PROTEINS: Structure, Function, and Bioinformatics 54:41–48 (2004)

Enriching the Sequence Substitution Matrix by Structural Information

Octavian Teodorescu,[1] Tamara Galor,[1] Jaroslaw Pillardy,[2] and Ron Elber[1]*
[1]*Department of Computer Science, Cornell University, Upson Hall 4130, Ithaca, New York 14853*
[2]*Cornell Theory Center, Cornell University, Upson Hall 4130, Ithaca, New York 14853*

ABSTRACT **A fundamental step in homology modeling is the comparison of two protein sequences: a probe sequence with an unknown structure and function and a template sequence for which the structure and function are known. The detection of protein similarities relies on a substitution matrix that scores the proximity of the aligned amino acids. Sequence-to-sequence alignments use symmetric substitution matrices, whereas the threading protocols use asymmetric matrices, testing the fitness of the probe sequence into the structure of the template protein. We propose a linear combination of threading and sequence-alignment scoring function, to produce a single (mixed) scoring table. By fitting a single parameter (which is the relative contribution of the BLOSUM 50 matrix and the threading energy table of THOM2) we obtain a significant increase in prediction capacity in the twilight zone of homology modeling (detecting sequences with <25% sequence identity and with very similar structures). For a difficult test of 176 homologous pairs, with no signal of sequence similarity, the mixed model makes it possible to detect between 40 and 100% more protein pairs than the number of pairs that are detected by pure threading. Surprisingly, the linear combination of the two models is performing better than threading and than sequence alignment when the percentage of sequence identity is low. We finally suggest that further enrichment of substitution matrices, combing more structural descriptors such as exposed surface area, or secondary structure is expected to enhance the signal as well. Proteins 2004;54:41–48.**

Key words: sequence alignment; threading; fitness function; sequence-to-structure matching; energy function; *Z*-score

INTRODUCTION

Annotation and classification of proteins rely on accurate and efficient comparison of pairs of proteins. An essential ingredient of the comparison algorithm is the substitution matrix, T. For a pair of amino acid types α and β in environments x and y the substitution matrix provides a score for their exchange between the two proteins $T \equiv T(\alpha, x|\beta, y)$. The score of an alignment (ignoring for the moment penalties for indels) is a sum over all substitution scores.

Environment consists of additional features (x,y) to the direct score of amino acid substitution, which we denote by $T(\alpha|\beta)$. For example, it may include (i) multiple sequence information,[1] (ii) secondary structure data,[2] (iii) exposed surface area,[3] and (iv) many other structural and functional fingerprints. Here we consider the information content of only a pair of proteins. Multiple sequence information [feature (i)] is not discussed here and can be added (in principle) once the scoring of a pair is optimized.

A class of environment features is the use of structural information. An alignment of a probe sequence into a shape of another protein is called threading and is usually associated with an energy function[4]; the energy measures the quality of sequence to structure fitness.[4] The amino acids are aligned into a known shape and three-dimensional interactions are scored, measuring protein stability. The sequence to sequence and sequence to structure alignments are done separately and have their own corresponding substitution matrices. For sequence alignment we have $T(\alpha|\beta)$ and for sequence to structure alignment we use $T'(\alpha|y)$.

It is interesting to note that one type of a substitution matrix dominates the scoring of sequence-to-sequence alignments in proteins (BLOSUM 50; Ref. 5), whereas there is no dominant scoring scheme (energy function) of matching sequences into structures. The BLOSUM 50 matrix was used as an example, because we have considerable experience in using it and comparing its results with threading approaches.[4] We anticipate a similar enhancement in recognition for other sequence-substitution matrices; however, we did not do calculations with other matrices. Part of the reason for the larger diversity of threading energy functions is the higher complexity of three-dimensional interactions compared with one-dimensional substitutions, making it more difficult to find the best choice. Another reason is the significant success of BLO-

The calculations were performed on Dell Edge cluster of the Cornell Theory Center funded by the tri-institutional grant.

Grant sponsor: National Science Foundation; Grant number: 9988519 to R.E. Grant sponsor: NSERC Canadian fellowship to O.T. Grant sponsor: a tri-institutional grant to Cornell and Rockefeller Universities and Memorial Sloan Kettering Cancer Center to J.P.

*Correspondence to: R. Elber, Department of Computer Science, Cornell University, Upson Hall 4130, Ithaca, NY 41583. E-mail: ron@cs.cornell.edu

Received 11 December 2002; Accepted 25 March 2003

SUM 50 in identifying evolutionary relationships compared with the much weaker sensitivity of stability energies.

Nevertheless, an interesting complementary relationship was observed in a number of studies.[4] At the twilight zone of similarity detection by sequence alignment it is possible to find remote evolutionary relationships by sequence-to-structure matching. Threading detects a significantly smaller number of similar protein pairs compared with sequence alignment; however, the set of hits in threading is not a subset of the sequence alignment hits. Therefore, threading alone is a potentially useful tool when sequence alignment fails to recover a signal.

Merging threading and sequence signals is done after separate alignments and scoring was performed. The raw scores or the statistical significance measures (e.g., the Z-scores[6]) are combined in an empirical formula[7] or in a neural net[8] to take advantage of the complementarities of the two techniques.

Here we propose another combination of sequence and structure signals at the level of the substitution matrix. A new substitution matrix, $M(\alpha|\beta,y)$, is defined as a linear combination of $T(\alpha|\beta)$ and $T'(\alpha|y)$:

$$M(\alpha|\beta,y) = \lambda T(\alpha|\beta) + (1 - \lambda)T'(\alpha|y) \qquad (1)$$

The parameter λ is a constant mixing term between zero and one that we optimize (see Methods). The new matrix is used in a dynamic programming algorithm to determine the optimal alignment.

Mixing the scoring of a structural factor and amino acid substitution score was done in the past in the context of secondary structure (and sequence alignment).[2] Here we extend that study to consider an alternative threading score. The hope is that the mixing will create positive consensus. That is, if the two measures agree that a partial alignment is good (even if the positive signal is rather weak for each measure), the combined signal may still be a match. At the same time when one of the scores strongly objects, the alignment is in doubt even if the second measure shows a positive signal. The hope is then to enhance the signal of true positives by consensus of the two measures and to reduce false signals by score conflicts.

If one of the signals is extremely strong and considered significant even alone, then the mixing is not necessarily beneficial. However, when both signals are not strong, then the proposed scheme may be helpful. We therefore propose the use of the mixed model for the twilight zone of detection for sequences that are (at least) lower than 25% sequence identity. In fact as is shown in Results, even sequences with only 25% sequence identity can carry a significant sequence-to-sequence signal. We therefore made the threshold for the twilight zone a bit tighter and consider only pairs of proteins that are structurally related (as measured by the structural alignment program CE[9]; see below) and have no significant sequence-to-sequence signal (defined as a Z-score < 2 for sequence alignment with the BLOSUM 50 matrix). Some of these pairs are found directly by threading alone; however, a considerable enhancement in detection is obtained when the mixed model is used.

The CE (Combinatorial Extension) program is a leading protocol for local structure alignment of two protein chains that has significant success in detecting remote structural relationships by overlapping the Cα positions of two proteins, minimizing the RMS distance between the two structures. In brief, CE uses a dynamic programming algorithm with empirically determined gap and extension penalty (or reward) to determine best matching local protein segments.[9] A Z-score determines the significance of the match.

In this article we compare the mixed model with direct sequence alignment, with direct threading experiment, and with PSI-BLAST.[10] We show that in the twilight zone of sequence similarity, the mixed model outperforms the other algorithms by wide margins.

METHODS

We consider the matching of a probe sequence S_i to a known protein with a sequence S_j and structural environment defined by the vector X_j. Similarly to the usual notion of amino acid sequence in which we describe the protein by a one-dimensional list of amino acids $S_i \equiv a_{1i},a_{2i}\ldots a_{ni}$, the vector X_j is a one-dimensional sequence of local structure descriptors. It is given by $X_j = x_{1j},x_{2j}\ldots x_{nj}$. The a_{kl} is one of the 20 amino acids, whereas the x_{kl} are finite set of local structural features that we use to describe the structural environment of an amino acid. These local features can be secondary structure, exposed surface area, number of contacts, etc. Extensions of the model below taking into account the different features mentioned above are quite obvious.

Here we rely on our previous experience in designing energy function for threading as implemented in the program LOOPP. LOOPP (Learning, Observing, and Outputting Protein Patterns) is a fold recognition program that emphasizes threading for annotation. In addition to prediction, LOOPP also learns energy parameters from native and decoy sets with the Mathematical Programming approach.[6] Source code and databases of LOOPP are available from http://cbsu.tc.cornell.edu/software/loopp.

One of the potential that we optimized for annotation is THOM2. In THOM2 the energy $[\epsilon_{i\alpha}(n,m)]$ of a structural site i is determined by the identity of the amino acid at the site, α, the number of neighbors to the site, n, and the number of neighbors to each of the direct neighbors of the site, m.[6] The total THOM2 energy, E_{THOM2}, is $E_{THOM2} = \Sigma_{i,n}\epsilon_{i,\alpha}(n,m)$. In summary, THOM2 describes the environment by two layers of contacts to the structural site. It is a two-dimensional table that provides a numerical value using two indices: (i) a type of an amino acid α and (ii) a pair of layers (n,m). The number of structural environments in THOM2 [16; possible combination of coarse grained (n,m) pairs] is comparable to the number of amino acid types, making the size of the THOM2 table comparable to that of BLOSUM 50.

In LOOPP we use dynamic programming algorithm[11] to find an optimal alignment of S_i against S_j and (separately)

TABLE I. The 176 Sequences of the Test Set†

1iaa	1hlm	1hxn	2pil	1ulo	1djs_A	3trx	1quw_A	2tnf_A
1fat_A	1evh_A	1www_X	1cww_A	1dyn_A	1b88_A	1irp	1bks_A	1qgh_A
3a3h	1hou_L	1cfv_L	1maj	1qtf_A	1a7v_A	1cdb	1tvd_A	1bbt_2
1piv_1	1tnn	1aud_A	1c8p_A	1myt	1igm_L	1msp_A	1c3k_A	1tnf_A
1qe5_A	2fsp	1hfd	1bab_B	3man_A	5mbn	1gc1_H	1cd9_A	1ppf_E
1fsl_A	1spg_A	1kac_B	1mig_L	1b6b_A	1tax_A	1itn	7tim_A	5tim_A
1eoe_A	1adl_A	1rhg_A	3sdh_A	1aag_L	1pot	1hlg_A	1hav_A	32c2_B
1fvw_L	1lou_A	1rho_A	1dc7_A	1fim	1cda_A	1gya	1rvv_A	1stm_A
1ita	1tof	1hmd_A	1lts_D	4fgf	2gst_A	1qft_A	256b_A	2hft
1iif_H	1b77_A	1cqk_A	1qqs_A	1qfw_L	1mba	1rlw	1rdx_A	1iai_I
2snv	1bpv	1vca_A	1aly	1cv8	1hib	1tim_A	1bfs	1cjx_A
3crd	1chg	1d7m_A	1ag4	1b49_A	1kb5_A	1cx1_A	35c8_H	1bcf_A
1irs_A	1ntr	3f58_H	2phy	1lh2	2bpa_2	1jhl_H	1nbc_A	1pdk_A
1aag_H	1d9k_B	1qab_A	1cd8	1wba	2fx2	1tgn	3fyg_A	1ryp_B
1rcy	2dlf_H	1pfc	1ilr_I	1eqy_S	1dqw_A	1kiq_B	1rml	1hlb
1jli	1qsv_A	1pls	2dhq_A	1isk_A	6ilb	1dbw_A	1gsa	1aut_C
1eod_A	2hfm_H	1e0s_A	2rhe	1vls	1il7	1e2a_A	1bla	1auo_A
1lxd_A	1nul_A	1tnm	2vhb_A	1svy	2hbg	2gmf_A	1igs	1b08_A
1neu	1ajw	1fxy_A	1qsm_A	1qa9_A	1uky	1vre_A	1bj7	1di0_A
1ece_A	1dvj_A	2acq	3hhr_B	1fgv_H				

†Each of the sequences has its own set of decoys and homologous structures. The goal is to identify as many as possible homologous pairs.

against X_j. We use both, local and global alignment, and in the final evaluation of the significance of the match we use the Z-score. The same alignment algorithm is used in the model described below.

The mixed model defines a combined alignment of a probe sequence S_i with another protein whose sequence S_j and structure X_j are known. A dynamic programming algorithm is used. At every step of building the dynamic matrix we consider local matches. The fitness of an amino acid a_k against the pair (b_l,x_l) is measured with a new (mixed) substitution matrix:

$$M(a_k|b_l,x_l) = \lambda T_{\text{BLOSUM-50}}(a_k|b_l) + (1-\lambda)T_{\text{THOM2}}(a_k|x_l) \quad (2)$$

The entries a_k and b_l are the types of the amino acids in positions k and l along the sequence. The entry x_l is the structural environment of position l. The (single) free parameter λ is the mixing parameter and was chosen empirically to be 0.125. In Results we explain how λ was computed and demonstrate that the model is not very sensitive to the choice of the mixing parameter. Ideally we may imagine λ being a parameter of the actual scores. For example, if the sequence-to-sequence signal is very high (e.g., =60% sequence identity), then we anticipate the mixing parameter to be one. However, in the present study we did not perform an optimization of λ for the complete range and consider its uses only for difficult-to-annotate sequences.

The matrix $T_{\text{BLOSUM 50}}$ is the BLOSUM 50 substitution matrix,[5] and T_{THOM2} is the threading energy of the THOM2 model.[6] The two matrices that we have used can be downloaded from the web http://cbsutest.tc.cornell.edu/var/loopp_testset/.

To evaluate the performances of the model and to examine different values of the mixing parameter, we have constructed a test set containing 176 sequences with lengths between 100 and 500 amino acids. The sequences are listed in Table I, and the complete test is available from http://cbsutest.tc.cornell.edu/var/loopp_testset/.

The test was constructed as follows. We started with the library of sequences and structures that we use in LOOPP. This library contains about 3900 proteins and provides a dense representation of the protein databank. We call it the prediction set. When we prepared the prediction set we removed closely related proteins but kept many homologous sequences and structures to increase the chance that a probe sequence will hit at least one of them.[6] It is the set that is used in all of the predictions made by the LOOPP server http://ser-loopp.tc.cornell.edu/loopp.html. To prepare the test set for this article, we select at random proteins from the prediction set. For each of the random selections we prepared an independent database using the same LOOPP set. The single-sequence search-database includes proteins with small length difference with the probe sequence. The search database has at least one homologue protein with low sequence identity (<25%) and three such pairs on the average. The CE program for structural alignment (see Introduction),[9] requiring a Z-score for structural comparison of at least 4.5, determines the homologue pairs. Typically, the set for a single sequence has 1000 decoy proteins (with CE Z-scores < 3.5). We selected only remote homologue pairs so that the local sequence alignments using the BLOSUM 50 matrix does not display a Z-score > 2.0 for any of them. Consequently, all the sequence alignment searches produced with the above-mentioned algorithm failed to place any pair in the list of top 10 ranking comparisons based on the Z-score.

As presented in the general description of our model, we base our ranking on the Z-score. The Z-score is a statistical measure of the quality of a fit of a sequence or a structure

into another sequence or a structure. We briefly described it and the algorithm used for its computation below. We consider the Z-score for sequence-to-sequence alignment. The same algorithm holds for other alignments as sequence-to-structure etc.

Let $\bar{S}_i$ and $\bar{S}_j$ be the extended aligned sequences (including deletions and insertions) found with dynamic programming.[11] For example, $\bar{S}_i = \bar{a}_1,\ldots,\bar{a}_k = a_1 - - a_2 a_3 a_5 \ldots a_n$ when a "−" denotes an insertion or a deletion, a^- is an extended amino acid that may be a gap, and k is the length of the alignment. Note that a^-_1 can be any of the characters $\{-, a_1,\ldots,a_n\}$. We denote the score of aligning S_i to S_j by Q_{ij} and it is given by $Q_{ij} = \sum_{i=1,\ldots,k} T(\bar{a}_{il}|\bar{b}_{jl})$. The Z-score for the ij pair is a dimensionless quantity defined by

$$Z = \frac{Q_{ij} - \langle Q \rangle}{\sqrt{\langle Q^2 \rangle - \langle Q \rangle^2}} \tag{3}$$

The average $\langle \ldots \rangle$ is defined over random sequences sampled with the same amino acid composition as the probe sequence. The random sequences are obtained by shuffling the amino acids of S_i to obtain a random sequence S_i^r that is aligned to S_j. The average score of S_j against the random sequences is given by $\langle Q \rangle = (1/R) \sum_{r=1R} Q^r_{ij}$. Hence, the Z-score is a measure of the average distance of the matched pair from random pairs. For the matching to be significant, this distance must be as large as possible. This means that the calculated Z-score must be significantly larger than the Z-score of a false positive.

Note that the absolute value of the Z-score is model dependent and it is therefore important to understand (at the least) the distribution of the Z-scores for false positives. This distribution can be computed numerically by calculating the values of Z-scores for wrong matches. Our test set contains a large number of wrong matches (decoy structures) and we used them to compute the corresponding Z-score distribution. If the model at hand is successful, the distribution of Z-scores for true positives will have little overlap with the distribution of false positives. We therefore consider both, the distribution of true and false positives.

To perform the alignment we need to assign a value for the gaps; i.e., the deletions or insertions. We use structural dependent gap penalties. For sequence alignment and threading the gap penalty depends on the number of neighbors. The values of the gap penalties are, however, different in the two cases. The LOOPP article[6] discusses both types of gap penalties assigned and their advantages versus a constant gap penalty. The gaps for the mixed model are mixed in a similar way to the energy values. The actual values for each of the substitution matrices (threading or sequence) are taken from Ref. 6.

The calculations used the LOOPP V2 program available at http://cbsu.tc.cornell.edu/software/loopp/index.htm. Besides the mixed model described here, LOOPP also includes modules for potential training by Mathematical Programming,[12,13] numerous scoring functions, and threading and sequence alignment algorithms.[6] The calculations described below were done on the Cornell Theory Center Dell Edge Cluster running Windows 2000. A single comparison of two proteins within the mixed model took from 1.5 to 30 s, depending on the sequence length. This includes the Z-score calculation with 100 randomly shuffled sequences. One annotation of a probe sequence (sweeping through the entire database) requires from 30 min to about 8 h for the longest sequence.

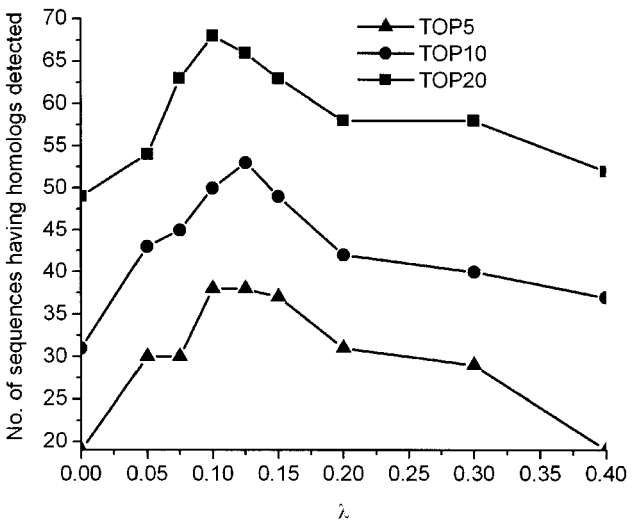

Fig. 1. The number of successful sequences assigned to their corresponding homologue proteins as function of the mixing parameter λ. The total number of potential hits is 176, so the success ratio is always <50% making the present test an intriguing case for future studies. Note that pure sequence alignment (mixing parameter equal 1) is not shown since the detection is negligible. The left hand side of the graph (λ = 0) corresponds to the pure THOM2 model.

RESULTS

The new substitution scheme involves just one independent parameter, the mixing value. We concentrate first on determining an optimal value for the mixing parameter using the new test set. We will address issues related to the model calibration using the Z-score later in this section.

We measure the performance of the substitution matrix in two ways. The first measure is how many of the probe sequences were identified? That is, how many probe sequences have at least one homologous protein found. In the second measure we count the number of homologous proteins that were detected (regardless of the fact that some of them may identify the same probe). It is important to use both measures because it is possible that a specific family is easy to detect and many homologous proteins are found for it, whereas other families remain undetected. On the other hand the second approach is potentially stricter, examining if we identify all homologues or only a fraction of them. The first test can be 100% successful, whereas some homologous proteins remain undetected.

Three levels of success are defined based on the Z-score ranking of the pair: the homologous protein is in the top 5, top 10, or top 20. For instance, a homologue pair is in the top 5 if there are four or less decoy structures found with a greater Z-score than the homologue-query pair.

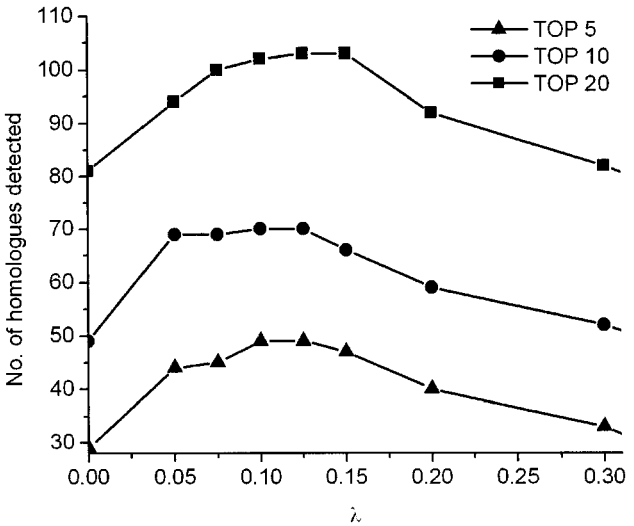

Fig. 2. The number of homologue pairs detected as function of the mixing parameter. The total number of homologues present is 740. The left-hand side corresponds to the THOM2 model and $\lambda = 1$ to the BLOSUM 50 matrix.

Figure 1 shows the performance of our model for different mixing parameters counting the number of query sequences that were identified.

We see an improvement of 40–100% over threading using the THOM2 scoring function, depending on the definition of the level of success. The added value of sequence similarity at the dynamic programming level is surprising because sequence alignment has no matches in the top 10. We have also used PSI-BLAST[10] (with BLOSUM 50 for a substitution matrix) to check the contribution of multiple sequence alignment on this test set. PSI-BLAST, which uses sequence information only, correctly identifies 14/15/16 sequences with homologues in the top 5/10/20. This is far below the performance of the threading model presented here. The threshold parameter for PSI-BLAST was 0.001, which is the default. We tried the threshold parameter 0.01 as well, and the results were similar. The number of iterations was set to the maximum of 3.

In Figure 2 we show the total number of homologues detected; the second measure of success in our study.

In this test we count for each sequence all the homologue proteins detected in the Top x ($x = 5$, 10, 20). Similar improvement to that presented in Figure 1 is seen again in Figure 2. In all plots a peak of performance is observed that is around values of a mixing parameter $0.05 < \lambda\ 0.15$. For the second ranking measure PSI-BLAST successfully identifies 15/17/18 homologue pairs in the Top 5/10/20, respectively. This is again well below the performance of the threading model (and of course, also of the mixed model).

Note that the properties of the BLOSUM 50 and the THOM2 matrices are very different, not only in absolute values but also in their variance.[5,6] BLOSUM 50 is strongly diagonal (prefer native amino acids, no change), whereas the THOM2 energy is not (THOM2 preference for native sequence is not so strong). The choice of the mixing parameter is therefore not trivial and requires the above experimentation. Moreover, the small value of the mixing parameter does not imply that the BLOSUM 50 contribution is small because the entries to the BLOSUM 50 matrix tend to have higher absolute values.

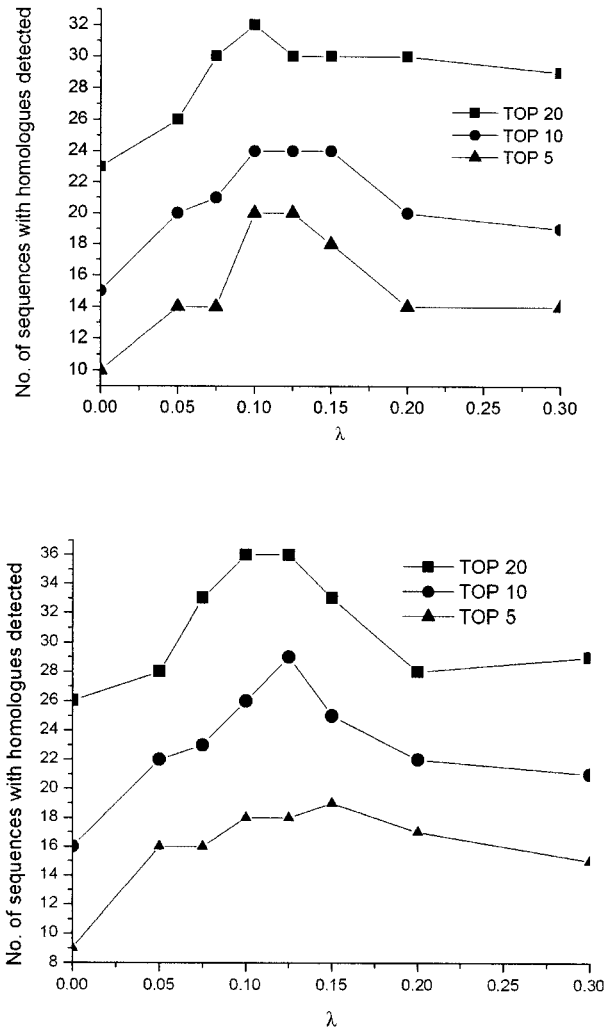

Fig. 3. The number of probe sequences that were identified, (the first measure of success; similar to Fig. 1), but for the two subsets of the test each containing only 88 sequences.

The improvement over both parent scoring schemes is significant up to about $\lambda = 0.35$. Beyond this value, the performance of the mixed model drops below the performance of threading. We select the value of 0.125, a value that was used in forthcoming studies.

To test the sensitivity of the parameter choice to the training set we split the test set in two and plotted the first measure (see Fig. 1), for both sets.

We see that the same feature is present for the two subsets independently and the 0.125 is still the preferred value if the training is done on these two independent sets (Fig. 3). We note that a Z-score of 4.0 in the mixed model corresponds to 10^{-3} chance that such a fit is for a decoy structure.

Typically, in our database we have <1000 structures with similar lengths (<20% difference) for each query sequence. Therefore, if we run with this length restriction, it is unlikely that a prediction with a Z-score of 4 and higher will be a false positive. If we combine the local and global threading as shown in Ref. 6, we can use Z-scores < 4 and still maintain high confidence level. In fact, global and local Z-scores for THOM2 that are above 3 (for global) and above 2 (for local) have false-positive probability of 10^{-4}. We also note that the mixing model has a slightly higher Z-score threshold for a given confidence compared with the pure threading model by roughly 0.3–0.4 in the region of interest of Z-scores of 2–4.[6] This is not surprising because scoring with sequence alignment (which is a part of the mixed model) typically has higher Z-scores for false positives.

It is useful to study also the distribution of true positive (in addition to false positives) to appreciate the degree of separation between the false and true predictions. In a good model we hope to have a very clear separation between the distributions, maximizing sensitivity and selectivity. In reality the separation is never perfect.

To obtain more comprehensive and independent statistics for false-positive Z-scores and true-positive predictions (correct high ranking by Z-score), we consider a larger set. We extend the earlier test set by relaxing the condition on sequence similarity (as defined by the threshold of 2 in the Z-score of the sequence alignment). Instead we include all the sequences and homologue structures having <25% sequence similarity.

Consequently, our test set expanded to 306 individual sequences, having on average eight structurally homologous proteins for each of the query sequences. In Figure 4, we show the distribution of the Z-scores of true positives for the three models: sequence alignment, threading by using the THOM2 potential, and the mixed model. The Z-scores are computed for local alignments. The probability density of the Z-scores is computed by binning. It is the probability of finding a Z-score between $Z - dZ/2$ and $Z + dZ/2$, divided by the box of size dZ. The size of the window used in the present calculation is 0.14.

We emphasize that 25% sequence identity still includes many related sequences that can be detected using sequence similarity measures (i.e., the raw score obtained with dynamic programming algorithm using the BLOSUM 50 matrix). Hence (perhaps surprisingly), the measure of sequence identity is too crude and is not a good indicator for the threshold of applicability for sequence alignment methods. This measure should be replaced (for example) with the more complex measure of the Z-score.

The high contribution of similar sequences in this study results in a comparable performance of the mixed and the sequence alignment models. The mixed model and the sequence alignment have a rather asymmetric distribution, with a long tail biased toward high Z-score values. We observe that the true positives are shifted toward larger Z-score values for the mixed model compared with threading using the THOM2 potential. However, the sets are nonoverlapping. Many of the sequences left undetected by sequence-to-sequence alignments are detected by the mixed model (and vice versa). To quantify the complementarities of the three models, we note that sequence alignment

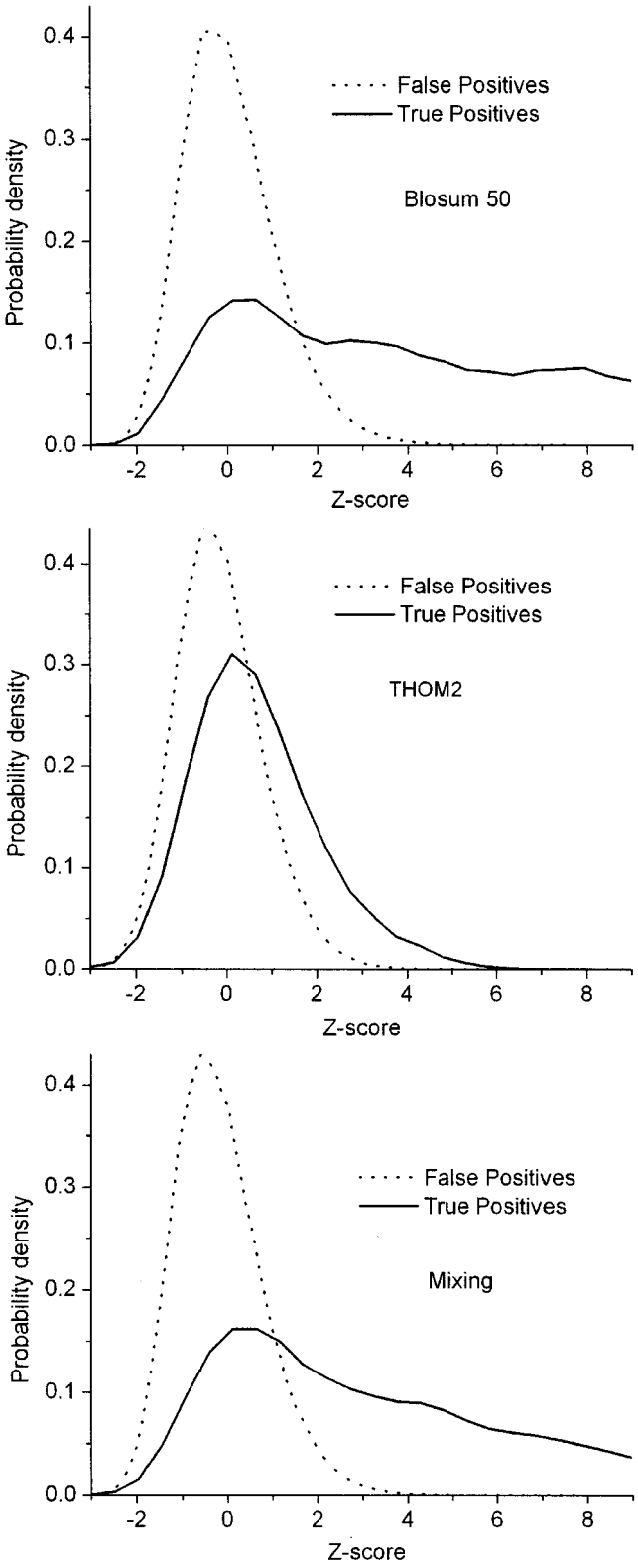

Fig. 4. The distribution of true and false positives for the three models using local alignments.

identifies 235 sequences with homologues in TOP 5 of the present test, whereas the mixed model and THOM2 identify 220 and 105 pairs, respectively. Ninety-four sequences are not detected by the mixed model or by THOM2. The mixed model detects 67 sequences that carry no signal from pure sequence or threading alignments. Twelve sequences are detected only by THOM2.

To appreciate the effect of variations in the mixing parameter we consider also a value of 0.3 instead of 0.125 for the mixing parameter, and we test the results on the larger test that includes many similar sequences. We have 236 pairs detected by the mixed model (71 of them detected *only* by the mixed model). With this mixing parameter, 58 sequences are detected only by sequence alignment and 15 sequences only by THOM2. We see little difference in the actual number of sequences detected when we shift the parameter from 0.125 to 0.3 (from 220 to 236, ~7% difference). What is of considerable interest is the significant number of sequences (67 and 71 sequences, respectively) undetected by both THOM2 and BLOSUM 50 and detected by the mixed model for both values of the mixing parameter.

It has been shown that a combination of local and global alignments brings added value to fold recognition.[6] Having determined a good performance value for the mixing parameter, we are ready to evaluate the behavior of the mixed model for global alignments. In Figure 5 we show the mixed model versus threading using THOM2 potential for both, number of sequences with homologues detected and the total number of homologues. We plot these values from top 5 to top 30. We see an important enhancement regardless of the definition of success. The same value of the mixing parameter and the same algorithm of combining the gap penalties as in the local alignment case have been used.

The global model shows a similar spectrum for the probability to detect a false positive above a given Z-score. We see in Figure 6 that the probability to have a false positive drops below 10^{-4} for Z-scores > 4. Generally speaking, the Z-scores of the global alignments are lower by ~0.5 than the Z-scores of local alignments for the same confidence level. The difference between the global THOM2 and the global mixed model is even lower than in the local alignment case for the Z-scores in the range of 2–4.

We also studied the Z-score probability densities for false and true positives in global alignments using threading and the mixed model. These densities behave similarly to local alignments and therefore not considered in details.

We have also checked two successful alignments of the mixed model and compared them to pure threading. We consider in detail the following two alignments: 1fgv_H (sequence) → 1kjc (structure), and 1cd8 (sequence) → 1mig_H (structure). The Z-scores for threading and for the mixed model have similar values (hence, the *threading* score is actually more significant). The first alignment was of length of 104 amino acids of which the mixed model correctly identified 73 pairs and the pure threading approach only 29 pairs. The second structural alignment was of 107 amino acids of which the mixed model identified correctly 49 amino acids and the threading approach only 37 pairs. It is known that alignments by threading (in general) are considerably worse than sequence alignment, and perhaps the mixed model is also a way of producing better alignments, maintaining (and enhancing) the ability to detect remote homologs.

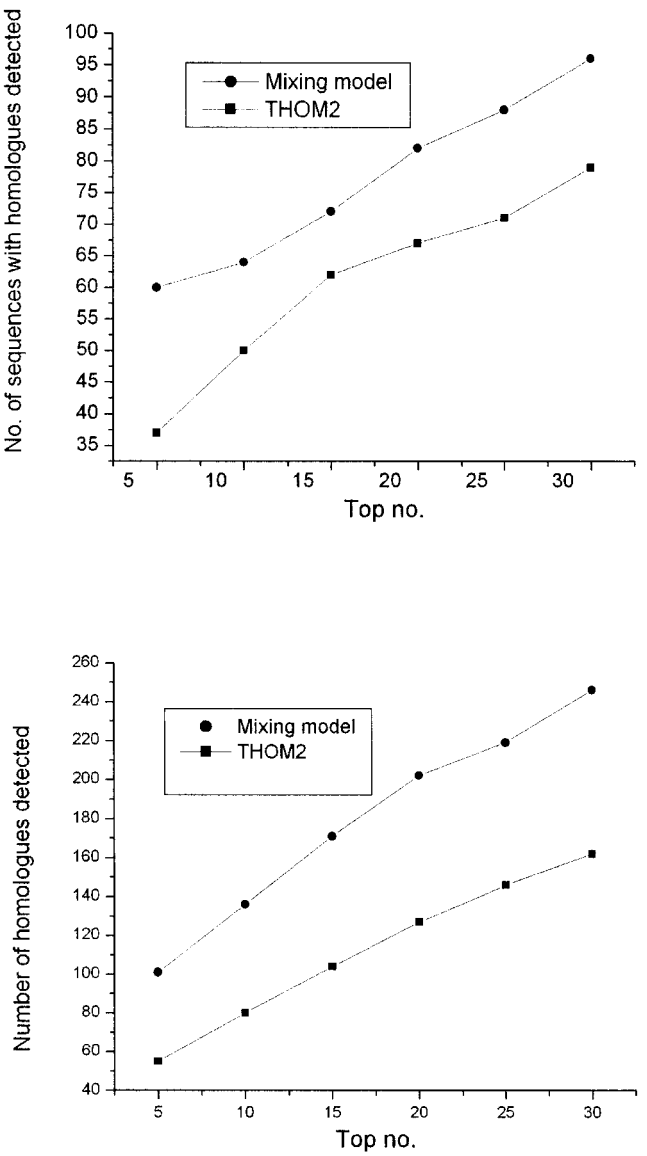

Fig. 5. A comparison of the mixed model and THOM2 for global alignments using the two difinitions of success: (i) the number of probes identified; (ii) the number of homologue proteins detected.

DISCUSSION

Algorithms for fold recognition rely on a wide range of scoring functions that test the similarity of two proteins. Measures for whole protein matches using sequence alignment, secondary structure prediction, and threading are combined together to a single score of significance. We call this approach: Combination of Global Scores (CGS). An alternative procedure of combining different measures of

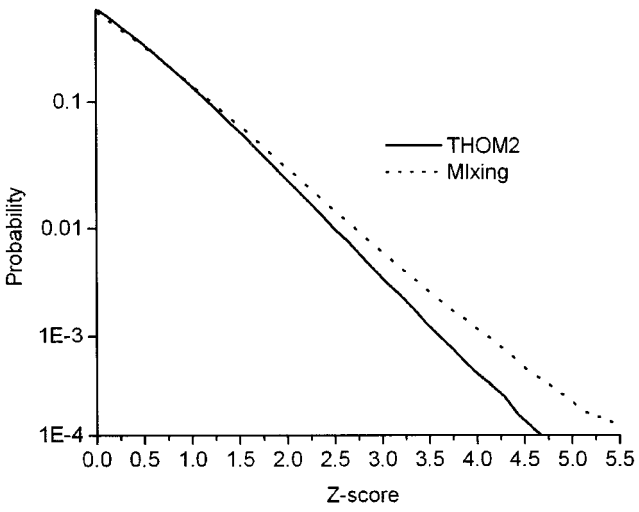

Fig. 6. Distribution of false positives for the global alignments of THOM2 and the mixed model.

similarity is at the scoring matrix level. Hence, we match an amino acid against another amino acid *and* its structural and functional properties. A single dynamic programming matrix is generated based on multiple scores. Because considerably more information is built into the alignment itself, and a single alignment is the end product, this annotation is more likely to be accurate compared with the (potentially) diverse alignments that were used in the generation of individual global scores. We call the approach that uses a single dynamic programming matrix: Combination of Local Scores (CLS).

From conceptual viewpoint we find the present approach more appealing than the CGS because a single alignment is used. From the perspective of performance, we have shown a case study in which CLS outperforms the individual scores.

For the measures used (sequence alignment based on BLOSUM 50[5] and threading based the THOM2 energy[6]) the CLS outperforms a CGS that we use in LOOPP. This is because the proteins examined have only negligible global signals from sequence alignment, leaving little to contribute to a CGS. At the same time the CLS increases the performance by up to a factor of 2. Nevertheless, this is one test and more tests will have to be done to establish the significance of the result.

For the future we expect to expand the CLS to include (at the least) also secondary structure prediction, and other measures that we worked on in the context of the LOOPP algorithm.

REFERENCES

1. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
2. An Y, Friesner RA. A novel fold recognition method using composite predicted secondary structures. Proteins Struct Funct Genet 2002;48:352–366.
3. Falicov A, Cohen FE. A surface of minimum area metric for the structural comparison of proteins. J Mol Biol 1996;258:871–892.
4. Meller J, Elber R. Protein recognition by sequence-to-structure fitness: bridging efficiency and capacity of threading models, Adv Chem Phys 2002;120:77–130.
5. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1989;89:10915–10919.
6. Meller J, Elber R. Linear optimization and a double statistical filter for protein threading protocols. Proteins Struct Funct Genet 2001;45:241–261.
7. Meller J, Elber R. unpublished results.
8. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
9. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 389–402.
11. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
12. Wagner M, Meller J, Elber R. Large-scale linear programming techniques for the design of protein folding potentials. Mathematical Programming 2003. Forthcoming.