Using Evolutionary Information for the Query and Target Improves Fold Recognition

Björn Wallner, Huisheng Fang, Tomas Ohlson, Johannes Frey-Skött, and Arne Elofsson
Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden

ABSTRACT In this study, we show that it is possible to increase the performance over PSI-BLAST by using evolutionary information for both query and target sequences. This information can be used in three different ways: by sequence linking, profile–profile alignments, and by combining sequence–profile and profile–sequence searches. If only PSI-BLAST is used, 16% of superfamily-related protein domains can be detected at 90% specificity, but if a sequence–profile and a profile–sequence search are combined, this is increased to 20%, profile–profile searches detects 19%, whereas a linking procedure identifies 22% of these proteins. All three methods show equal performance, but the best combination of speed and accuracy seems to be obtained by the combined searches, because this method shows a good performance even at high specificity and the lowest computational cost. In addition, we show that the E-values reported by all these methods, including PSI-BLAST, underestimate the true rate of false positives. This behavior is seen even if a very strict E-value cutoff and a limited number of iterations are used. However, the difference is more pronounced with a looser E-value cutoff and more iterations. Proteins 2004;54:342–350. © 2003 Wiley-Liss, Inc.

Key words: sequence alignment; fold recognition; profile–profile; sequence linking; intermediate sequence searches; PSI-BLAST; E-value

INTRODUCTION

As the genome projects proceed, we are presented with an exponentially increasing number of protein sequences but with only a very limited knowledge of their structure or function. Because the experimental determination of structure or function is a nontrivial task, the quickest way to gain some understanding of these proteins and their genes is by relating them to proteins or genes with known properties. This can be done by searching for homologous proteins. When one chooses a method for the detection of homologous proteins, there are several factors to consider. The method should be fast enough, it should detect as distantly related proteins as possible, and clearly separate correct and incorrect hits (i.e., show a high specificity). Sometimes, these goals are not contradictory; e.g., it has been shown[1] that full Smith–Waterman[2] alignments are more sensitive than BLAST.[3] However, using the fast BLAST algorithm in PSI-BLAST makes it possible to use multiple-sequence information from a large database and thereby perform better than standard Smith–Waterman alignments, even using less computer time. It would obviously be possible to use Smith–Waterman alignments instead of BLAST in a PSI-BLAST-like procedure, but that would be intolerably slow for most practical purposes.

During the last few years, it has been shown that methods that use multiple sequences (i.e., evolutionary information) are superior to methods that only use single sequences.[4] PSI-BLAST[3] is arguably the most efficient method to detect related proteins being relatively fast, detecting distantly related proteins and showing a reliable specificity. However, for detecting distantly related proteins, slower methods that clearly perform better than PSI-BLAST do exist.[5,6]

One limitation to the PSI-BLAST procedure is that it only uses multiple-sequence information from either a query protein or the target proteins, but not from both. In this study, we examine some methods to combine the multiple-sequence information from both query and target. In principle, this can be done in at least three different ways: by using profile–profile alignments, sequence linking, or combined profile–sequence and sequence–profile searches, which we refer to as "combined searches."

Profile–profile alignments can be implemented in several different ways[7–11] and have shown a good performance. However, the alignment procedure is at least 20 times slower than traditional sequence–profile alignments, as two vectors of 20 amino acid frequencies have been compared to each other. In addition, the speed of sequence–profile searches can be increased by the use of heuristic search methods, but we cannot see how to use similar heuristic algorithms in profile–profile alignments. Finally, there is a cost to build the profiles for all the target sequences. Therefore, the computational cost of profile–

profile alignment is often too expensive for large database searches.

It is also possible to increase the detection of distantly related proteins by using linking, sometimes called intermediate sequence searches, between distantly related proteins.[1,4,12–17] Here, protein A and B can be related either directly or through an intermediate (linking) protein C. Venclovas has used this approach for detecting templates in distant homology modeling successfully in CASP4[15] and CASP5.

Finally, several fold recognition methods use a combination of sequence–profile and profile–sequence searches (e.g., 3D-PSSM performs three alignments for each sequence-template pair and by using the best of these increases the performance by 20%).[18]

Therefore, using similar methods as in earlier studies,[1,4] we decided to examine different methods that use evolutionary information for both the query and the target. We show that all three methods increase the performance over PSI-BLAST. In addition, we have studied how well the reported E-values agree with the observed E-values for the different methods.

RESULTS AND DISCUSSION

The performance of fold recognition methods can be examined by the use of several different measures. Traditionally, the number of "correctly identified" proteins has been used, but such measures do not take into account the reliability of the predictions and, therefore, might give misleading results. For instance, the number of correctly identified superfamily-related proteins increases by the use of a loose E-value cutoff in PSI-BLAST, but this increase is only obtained by a significant loss of specificity; data not shown. Instead, by studying the correct and incorrect predictions for a given score, the sensitivity of a method at a given specificity can be examined. Here, we use "spec-sens" plots (i.e., at any given score value we plot the specificity versus the sensitivity).[4] Alternatively, it is possible to use receiver operating characteristic plots (ROC), where the number of correctly identified proteins is plotted against the number of wrong predictions. In addition, we have selected a single point in the spec-sens curve that can be illustrative for comparison purposes. For family-related proteins, we chose to measure the sensitivity at 99% specificity, and for superfamilies, we chose 90% specificity. These two values are referred to as "family sensitivity" and "superfamily sensitivity'" below. Because the fold level curves only rarely reaches over 50% specificity, we did not find it particularly useful to include a fold sensitivity measure.

Tuning PSI-BLAST

There are two reasons why it is important to obtain the best possible performance from PSI-BLAST. First, it is an important baseline to use when we compare the performance of other methods. Second, we use the profiles created by PSI-BLAST in the other methods. Two alternative methods exist to use the profiles: a profile can be made either for the query protein or for each of the targets. We

TABLE I. Family Sensitivity at 99% Specificity Depending on the Number of Iterations and the E-Value Cutoff[†]

E-value cutoff	10^{-1}	10^{-2}	10^{-3}	10^{-5}	10^{-10}
Iteration 1	42	42	42	42	42
Iteration 2	61	59	58	56	52
Iteration 3	67	66	64	61	55
Iteration 5	58	69	68	64	56
Iteration 10	29	67	67	64	56

[†]Values are percents.

TABLE II. Superfamily Sensitivity at 90% Specificity Depending on the Number of Iterations and the E-Value Cutoff

E-value cutoff	10^{-1}	10^{-2}	10^{-3}	10^{-5}	10^{-10}
Iteration 1	1	1	1	1	1
Iteration 2	5	5	4	4	2
Iteration 3	9	8	7	5	3
Iteration 5	13	12	10	8	5
Iteration 10	16	15	13	9	5

Values are percents.

refer to the first scenario as profile–sequence searches and the second as sequence–profile searches. In a profile–sequence search, all sequences in a database are matched to the profile of a query protein, whereas in a sequence–profile search, the query sequence is matched against the profiles from the target proteins. Sequence–profile searches can be performed by using the IMPALA program,[19] whereas profile–sequence results can be obtained directly from PSI-BLAST. We only observed marginal differences in the performance between these two methods; therefore, we only show the results from PSI-BLAST in this section. Results from IMPALA, as well as sequence–profile and profile–sequence searches using PALIGN, are available from our web site.

To tune PSI-BLAST, there are four important parameters to consider: the choice of substitution matrix, gap penalties, E-value cutoff, and the number of allowed iterations. We have examined a large set of values for these parameters. A number of gap penalties and substitution matrices were studied, but we could not find any set of parameters that performed better than the default ones (i.e., the BLOSUM62 matrix with gap costs of -11 and -1). However, it should be noted that there were only marginal differences in performance for a large number of different parameters. Furthermore, we examined E-value cutoffs, from 10^{-0} to 10^{-50} and between 1 and 10 allowed iterations; see Tables I and II and Figure 1. The best family sensitivity was obtained by using an E-value cutoff between 10^{-2} and 10^{-4} and between 5 and 10 iterations. The best superfamily sensitivity was obtained by using 10^{-1} and 10 iterations, but this performance was only obtained by a large loss in family sensitivity due to a number of high scoring false positives.

From Tables I and II, it can be noted that the sensitivity increases quite rapidly for the first few iterations. At the first iteration, the superfamily sensitivity is only 1%, but after the next it is raised to 5% (using a cutoff of 10^{-2}), and
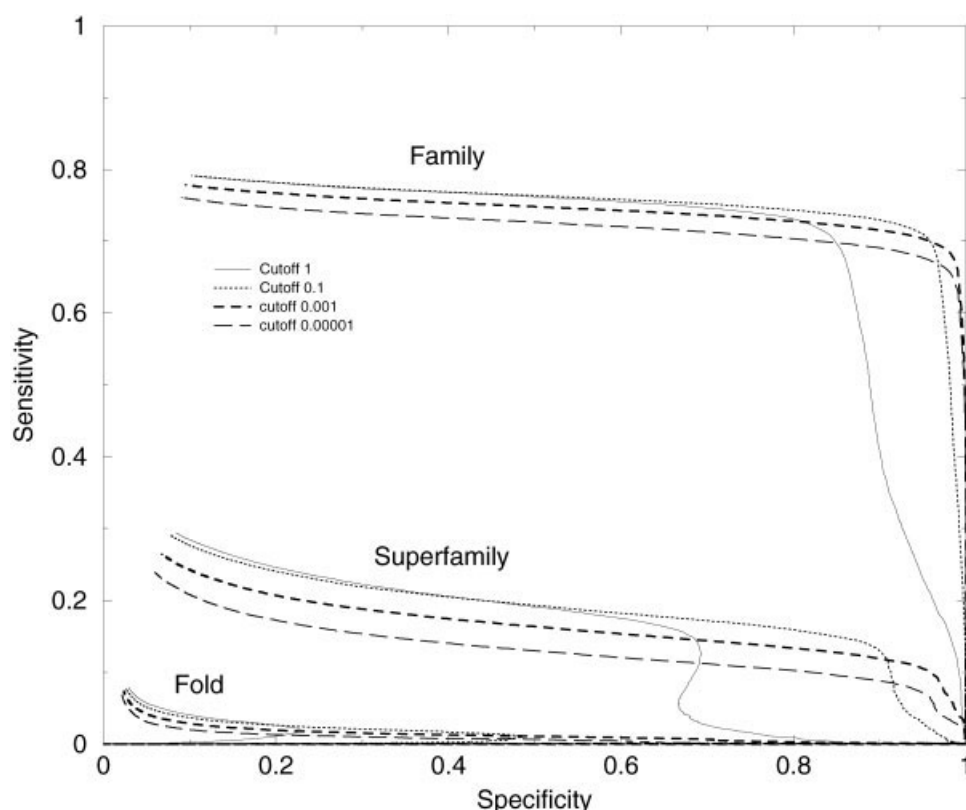
Fig. 1. Specificity versus sensitivity plot for PSI-BLAST using four different E-value cutoffs and 10 iterations. In the top four curves, the comparisons are made on family levels, in the next four on the superfamily level, and in the final four on the fold level.

even at the third iteration, the sensitivity almost doubles. The family sensitivity also increases significantly during the first three iterations, from 42 to 66%. If a loose E-value cutoff (10^{-1}) is used, a clear drop in family sensitivity is obtained after only a few iterations, whereas this is not seen during 10 iterations with a stricter cutoff. In conclusion, family sensitivity reaches a peak after a few iterations, whereas the superfamily sensitivity reaches its peak closer to 10 iterations, and the best overall performance is obtained by using an E-value cutoff close to the default (5×10^{-2}). Taking all this information into account, we decided to use 10 iterations, an E-value cutoff of either 10^{-2} or 10^{-3}, BLOSUM62, and default gap penalties as the basis for the further studies.

Using Evolutionary Information for Both the Query and Target

In profile searches, evolutionary information for the query protein or from the target sequences is used, but not from both simultaneously. There are several different methods to use evolutionary information from both queries and targets. Here, we examine three fundamentally different methods: profile–profile alignments, sequence linking and "combined searches." The performance of these methods will be compared with each other as well as with the performance of PSI-BLAST.

Profile–Profile Alignments

It is possible to align two profiles against each other, and it has been shown that such methods perform better than PSI-BLAST.[7,9] Several different profile–profile implementations exist,[7–11] and the exact details on how to best perform profile–profile alignments are not well studied but beyond the scope of this article. We have implemented the prof_sim profile–profile search algorithm developed by Yona and Levitt[7] and used the profiles obtained after 10 iterations of PSI-BLAST using either a 10^{-2} or 10^{-3} E-value cutoff. Some limited optimization of gap penalties and other parameters was tried, but we found that the default parameters worked quite well. However, we observed that if the amino acid frequencies were back calculated from the log profiles, a higher performance was obtained than if the amino acid frequencies were used directly; data not shown.

Approximately 5–10% more superfamily-related proteins can be recognized by using profile–profile alignments compared with PSI-BLAST; see Figure 2 and Table III. By using the 10^{-3} profiles, the superfamily sensitivity is increased from 13 to 17%, whereas the family sensitivity remains almost the same (68% vs 67%), and by using the 10^{-2} profiles, the superfamily sensitivity is 19%, but the family sensitivity drops to 64%. From Figure 2, it is obvious that the sensitivities for both family and superfam-
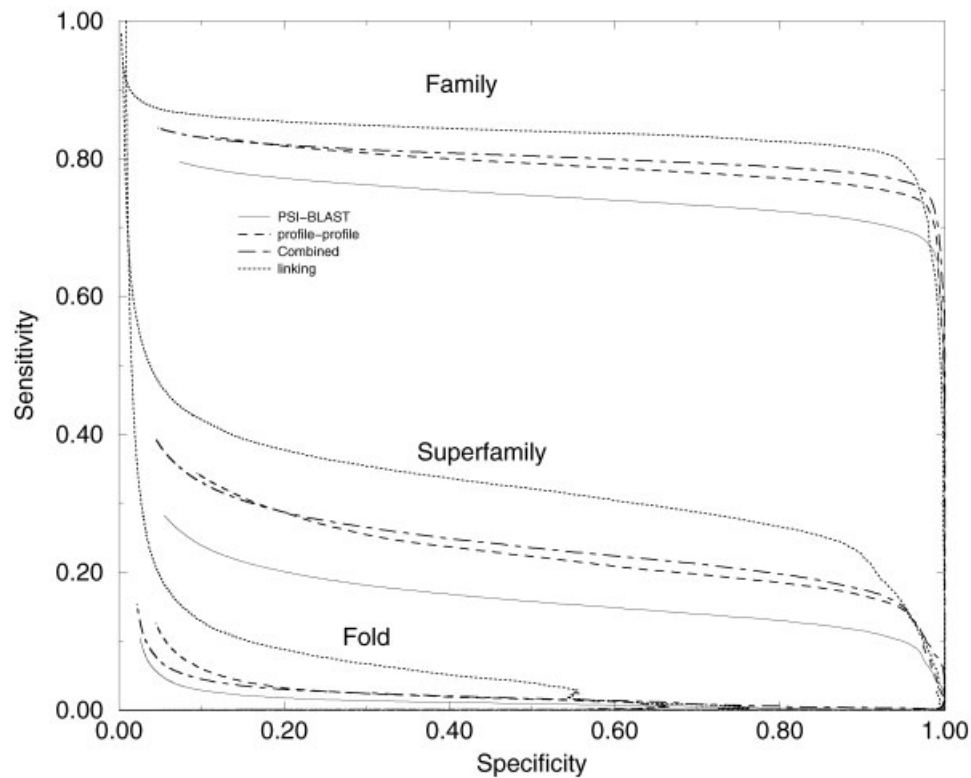
Fig. 2. Specificity versus sensitivity plot for different methods using evolutionary information for both the query and target sequences using an E-value cutoff of 10^{-3}. The curves describe different types of relationships, fold (bottom), superfamily (middle) and family (top). For comparison, the PSI-BLAST curve using 10 iterations and a 10^{-3} E-value cutoff is also shown.

TABLE III. Performance of Methods That Use Evolutionary Information for Both the Query and Target Sequence

Method	E-value	Family specificity (%)	Superfamily specificity (%)
PSI-BLAST	10^{-2}	67	15
PSI-BLAST	10^{-3}	67	13
Profile–Profile	10^{-2}	64	19
Profile–Profile	10^{-3}	68	17
Linking	10^{-3}	61	22
Combined Max	10^{-2}	67	11
Combined Ave	10^{-2}	72	20
Combined Min	10^{-2}	70	20
Combined Max	10^{-3}	68	10
Combined Ave	10^{-3}	71	18
Combined Min	10^{-3}	67	17

ily-related proteins are improved, but only below 98% specificity. In total, this increase in performance is comparable with the increase reported by Yona and Levitt.

Sequence Linking

Another popular method to include additional evolutionary information is to use sequence linking,[1,4] also referred to as intermediate sequence search.[13] Here, a relationship between protein A and protein B can be detected through protein C. In the first linking studies, single-sequence search methods were used[1,4,12–14]; however PSI-BLAST has also been used lately.[15,17] In linking, there are many options to consider, such as the size of the databases, what proteins should be considered for linking, etc. A full study of all parameters would be extremely time-consuming. However, after some limited optimization, we found that the method described in Figure 3 worked quite well. For the linking method, we only used the 10^{-3} profiles because the 10^{-2} profile decreased the performance significantly.

The linking methods obtained the highest sensitivity of all methods (22%) for the superfamily-related proteins at 90% specificity, but this was only obtained at the cost of a decreased family sensitivity (61%); see Table III. In Figure 2, it can be seen that the linking procedure showed the highest sensitivities of all methods at specificities below 97%. This finding indicates that this could be the best method to use if other methods (e.g., manual inspection) that could filter out high scoring false positives existed. Furthermore, it is likely that some variation of the linking procedure could be introduced to improve the specificity.

Combined Searches

In several studies, it has been shown that combining the results from several independent searches increases the performance of fold recognition methods. The combinations might be performed by selecting the alignment that gave the highest score,[18] including information about the
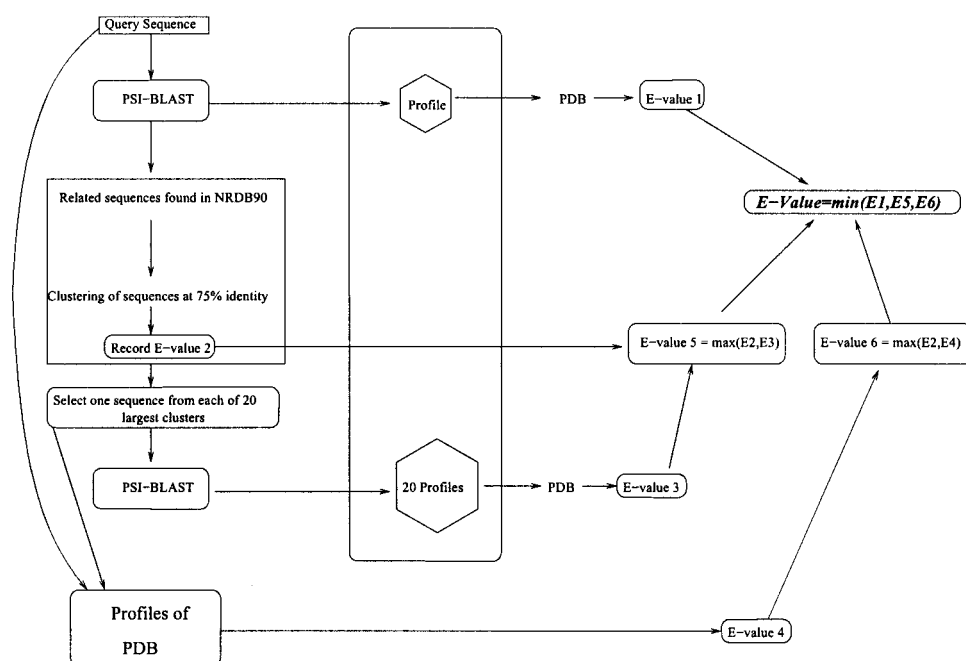
Fig. 3.   A description of the linking method used in this study. PSI-BLAST is run on a query sequence for 10 iterations using an E-value cutoff of 10^{-3}. The proteins that are found with this search are then clustered by using a 75% sequence identity cutoff. From each of the 20 largest clusters, one linking sequence is used for another PSI-BLAST run. Profiles for all proteins in the database (PDB) are also used to search for relationships either directly to the original sequences or to one of the linking sequences. Finally, the lowest E-value from the direct matches or from one of the linking methods is reported.

rank[8] or performing a comparison between all final models.[20] A simple combined score between two or more methods can be obtained by using the highest, lowest, or average score. Here, we show results for the combination of IMPALA and PSI-BLAST. At the accompanying web site, results for the sequence–profile and profile–sequence alignments using the PALIGN program are available as well as other combinations. The results for the different combination of two methods are virtually identical, although the respective E-values are calculated completely differently.

Using the average E-value from two 10^{-3} profile searches increases the family and superfamily sensitivities to 71% and 18%, respectively, whereas the 10^{-2} profiles gives an even better performance, 72% and 20%; see Table III. In Figure 2, it can be seen that the performance is almost identical to the performance of the profile–profile method, with a slightly better sensitivity at high specificities. It should be remembered that if the target profiles are precalculated, this method is only slightly slower than PSI-BLAST and much faster than the profile–profile method. As an alternative to the average E-value, the maximum (worse) or minimum (best) E-value can be used. Using the minimum E-value gave a similar performance to the average value, whereas the maximum value provided no significant improvement over PSI-BLAST; see Table III. The improvement seen here is similar to what was observed during the development of 3D-PSSM, where the minimum E-value was used. There, when a combination of sequence–profile and profile–sequence searches was used,

the number of correctly identified protein domains increased by 20%.[18]

E-Values

An E-value should describe the expected number of hits to unrelated proteins. Only for ungapped local alignments can this E-value be calculated analytically, and in all other cases, this value is obtained from fitting a function to the distribution of observed data. Obviously, a number of assumptions are made when calculating the E-values. Here, we actually obtain E-values in several different ways. IMPALA and PSI-BLAST use precalculated values, the PALIGN program calculates them from all observed values as in FASTA,[21] the combined method uses an average value from several different E-values, and the linking method uses the lowest out of several E-values. All these different methods produce something that we can refer to as an E-value, but we actually do not know how this E-value correlates with the observed error rate. Therefore, we wanted to study how the different reported E-values agree with the observed ratio of false hits. The observed ratio of false hits was simply calculated from all hits where the SCOP fold classification between query and target sequence disagreed. It should also be remembered that we have ignored all hits between a number of SCOP folds that earlier were reported to be related.[22]

In Figure 4, reported E-values are plotted against observed E-values (i.e., the rate of false positives). Here, it can be seen that all methods, excluding BLAST, are significantly less specific than what should be assumed
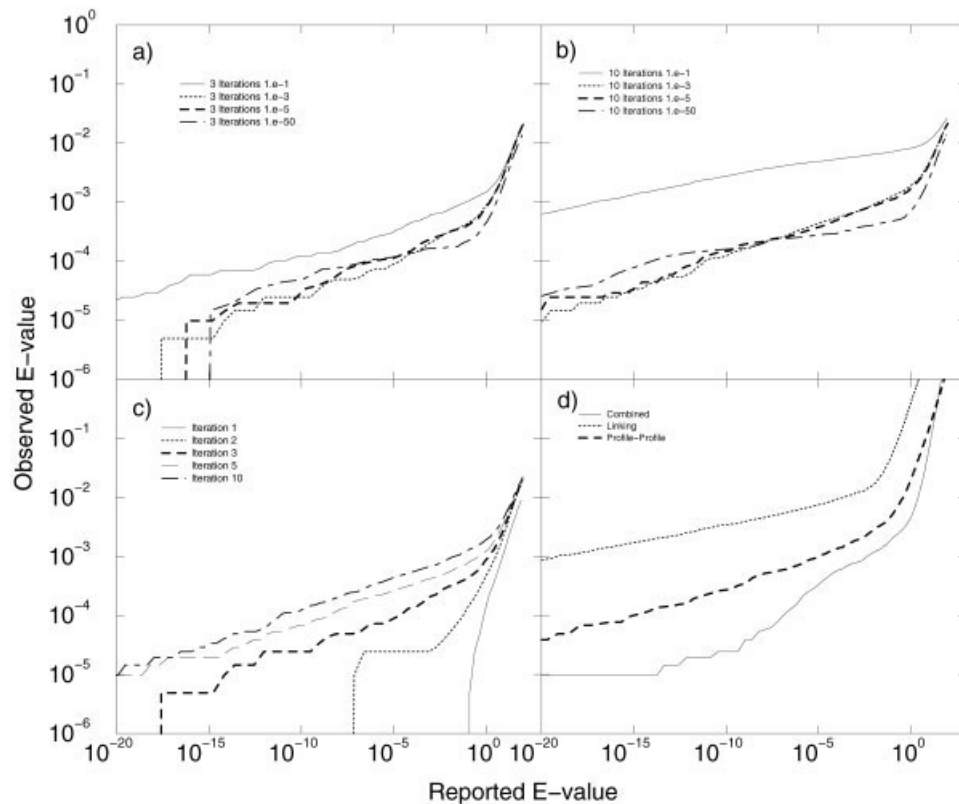
Fig. 4. Reported versus observed E-value (rate of false positives) of different methods. In (**a**) and (**b**), 3 or 10 iterations of PSI-BLAST were used by using different E-value cutoffs. In (**c**), an E-value cutoff of 10^{-3} was used, and the results were studied after 1–10 iterations. Finally, in (**d**), the observed/reported E-values from the three methods that use sequence information for both the query and target sequences are shown.

from the reported E-value. For high E-values (above $\sim 10^{-3}$), the observed ratio is quite correct, but at lower values, the observed ratio is much higher. In Figure 4(a) and (b) it can be seen that using an E-value cutoff of 10^{-1} creates more false positives than a stricter E-value but cutoffs of 10^{-3}, 10^{-5}, or even 10^{-50} does not show any significant differences. In Figure 4(c), it can be seen that the difference between reported and observed E-values increase fast for the first few iterations, but after approximately five iterations, there are only minimal changes. The observed error rates for PSI-BLAST, profile–profile searches, and combined searches are quite similar, whereas the linking method has a much higher error rate; see Figure 4(d). This is in agreement with the observation that our linking method did not perform that well at high specificities.

Because similar differences between reported and observed E-values are observed for all methods, it is unlikely that the origin of these errors are due to how E-values were calculated or any other method-specific feature. Furthermore, it can be noted that the difference increases with the use of a looser E-value cutoff or an increased number of PSI-BLAST iterations. One possible explanation could be that a number of SCOP folds actually were distantly related. However, we do not believe this is the case because we have excluded fold pairs that earlier were reported to

be distantly related and we could not detect any pair of folds that dominated the high scoring false hits. For instance, PSI-BLAST using either the 10^{-2} or 10^{-3} profiles resulted in 189 and 148 false hits with a reported E-value below 10^{-5}, respectively. However, only 54 of these were shared between the two sets, and only in two cases did the query protein detect a high scoring target protein when the target protein was used as query.

An alternative explanation for this behavior could be sequence drift. Here, if one incorrect sequence is found with an E-value larger than the cutoff, this sequence is included in the next PSI-BLAST iteration. Once this sequence is included in the profile, it is quite likely that additional incorrect sequences will be found. On average, three false hits were observed for each profile, which is almost identical to the average number of sequences per family in our test set.

We believe that if this problem could be avoided, a significant increase in performance could be obtained for all profile-based methods. However, the use of a stricter cutoff does not seem to help. We could imagine that it is possible that the use of secondary structures predictions[23] or structural evaluations[24,25] might help. However, some limited trials to use this information have not been very successful.

TABLE IV. Description of the Test Set

Description	Number
Number of protein domains	4,972
Number of different families	1,543
Number of different superfamilies	905
Number of different folds	579
Number of pairs on family level	52,532
Number of pairs on superfamily level	101,954
Number of pairs on fold level	125,090

CONCLUSIONS

In this study, we show that at least three (slower) methods exist to increase the performance over what is obtained by standard PSI-BLAST searches. All these methods use evolutionary information both from the query and target sequences. These methods are profile–profile searches, linking, and combined sequence–profile/profile–sequence searches. The highest sensitivity was obtained by the linking procedure, but it showed a lack of performance at high specificities. The performances of the profile–profile methods and the combined searches are very similar, but the combined method is significantly cheaper in computational cost and can, therefore, be recommended. Furthermore, it was shown that the E-values reported by PSI-BLAST, or any of the other methods, underestimates the number of false positives significantly.

MATERIALS AND METHODS

Data Set

To compare the performance of different fold recognition methods, it is of greatest importance to use a large and well-created benchmark. Several studies[1,4,26–28] have shown that a useful benchmark can be created by using SCOP[29] as a standard for classifying protein domains into families of similar fold or of evolutionary relationship. SCOP is a database in which protein structures are classified into a hierarchical classification: class, fold, superfamily, and family. Two domains that are classified into the same fold have the same secondary structure elements in a similar topological arrangement, whereas two domains that belong to the same superfamily should have a likely common evolutionary origin and family-related proteins are clearly homologous.

We created a benchmark starting from the pdb75 data set of SCOP version 1.57. This data set contains a subset of SCOP in which no protein domains have more than 75% sequence identity to any other member of the data set. This resulted in a set of 5769 domains. This set was further reduced to 4972 proteins by only including domains from SCOP class a to e (i.e., ignoring membrane proteins, small proteins, coiled–coiled proteins, low-resolution structures, peptides, and designed proteins). Some statistics from the final set are summarized in Table IV. All proteins were matched to all other proteins, and for each pair, the folds and families, according to SCOP, were recorded. At each level of comparison, the test set includes between 50,000 and 125,000 related pairs. Because the performance differences between methods are quite small, we think it is necessary to use a set of this size to be able to draw clear conclusions about small differences in performance. The set is available from the accompanying web site.

A noticeable difference in this study from earlier studies is that we have used a significantly larger set. We have used a much less stringent homology cutoff to include proteins in this set. The reason to include more proteins is that thereby we hoped to be able to obtain a more detailed picture of the difference in performance. Because all methods should detect all protein domain pairs with >30% sequence identity, this will only raise the bar for all methods and should not affect their relative performance. The performance on superfamily and fold levels is not affected at all, because there are no proteins from two different families with >30% sequence identity.

Comparisons

Because SCOP is a hierarchical database, a comparison can be done at different levels of the database. When proteins are studied on the superfamily level, all proteins that belong to the same family are ignored. Proteins that belong to the same family are much easier to detect than proteins that belong to different families but the same superfamily. PSI-BLAST finds 86% of the proteins that belong to the same family but only 8% of the proteins that belong to the same superfamily. Less than 1% of the fold-related proteins could be found. To not be biased by dubious assignments in SCOP, all proteins that belong to the same fold were ignored when annotating as false matches, as in earlier studies.[4,30] In addition, we have ignored hits between a few SCOP families that earlier had been reported to be evolutionary related[22]; see http://www.sbc.su.se/~arne/psicomp/ for a detailed description.

We have used specificity-sensitivity plots in an identical way as in our earlier studies.[4,26,28] The main advantage of this method is that it describes the ability of a method to find all pairwise matches in the benchmark. The sensitivity is the method's ability to find all related members in a fold/superfamily/family. In other words:

$$SENS(score) = TP(score)/(TP(score) + FN(score))$$

(1)

where $TP(score)$ is the number of correctly identified protein pairs that have a score above $score$, and $FN(score)$ is the number of related pairs with a score less than $score$ (i.e., $FN + TP$ is total number of pairs that could be detected). The specificity measures the probability that a pair of sequences with a score greater than a certain threshold belong to the same fold/superfamily/family. The specificity is defined as:

$$SPEC(score) = TP(score)/(TP(score) + FP(score)) \quad (2)$$

where $FP(score)$ is the number of false hits that have a score above $score$ and TP is defined as above. The sensitivity is plotted as a function of specificity, each point in the plot corresponding to a certain score.

To simplify the comparisons, we have selected two points from the specificity-sensitivity curve that we think

are representative for the performance. These two points are the family sensitivity at 99% specificity and the superfamily sensitivity at 90% specificity.

Creation of Profiles

All profiles in this study were generated by using PSI-BLAST version 2.2.2 and the nrdb90 database from EBI.[31] We have used two sets of programs for sequence–profile and profile–sequence alignments: the NCBI suite of programs and the PALIGN[32] programs. The NCBI programs (PSI-BLAST and IMPALA) apply the BLAST2 heuristic search algorithm and calculate E-values from precalculated distributions of scores. In contrast, PALIGN uses a complete local Smith–Waterman search algorithm and E-values calculated as in FASTA.[21] When using the same profile, the performance is virtually identical between the two different methods (see Additional Comparisons).

Alignments

Here we use only local alignments[2] instead of global[33] or local–global[34] alignments used by many other fold recognition methods. The reasons are that we obtained better specificity using local alignments and that a comparison with PSI-BLAST (which uses local alignments) is simplified by only using local alignments.

Sequence–Profile and Profile–Sequence Alignments

Profile information can be used in two different ways. Either a profile created from a query sequence can be used to search against a set of target sequence, or alternatively, the query sequence can be used to search a set of profiles created for all of the target sequences. We refer to the first method as profile–sequence searches and the second as sequence–profile searches. Profile–sequence searches can be done either directly in PSI-BLAST or by using the profile for a separate search later, whereas sequence–profile searches can be performed using IMPALA[19] or PALIGN. In this article, we only show the results for the profile–sequence search because the results for the sequence–profile method are very similar (see web site). The computational time for profile–sequence and sequence–profile searches depends on the size of the database. If you want to find all related proteins for one query protein, it is much more efficient to use profile–sequence searches because you only need to make one profile. However, if the database is limited in size, as is the case in fold recognition, it could be more efficient to use sequence–profile searches because the profiles can be precalculated.

Profile–Profile Alignments

The profile–profile alignment algorithm described by Yona and Levitt was implemented into the PALIGN package. The frequency profiles were calculated from the final PSI-BLAST profiles because it was shown that this resulted in a higher performance than using the frequencies directly.

Linking

In linking methods, similarity between two proteins is detected by finding a neighbor to both these proteins. This scheme has been used before with more or less automated methods.[4,30,31] Here, we have tried several approaches to automate the linking method and obtained the best results by using this approach. After some optimization, we found that the scheme described below performed quite well. Other schemes, such as using BLAST and several steps of linking, were tried, but the results were not that good.

The linking methodology used here is similar to the one used in our earlier study,[4] with the exception that we use PSI-BLAST instead of BLAST. First, PSI-BLAST is run on a query sequence for 10 iterations using an E-value cutoff of 10^{-3}; see Figure 3. The detected proteins are then clustered by using a 75% sequence identity cutoff. This profile is also used to search the SCOP database directly, generating E-value E1. From each of the 20 largest clusters, one sequence is used for another PSI-BLAST run. These 20 profiles are used to search SCOP again, and a combined E-value, taking the maximum (worse) of the two E-values E2 and E3, is recorded as E5. Finally, profile–sequence searches against the original and the 20 intermediate sequences are performed, generating E-value E6. Finally, the lowest E-value of the three E-values (E1, E5, and E6) is reported. This protocol is quite similar to the IPS protocol used by Li et al.,[17] with the main difference that we only create a single set of intermediate profiles. Obviously, this search protocol is at least 20 times slower than a standard PSI-BLAST search.

Combined Method

When two or more methods were combined, each method produced an E-value and a combined score for each query-target pair. We have used three different methods to calculate a single score from a combination of several E-values using the lowest, the highest, or the average E-value. The average E-value is calculated by:

$$Average = e^{\frac{\sum \log(E_i)}{N}} \qquad (3)$$

where E_i is the E-value of method i and N is the number of methods that are combined.

Availability

The PALIGN program package is freely available under a GPL license from http://www.sbc.su.se/~arne/palign/

Additional Comparisons

In this study, >200 different sets of parameters and alignment methods were tested. Many of the results from these comparisons are available from http://www.sbc.su.se/~arne/psicomp/. These include different gap penalties and other parameters for the different methods, different alignment techniques, and different linking methods. In general, most sets of parameters did not perform as well as the methods discussed here; however, it is possible to combine several methods to obtain equal or even slightly better results.

REFERENCES

1. Abagyan R A, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
2. Smith T, Waterman M. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
3. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
4. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. J Mol Biol 2000;295:613–625.
5. Bujnicki J, Elofsson A, Fischer D, Rychlewski L. Livebench: continuous benchmarking of protein structure prediction servers. Protein Sci 2001;10:352–361.
6. Bujnicki M, Elofsson A, Fischer D, Rychlewski L. Livebench-2: large-scale automated evaluation of protein structure prediction servers. Proteins 2001;45 Suppl 5:184–191.
7. 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.
8. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. In: Altman R, Dunker A, Hunter L, Klien T, editors. Pacific Symposium on Biocomputing. World Scientific 2000;5:116–127.
9. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
10. Larsson B. Development of hmms that use multiple sequence alignments. Master's thesis. Stockholm University, 1999.
11. von Öhsen N, Zimmer R. Improving profile–profile alignments via log average scoring. In Gascuel O, Moret BME, editors. Algorithms in Bioinformatics, First International Workshop, WABI, 2001. Springer-Verlag: New York, 2001;2149:11–26.
12. Holm L, Sander C. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins 1997;28:72–82.
13. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:249–254.
14. Salamov A, Suwa M, Orengo CA, Swindells MB. Genome analysis: assigning protein coding regions to three-dimensional structures. Protein Sci 1999;8:771–777.
15. Venclovas C. Comparative modeling of casp4 target proteins: combining results of sequence search with three-dimensional structure assessment. Proteins 2001;45 Suppl 5:47–54.
16. Koretke K, Russell R, Lupas A. Fold recognition without folds. Protein Sci 2002;22:1575–1579.
17. Li W, Jaroscewski, Godzik A. Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Eng 2002;15:643–649.
18. Kelley L, MacCallum R, Sternberg M. Enhanced genome annotation using structural profiles in the program 3d-pssm. J Mol Biol 2000;299:523–544.
19. Schaffer A, Wolf Y, Ponting C, Koonin E, Aravind L, Altschul S. Impala: matching a protein sequence against a collection of psi-blast-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
20. Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural network based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2365.
21. Pearson WR, Lipman DJ. Improved tools for biological sequence analysis. Proc Natl Acad Sci USA 1988;85:2444–2448.
22. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 2001;313:903–919.
23. Errami M, Geourjan C, Deléage G. Detectionof unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 2003;19:506–512.
24. Jones D. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
25. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
26. Rice D, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
27. Brenner SE, Chothia C, Hubbard T. Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
28. Hargbo J, Elofsson A. A study of hidden markov models that use predicted secondary structures for fold recognition. Proteins 1999;36:68–87.
29. Murzin A, Brenner S, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
30. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
31. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–429.
32. Elofsson A. A study on protein sequence alignment quality. Proteins 2002;46:330–339.
33. Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
34. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947–955.