46	0.19
$\delta = \pm 1^{a}$	0.69	0.42
$\delta = \pm 2^{a}$	0.80	0.57
$f^{b,c}$	2.4	0.90
Contact order^{c}	32.0	19.0

[a] Average fraction of contacts predicted within $\delta = \pm m$ residues of a native contact.

[b] Ratio of the average number of predicted contacts/number of residues in the protein.

[c] Average contact order, i.e., the average residue spacing between predicted contacts.

Does the full methodology impart any advantages over simply using the most sensitive of the sequence profile methods? In column 3 of Table IIB, we compare the results for the easy set with those obtained just using the most sensitive of sequence profiles, the distant sequence profiles from the PROSPECTOR_3H$_{rev}$ distant profile, $m = 1$ pass. Over all RMSD ranges, the entire threading protocol generates significantly better results over all RMSD ranges, at the cost of marginally lower average coverage. This clearly shows that the full methodology is better than the simple sequence profiles (not to mention that many of the target/template pairs do not have a significant Z-score to confidently match the target to the template based on the PROSPECTOR_3H$_{rev}$ distant profile, $m = 1$ pass). Here, we just use those targets identified as belonging to the easy set and use the same number of top distant sequence profile templates as are assigned in the easy set. Overall, on the basis of the number of targets identified, the average RMSD and the average coverage, improvement over using the full methodology is evident.

Contact prediction accuracy

The next question we address is the average accuracy of the predicted contacts for the easy targets. Table III shows that the average side chain contact prediction accuracy is 46%, with an average predicted number of contacts per residue of 2.4 and an average contact order of 32 residues. Furthermore, the average accuracy of contacts predicted within ±1 residue is 70%. Thus, even though the average

TABLE IVA. Average RMSD and Length Distribution of Continuous Fragments in Easy Set Templates[a]

Criterion	Fraction of fragments	Average RMSD	Average length	Relative RMSD average Z-score
RMSD < 1 Å	0.27	0.62	11.8	−2.83
RMSD < 2 Å	0.6	1.1	15.6	−2.81
RMSD < 3 Å	0.81	1.5	16.5	−2.61
RMSD < 4 Å	0.92	1.7	17.0	−2.49
RMSD < 5 Å	0.96	1.8	17.2	−2.43
RMSD < 6 Å	0.98	1.9	17.3	−2.40

[a] Calculated for all continuous alignments five residues or longer in length.

TABLE IVB. Average RMSD and Length Distribution of Continuous Fragments in Medium Set PROSPECTOR_3H$_{rev}$ Distant Profile m = 1 Pass Alignments[a]

Criterion	Fraction of fragments	Average RMSD	Average length	Relative RMSD average Z-score
RMSD < 1 Å	0.21	0.50	8.8	−2.64
RMSD < 2 Å	0.40	0.98	8.8	−2.17
RMSD < 3 Å	0.62	1.5	8.8	−1.63
RMSD < 4 Å	0.85	2.1	8.9	−1.15
RMSD < 5 Å	0.96	2.3	9.0	−0.98
RMSD < 6 Å	0.98	2.4	9.1	−0.93

[a] Calculated for all continuous alignments between 7 and 12 residues in length; a size regime chosen because the fragments are most prevalent in this length regime.

sequence identity is low (22%), they tend to have consistent consensus alignments that can be used to extract predicted side chain contacts of acceptable accuracy.

Accuracy of local fragments

The next issue we explore is the accuracy of the local fragments predicted for the easy set. If we focus on continuous fragments of at least five residues in length, then the average continuous fragment predicted has a length of 17.6 residues, with an average RMSD of 2.0 Å from native. Moreover, on average, about 87% of a target chain is covered by a minimum of at least one continuous fragment that is at least five residues in length. The resulting cumulative RMSD distribution is shown in Table VIA. We also include an estimate of the statistical significance of the set of selected fragments relative to a random collection of PDB fragments. This can be assessed by the average Z-score of the relative RMSD (which would be zero for a random set).[61] Interestingly, more than 90% of the selected fragments have a RMSD from native of ≤4 Å with acceptable statistical significance. This would again suggest that a fragment assembly algorithm that employs these reasonably long fragments might be successful in improving the initial alignments. This idea forms the basis of the TASSER assembly algorithm, which in general results in a systematic improvement in the threading based alignments, especially for these easy set proteins.[62]

Properties of the Medium Set

As summarized in Table IIB, there are 606 proteins assigned to the medium set. Of these, 505 (87%) targets have acceptable top scoring templates with good structural alignments and 56% average coverage. Roughly 31% of these have good threading based alignments, but here the average coverage is only 40%. The PROSPECTOR_3H$_{rev}$ distant profile, m = 1 pass, distant sequence profile provides only 15% of targets having the top five templates with a RMSD below 6.5 Å, but the average coverage is 65%. (Actually, if we examine those proteins of higher coverage, essentially the same number is found in the medium set but the coverage is better by about 12% for the PROSPECTOR_3H$_{rev}$ distant profile, m = 1 pass provided alignments.) Thus, while reasonable templates have been identified in most cases, often the predicted alignment needs improvement.

Contact prediction accuracy

Reflecting the shorter than average coverage and lower than average accuracy of the medium set compared to the easy set, as shown in Table III, column 3, the average contact prediction accuracy from the templates is 19%, with many fewer contacts predicted (0.9/residue) and a much lower contact order of 19 residues. Thus, techniques need to be developed that improve the quality of the alignments of the medium set proteins. We do note that neural network approaches to contact prediction give about 14–16% accuracy.[63]

Accuracy of local fragments

We found that there are too few local fragments provided by the medium set templates to generate significant results. Interestingly, results of comparable average accuracy but greater coverage are provided by the PROSPECTOR_3H$_{rev}$ distant profile, m = 1 pass set of alignments. Thus, for the medium targets we used the top scoring templates from the PROSPECTOR_3H$_{rev}$ distant profile. The resulting cumulative RMSD distribution, average RMSD, length and relative RMSD Z-score are summarized in Table IVB. (This is in contrast to the contact predictions that are on average worse by about 5% using the PROSPECTOR_3H_{rev} distant profile, $m = 1$ pass set of alignments as compared to the medium set templates). Here, the average length of the continuous aligned fragments is about nine residues, with the most accurate non-trivial predictions found for fragments between seven and 12 residues in length. These have an average alignment length of 9.1 residues, an average RMSD of 2.5 Å, and on average 79% of residues belonging to a minimum of at least one continuous fragment. In contrast to the easy set, the average accuracy of the continuous fragments at a given RMSD threshold is less. What clearly distinguishes the easy from the medium set is the fact that, in the easy set, these continuous fragments are longer and cover more of the molecule. PROSPECTOR_3 works by stringing together these continuous fragment alignments. When such fragments are longer and cover more of the molecule, high Z-scores and confident fold identifications result; when this is not the case, there are still significant predicted chunks that could be used in fragment assembly.

Composite Results of the Easy/Medium Sets

Combining the easy and medium sets, 63% (927/1482) of the benchmark targets have an acceptable template on the basis of the PROSPECTOR_3 provided alignment (with a RMSD below 6.5 Å and over 80% average coverage). Furthermore, 91% (1348/1482) of the proteins have a good structural alignment with a RMSD below 6.5 Å with 72% average coverage. These results are consistent with the observation that the PDB is a covering set of single-domain proteins. Thus, PROSPECTOR fails to identify related folds for about 10% of the target sequences, and the alignment accuracy needs to be improved for one-third of the targets. Interestingly, there are only nine proteins in the hard set for which no global template is predicted.

A summary of all results, templates and alignments may be found on our website for the easy and medium set results at http://www.bioinformatics.buffalo.edu/threadingbenchmark/easy and http://www.bioinformatics.buffalo.edu/threadingbenchmark/medium respectively.

Comparison to Results of PSIBLAST

If PSIBLAST[20] is run against our template library and all targets with an E-value less than 0.01 are selected, then 487/1491 targets satisfy this criterion. A total of 462 target proteins in the PSIBLAST alignment have a RMSD below 6.5 Å. As shown in Table V, column 3, the average coverage of these proteins is only 77%. In contrast, in the easy set alone, 761 proteins with good global alignments were selected. Of the 487 targets identified by PSIBLAST, 448 belong to the easy set and 11 to the medium set. If we select the more accurate regions of the PROSPECTOR_3ϑ alignments (see Table V, column 2) that are obtained from regions of the PROSPECTOR_3ϑ alignments having a RMSD to native of less than 5 Å, then the coverage is essentially the same as in PSIBLAST (see Table V, column 3), but the accuracy of the PROSPECTOR_3ϑ alignments is significantly better over all RMSD ranges. Thus, on the basis of both accuracy and fold recognition ability, consistent with the results of CASP5, we conclude that the PROSPECTOR_3 belongs to the new generation of improved threading algorithms.[13,35]

TABLE V. Comparison of Targets Identified Assigned by PSIBLAST to Easy/Medium Sets

Selection criterion	Easy/medium sets[a,b]	PSIBLAST results[b]
RMSD < 1 Å	88/1.2/0.82	18/0.81/0.64
RMSD < 2 Å	383/1.8/0.7	170/1.5/0.77
RMSD < 3 Å	454/1.9/0.75	348/2.0/0.78
RMSD < 4 Å	459/2.0/0.75	414/2.2/0.78
RMSD < 5 Å	459/2.0/0.75	436/2.3/0.77
RMSD < 6 Å	459/2.0/0.75	453/2.5/0.77
RMSD < 6.5 Å	459/2.0/0.75	462/2.5/0.77
% < 6.5 Å[b]		95%

[a]Listed in columns 2–3 are the numbers of target proteins whose best of top two templates has an RMSD below the threshold specified in column 1, the average RMSD and the average coverage. Listed in columns 2 are the results from the easy/medium proteins whose targets have an E-value less than or equal to 0.01 as identified by PSIBLAST and those residues have a local RMSD below 5 Å.

[b]Percent of targets that with E-value < 0.01 that have an RMSD < 6.5 Å.

Comparison to Other State-of-the-Art Fold Recognition/Sequence Based Methods

We next compare the results of PROSPECTOR_3ϑ with those of other state-of-the-art methods. To make the comparison truly informative, each of the methods must be compared for the same set of targets against the same template library. Otherwise details of the fold library (completeness, presence of homologs, fold space coverage) will be confused with the actual performance of the method itself. Such a consistent comparison is not possible using EVA[64] or LIVEBENCH,[65] in which different template libraries are used for different methodologies. It is our intention to include PROSPECTOR_3ϑ in LIVEBENCH. To comprehensively and fairly benchmark PROSPECTOR_3ϑ against other state-of-the-art algorithms, we have run one of the best threading algorithms (PROSPECT),[35] one of the best of the sequence methods (SAM-T99),[25] and a newly developed fold recognition algorithm that uses sequence profiles and secondary structure (SPARK)[60] for 1482 targets against the same template library as was used in the evaluation of PROSPECTOR_3ϑ.

There remains the issue of what cut-off of target/template score significance should be used for PROSPECT, SAM-T99 and SPARK. In any threading/sequence method, an acceptable accuracy criterion for fold identification/alignment quality must be specified. Here, for both SAM-T99 and SPARK, we opted for a cutoff that gives roughly 79% accuracy in the sense that the best of the top five structures has a RMSD less than 6.5 Å. The problem with PROSPECT is that this will identify too few templates. Thus, we reduced the Z-score cutoff to a value for which 69% accuracy is obtained. This was compared against the easy set from PROSPECTOR_3ϑ, for which we attained 87% accuracy for good fold identification.

The resulting comparison against the easy set of protein results from PROSPECTOR_3ϑ is shown in Table VI, with

TABLE VI. Comparison of Easy Targets with PROSPECT, SAM-T99, and SPARK

Selection criterion	Easy[a]	PROSPECT[b]	SAM-T99[c]	SPARK[d]
Total	877/4.1/0.85	607/5.0/0.91	765/5.0/0.86	734/5.2/0.90
RMSD < 2 Å	150/1.5/0.89	112/2.3/0.94	74/1.6/0.87	43/1.6/0.95
RMSD < 3 Å	391/2.1/0.89	222/2.9/0.93	228/2.3/0.87	154/2.3/0.94
RMSD < 4 Å	582/2.6/0.86	311/3.3/0.92	383/2.8/0.88	319/2.9/0.92
RMSD < 5 Å	689/2.9/0.86	311/3.3/0.92	506/3.2/0.87	453/3.4/0.91
RMSD < 6 Å	751/3.1/0.86	397/3.8/0.92	581/3.5/0.87	532/3.7/0.91
RMSD < 6.5 Å	761/3.1/0.86	418/3.9/0.92	602/3.6/0.87	554/3.8/0.91
% < 6.5 Å[d]	87%	69%	79%	79%

[a] For PROSPECTOR_3ϖ, easy targets were considered.

[b] For PROSPECT,[35] of 1482 targets, up to the top five target template pairs with a Z-score > 3 were considered.

[c] For SAM-T99,[28] of 1482 targets, up to the top five target template pairs with a Z-score > 9.5 were considered.

[d] For SPARK,[60] of 1482 targets, up to the top five target template pairs with a Z-score > 6.6 are considered. Percent of targets satisfying the selection criteria that have an RMSD < 6.5 Å.

the corresponding Z-score cut-offs shown in the footnote. On the basis of the number of templates identified at a given RMSD threshold and the overall accuracy, we concluded that PROSPECTOR_3ϑ performs the best. It is followed by SAM-T99, then SPARK and then PROSPECT. PROSPECT does, however, give slightly higher coverage than the alternative approaches. Indeed, at a RMSD threshold of 4 Å, PROSPECTOR_3ϑ identified 199 more target/template pairs than did the closest competing method, SAM-T99. At a RMSD threshold of 6.5 Å, 159 more good target/template pairs were identified compared to SAM-T99. Of course, because in principle all targets actually have good templates in the template library,[46] the fact that we still fail to identify roughly half of these templates as belonging to the easy set indicates that all methods need significant improvement.

There is an interesting point that emerges from this analysis. Different threading algorithms recognize different target/template pairs in the high confidence regime (viz. easy proteins). Thus, we should apply each threading algorithm to identify easy targets using the same methodology as that used in PROSPECTOR_3ϑ and build a composite algorithm that basically tiles fold space using the union of all easy targets. We are now in the process of incorporating such an approach to identify accurate templates in our TASSER[62] protein fragment assembly/refinement approach.

Application to Small ORFs in the *M. genitalium*, *E. coli* and S. cerevisiae Genomes

To examine whether or not the results presented here for the PDB benchmark also hold for the small ORFS (200 residues or smaller) in genomes, we next applied the PROSPECTOR_3 methodology to the *M. genitalium*,[47] *E. coli*,[48] and *S. cerevisiae*[1] genomes. In contrast to the PDB benchmark described above, homologues were allowed in this study. We present an overview of the essential results to establish whether they are entirely consistent with the PDB benchmark; more detailed analysis of all of the ORFs in these genomes can be found elsewhere.[66]

TABLE VIIA. Distribution of Secondary Structure Class for Easy & Medium Proteins of *M. genitalium*, *E. coli*, and *S. cerevisiae* Genomes[a]

Fold class	*M. genitalium*	*E. coli*	*S. cerevisiae*
α	23.5%	23.2%	29.1%
β	15.7%	16.7%	20.1%
α/β	58.8%	58.8%	49.0%
Small	2.0%	1.3%	1.8%

[a] For all ORFs ≤ 200 residues in length.

M. genitalium Genome Threading Results

The M. genitalium[47] genome has 128 ORFs of less than 200 residues, of which 93 were assigned to the easy set with an average coverage of 87% and 35 were assigned to the medium set, with an average coverage of 54%. As in the PDB benchmark, there were few (or in this case, no) hard targets. A slightly larger percentage was classified as easy targets (73%) compared to the PDB benchmark (59%), but this reflects the fact that homologous proteins were not excluded. Of the ORFs in the easy set, 59 were less than 150 residues in length. In contrast, using the previous version of PROSPECTOR,[37] we were only able to assign 35 ORFs of less than 150 residues in length to templates.[67] This is due partly to the somewhat smaller template library that was used a year ago (about 3000 templates compared to the 3575 templates used now), but more importantly, it reflects the improvement in our ability to recognize templates. As others have observed, and as shown in Table VIIA, column 3, α/β proteins represent the dominant secondary structure class in this genome, with the Rossmann fold predicted to be the most dominant fold type, and the α/β plaits predicted to be the second most dominant (see Table VIIB, column 2). The resulting fold predictions can be found at http://www.bioinformatics.buffalo.edu/threading/mgen/easy and http://www.bioinformatics.buffalo.edu/threading/mgen/medium respectively.

E. coli Genome Threading Results

In the *E. coli* genome,[48] there are 1360 ORFs of less than 200 residues, of which PROSPECTOR_3 assigned 829

TABLE VIIB. Top Five Abundant Topologies (CATH) in the *M. genitalium, E. coli,* and *S. cerevisiae* Genomes for all ORFs ORFs ≤ 200 Residues in Length

M. genitalium (%)[a]	*E. coli*[a]	*S. cerevisiae*[a]
Rossmann fold 3.40.50 (13.0)	Rossmann fold 3.40.50 (18.3)	Rossmann fold 3.40.50 (8.6)
αβ plaits 3.30.70 (13.1)	Arc repressor mutant subunit A 1.10.10 (8.7)	Double stranded RNA binding domain 3.30.160 (6.8)
OB fold 2.40.50 (6.6)	αβ plaits 3.30.70 (7.2)	αβ plaits 3.30.70 (6.5)
Nucleotidyltransferase domain 5 3.30.420 (4.9)	Immunoglobulin-like 2.60.40 (5.5)	Glutaredoxin 3.40.30 (4.4)
Ribosomal RNA binding protein S15 1.10.287 (4.9)	Aminopeptidase 3.40.630 (3.2)	SH3 type barrels 2.30.30 (4.4)

[a] Percentage of the total number of detected CATH domains is indicated in parentheses.

(61%) ORFs to the easy set, with an average coverage of 82% and 521 ORFs (38%) to the medium set, with an average coverage of 51%. Only a very small number (10) was assigned to the hard set. These results are very similar to the PDB benchmark results, with a slightly larger percentage of targets assigned to the easy set in *E. coli* due to the fact that homologues were permitted. Once again, in this real world case, there were very few putatively New Fold proteins.

The resulting distribution of proteins of different secondary structure classes is shown in Table VIIA, column 3, whereas in *M. genitalium*, α/β proteins are dominant. As shown in Table VIIB, column 3, the most populated fold is the Rossmann fold, comprising about 18% of all assigned ORFs. This is followed by the Arc repressor mutant fold, which is predicted to be adopted by about 9% of the ORFs. Next is the immunoglobulin-like fold (a β-protein) predicted to be adopted by about 6% of the ORFs. All fold predictions may be found on our website at http://www.bioinformatics.buffalo.edu/threading/ecoli/easy and http://www.bioinformatics.buffalo.edu/threading/ecoli/medium respectively.

S. cerevisiae Genome Threading Results

In this genome,[1] there are 1496 ORFs of 200 residues in length or smaller. Of these, 793 (53%) are assigned to the easy set, with an average alignment coverage of 75%, and 682 (45%) are assigned to the medium set, with an average coverage of 65%. Thus, 1475/1496 ORFs probably have their fold identified; again, there are very few (21) hard targets (i.e. there are very few putative New Folds). The distribution of secondary class type is shown in Table VIIA, column 4, and the top scoring topologies are shown in Table VIIB, column 4. The α/β proteins still dominate, with the Rossmann fold being most populated, but relative to *M. genitalium* and *E. coli* genomes, the relative population of ORFs adopting the Rossmann fold is significantly reduced. New entries into the most populated folds, the Double Stranded RNA binding domain is the second most common, and the Glutaredoxin fold ranks fourth. All fold predictions may be found on our website at http://www.bioinformatics.buffalo.edu/threading/yeast/easy and http://www.bioinformatics.buffalo.edu/threading/yeast/medium.

Some Observations About Genome Scale Threading

From Table VIIA, it is apparent that the distributions of secondary structure types for these three genomes are quite similar, with that of the *M. genitalium* and *E. coli* genomes being very close and *S. cerevisiae* having a relative reduction of about 10% in α/β proteins, which nevertheless is the dominant secondary structure type in all three genomes. The next most populated secondary structure class is helical proteins. Also, as shown in Table VIIB, the Rossmann fold is the most prevalent single fold type in all three genomes, with α/β plaits also highly populated. For all but a very few proteins in all three genomes, as was the case in the PDB benchmark, it is quite likely that their topology has been identified. This is further evidence supporting the observation that the PDB is a covering set of single domain protein structures.[46] Nevertheless, based on the PDB benchmark results, only about 60% of the ORFs considered are likely to have good alignments. Additional techniques that build entire chains and refine the results are required for full assignment.[62,67–70]

Discussion

Recently, we have demonstrated that, for single domain proteins, the PDB is complete at the level of low-to-moderate-resolution structures.[46] These results strongly suggest that the most promising means of solving the protein folding problem is developing threading algorithms that are capable of detecting proteins with similar folds, whether or not the target and template sequences are evolutionarily related. While this has been demonstrated in principle, in practice the problem is developing threading algorithms that can detect such fold relationships. In this spirit, here we have described and benchmarked the PROSPECTOR_3 series of threading algorithms that employ a variety of scoring functions. Our approach is based on the idea that, for imperfect algorithms, different scoring functions will match different pairs of target/template proteins; if one can establish reliability criteria, then one can combine a series of such scoring functions to provide a larger number of accurate predictions. Thus, we have developed criteria to define the easy, medium and hard sets of proteins. Easy proteins are very likely to have accurate alignments of high coverage. Medium proteins contain information about their fold, but the detailed alignment can be in error, and hard proteins are those that, from the point of view of the algorithm, fail to have a known template structure assigned.

Compared to our previous generation of PROSPECTOR, not only has the algorithm been improved by our ability to do reasonably successful classification of the reliability of our results, but our predictive ability has been improved by a number of factors. Following the suggestion of Karplus,[30] using both the forward and reversed sequences, we have enhanced the sensitivity of fold recognition. We have introduced a variety of more specific pair potentials and used variants that have different dependencies on secondary structure prediction accuracy. By identifying structurally similar regions in multiple templates, we can identify the highly accurate regions of the alignments. By reducing the terminal gap penalties to zero (not a new idea[71,72]), we can increase our ability to find related topologies as well as generate significant alignments. By establishing rigorous Z-score thresholds for significant results, we found that, even if the apparent global RMSD is high, in almost all cases it reflects differences in packing angles of secondary or supersecondary elements, especially at the N- and C-termini. Continuous aligned regions provide rather accurate native-like fragments that can be used in fold assembly algorithms. The last two observations have motivated the development of the TASSER fold threading/assembly/refinement algorithm.[62]

On an encouraging note, the current version of PROSPECTOR has the ability to identify good template structures for over 90% of a set of nonhomologous or weakly homologous sequences for the PDB benchmark. Unfortunately, the alignments are good for only two-thirds of all target sequences, suggesting that algorithms that improve alignment accuracy are needed for the remaining cases. This might be achieved by first running PROSPECTOR and then refining the subsequent alignment,[62] or by using alternative threading approaches and developing fold reliability classification schemes such as those used here (that define the easy, medium and hard sets) to 'tile' target sequence space (viz. to combine different threading algorithms, each of which has distinct, partially overlapping sets of high reliability target sequence predictions in order to generate more accurate composite results). Both approaches are currently being explored.

Application of PROSPECTOR_3 to the M. genitalium,[47] E. coli[48] and S. cerevisiae[1] genomes yields comparable results to those from the PDB benchmark, suggesting that the PDB benchmark is representative of what can be expected when PROSPECTOR_3 is applied on a large scale to typical target sequences, that is, very few proteins assigned to the hard set and about 60% of the proteins assigned to the easy set. As others have found, α/β proteins are the dominant secondary structure class in these genomes,[2,4,73,74] with the Rossmann fold being the most populated. Clearly, these threading results provide a useful starting point, both for functional annotation based on the threading template identification/target sequence alignments[18] and for fragments that will be used in subsequent fragment assembly algorithms.[62] Such efforts are currently underway for these three genomes.

At this juncture, there are a number of possible directions in which to further improve PROSPECTOR_3. One possibility involves the use of profile–profile approaches to generate better initial alignments;[29] alternatively, other threading algorithms could be used to provide these initial alignments as well as a set of predicted seed contacts that will be used in subsequent threading iterations. We can also use distance geometry to combine the set of templates[75,76] and to build consensus alignments to be used either in contact prediction or subsequent folding studies. Another issue that will be immediately addressed is the benchmarking of proteins between 200 and 300 residues in length to determine the algorithm's performance on larger proteins and to assess whether our ability to classify targets into easy and medium proteins holds. As the targets get larger, our structural alignment studies suggest[46] that it will take approximately five templates to generate 90% alignment coverage. Techniques that combine multiple templates by aligning different regions will have to be developed. Overall, though, by the careful benchmarking of PROSPECTOR_3 on a comprehensive and representative set of all protein domains below 201 residues in length, we now have a very good idea of the strengths and weaknesses of this algorithm. Most encouragingly, these insights seem to carry over when PROSPECTOR_3 is applied to genomes. Thus, while much remains to be done, there is encouraging progress in fold recognition. Another significant conclusions is that PROSPECTOR_3 is among the next generation of threading algorithms that significantly outperform PSIBLAST.

ACKNOWLEDGMENTS

This research was supported in part by Grant No. GM-48835 of the Division of General Medical Sciences of the National Institutes of Health. Stimulating and very useful discussions with Dr. Adrian K. Arakaki are gratefully acknowledged. We thank Drs. Y. Xu, R Hughey, H Zhou and Y. Zhou for providing us with the PROSPECT, SAM-T99, and SPARK algorithms respectively.

RERERENCES

1. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, Stocker S, Weil B. MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2000;28(1):37–40.
2. Balasubramanian S, Schneider T, Gerstein M, Regan L. Proteomics of *Mycoplasma genitalium*: identification and characterization of unannotated and atypical proteins in a small model genome. Nucleic Acids Res 2000;28(16):3075–3082.
3. Buchanan SG. Structural genomics: bridging functional genomics and structure-based drug design. Curr Opin Drug Discov Devel 2002;5(3):367–381.
4. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW. Functional and structural genomics using PEDANT. Bioinformatics 2001;17(1):44–57.
5. Moxon ER, Hood DW, Saunders NJ, Schweda EK, Richards JC. Functional genomics of pathogenic bacteria. Philos Trans R Soc Lond B Biol Sci 2002;357(1417):109–116.
6. Skolnick J, Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol 2000;18(1):34–39.
7. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its

importance for gene function analysis. Nat Biotechnol 2000;18(3): 283–287.

8. Fetrow JS, Siew N, Di Gennaro JA, Martinez-Yamout M, Dyson HJ, Skolnick J. Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? Protein Sci 2001;10(5):1005–1014.
9. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306(5):1191–1199.
10. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93–96.
11. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci U S A 1998;95(26):15189–15193.
12. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 2002;11(4):723–738.
13. Skolnick J, Zhang Y, Arakaki A, Kolinski A, Boniecki M, Szilagyi A, Kihara D. A unified approach to protein structure prediction. Proteins 2003;CASP5 Suppl:469–479.
14. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31(14):3982–3992.
15. Williams MG, Shirai H, Shi J, Nagendra HG, Mueller J, Mizuguchi K, Miguel RN, Lovell SC, Innis CA, Deane CM, Chen L, Campillo N, Burke DF, Blundell TL, de Bakker PI. Sequence-structure homology recognition by iterative alignment refinement and comparative modeling. Proteins 2001;Suppl 5:92–97.
16. Nagendra HG, Harrington AE, Harmer NJ, Pellegrini L, Blundell TL, Burke DF. Sequence analyses and comparative modeling of fly and worm fibroblast growth factor receptors indicate that the determinants for FGF and heparin binding are retained in evolution. FEBS Lett 2001;501(1):51–58.
17. Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, Fetrow JS. Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 2001;134(2-3):232–245.
18. Arakaki AK, Zhang Y, Skolnick J. Large scale assessment of the utility of low resolution structures for biochemical function assignment. Proc Natl Sci USA 2003, Submitted for publication.
19. Wojciechowski M, Skolnick J. Docking of small ligands to low-resolution and theoretically predicted receptor structures. J Comput Chem 2002;23(1):189–197.
20. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23(11):444–447.
21. 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(12):1000–1011.
22. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29(14):2994–3005.
23. Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS, Sowdhamini R, Srinivasan N. SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res 2002;30(1):289–293.
24. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 2002;30(1):268–272.
25. Madera M, Gough J. A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res 2002;30(19):4321–4328.
26. Pearl FM, Martin N, Bray JE, Buchan DW, Harrison AP, Lee D, Reeves GA, Shepherd AJ, Sillitoe I, Todd AE, Thornton JM, Orengo CA. A rapid classification protocol for the CATH Domain Database to support structural genomics. Nucleic Acids Res 2001;29(1):223–227.
27. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28(1):257–259.
28. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001;17(8):713–720.
29. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315(5):1257–1275.
30. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14(10): 846–856.
31. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001;Suppl 5:86–91.
32. Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322(1):65.
33. Jones DT, Miller RT, Thornton JM. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 1995;23(3):387–397.
34. Miller RT, Jones DT, Thornton JM. Protein fold recognition by sequence threading: tools and assessment techniques. Faseb J 1996;10(1):171–178.
35. Xu D, Crawford OH, LoCascio PF, Xu Y. Application of PROSPECT in CASP4: characterizing protein structures with new folds. Proteins 2001;Suppl 5:140–148.
36. McGuffin LJ, Jones DT. Targeting novel folds for structural genomics. Proteins 2002;48(1):44–52.
37. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42(3):319–331.
38. Marchler-Bauer A, Bryant SH. Measures of threading specificity and accuracy. Proteins 1997;Suppl 1:74–82.
39. Marchler-Bauer A, Bryant SH. A measure of progress in fold recognition? Proteins 1999;Suppl 3:218–225.
40. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296(5):1319–1331.
41. Lathrop RH. An anytime local-to-global optimization algorithm for protein threading in theta (m2n2) space. J Comput Biol 1999;6(3-4):405–418.
42. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19(7): 874–881.
43. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10(11):2354–2362.
44. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51(3):434–441.
45. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58(Pt 6 Pt 1):899–907.
46. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003, To appear.
47. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. The minimal gene complement of *Mycoplasma genitalium*. Science 1995;270(5235):397–403.
48. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12 [comment] [see comments]. Science 1997;277(5331):1453–1474.
49. 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(3):443–453.
50. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38(1):3–16.
51. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6(3):676–688.
52. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I,

Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31(1): 365–370.

53. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res 2002;30(1):42–46.
54. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276(1):71–84.
55. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 1994;24:307–331.
56. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31(13):3497–3500.
57. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17(1):49–61.
58. Levin JM. Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng 1997;10(7):771–776.
59. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404-405.
60. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2003, To appear.
61. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001;59(5):305–309.
62. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 2003, Submitted for publication.
63. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2001;Suppl 5:157–162.
64. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001;Suppl 5:192–199.
65. Rychlewski L, Fischer D, Elofsson A. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins 2003;53 Suppl 6:542–547.
66. Kihara D, Skolnick J. Microbial Genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_3Q. Proteins 2003, Submitted for publication.
67. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. *Ab initio* protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. Proc Natl Acad Sci U S A 2002;99(9): 5993–5998.
68. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to *ab initio* protein structure prediction. Biophys J 2003;85(2):1145–1164.
69. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in *ab initio* protein structure prediction. Proteins 2001;Suppl 5:119–126.
70. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and *ab initio* protein structure prediction. Protein Sci 2002;11(8):1937–1944.
71. Morgenstern B, Atchley WR, Hahn K, Dress A. Segment-based scores for pairwise and multiple sequence alignments. Proc Int Conf Intell Syst Mol Biol 1998;6:115–121.
72. Chao KM, Hardison RC, Miller W. Recent developments in linear-space alignment methods: a survey. J Comput Biol 1994;1(4): 271–291.
73. Homma K, Nishikawa K. [Protein structure information provided by the GTOP database and its applications]. Tanpakushitsu Kakusan Koso 2002;47(8 Suppl):1076–1082.
74. Frishman D. Knowledge-based selection of targets for structural genomics. Protein Eng 2002;15(3):169–183.
75. Havel TF, Crippen GM, Kuntz ID, Blaney JM. The combinatorial distance geometry method for the calculation of molecular conformation. II. Sample problems and computational statistics. J Theor Biol 1983;104(3):383–400.
76. Xia Y, Huang ES, Levitt M, Samudrala R. *Ab initio* construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300(1):171–185.