doi:10.1006/jmbi.2001.5293 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **315**, 1257–1275

JMB

Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based on Information Theory

Golan Yona* and Michael Levitt

Department of Structural Biology, Fairchild Bldg. D-109 Stanford University CA 94305, USA

**Corresponding author*

This paper presents a novel approach to profile-profile comparison. The method compares two input profiles (like those that are generated by PSI-BLAST) and assigns a similarity score to assess their statistical similarity. Our profile-profile comparison tool, which allows for gaps, can be used to detect weak similarities between protein families. It has also been optimized to produce alignments that are in very good agreement with structural alignments. Tests show that the profile-profile alignments are indeed highly correlated with similarities between secondary structure elements and tertiary structure. Exhaustive evaluations show that our method is significantly more sensitive in detecting distant homologies than the popular profile-based search programs PSI-BLAST and IMPALA. The relative improvement is the same order of magnitude as the improvement of PSI-BLAST relative to BLAST. Our new tool often detects similarities that fall within the twilight zone of sequence similarity.

Keywords: profile-profile comparison; PSI-BLAST; structural alignment; remote homologies

Introduction

Traditionally, database searches have been the means by which new protein sequences are analyzed. The query sequence is compared with each individual sequence of the database, one at a time, in what is termed a pairwise comparison. Significant sequence similarities with database sequences may imply a common evolutionary ancestry (homology) between the corresponding proteins. Homologous sequences usually have similar fold and close or related biological function,[1,2] therefore, detecting homology can help to assign a putative function to a new protein sequence.

Pairwise sequence comparison algorithms are useful to detect similarities between sequences that have not diverged greatly, beyond the twilight zone of sequence similarity (defined as 20%-30% sequence identity[3]). However, in the coarse of evolution, sequences may have changed significantly due to mutations and insertions, and in many cases proteins may still have the same fold and close biological function without a significant sequence similarity.[4–6] Those similarities are usually missed by pairwise sequence comparisons.

A great deal of work has been done to develop tools that can detect weak relationships among sequences. Among these are search programs that use statistical representations of protein families, such as profiles[7] and hidden Markov models (HMM)[8] and the latest generation of powerful iterated search programs such as PSI-BLAST[9] and SAM-T98,[10] that use significant hits detected in the first iteration to create a profile or HMM. The model is then used to search the database again, repeatedly, until no more new hits are detected. The underlying idea in all these techniques is that by integrating the information from multiple, related sequences, one can achieve a concise, robust and powerful statistical representation of a protein family. These tools were able to enhance the ability to detect relationships between distantly related proteins and became the standard means by which sequences are analyzed these days.

Nonetheless, even iterative tools such as PSI-BLAST may miss weak sequence similarities. Moreover, these tools are sensitive to parameter tuning. For example, using a permissive threshold for inclusion in a profile may cause the inclusion of unrelated sequences in the profile and lead to diversion from the original query sequence. There-

Present address: G. Yona, Department of Computer Science, Cornell University, Ithaca, NY 14853-7501, USA.

Abbreviations used: HMM, hidden Markov model.

E-mail address of the corresponding author: golan@cs.cornell.edu

0022-2836/02/051257–19 $35.00/0

fore, users are often forced to use more stringent thresholds at the cost of sensitivity.

There are other related methods that can be quite effective in the twilight zone of sequence similarity. These methods are especially tuned to detect similarities with sequences of known structures, and most of them integrate additional information (e.g. predicted secondary structure, structural information) in the process. Such methods were applied successfully in the latest CASP (critical assessment of structure prediction) meeting†. For example, PDB-BLAST[11] is a variation on the PSI-BLAST program. PSI-BLAST is used to collect the protein sequences belonging to a particular family. Using the supplied query sequence, PSI-BLAST runs for five iterations. A profile is generated from this collection but in a different way than done in PSI-BLAST. The sequence profile is then saved and used to scan the database of proteins with known structure. SAM-T99[12] is a variation on SAM-T98 that builds a multiple alignment by iterated search using hidden Markov models. It uses the alignment to predict secondary structure (using various methods) and to build a HMM that is then used to search the PDB for similar proteins. A library of HMMs built by similar methods from PDB sequences is used to score the target sequence. INBGU[13] is a combination of five methods that exploit sequence and structure information in different ways to produce one consensus prediction. It uses predicted *versus* observed secondary structure and sequence profiles for both the query and for the folds in the library. GenTHREADER[14] uses a combination of various methods including sequence alignment with structure based scoring functions. It uses a neural network based jury system to calculate the final score for the alignment. 3D-PSSM[15] is based on a threading approach using 1D and 3D profiles coupled with secondary structure and solvation potentials. The current version of this server uses a fold library, which is automatically updated every week.

Although it seems that PSI-BLAST and similar procedures already exploit the maximum information that is encoded in the sequence, we believe that this information has not been fully utilized yet. All the aforementioned methods compare a sequence to a model as encoded in a template; their power stems from the ability of such a model to better discern related from unrelated proteins. Our approach takes this idea one step further and compares two models. Specifically, we use the profile representation (as is generated by PSI-BLAST) as a statistical model of a protein family, and compare profiles of different protein families, in search of possible remote kinship.

The choice of the model can greatly affect the effectiveness of the method. For example, a statistical model for the sequences such as a HMM makes certain assumptions on the origin and diversity of the sequences that are not always justified. Methods that make minimal assumptions about the nature of sequences are desirable, as such methods can be more robust. In our choice of the model and the statistical similarity measure we have tried to follow this guideline.

Several procedures to compare profiles have been reported in the literature. Gotoh[16] proposed an iterative alignment method to align two groups of biological sequences, including profile-based operations. However, his method is based on optimizing a weighted sum-of-pairs score, and essentially compares pairs of sequences, with overall computation time that is proportional to the product of the numbers of sequences in the two groups. Pietrokovski[17] compared profiles that were generated from multiple alignments of protein families in the Blocks database,[18] but his method does not allow gaps in the alignment. Lyngso *et al.*[19] used the co-emission probability of two profile hidden Markov models to measure their similarity. Despite the mathematical elegance of their approach, the metrics that they propose are overly sensitive to the differences between the probability distributions and to the size of the training data. In other words, their metrics emphasize the differences rather than the similarity of the two models. Therefore, it is not clear whether the method can detect subtle similarities between protein families. The most similar effort to ours was done by Rychlewski, Godzik, and colleagues.[20,21] Their profile-profile comparison procedure is based on a dynamic programming algorithm. Their algorithm uses the correlation of probability distributions as a measure of similarity between profile columns. Their method (called FFAS) was applied successfully in the last CASP meeting.

Our algorithm for profile-profile comparison is based on the classical dynamic programming algorithm first used for protein sequences almost 30 years ago[21] with the modification to allow for local similarities.[22] The novel ingredient in our procedure is the definition of profile similarity scores. Our scores are based on a powerful, information theory based, measure of similarity between probability distributions. The similarity measure is based only on the observed distributions and therefore provides a model independent criterion for comparing two statistical sources. The similarity score of two columns in two different profiles is defined as a combination of their statistical similarity and the significance of the statistical similarity. A transformation is then applied to these scores, so as to make them suitable for detecting local sequence similarities.

The paper is organized as follow: we first describe the methods and the similarity measures that we use. Then we evaluate the performance of the new tool by testing it on a large set of protein families. We end with selected examples that demonstrate the power of the profile-profile comparison method.

† http://predictioncenter.unl.gov/casp4/Casp4.html

Methods

Definition of a profile

A profile is a representation of a group of related protein sequences, usually based on a multiple alignment of those sequences (reviews on algorithms for multiple alignment[23–25]). Once the multiple alignment is defined, the profile is constructed by counting the numbers of each amino acid at each position along the multiple alignment. These counts are transformed into probabilities by normalizing the counts by the total number of amino acids and gaps observed at that position. These empirical probabilities reflect the likelihood of observing any amino acid k at position i. Since the counts are based on a finite set of sequences it can happen that not all 20 amino acids are observed at each position. Therefore, pseudo counts are introduced so that no amino acid has a zero probability to occur at position i. For more information on profile generating techniques, see Gribskov & Veretnik.[26]

Iterative search procedures such as PSI-BLAST[9] can also be used to generate multiple alignments and profiles. PSI-BLAST, arguably the most popular search method today, is an iterative version of BLAST, with a position-specific scoring matrix, which is generated from significant alignments found in round i and used in round $i + 1$.

Once the probability distributions have been calculated for each position along the multiple alignment the profile is defined as a series of probability distributions (one per each position) $\mathbf{P} = \mathbf{p_1}\mathbf{p_2}\cdots\mathbf{p_n}$ where n is the length of the multiple alignment, and $\mathbf{p_i}$ is a probability distribution over the 20 amino acids at position i. One can think of the profile as a (k,i) matrix of 20 rows and n columns. Each probability distribution is one column in the matrix representation and hence is called profile column.

The data set

There are several publicly available classifications of protein architectures including SCOP,[27] CATH[28] and FSSP/DALI.[29] These classifications provide excellent sets for testing protein sequence and structure comparison algorithms. Whereas SCOP is built by the careful manual curation of Dr Alexei Murzin, both CATH and FSSP are built more or less automatically from structural alignments. CATH has a rather simple hierarchy with just four fold classes and a few tens of architectures in each class. SCOP has a much more complicated hierarchy with sevenfold classes, some containing over a 100-folds. The FSSP classification is automatic with a hierarchy built by the Z-score similarity of proteins in each branch of the tree. While the CATH and FSSP classifications use protein chains as the object of interest, SCOP breaks proteins into domains as required, thus eliminating the problem of placing multi-domain proteins in the hierarchy. Our choice of the SCOP classification was motivated by the high quality of this database, the use of domains instead of complete protein chains, and our extensive experience with this database for sequence and structure classifications.[30–32]

We use the SCOP 1.50 classification of protein structures; This manually curated database contains 23,780 protein domains classified into 1287 protein families, 814 superfamilies, 545 folds and seven classes. Each of the 1287 families is represented by a profile that is generated using PSI-BLAST. We first select as a seed the sequence whose average distance from all the other members in the family is the smallest. Then we use PSI-BLAST to run this seed sequence against the sequences in the family, after eliminating identical entries. Families for which there is only one member, or for which PSI-BLAST failed to generate a profile, were represented by a profile generated directly from the seed sequence by using probabilities derived from the original BLOSUM62 frequency matrix.[33]

It is well known that a general PSI-BLAST search is very sensitive to the program parameters (threshold for inclusion in the profile, number of iterations). An iterative PSI-BLAST search may overestimate the statistical significance of matches with unrelated sequences, as a result of integrating unrelated sequences into the profile. Our PSI-BLAST results are somewhat cleaner than those from a typical PSI-BLAST search. The seed sequence is searched only against the other members of the family (the ''database''). Since there are no unrelated sequences in the database, there is no danger that false positives will be included in the profile. Therefore, our procedure creates a clean profile that reliably represents the protein family and is less error-prone than profiles that are generated by an iterative PSI-BLAST search against a large sequence database.

For parameter optimization and performance evaluation we used a subset of 456 families. Those are all families within superfamilies that contain at least two other families (the largest superfamily, the ''Winged helix'' DNA-binding domain, contains 21 families). Approximately one-quarter of those families (120 families, each with at least six members after eliminating redundancy, and with a seed sequence longer than 50 amino acids) were chosen as the training set for parameter optimization (see below).

Profile-profile comparison

The profile-profile comparison is performed using dynamic programming algorithm, and the alignment is assigned a score that accounts for matches, insertions and deletions, much in the same way sequence-sequence alignment is calculated. The differences are in the scoring scheme. Unlike sequence-sequence comparison, where a

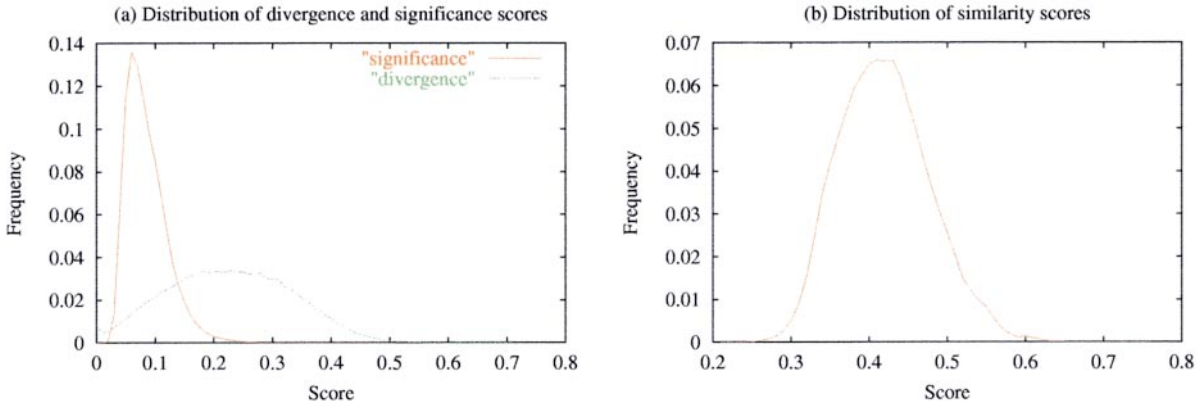

Figure 1. (a) **Distributions of** divergence scores **and** significance scores **(a) and of** similarity scores **(b)**. All distributions are based on the largest 100 families in the SCOP 1.50 database. For each family a profile was generated (see Methods) and the divergence score D^{JS}, the significance score S and the similarity score were calculated for every pair of columns along the profile, giving a total of 2986,151 column pairs.

scoring matrix like BLOSUM62 gives the score for different pairs of aligned amino acids, profile-profile comparison is more complicated. The core of our procedure is the definition of profile similarity scores, and the parameters used to quantify this measure of similarity. Our scheme has been described before, but in much less detail.[34,35]

The divergence score

Given two profiles $\mathbf{P} = \mathbf{p_1}\mathbf{p_2}\mathbf{p_3}\cdots\mathbf{p}_n$ and $\mathbf{Q} = \mathbf{q_1}\mathbf{q_2}\mathbf{q_3}\cdots\mathbf{q}_m$, where n and m are the lengths of the profiles (the number of positions or columns) and $\mathbf{p}_i$, $\mathbf{q}_j$ are probability distributions over the 20 letter alphabet of amino acids, we define the match score between two columns $\mathbf{p}_i$ and $\mathbf{q}_j$ based on their statistical similarity.

A commonly used measure of statistical similarity between two arbitrary probability distributions $\mathbf{p}_i(x)$ and $\mathbf{q}_j(x)$, is the Kullback-Leibler (KL) divergence[36] defined as:

$$D^{KL}[\mathbf{p_i}||\mathbf{q_j}] = \sum_k p_{ik} \log_2 \frac{p_{ik}}{q_{jk}}$$

This measure has the disadvantages of being asymmetric and unbounded. A better measure of statistical similarity is the Jensen-Shannon (JS) divergence between probability distributions.[37]

Given two (empirical) probability distributions $\mathbf{p}$ and $\mathbf{q}$, for every $0 \leqslant \lambda \leqslant 1$, the λ-JS divergence is defined as:

$$D^{JS}_{\lambda}[\mathbf{p}||\mathbf{q}] = \lambda D^{KL}[\mathbf{p}||\mathbf{r}] + (1-\lambda) D^{KL}[\mathbf{q}||\mathbf{r}]$$

where:

$$\mathbf{r} = \lambda\mathbf{p} + (1-\lambda)\mathbf{q}$$

can be considered as the most likely common source distribution of both distributions $\mathbf{p}$ and $\mathbf{q}$, with λ as a prior weight. Without *a priori* information, a natural choice is $\lambda = 1/2$. We call the corresponding measure the divergence score and denote it by D^{JS}. This measure is symmetric and ranges between 0 and 1, where the divergence for identical distributions is 0. In Figure 1(a) we plot the distribution of D^{JS} for actual amino acid distributions in profiles of groups of related proteins.

An attractive feature of the D^{JS} divergence measure is that it is proportional to minus logarithm of the probability that the two empirical distributions represent samples drawn from the same (''common'') source distribution.[38] This aspect of the similarity measure makes it appealing in the context of protein sequence comparison, since by comparing profiles we wish to detect an evolutionary relationship, i.e. a common ancestry. The common source distribution as defined above is the source distribution most likely to produce the two distributions that are actually observed for the two profiles that are being compared. A small value of D^{JS} indicates that the two profile columns are closely related, and may be well approximated by the distribution of the common source†.

The significance score

While a statistical measure estimating the probability that two distributions represent the same

† This in itself is not enough to determine common ancestry, the same way that similarity or identity of two amino acids along a sequence alignment is not enough to determine kinship. Only the overall similarity score of the alignment may indicate such kinship. In that case a profile that consists of the common source distributions (along the alignment) can be considered as a faithful representation of the common ancestry.

source distribution seems appropriate for the comparison of profiles, a major ingredient is ignored; the *a priori* probability of the source distribution. Imagine that the two particular empirical distributions each resembles the overall distribution of amino acids in the database (i.e. the distribution of the common source is similar to the background distribution). In this case, should they be considered significantly similar? Obviously not, as they both match the distribution expected for random profiles. Clearly this match is not as significant as a match of two probability distributions that both resemble a unique distribution (i.e. a distribution different from the overall distribution of amino acids in the database). In other words, the similarity of two random distributions is not as significant as the similarity of two unique distributions.

To assess the significance score S of a match we measure the JS divergence of the (common) source distribution, $\mathbf{r}$, from the base (background) distribution, $\mathbf{P_0}$ (defined as the overall amino acid distribution in a large sequence database, such as SWISSPROT + TrEMBL[40]):

$$S = D^{\mathrm{JS}}[\mathbf{r}||\mathbf{P_0}]$$

This measure reflects the probability that the source distribution, $\mathbf{r}$, could have been obtained by chance. The higher it is, the more distinctive is the common source distribution, and the lower is the probability that it could have been obtained by chance. The distribution of significance scores is shown in Figure 1(a).

Combining the divergence score and the significance score in a single similarity score

Because our method uses dynamic programming to compare profiles, a match between two columns needs to be assigned a single score†. This score should ideally reflect both the divergence score and the significance score. Therefore, we define the similarity score of two probability distributions $\mathbf{p}$ and $\mathbf{q}$ to be:

$$\begin{aligned} Score(\mathbf{p},\mathbf{q}) &= \frac{1}{2}(1-D)(1+S) \\ &= \frac{1}{2}(1-D^{\mathrm{JS}}[\mathbf{p}||\mathbf{q}])(1+D^{\mathrm{JS}}[\mathbf{r}||\mathbf{P_0}]) \end{aligned}$$

With this expression, the similarity score of two similar distributions ($D \to 0$) whose common source is far from the background distribution ($S \to 1$), tends to one. On the other hand, the similarity score of two dissimilar distributions ($D \to 1$) whose most likely common source distribution resembles the background distribution ($S \to 0$) tends to zero. This scoring scheme also distinguishes two distributions that each are similar to the background distribution ($D \to 0$ and $S \to 0$, giving $Score = 1/2$) from two dissimilar distributions, but whose common source is similar to the background distribution ($D \to 1$ and $S \to 0$, giving $Score = 0$). The distribution of similarity scores observed in protein families is plotted in Figure 1(b).

† One may think of a different framework, in which these two different scores are treated separately. For example, use the divergence scores to assess the overall score of the alignment, and use the significance scores to assess the overall significance of the alignment. However, in such scenario, an alignment that is optimized for the divergence score is not necessarily optimized for the significance score.

Score transformation - creating a local alignment scoring scheme

The profile similarity scores defined in the previous section range from zero to one. For local alignments, the similarity function $Score(a,b)$ must satisfy two requirements: (i) $mean(Score(a,b)) < 0$ and (ii) $max\{Score(a,b)\} > 0$. The first requirement guarantees that the average score of a random match is negative (otherwise, an extension of a random match would tend to increase its score, contradicting the idea of local similarity). The second condition means that a match with a positive score is possible. These criteria are satisfied by all standard scoring matrices, such as the BLOSUM and the PAM matrices.

Our profile similarity scores must be adjusted to meet these requirements. A simple transformation would be to subtract a constant offset from all similarity scores. Here we tested this transformation (named shift transformation) extensively. We also tested a more elaborate transformation (named mass-conserving transformation), which has been described elsewhere.[35]

The shift transformation

The average similarity score calculated for a large set of profile column pairs helps determine the constant offset for the shift transformation. Based on the largest 100 families in the SCOP database, this average is 0.42. Such a shift would give a scoring scheme whose average is zero. However, the optimum average may well be less than zero, as is the case for BLOSUM62 matrix, where the average is −0.95. It has been shown that the statistical properties of matches between random sequences are very sensitive to the exact average score value.[40,41] Matches that are scored using a scoring function with zero average may have unstable properties so that small fluctuations may greatly affect the selectivity of the method. Therefore, the shift should be higher than 0.42 to insure selectivity, but it should not be so much higher as to reduce sensitivity.

To limit the search space we first calculated four different distributions of similarity scores for a large sample of profile columns (see Figure 2(a);

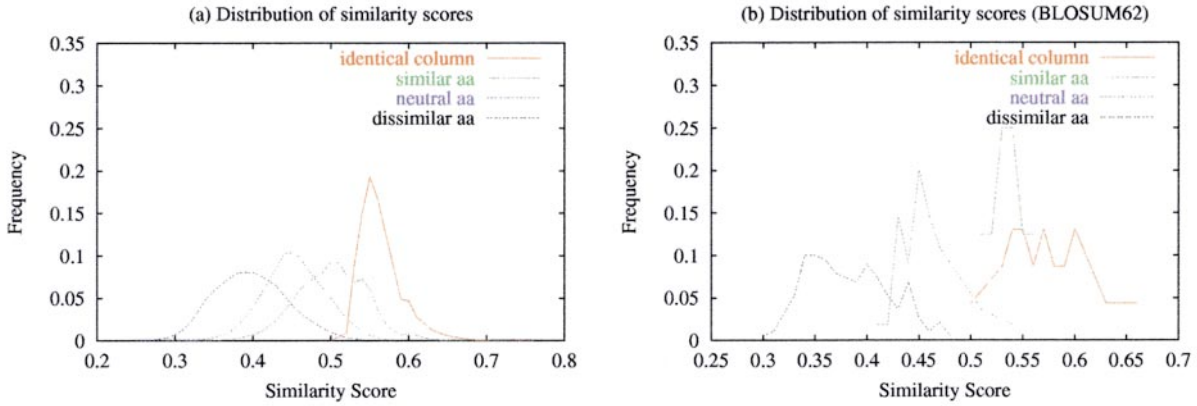

Figure 2. (a) **Distribution of** similarity scores **for different column types**. The distributions are based on the largest 100 families in SCOP 1.50 database. The pairs of profile columns are divided into four categories depending on the nature of the seed amino acids. The categories are: (1) identical columns (a column with itself); (2) different columns with similar or same seed amino acids; (3) different columns with neutral seed amino acids; and (4) different columns with dissimilar seed amino acids. (b) **Distribution of** similarity scores **for different columns of BLOSUM62**. The distributions for the BLOSUM62 matrix are derived by first converting the frequencies to probability values. Each column, which corresponds to the replacement probability of a particular amino acid with all others, is akin to a column in a profile and different columns can be compared using our similarity score. Four distributions are plotted as for (a). In this case, the ''similar aa'' category does not include the pairs with the same amino acid; these are counted in the ''identical column'' category.

the four different distributions correspond to different column types). The same four distributions were calculated from the original BLOSUM62 frequency matrix[33] (Figure 2(b)).

As Figure 2 shows, around 0.5 there is a clear distinction between distributions of identical columns (red line) and distributions with dissimilar seed amino acids (black line). All identical columns scored at least 0.5, as well as a substantial part of the distributions with similar seed amino acid. The same is true for similarity scores of distributions derived from the BLOSUM62 matrix. In addition, distributions with mutually neutral seed amino acids peak at a similarity score around 0.45. In local alignments one would expect a neutral amino acid to score close to zero. This indicates that the shift should lie between 0.42 and 0.5.

To derive a sensitive scoring scheme, the shift parameter should be defined more precisely. In addition, gap penalties can make the difference between a sensitive scoring scheme and a mediocre one, so they also need to be optimized. In Optimization of Parameters, below, we embarked upon extensive tests to define the best set of parameters, and the best score transformation technique.

Optimization of Parameters

In search of maximal sensitivity

To optimize the parameters we used a test set from the SCOP database (see above). This set comprises of 120 families, each of which has at least two related families within the same SCOP superfamily. We refer to family-family relationships within the same superfamily as true relationships and to all other relationships as false relationships. Given a specific set of parameters, for each family in the test set we calculate the profile-profile similarities with all 1287 families. For each family the results are sorted and we count the number of true family-family relationships detected before the first false relationship is detected. Finally, the results are summed over all families in the test set to give the total number of true family-family relationships detected.

A total of $6 \times 4 \times 4$ sets of parameters were tested, with shift value between 0.43 and 0.5 (values of 0.43, 0.44, 0.45, 0.46, 0.47, 0.5), a gap opening penalty between 1 and 4 (values of 1, 2, 3, 4), and gap extension penalty between 0.1 and 0.4 (values of 0.1, 0.2, 0.3, 0.4). In keeping with the BLOSUM62 matrix and the default gap parameters used by PSI-BLAST we set the gap opening penalty parameter to start at the maximal match value (here 1.0), and the gap extension penalties to be one order of magnitude smaller. The results, for selected sets of parameters, and for the scoring scheme based on the mass conserving transformation, are given in Figure 3.

The shift transformation did better than the mass-conserving transformation, performing very well for several sets of parameters. These sets all have a shift parameter of 0.45, and have gap penalty pairs (opening, extending) of (2,0.2), (2,0.3), (2,0.4), (3,0.1), (3,0.2), (3,0.3) and (3,0.4). The best set of parameters was selected using a second test described next.

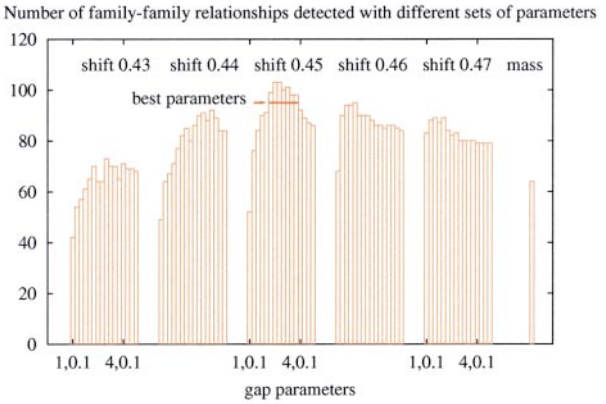

Figure 3. Performance for different sets of parameters. For each set of parameters, we give the number of true family-family relationships that are detected before the first erroneous relationship (see the text for details). Parameter sets with a shift value of 0.5 performed worse than the others and were omitted for greater clarity.

In search of maximal accuracy

With our profile-profile procedure we wish not only to detect as many relationships as possible (maximum sensitivity). We would also like to produce reliable alignments with the highest accuracy. To assess the quality of profile-profile alignments we compared them with structural alignments. The structural alignments are calculated between the structures of the family seeds (the same seeds that were used to generate the profiles) using two methods, *Structal*[30] and *CE*.[41]

Clearly, the test set of family-family alignments should be the same for each set of parameters tested. However, alignments are not necessarily reported at the same level of significance by different sets of parameters. Indeed, some alignments reported as significant with one parameter set may not be significant with another set. To create a consistent set of alignments that are considered significant by all sets tested, one needs a clear definition of significance. We elaborate on this below. Using such statistical estimates, we are able to associate with each raw alignment score an *E*-value (expectation value), which is the number of times such a score would be expected to be obtained by chance. In our final set of test alignments, we consider only alignments that are reported with *E*-value $\leqslant 1$, and take the intersection of the alignments of all parameter sets as a common subset of 82 alignments.

Several different indices are used to measure the quality of a profile-profile alignment with respect to a structural alignment (see Figure 4). (a) N_{aligned} is the overlap index, which is the total number of positions in the query sequence that are aligned by both methods, excluding gaps. (b) Q_{shift} is the shift index, which is the average shift of positions along the aligned region. (c)$N_{\text{agreement}}$ is the quality index, which is the number of positions in both alignments that are in agreement (number of pairs that are aligned identically in both alignments). This defines the ''correctly aligned'' positions. (d) Q_{modeler} is the quality of the alignment from the modeler's point of view;[42] it is the fraction of correctly aligned positions in the profile-profile alignment (number of correctly aligned positions divided by total number of aligned positions). If we were to build a 3D model based on profile-profile alignment, this index indicates the fraction of the model that would ''agree'' with the structure of the target when it becomes known. (e) $Q_{\text{developer}}$ is the quality of the alignment from the developer's point of view;[42] it is the fraction of correctly aligned positions in the structural alignment. All these indices are shown in Figure 4 for the seven best sets of parameters (see above).

Since the average length of the profile-profile alignment tends to decrease with increasing the gap opening penalty, Q_{modeler} clearly increases, while $Q_{\text{developer}}$ decreases (see opposing trends seen in Figure 4(d) and (e)). A better quality index should take into account those parts of the structural alignment that are missed by profile-profile alignment as well as parts that are added by the profile-profile alignment. We define Q_{combined} as the fraction of correctly aligned positions divided by the total number of positions that are aligned by either *Structal* or *prof_sim* (Figure 4(f)).

Based on these graphs one set of parameters stands out; that with gap opening penalty of 2 and gap extension penalty of 0.2. Similar results were obtained when we used *CE* instead of *Structal*. It should be noted that the differences in performance between all seven candidate sets of parameters both in terms of accuracy and sensitivity are relatively marginal, and specifically five sets of parameters provide overall good performance: (2,0.2), (2,0.3), (2,0.4), (3,0.1) and (3,0.2). With larger test sets, the parameter pairs (3,0.1) and (2,0.3) performed slightly better in terms of overall sensitivity (number of true family-family relationships detected), but produced shorter alignments. We recommend using the set (2,0.2) because it is almost as sensitive as the best of these sets, and it produces relatively longer alignments with overall slightly better accuracy.

Statistical Significance of Profile-Profile Matches

In order to distinguish true similarities from random matches one need to use a statistical measure that estimates the probability that a particular match could have been obtained by chance. Though statistically significant similarity is neither necessary nor sufficient for a biological relation-

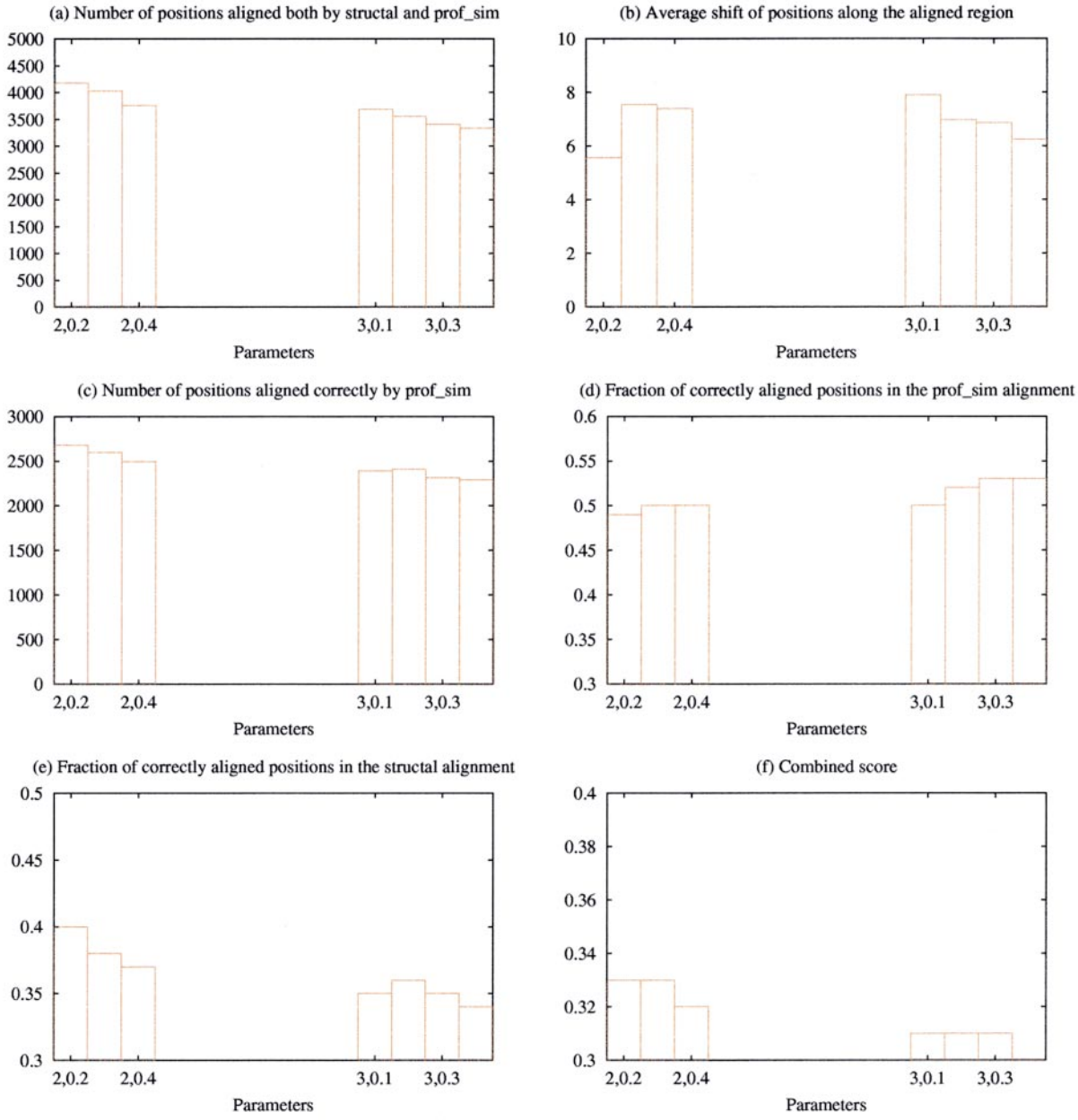

Figure 4. Accuracy of profile-profile alignments. For each candidate set of parameters (see Optimization of Parameters) we measured the quality of the profile-profile alignments with respect to structural alignments that were generated using *Structal*. Several indices of quality, described fully in the text, were used: (a) N_{aligned}; (b) Q_{shift}; (c) $N_{\text{agreement}}$; (d) Q_{modeler}; (e) $Q_{\text{developer}}$; (f) Q_{combined}.

ship, it may give us a good indication of such relationship.

Empirical studies[43,44] have shown that the distribution of local gapped similarity scores can be well approximated by the extreme value distribution[45] (though some correction factors may apply[46])†. To assess the significance of profile-profile matches we established two baseline empirical distributions.

The first is based on matches of profiles of unrelated families (families that belong to different SCOP classes and do not share significant structural similarity); it can be used to assess the significance of a match for any two given profiles, without further computations. This distribution is shown in Figure 5(a). Effectively, this distribution

† Local ungapped alignments were studied intensively and characterized mathematically and were shown to follow the extreme-value distribution.[40,47,48] However, introducing gaps in the alignments greatly complicates their mathematical tractability, and rigorous results have been obtained only for local alignments without gaps. Nevertheless, recent studies strongly suggest that the score of local gapped alignments can be characterized in the same manner as the score of local ungapped alignments.

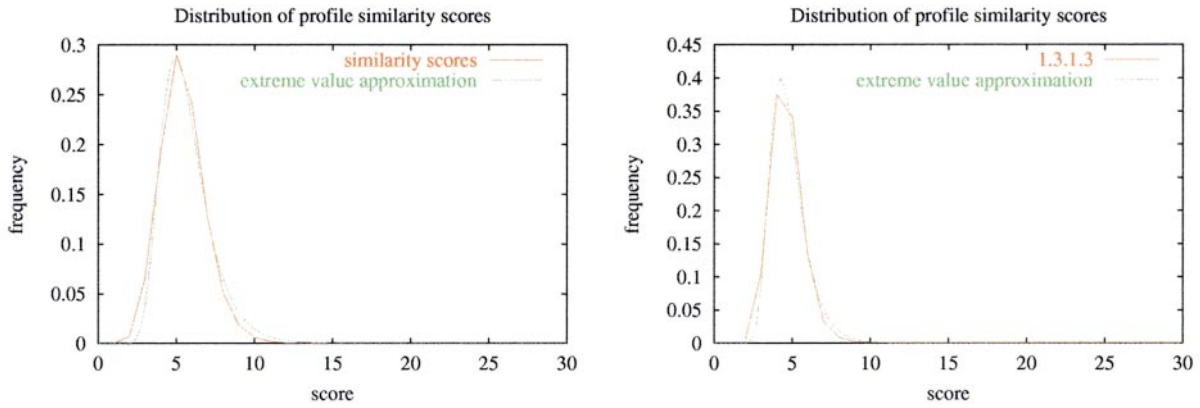

Figure 5. (a) Distribution of profile similarity scores of random profiles. The distribution is based on a large set of profile-profile similarity scores of unrelated families. **(b) Distribution of profile similarity scores for a specific SCOP family** (SCOP ID 1.3.1.3, the two-domain cytochrome *c*; the distributions for other families resemble this distribution). Both distributions follow an extreme value distribution.

reflects the distribution of similarity scores of random profiles. An extreme-value distribution was fitted to this distribution, so as to estimate the statistical significance (*E*-value) for any raw similarity score.

The second distribution is based on matches of profiles of a specific family; it provides a better approach to assess the significance of matches with the particular profile. When calculating the similarity score of a given profile with all other profiles, one gets hundreds of scores from profiles that are unrelated to the query profile. These profiles are effectively random, and the corresponding scores provide a reliable baseline distribution. We derive for each profile the similarity scores with all other profiles and fit an extreme-value distribution to the resulting distribution of scores, after excluding high-scoring profiles (see Figure 5(b)). Based on the theoretical fit we estimate the *E*-value of raw similarity scores. This approach, which allows for any unusual properties of the query profile (e.g. unusual amino acid composition) is self-calibrated, and has proved to be more accurate and robust.[6,49,50]

Performance Evaluation

To evaluate the sensitivity and selectivity of our algorithm in detecting weak relationships between protein families, we have checked it against relationships implied by the SCOP classification. Our test set is composed of 456 families that have at least two related families within the same superfamily (see above). This gives a total number of 2492 family-family relationships.

The definition of a protein superfamily in the SCOP database is a collection of protein families that are considered to be distantly related through evolution. Proteins that belong to the same superfamily have similar structures and usually have close or related biological functions,[51,52] but sequence similarity is often not detectable. Moving up the SCOP hierarchy, protein superfamilies that belong to the same fold are expected to share substructures. In some cases even proteins of different folds within the same class may show some structural similarity. Even belonging to different classes does not necessarily imply that the proteins cannot be structurally similar as most SCOP classes share secondary structure elements. Out of the seven classes in SCOP (all-alpha, all-beta, alpha/beta, alpha and beta, multi-domain proteins (alpha and beta), membrane and cell surface proteins, and small proteins), only two classes (all-alpha and all-beta) are not expected to share any secondary structure elements and therefore are not expected to be similar structurally. Because similar secondary structure will affect the allowed amino acids in a particular region, these preferences may be detected by profile-profile comparison.

Two quality measures are used to assess the effectiveness of parameters and evaluate the performance based on this hierarchy. The first, which is applied to each family individually, is the number of true family-family relationships that are detected before the first false connection occurs (as above). The definition of a false connection is subtle. One may say that any relationship besides true family-family relationship is false. Some may argue that relationships between families that belong to the same fold may as well be true. We define a relationship between two protein families to be a true relationship if both families belong to the same superfamily, a possible relationship if both families belong to the same SCOP class, an error if one protein is all-alpha and the second is all-beta, and suspicious otherwise. For each family in the test set we calculated the profile-profile similarities with all other 1286 families. The results are sorted by significance (*E*-value) and we count the number of family-family relationships that are detected before the first possible, suspicious and

Table 1. Performance evaluation results

Type of first false-connection	Number of true family-family relationships detected by:			
	Gapped-BLAST	IMPALA	PSI-BLAST	*prof_sim*
Different superfamily, same class (''possible'' relationship)	163	168	189 (205)	231
Different class (''suspicious'' relationship)	174	189	205 (221)	253
Alpha ↔ Beta (''error'' relationship)	709	810	690 (694)	1586

Methods compared: Gapped-BLAST, PSI-BLAST, IMPALA and profile-profile similarity (*prof_sim*). For each method, the number of true relationships that are detected before the first false connection occurs, is given. Results are given for the following types of false connections: possible, suspicious and error (see the text for details).

error relationship is reported†. The results, summed for all families in the test set are summarized in Table 1.

For comparison we provide the same numbers for Gapped-Blast, PSI-BLAST and IMPALA.[53] For PSI-BLAST we use the family profiles to search the set of 1287 seed proteins of all 1287 protein families in our benchmark. For each input profile, we run PSI-BLAST only for one iteration to score the seed sequences. This is essentially a single iteration BLAST search (the iterative phase of PSI-BLAST is in creating the profile, as described above). Therefore, the statistical estimates should be as reliable as in a standard BLAST search. IMPALA is different, since it compares a sequence with a library of profiles; it uses statistical estimates that are better than those used by PSI-BLAST as well as a rigorous dynamic programming algorithm. Our library of profiles contains all 1287 family profiles (see above). For each family in the test set the query sequence is selected to be the same seed sequence that was used to generate the profile for this family.

When comparing the performance of our method to that of PSI-BLAST we tried to focus on the comparison method, while keeping all other components of the evaluation procedure identical for all methods compared. In particular, the input (the profile) was fixed. Iterative PSI-BLAST may introduce other, related or unrelated sequences in the profile, and the input profile is changing from one iteration to the next one, thus violating the setup of the experiment, and creating a bias. Under this setting it is harder to compare PSI-BLAST and *prof_sim* and draw decisive conclusions from the results, since the two methods do not use the same input models anymore. Nevertheless, we have also tried iterative PSI-BLAST with up to ten iterations. There is an improvement in PSI-BLAST performance (see third column in Table 1, numbers in parentheses), however, *prof_sim* is still more sensitive, even without incorporating the information from the sequences that were not part of the original input profile and were integrated into the PSI-BLAST profile in subsequent iterations. Using *prof_sim* with the updated profile that is generated in the last iteration of PSI-BLAST before convergence is expected to improve the performance of *prof_sim* as well.

As Table 1 indicates, our program, *prof_sim*, detects more true similarities in all tests (i.e. for all possible types of first false-connection). Especially interesting is the case where the first false connection is a connection between an all-alpha protein and an all-beta protein. The number of true family-family relationships that are detected before the first such erroneous connection is reported is more than twice the number of such relationships that are detected by PSI-BLAST under the same test. This fact clearly indicates that beyond pure sequence similarity, profile-profile similarity reflects the similarity in the secondary structure content of protein families. The preferences for specific types of amino acids in every position that are encoded in the profile representation, are strongly correlated with the characteristic secondary structure in that position. Different secondary structures are likely to induce different preferences which differ markedly between secondary structures such as beta strands and alpha helix, thus resulting in a low or negative similarity score for the corresponding distributions. This information is not accounted for when comparing a sequence with a sequence or a profile with a sequence.

The *E*-value that is associated with a similarity score is usually used to set a threshold above which similarities are suspected to reflect true relationships. Table 2 lists the number of true relationships that are detected by each of the four methods tested, with *E*-value $\leqslant 0.1$ (one random match per ten searches, on average). As Table 2 suggests, compared to all other methods, more true relationships are detected by *prof_sim* with *E*-value $\leqslant 0.1$, whereas less errors and suspicious connections are reported (e.g. 166 true relationships and 21 suspicious connections are reported with *prof_sim* as opposed to 146 true relationships and 31 suspicious connections with PSI-BLAST).

The second quality measure is the receiver operating characteristic (ROC), which is a common way of assessing sensitivity and selectivity. Given a

† Our tests show that the results are almost identical if we consider an erroneous match between families as being to a different fold rather than to a different superfamily

Table 2. Performance evaluation results

Relationship-type	Number of relationships with *E*-value $\leqslant 0.1$ detected by:			
	Gapped-BLAST	IMPALA	PSI-BLAST	*prof_sim*
Same superfamily (true relationship)	116	115	146	166
Same fold (''possible'' relationship)	0	0	3	1
Same class (''possible'' relationship)	18	14	17	14
Different class (''suspicious'' relationship)	31	20	31	21
Alpha $\leftrightarrow$ Beta (''error'' relationship)	1	1	1	0
Total (with *E*-value $\leqslant 0.1$)	166	150	198	202

For each method, the number of true, possible, suspicious and error relationships that are detected with *E*-value $\leqslant 0.1$ is reported.

sorted list of hits, the ROC curve plots the number of true positives that are detected as a function of the number of false positives. In Figure 6 we plot the ROC50 curves for BLAST, IMPALA, PSI-BLAST and *prof_sim*. To plot this curve for a specific method, we first sort the scores of all pairwise family-family comparisons by their *E*-value and then count the number of true positives that are detected until 50 errors occur. A true positive is defined as a match between families within the same superfamily; all other types of connections are defined as false positives. The idea behind this plot is that in scanning a database search results one may be willing to overlook few errors, if additional meaningful similarities can be detected. The area under the curve can be used to compare the overall performance of different methods. The distribution of connections and the area under the ROC curves are given in Table 3. From those results it is clear that the relative improvement of *prof_sim* with respect to PSI-BLAST is as significant as the relative improvement of PSI-BLAST with respect to BLAST.

Note that the *E*-value at which the 50th error occurs is about 0.1 for BLAST, PSI-BLAST (Table 3), and 0.14 for IMPALA and *prof_sim*. By definition, a search with a threshold *E*-value of 0.1 will yield one erroneous similarity per ten searches, on average. Since we repeat the search 456 times, we may expect as many as 46 erroneous similarities (distributed between the 456 searches). Indeed, the results for BLAST, IMPALA, PSI-BLAST and *prof_sim* are fairly consistent with this simple argument. BLAST and PSI-BLAST report total of 50 supposedly false positives with *E*-value $\leqslant 0.1$, a little bit more than expected, but within a reasonable margin (Table 3). The statistical estimates of IMPALA and *prof_sim* are more conservative. The 50th error is detected with *E*-value of 0.14, while with this *E*-value we should have expected $456 \times 0.14 = 64$ false positives. The advantage of this is that it decreases the chances of detecting chance similarity with *E*-value <0.1. The agreement of the statistical

Table 3. Performance evaluation results

Relationship-type	Number of relationships detected by:				
	Gapped-BLAST	IMPALA	PSI-BLAST	*prof_sim*	*Structal*
Same superfamily (true relationship)	116	120	146	173	355
Same fold (''possible'' relationship)	0	0	2	1	45
Same class (''possible'' relationship)	18	21	17	18	5
Different class (''suspicious'' relationship)	31	28	30	30	0
Alpha $\leftrightarrow$ Beta (''error'' relationship)	1	1	1	1	0
Total	166	170	196	223	405
E-value	0.1	0.14	0.1	0.14	4.93e-07
ROC area	5155	5322 (3.3%)	6335 (23%)	7266 (41%)	13667 (165%)

For each method, we report the number of true, possible, suspicious and error relationships that are detected until 50 false connections occur (a false connection is everything but a true relationship). Also given are the *E*-value at which the 50th error occurs, and the area under the ROC plot, with the relative improvement with respect to BLAST in parentheses. The last column lists the number of relationships that are detected with *Structal*.

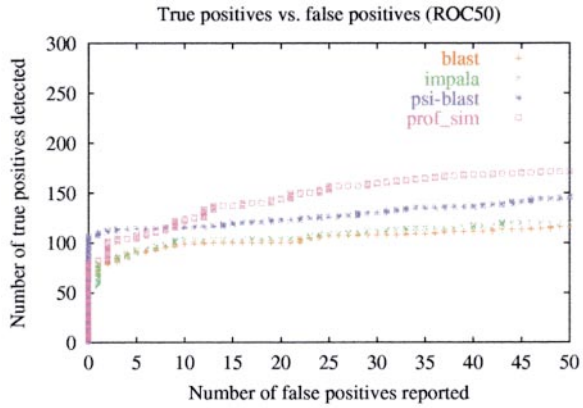

Figure 6. ROC50 curves. A true positive is defined as a connection between families within the same superfamily. Note that the relative improvement of *prof_sim* with respect to PSI-BLAST is comparable to the relative improvement of PSI-BLAST with respect to BLAST (see also Table 3).

estimates of IMPALA and *prof_sim* indicates that our statistical estimates are reliable.

There is another factor to consider when assessing the performance. In the SCOP database, there is a great variation in the number of families in a superfamily, and larger superfamilies may dominate the evaluation results. To verify that our results are not biased toward the larger superfamilies, we repeated the calculations with normalized counts. Two normalization schemes were used. In the first scheme, we normalize the counts (number of families detected) by the total number of related families in the superfamily. The second scheme, we normalize the counts by the total number of pairwise relationships in the superfamily. With the first normalization each superfamily with n families can add n at the most to the total count. With the second normalization, each superfamily can add one at the most to the total count. Without any normalization each superfamily can add up to $n(n-1)$ to the count. All schemes gave the same qualitative results: the relative improvement of *prof_sim* over PSI-BLAST is at least as large as the relative improvement of PSI-BLAST over BLAST.

It did not escape our attention that a few supposedly false positives are ranked by *prof_sim* higher than by PSI-BLAST. Consequently, when averaged over all families, PSI-BLAST (Figure 6) detects more true similarities than *prof_sim* up to eight false positives, but then *prof_sim* takes the lead and performs much better. We took a closer look at the first ten false positives. These similarities are intriguing. Four of these are between families that are classified to the same class and three of these are supported by significant structural similarities (when using the profile-profile alignment). Another four agree on about 50% of their secondary structure elements at the level of the amino acid. Therefore we strongly believe that some of these similarities actually reflect true similarities (for a more detailed analysis, see below).

The last column in Table 3 lists the number of family-family relationships that are detected with *Structal* using the same criteria. These numbers provide an upper bound on the maximal sensitivity that we should expect with sequence-based methods. Note that most allegedly false structural similarities are between families with the same fold, with a few from the same class. This is not surprising as structural similarities are common between families that adopt the same fold, and even between families from the same class. These similarities are rarely due to a common evolutionary origin, yet they may imply related biological functions. By relaxing the definition of a false positive, many more significant structural similarities are detected between families of the same superfamily, fold and class. For example, when a true positive is defined as a relationship between families from either the same superfamily or the same fold, 613 relationships are detected from the first type and 215 relationships are detected from the second type before the 50th error occurs (most of those errors are between families within the same class). The improvement in performance for the sequence-based methods is not as significant.

Alignment accuracy

Our goal in developing the profile-profile comparison algorithm was to devise a sensitive as well as an accurate algorithm, where accuracy is measured in terms of alignment accuracy. Optimizing the accuracy of sequence alignment with respect to structural alignments is especially important since one can expect that such optimized alignments will provide good initial seeds for reliable 3D models, and they can help direct site-specific experiments.

We compared the accuracy of the *prof_sim* alignments with the accuracy of BLAST, IMPALA and PSI-BLAST, using the same methodology that is described above. The results are shown in Figure 7. Note that *prof_sim* performed better in almost all tests, producing more accurate alignments, with smaller shift value, larger coverage and a higher percent of correctly aligned residues.

Detecting Interesting Similarities between Protein Families

We were curious to see what kind of similarities are missed by PSI-BLAST and detected by *prof_sim* and *vice versa*. Of all true family-family relationships that are detected by PSI-BLAST with E-value $\leqslant 1$, 26 are missed by *prof_sim* (i.e. reported with E-value > 1). Of these, ten are reported by

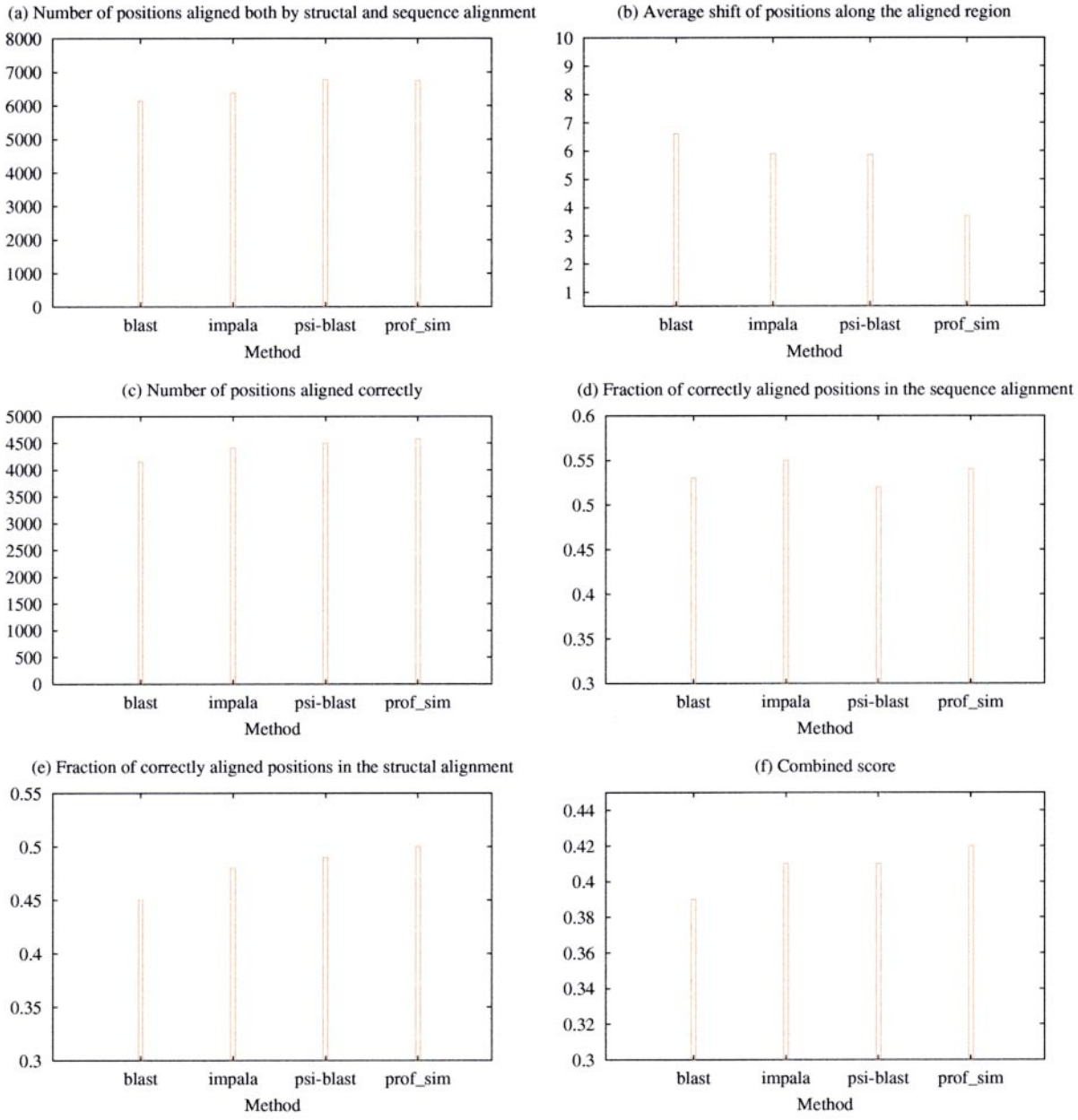

Figure 7. Accuracy of sequence-based alignments. For each sequence-based method (BLAST, IMPALA, PSI-BLAST and *prof_sim*) we measured the quality of the alignments with respect to structural alignments that were generated using *Structal*. Several indices of quality were used: (a) $N_{aligned}$; (b) Q_{shift}; (c) $N_{agreement}$; (d) $Q_{modeler}$; (e) $Q_{developer}$; (f) $Q_{combined}$. See Optimization of Parameters for details.

PSI-BLAST with *E*-value $\leqslant 0.1$ (the threshold defined by the ROC50 curve; see Table 3). The most significant hit of those is between SCOP families 1.110.1.2 and 1.110.1.1 (ARM repeat superfamily). PSI-BLAST reports this similarity with *E*-value of 2e-05 (see Figure 8). Our method, *prof_sim*, reports a slightly longer similarity with *E*-value of 5.8. The percent identity in this alignment (12 %) is much lower than the percent identity in the PSI-BLAST alignment (24 %). Strikingly, this alignment corresponds to a better match in terms of structure: aligning both structures based on the PSI-BLAST and the *prof_sim* alignments, gives RMS values of 5.7 Å and 4.1 Å, respectively (Figure 8). This indicates that *prof_sim* produces alignments that are driven by structural similarities more than other sequence-based methods. The second most significant miss is between SCOP families 1.97.4.3 and 1.97.4.2 (terpenoid cylases/protein prenyltransferases superfamily). It is reported by PSI-BLAST with *E*-value of 9e-04 and corresponds to structural similarity with RMS value of 7.0 Å. Again, *prof_sim* reports a different (tough shorter) alignment, with *E*-value 7.1, but with RMS value of 4.1 Å. The

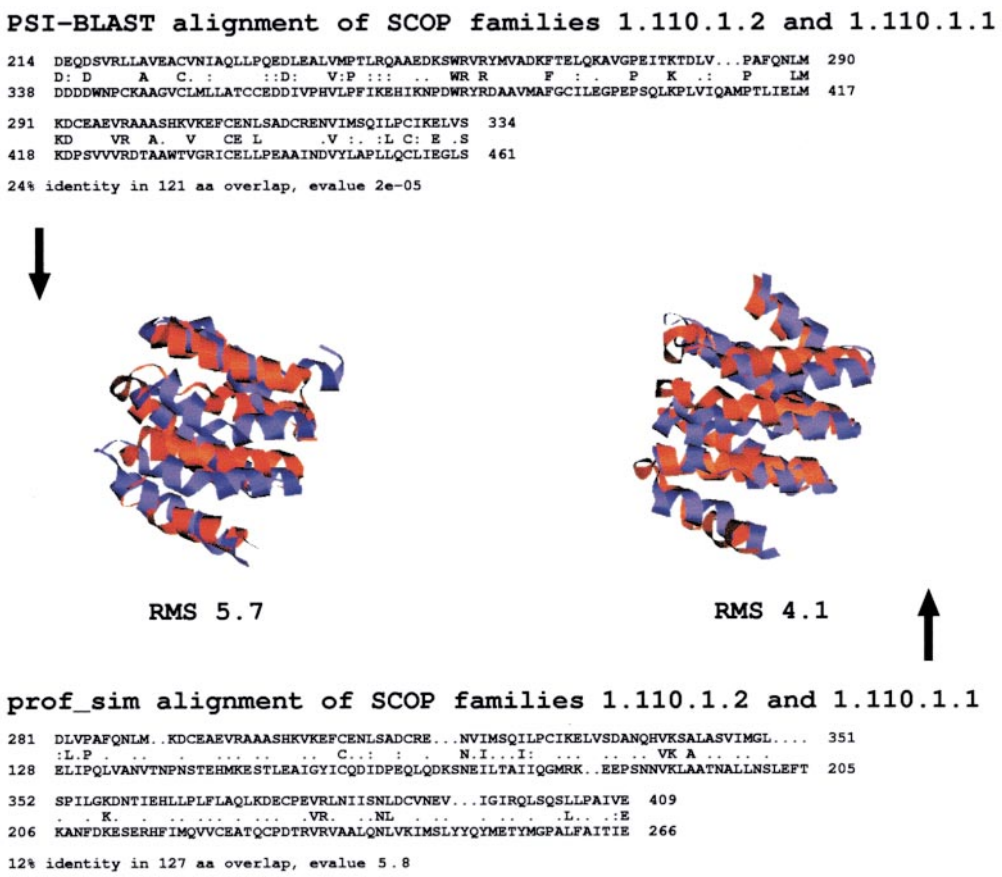

Figure 8. PSI-BLAST and *prof_sim* alignment of SCOP families 1.110.1.2 and 1.110.1.1. The SCOP identifiers of the seeds are d1b3ua and d1qgra, respectively. The *prof_sim* alignment is more accurate in terms of structure (lower RMS).

other similarities are reported by PSI-BLAST with *E*-values between 0.001 and 0.1. Our program reports similarities between the other eight family pairs with *E*-values that range between 1.3 and 10.8. In three cases *prof_sim* reports a fragment of the same alignment reported by PSI-BLAST, and in the other five cases *prof_sim* reports an alignment which differs from the PSI-BLAST alignment (though usually shorter). However, the improvement in the RMS value is significant. In five out of the eight pairs, the RMS values of the PSI-BLAST alignments ranges between 6.7 and 13.2 while the *prof_sim* alignments have RMS values between 2.06 and 6.4 (improvement ranges between 1.5 Å and 10.5 Å). In one case both alignments have high RMS values (though *prof_sim* alignment improves the RMS from 13.2 to 12.45), and in two other cases both have low and almost the same RMS values (within 0.22 Å), in one case the PSI-BLAST alignment is longer, and in the other case the *prof_sim* alignment is longer.

There are 95 true family-family relationships for which *prof_sim* reports an alignment with *E*-value $\leqslant 1$ while PSI-BLAST miss them. Forty alignments are reported with *E*-value $\leqslant 0.14$ (the threshold as defined by the ROC50). The most significant hit has an *E*-value of $2.7e-04$. Of these, 18 are reported by PSI-BLAST with *E*-value between 1 and 10, and in three more cases PSI-BLAST reports shorter similarities with *E*-values 42, 49 and 259. All the rest, 19 pairs, are not detected with *E*-value $\leqslant 500$. Five examples are shown in Figure 9; they correspond to alignments with RMS values between 1.7 Å and 5.9 Å.

In all other cases both methods reported a similarity usually with a different significance value. Those differences are expected due to the different statistical estimates or different alignments. In most cases the same or almost the same alignments are detected with *prof_sim* but because of the more conservative statistical estimates of *prof_sim* they are sometimes reported as less significant. In other cases *prof_sim* reports a different alignment.

An interesting case of homology that is not easily detected by sequence analysis methods is the actin/kinase/hsp70 homology. These three families share the same fold and are considered as homologous families.[54] In SCOP, the actins and heat shock proteins are classified into the same family (designated 3.50.1.1). The mammalian type I

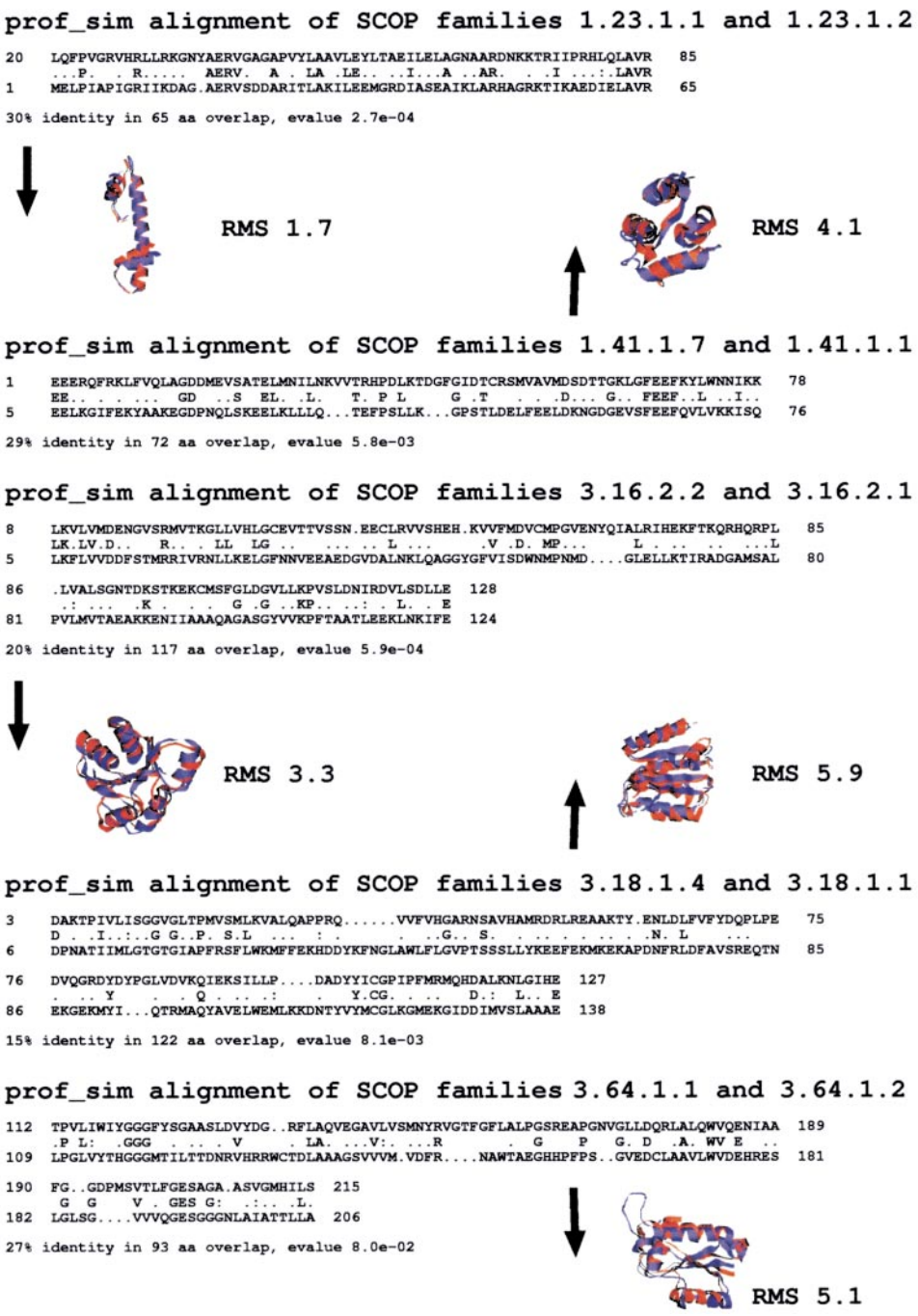

Figure 9. Significant similarities that are detected by *prof_sim* but missed by PSI-BLAST. The SCOP identifiers of the seeds: d1aoic (1.23.1.1); d1a7w (1.23.1.2); d1aj5a (1.41.1.7); d4icb (1.41.1.1); d1dcfa (3.16.2.2); d3chy (3.16.2.1); d1cqxa3 (3.18.1.4); d1fnc_2 (3.18.1.1); d1maac (3.64.1.1); and d1jkmb (3.64.1.2).

hexokinase is classified into family 3.50.1.2 and the glycerol kinase is classified to 3.50.1.3.

None of these similarities are detected by BLAST. PSI-BLAST detects the similarity between 3.50.1.1 and 3.50.1.3 with *E*-value of 4.9 (third match after two false positives). *prof_sim* detects this similarity with *E*-value of 0.31. This is the first match and there are no false positives. Moreover, the *prof_sim* alignment is longer than the PSI-BLAST alignment by 33 amino acid pairs.

When the seed sequence of family 3.50.1.3 is used as a query, PSI-BLAST misses the similarity with 3.50.1.1 (up to *E*-value of 500), while *prof_sim* detects a similarity with *E*-value of 1.0 (it is the second match after one false positive). Interestingly, the false positive (d2dik_3, family 4.121.1.5) is classified as a member of another class (class 4 of a + b proteins with segregated alpha and beta regions, while family 3.50.1.3 consists of a/b proteins), but both share similar secondary structure

elements, and to some extent, similar arrangements of these elements. None of the methods detect the similarity with family 3.50.1.2 as significant. Further tuning and integration of additional information (see Discussion) is probably needed to detect this similarity.

Analysis of errors

The SCOP classification is an excellent resource, which is based on extensive expert knowledge. Nonetheless, one should keep in mind that the definitions of the family/superfamily/fold levels in the SCOP hierarchy are based in large part on observations made by human experts, rather than on a quantitative measure. One may argue that this hierarchy is biased by human perceptions, and does not truly reflect nature hierarchy. In other words, the definitions of domains, folds and classes do not necessarily conform to ''Nature's definitions''. Therefore, any assessment using SCOP may be biased due to errors in that reference classification.

To see if some interesting similarities might have been missed by SCOP experts, we checked the most significant similarities between families from different folds. Most of the suspicious connections are due to similarity along a relatively small substructure (few common secondary structure elements). Such similarities are often observed between proteins from different SCOP classes and are not necessarily false similarities. Such similarities are usually too short to maintain the same topology, and within the context of the whole protein structure they may get different interpretations.

Out of the first 50 errors (see Table 3), nine are relatively short fragments (average length of 30) with very high agreement in secondary structure content along the aligned residues (at least 70% per alignment, and 83% on average). Additional 11 alignments (with average length of 88) agree on more than 50% of the aligned residue in terms of the secondary structure content. Therefore, we believe that at least some of the supposedly false positives are indeed true relationships.

Discussion

With the rapid increase in the number of sequences that cannot be characterized by existing tools for sequence analysis, there is a growing need for more powerful tools that can detect weak but significant sequence similarities. The latest generation of iterated search methods, such as PSI-BLAST improved sensitivity of sequence analysis significantly.[55] Yet, the vast majority of remote relationships cannot be detected by these methods.

Here, we have presented a new tool for profile-profile comparison that can be used to detect such weak relationships between protein families. What distinguishes sequence-sequence comparison from profile-profile comparison is the scoring scheme used to assess the similarity/dissimilarity of two ''atomic'' objects in the alignment (pairs of amino acids in sequence-sequence alignment and pairs of probability distributions in profile-profile comparison). The power of the profile-profile comparison lies in the definition of those similarity scores. The information that is coded in the multiple alignment that was used to generate the profile can decipher ambiguities or insignificant similarities that are frequently observed in sequence-sequence comparison, and a sensitive scoring scheme that is wisely designed can help detect those subtle similarities. Our scoring scheme was designed to obtain maximal sensitivity by using an information theory based measure of similarity between probability distributions. The significance of the profile-profile similarity scores is assessed using the same statistical framework as for sequence analysis, and extreme value distributions are fitted to the empirical distributions.

A collection of protein families from the SCOP database is used as a benchmark. This benchmark provides extensive test of the selectivity, sensitivity and accuracy of the new method we propose. Our tests show that this tool outperforms pairwise comparison algorithms such as BLAST, as well as the more powerful iterated PSI-BLAST (surprisingly, IMPALA did not perform as well as PSI-BLAST). The relative improvement of *prof_sim* over to PSI-BLAST is the same order of magnitude or even larger than the relative improvement of PSI-BLAST over BLAST. Therefore, we believe that the profile-profile measure of similarity can be used to detect weak relationships between protein families that have diverged much. Our method for parameter selection is based on a two-phase optimization procedure. This procedure is especially relevant where there are several sets of parameters that perform fairly well using one criterion (such as classification quality), but some may perform poorly when considering some other criterion (such as alignment accuracy). As our method was also optimized to produce alignments in a very good agreement with structural alignments, it had higher accuracy than all other methods, and can be expected to produce better three-dimensional homology models.

In this work we compare our method to PSI-BLAST. Of all publicly available software, PSI-BLAST is an advanced method and is probably the most popular program used today to compare protein sequences. There are other methods that are used *via* publicly accessible web-servers. The most notable of these include: PDB-BLAST and FFAS from Godzik's group;[11] SAM-T99 from Karplus' group;[12] INBGU from Fischer's group;[13] GenTHREADER Jones';[14] and 3D-PSSM from Sternberg's group.[15]

While these servers are ideally suited for use in blind prediction schemes like critical assessment of structure prediction (CASP), they cannot be easily compared with the present method as there is no way to know exactly what known relationships are built into the server. Therefore, it is hard to predict

how well our method would perform in CASP. Moreover, in the CASP meeting, methods were ranked based on the quality of the model they generated, and many methods used optimization procedures to optimize the position of side-chains and gaps, or were manually calibrated.

Unfortunately, our method was not available in time to be tested at CASP4, which closed on September 2001. We recognize that fact that our use of PSI-BLAST and profile comparison is in fact rather close to the service provided by the FFAS server.[11] This server performed well at CASP and we believe that our method performs as well†. We intend to convert our *prof_sim* method to a web-based server so that it can be tested both at CASP5 to be held in 2002 as well as by the automatic server testing schemes like LiveBench.[13]

We have integrated our method in a large-scale effort to map the protein space and create hierarchical organization of protein families and superfamilies (the BioSpace system[35]). In this study a multi-stage analysis was carried that first identified structure-based clusters (clusters with structural representatives) and sequence-based clusters (clusters with no structural representative). Clusters are compared using either a structure metric (when 3D structures are known) or our profile-profile metric. These scores are used to define a unified and consistent metric between all clusters, and clusters are organized in a meta-organization of super-clusters, using the unified metric. Many of the super-clusters contain both structure-based clusters and sequence-based clusters, due to cluster similarities that were detected only by the profile-profile metric. Based on this meta-organization we can infer plausible conformations for families, which were not structurally characterized, and are clustered into the same super-cluster as a structure-based cluster. In some cases we can also provide hints about the possible functionality of the family. We, therefore, believe that this tool can extend structure and function prediction beyond what is possible with current means of sequence analysis.

The disparity in performance between sequence-based methods and the structure-based method (as in Table 3) challenges for future developers of sequence-based methods. Further improvements of the profile-profile comparison method should consider the topology of the protein, if its structure is known, or the propensity of amino acids to be part of loops, alpha helices or beta strands. The propensities of amino acids to form loops would be especially useful as it can be used to assign variable gap penalties. These enhancements will be integrated in future versions of our method (Yona & Yeh, work in progress).

† We could not implement their method, since we could not reproduce the parameters used by Rychlewski *et al.*[11] However, preliminary results suggest that the correlation scores used in FFAS to compare probability distributions are less sensitive than our measures, which are based on information theory principles. Specifically, we calculated the distributions of correlation scores of profile columns for different column types (the same way as in Figure 2). According to these distributions (not shown) the correlation scores are less successful in distinguishing related columns from columns which are likely to be unrelated. In general the distributions highly overlap, and the tail of the second distribution falls well within the first distribution. We believe that this may affect the performance significantly.

Acknowledgments

We thank Patrice Koehl for providing us with his programs that generate structural alignments based on sequence alignments. G.Y. is supported by a Burroughs-Welcome Fellowship from the Program in Mathematics and Molecular Biology (PMMB). This work was supported by DOE award DE-FG03-95ER62135 (to M.L.).

References

1. Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56-68.
2. Hilbert, M., Bohm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins: Struct. Funct. Genet.* **17**, 138-151.
3. Doolittle, R. F. (1992). Reconstructing history with amino acid sequences. *Protein Sci.* **1**, 191-200.
4. Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12**, 861-867.
5. Pearson, W. R. (1997). Identifying distantly related protein sequences. *Comp. App. Biosci.* **13**, 325-332.
6. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.
7. Gribskov, M., Mclachlen, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355-4358.
8. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1996). Hidden Markov models in computational biology: application to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
10. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846-856.
11. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232-241.
12. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. & Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins: Struct. Funct. Genet.* **37**, 121-125.

13. Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. (2001). LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.* **10**, 352-361.
14. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.
15. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499-520.
16. Gotoh, O. (1993). Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.* **9**, 361-370.
17. Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* **24**, 3836-3845.
18. Henikoff, J. G., Henikoff, S. & Pietrokovski, S. (1999). New features of the Blocks Database servers. *Nucl. Acids Res.* **27**, 226-228.
19. Lyngso, R. B., Pedersen, C. N. S. & Nielsen, H. (1999). Metrics and similarity measures for hidden Markov models. In *The Proceedings of ISMB 1999*, pp. 178-186, AAAI Press, Menlo Park.
20. Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold. Des.* **3**, 229-238.
21. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.
22. Smith, T. F. & Waterman, M. S. (1981). Comparison of biosequences. *Adv. App. Math.* **2**, 482-489.
23. Waterman, M. S. (1995). *Introduction to Computational Biology*, Chapman & Hall, London.
24. Setubal, J. C. & Meidanis, J. (1996). *Introduction to Computational Molecular Biology*, PWS Publishing Co., Boston.
25. Taylor, W. R. (1996). Multiple protein sequence alignment: algorithms and gap insertion. *Methods Enzymol.* **266**, 343-367.
26. Gribskov, M. & Veretnik, S. (1996). Identification of sequence patterns with profile analysis. *Methods Enzymol.* **266**, 198-211.
27. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
28. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.
29. Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231-234.
30. Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA,* **95**, 5913-5920.
31. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* **7**, 445-456.
32. Brenner, S. E., Koehl, P. & Levitt, M. (2000). The astral compendium for protein structure and sequence analysis. *Nucl. Acids Res.* **28**, 255-256.
33. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.
34. Yona, G. & Levitt, M. (2000). A unified sequence-structure classification of proteins: combining sequence and structure in a map of protein space. In *The proceedings of RECOMB 2000*, pp. 308-317, ACM Press, New York.
35. Yona, G. & Levitt, M. (2000). Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins. In *The Proceedings of ISMB 2000*, pp. 395-406, AAAI Press, Menlo Park.
36. Kullback, S. (1959). *Information Theory and Statistics*, John Wiley and Sons, New York.
37. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Info. Theory,* **37**, 145-151.
38. El-Yaniv, R., Fine, S. & Tishby, N. (1997). Agnostic classification of markovian sequences. *Advan. Neural Informat. Proc. Syst.* **10**, 465-471.
39. Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.
40. Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA,* **87**, 2264-2268.
41. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739-747.
42. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L., Jr. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct. Funct. Genet.* **40**, 6-22.
43. Smith, T. F., Waterman, M. S. & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucl. Acids Res.* **13**, 645-656.
44. Waterman, M. S. & Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA,* **91**, 4625-4628.
45. Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.
46. Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460-480.
47. Dembo, A. & Karlin, S. (1991). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d variables. *Ann. Prob.* **19**, 1737-1755.
48. Dembo, A., Karlin, S. & Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.* **22**, 2022-2039.
49. Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.
50. Yona, G., Linial, N. & Linial, M. (1999). ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Struct. Funct. Genet.* **37**, 360-378.
51. Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A. *et al.* (1998). Protein folds and functions. *Structure,* **6**, 875-884.
52. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.
53. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA:

matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics,* **15**, 1000-1011.
54. Holmes, K. C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* **3**, 53-59.
55. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.

Edited by B. Honig

(Received 8 November 2000; received in revised form 18 May 2001; accepted 21 November 2001)