Identification of related proteins on family, superfamily and fold level

Erik Lindahl *
and
Arne Elofsson†

November 2, 1999

† To whom correspondence should be addressed
Fax: +46-8-15 3679
Tel: +46-8-16 1553
Email: arne@biokemi.su.se

Running Title:
Identification of related proteins.

Abbreviations: HMM, hidden Markov model; Scop, the Structural Classification of Proteins database; family, protein domains that are closely related having a common origin according to Scop; superfamily, protein domains of probable common origin according to Scop; fold, protein domains that have major structural similarities according to Scop.

Keywords: BLAST2; SSEARCH, PSI-BLAST; Sequence comparison; Scop; benchmark; hidden Markov model; fold recognition; threading; fold; superfamily; family.

*Theoretical Physics, Royal Institute of Technology, SE-100 44 Stockholm, Sweden. E-mail:erik@theophys.kth.se
†Department of Biochemistry, Stockholm University, SE-106 91 Stockholm, Sweden E-mail:arne@biokemi.su.se

Summary

Proteins might have considerable structural similarities even when no evolutionary relationship of their sequences can be detected. This property is often referred to as the proteins sharing only a "fold". Of course, there are also sequences of common origin in each fold – called "superfamily", and in them groups of sequences with clear similarities, designated "family". Developing algorithms to reliably identify proteins related at any level is one of the most important challenges in the fast growing field of bioinformatics today. However, it is not at all certain that a method excellent at finding sequence similarities performs well at the other levels, or vice versa.

In this review we have compared the performance of various search methods on these different levels of similarity. As expected, we show that it becomes much harder to detect proteins as their sequences diverge. For family related sequences the best method gets 75% of the top hits correct. When the sequences differ but the proteins belong to the same superfamily this drops to 29%, and in the case of proteins with only fold similarity it is as low as 15%. We have made a more complete analysis of the performance of different algorithms than earlier studies, also including threading methods in the comparison. By this a more detailed picture emerges, showing multiple sequence information to improve detection on the two closer levels of relationship. We have also compared the different methods of including this information in prediction algorithms.

For lower specificities the best scheme to use is a linking method connecting proteins through an intermediate hit. For higher specificities better performance is obtained by PSI-BLAST and some procedures using hidden Markov models. We also show that a threading

method, THREADER, performs significantly better than any other method at fold recognition.

Introduction

As the genome projects proceed we are presented with an exponentially increasing number of protein sequences without any knowledge of their structure or biochemical function. Since structure and function determination is a nontrivial task even for a single protein, the best way to gain understanding of all these sequences is if we can relate them to other proteins with known properties by searching databases. Improving these algorithms is one of the fundamental challenges in bioinformatics today. By determining how sequences are related to known proteins we can make predictions of their structural, functional and evolutionary features. The relationships between proteins span a broad range, from the case of almost identical sequences to apparently unrelated sequences sharing only rough 3d-structure. This poses different challenges to the detection algorithms used – a method excellent at finding sequence similarity might not perform very well in the case of only structural relationship, or vice versa.

In this work we have compared the performance of various recognition methods on different levels of similarity, extending earlier studies by (i) completely separating relationship levels, (ii) including more categories of sequence recognition methods and (iii) including fold recognition (threading) methods. This has resulted in new insight how to best utilize evolutionary and structural information as well as ideas on the relative merits of different methods.

Using Scop as a similarity benchmark

During the last few years several excellent papers have changed the view of the best methods to detect relationship between proteins. These papers differ in detail but have one common nominator, they use the Scop classifications by Murzin *et al.* (1995), to create a benchmark used in evaluating the performance of different recognition methods. Scop is a hierarchical scheme where each protein domain is classified into a family which in turn belongs to a superfamily that is a subclassification of the fold category. The Scop database is to a large extent hand tuned by Alexei Murzin, giving the following explanations to the levels: Proteins sharing family have a "Clear evolutionary relationship", those within a superfamily are of "Probable common evolutionary origin" while the fold level is characterized by "Major structural similarity". This manual classification of the proteins makes Scop independent of any specific sequence or structure comparison algorithm and thereby ideal for comparison between such methods, further Scop is considered to be of a very high quality.

Three earlier studies (Abagyan & Batalov, 1997; Brenner *et al.*, 1998; Arvestad *et al.*, 1999) of sequence comparison methods have resulted in a rather clear picture with some important conclusions: (i) The common method of describing similarity as fraction identical residues should be abandoned. (ii) The exact choice of parameters such as gap-penalties are crucial when choosing the best methods, and (iii) heuristic methods such as FASTA (Pearson & Lipman, 1988; Pearson, 1995) and BLAST2, Altschul *et al.* (1997), do not perform as well as methods using the optimal alignments. These studies differ in the hierarchical level of Scop used: Brenner *et al.* (1998) chose the superfamily classification, Abagyan & Batalov (1997) the fold level, while Arvestad *et al.* (1999) studied both fold and family

classifications. In practice, however, all three studies compared the power of methods to identify relationships within a family. This is because the hits on e.g. fold level include pairs on superfamily and family level. For obvious reasons it is much easier to identify proteins within a common family; therefore this part will dominate the correctly identified pairs. In several other studies the Scop classification has also been used to compare different fold recognition methods (Rice & Eisenberg, 1997; Hargbo & Elofsson, 1999). In these cases all hits within a family (or superfamily) were ignored. To extend our understanding of different search methods we have made a complete separation of different levels in the classification. We have thus discarded family hits when studying superfamilies and ignored both family/superfamily hits when judging fold level performance. This makes it possible to distinguish methods performing well on different levels in the Scop hierarchy.

Evolutionary information improves detection

Even within a group of proteins with a common origin, a single pair of sequences can differ quite substantially in composition. For many years it has been assumed that the use of evolutionary information to create a multiple sequence alignment helps detecting such distant relationships. However, it is only recently this was clearly shown to be true using a large and comprehensive benchmark (Park *et al.*, 1997; Park *et al.*, 1998). In the latter of these studies it was shown that including evolutionary information detects three times as many remote homologues at the same false positive rate. They also concluded the best way of employing multiple sequence information was to use it in the iterative SAM-T98 hidden Markov model. In a similar study, not using Scop but CATH, Orengo *et al.* (1997), as the

reference databases, Salamov *et al.* (1999a) found results similar to the ones by Park *et al.* (1998).

There are several quite different ways of using evolutionary information, see figure 1. One possibility, used in e.g. PFAM, Sonnhammer *et al.* (1997), is to start from a family already aligned and then search for more members belonging to the same set. PSI-BLAST and the hidden Markov models used in Park *et al.* (1998) utilize an iterative approach that starts from a single sequence, find all related sequences and create a multiple sequence alignment. From this alignment a new iteration is started and this procedure is repeated until it converges. A third alternative is to consider two proteins to be related they are identified by the search algorithm directly or if they both are found to be related to a third protein domain (Holm & Sander, 1997; Park *et al.*, 1997; Abagyan & Batalov, 1997; Salamov *et al.*, 1999b; Arvestad *et al.*, 1999). All these methods have different strengths; the direct search method is clearly fastest with a runtime linear to the size of the database, while the iterative method should be at least n times slower, n being the number of iterations made. Often the method is even slower as for the iterative searches a larger database is used. The last method should only be a factor two slower, but also in this case a larger database for the intermediate search is common, making it substantially slower than the direct approach.

Fold recognition

There are many proteins with similar structure where no obvious homology has been detected. Methods developed to identify this *structural* relationship are often referred to as fold recognition (or threading) methods. They can roughly be divided in two categories; pre-

diction based methods, (Sheridan *et al.*, 1985; Fischer & Eisenberg, 1996; Rice & Eisenberg, 1997; Di Francesco *et al.*, 1997; Rost *et al.*, 1997; Hargbo & Elofsson, 1999) and structural methods, (Bowie *et al.*, 1991; Jones *et al.*, 1992; Flöckner *et al.*, 1997). Besides these two categories it is of course possible to use purely sequence based methods even for fold recognition, Karplus *et al.* (1997), or combine several approaches, Elofsson *et al.* (1996).

The structure based methods differ from all others described here since they do not directly use any sequence information to detect whether two proteins share a fold or not. Instead they create an energy function describing how well a probe sequence matches a target fold. The energy function is often obtained from a database of known protein structures and may for instance describe the environment of each residue, Bowie *et al.* (1991), or the probability of finding two residues at a certain distance from each other, (Jones *et al.*, 1992; Flöckner *et al.*, 1997).

Proteins having a similar fold by definition also have very similar secondary structure, meaning that even when amino acid composition are unrelated the secondary structure should largely be the same within a fold. Since secondary structure can be predicted with an accuracy of more than 70% today, Rost & Sander (1993), several attempts have been made to use this information to improve fold recognition methods (Fischer & Eisenberg, 1996; Rice & Eisenberg, 1997; Rost *et al.*, 1997; Hargbo & Elofsson, 1999). These methods add a positive score to the sequence alignment score if the predicted secondary structure for a certain residue agree with the secondary structure state of the residue.

Every two years, starting in 1994, the CASP conference has been organized to evaluate the ability to blindly predict the structure of proteins, Moult *et al.* (1997). The blind predictions

was deemed necessary to evaluate the different methods, as it was considered difficult to avoid creating a biased benchmark. One important outcome from the CASP process regarding fold recognition is that several groups using fundamentally different methods consistently perform very well. An extreme example is the excellent predictions by Murzin in CASP2 where no fold recognition methods were used but mainly biochemical knowledge (Murzin & Bateman, 1997). However, a complication in CASP when evaluating fold recognition methods is the mix of fold and (easier) superfamily level targets.

It is thus our belief that a complementary way of assessing fold recognition methods is to use a complete benchmark while simultaneously separating the levels of similarity by ignoring hits also present in lower levels. Unfortunately, few fold recognition methods are publicly available, and other are still very time consuming. Therefore we have only used three different methods, all showing some success in the latest CASP process: THREADER, Jones *et al.* (1992), SAM-T98, Karplus *et al.* (1997) and ssHMM, Hargbo & Elofsson (1999). The results from the fold recognition methods were compared with results from standard sequence alignment methods.

Results

The results of the all against all comparison of the 976 protein sequences are summarized as spec-sens curves in figures 2 - 4 and top ranks in tables III - V.

Starting on family level, table III shows that the best method, the linking algorithm, finds 75% of the sequences in top rank and that all methods except THREADER find more than 65% sequences in first place. Figure 2 shows that at 100% specificity the best sensitivity is obtained by HMMER-PSIBLAST with 40%, followed by PSI-BLAST with 37%, while at lower specificities BLAST-LINK clearly performs best with a sensitivity of 54% at 50% specificity.

For new protein domains the results are rather similar to the behavior of the all against all comparisons. The different search methods find between 57% and 71% of the proteins in first rank. The best performance was obtained by SSEARCH and the worst by BLAST-LINK. Since the is only a few more domains detected it is doubtfull if it significant.

As expected, protein domains related solely on superfamily level are harder to detect. For these pairs only 29% of the superfamily related proteins were found in first rank using BLAST-LINK, see table IV. Almost as many pairs were found using SAM-PSIBLAST. At a specificity higher than 30% the best results are obtained using PSI-BLAST that finds 4% of the pairs at 100% specificity while all other methods detect less than 2%, see figure 3. At 50% specificity PSI-BLAST detects almost 10% of the superfamily related pairs.

In the set of new protein domains the performance is worse than for the all against all comparison set, with only a few methods performing as good as on the complete set. However, both the SAM based methods, SAM-HSSP and SAM-PSIBLAST perform very

well detecting 30% and 23% of the proteins in first rank respectively.

Proteins related only on the fold level are even more difficult to detect, see table V, where it can be seen that the best method, THREADER, only finds 15% of the related pairs in first rank. Still, THREADER performs significantly better than all other methods, finding twice as many proteins in first rank and also a higher specificity, see figure 4. In figure 4 it can also be seen that no method is capable of reliably detecting proteins related on the fold level as even THREADER only finds a handful of the pairs at a specificity higher than 5%, i.e. when 19 false positives are found for each true positive. The use of secondary structure constraints does not significantly alter these results, neither did the optional sequence shuffling step designed to find false hits, data not shown.

The relative performance of THREADER compared to the other methods is not only reflected but also strengthened on the set of new protein domains, see table VI. The performance of THREADER is clearly best detecting 42% of the target correctly in the first place and 65% in the top 5 ranks. The performance of the other methods is as low for this set as for the complete set., i.e. only a few percent of the targets are detected in the first rank.

Discussion

40% of family level pairs but only 4% of superfamily pairs can be detected reliably

The most common use of sequence comparison methods is to search in databases to find proteins belonging to the same family, i.e. those with similar function and clear evolutionary relationship. Both the results from ranking, see table III, and from the spec-sens curves, see figure 2, indicate that the best performance is obtained by BLAST-LINK. The exception is at specificities above 97%, where PSI-BLAST and HMMER-PSIBLAST perform better. It should be noted that all methods using sequence information are good at identifying the correct family members, while THREADER performs significantly worse. Using any of these methods about half of the members are identified at a specificity of 50% and more than two-thirds of the members are correctly identified in first rank. Even if all these methods perform well, there are differences that are significant and very important mainly for automatic assignments.

To identify proteins sharing superfamily but not family is much more difficult than identifying members of the same family as can be seen in Table IV and figure 3. From figure 3 it is clear that the best sensitivity is obtained by PSI-BLAST, followed by SAM-HSSP and BLAST-LINK. Even though PSI-BLAST performs excellent, at high specificities there is still a long way to go until superfamily relationships can be detected reliably, as PSI-BLAST and SAM-HSSP only find about 4% of the possible matches at 100% specificity.

As expected the performance of SSEARCH is slightly better than for BLAST2 at the

family and superfamily levels, see Tables III and IV. However at high specificities BLAST2 performs slightly better than SSEARCH, showing that scoring scheme used in BLAST2 is better tuned for our benchmark.

Multiple sequence information helps finding evolutionary related proteins

Including multiple sequence information can in principle be done in three different ways, see figure 1. The multiple sequence information can be used: (a) in an iterative fashion as in PSI-BLAST, (b) with only a single iteration as when using the multiple sequence alignments from HSSP to create a hidden Markov model or (c) by linking sequences through an intermediate protein. We have used all three methods, while Park *et al.* (1998) only compared (a) and (c). Besides these fundamentally differences between the methods there are different algorithms that can be used. The differences occur from how the multiple sequence alignments are made and how the information is coded into a hidden Markov model (HMM) or a profile method, such as PSI-BLAST.

Park *et al.* (1998) showed that by using multiple sequence information three times as many proteins could be detected at a false detection rate of 1/100000. In figure 2 we show a similar tendency, i.e. we find significantly more pairs using multiple sequence information. The improvements we find are slightly lower than the ones obtained by Park *et al.* (1998), probably due to that they compared on a superfamily level while our comparison is based on the family level. Another factor complicating the comparison with Park *et al.* (1998) is that they only show the part of the curves corresponding to sensitivities above approximately

80%, and in that region the multiple sequence information is more important then at lower specificities.

So how is multiple sequence information best put to use? From the study of Park *et al.* (1998), it is suggested that the best method should be an iterative one. Our results give a more detailed and somewhat different picture, showing that in general a linking method performs best at low specificities. The reason why we receive better performance using a linking method, we believe, is due to that Park *et al.* (1998) using FASTA for the linking, while we employed BLAST2. The problem using FASTA is that the expectation value depends on the size of the database and in a linking method the size of the intermediate database is much larger than in the final database. In Park *et al.* (1997) the problem of using fasta was solved by using two different cutoffs (E=0.081 for the first search and E=0.0006 for the second), while by using BLAST2 we only need a single cutoff. The results from this study, Park *et al.* (1997) and Salamov *et al.* (1999a) all agree in that at high specificities PSI-BLAST appears to be better than the intermediate search methods as it reliably detects more family and superfamily pairs, see fig 3 and 2. It seems as if at higher specificity the tuning of any method and the exact way of using multiple sequence information is very important, and that PSI-BLAST is very well tuned for this. Among the hidden Markov models HMMER-PSIBLAST performs best on the family level targets, while SAM-HSSP is superior on the superfamily level targets. Comparing our results from SAM-T98 with those of Park *et al.* (1998), it can be seen that an improvement appears to be obtained by using the method in an iterative way. However, using SAM-T98 iteratively is not computationally realistic for large scale sequence searching. It can also be seen that the aid from multiple

sequence information is not that big at lower specificities, for instance BLAST2 finds 42% of the family related protein at 50% specificity, PSI-BLAST 47%, the best HMM method 51% and BLAST-LINK 54%.

SAM-T98 is better at superfamily recognition while HMMER is better at family recognition

There are two different packages available that use HMMs for protein recognition, the HMMER-2.1 package by Eddy (1998), and the SAM-T98 package by Karplus and coworkers, (Krogh *et al.*, 1994; Karplus *et al.*, 1997; Karplus *et al.*, 1998). These two packages differ both in the way that a HMM is designed, where HMMER uses an architecture not allowing transitions from a gap to an insertion, and in the way expectation values are calculated. It should be noted that the present HMMER-2.1 performs significantly better than the previous HMMER-1.8 which didn't use expectation values and an architecture similar to the one in SAM-T98 (data not shown).

It was a bit surprising to us that the two HMM methods perform this different, especially at high specificities, see figure 2 and figure 3. HMMER-PSIBLAST and HMMER-HSSP match the performance of BLAST-LINK and PSI-BLAST on the family level and at high specificities the performance of HMMER-PSIBLAST is excellent. Using SAM-T98 with the same multiple sequence alignments it is impossible to reach very high specificities, indicating that some few false positives get unrealistically high scores. At the same time, from figure 3 and table IV, it is clear that the performance of SAM-T98 is significantly better than HMMER for superfamily recognitions, using the same multiple sequence alignments. It can be

seen that SAM-T98 finds 30 to 50% more pairs in first rank than HMMER does using the same multiple sequence information.

When we only use proteins that are submitted to PDB after January 1 1998 the results are similar, i.e. it shows that SAM-T98 is better at superfamily recognitions than HMMER, however the difference on family level targets is neglectable.

Using more sequences for building the HMM is not always better

If one decides to use a specific multiple sequence alignment method, the most important questions to answer are what database should be used and how should the multiple sequence alignment be done. Trying to answer these questions we used two different databases for the creation of multiple sequence alignments, either the HSSP database (Sander & Schneider, 1991) or by using PSI-BLAST to create a multiple sequence alignment, see table II. The main difference between these two sets is that the PSI-BLAST alignments are being created from a much larger database including both TREMBL and SWISSPROT (Bairoch & Apweiler, 1996), while HSSP was created from SWISSPROT only. In average each sequence was aligned to 25 sequences in HSSP and to 97 sequences using PSI-BLAST. As seen in table IV the performance of SAM-T98 and HMMER were slightly increased by using the PSI-BLAST set at the superfamily ranking, while a much better sensitivity is reached using SAM-HSSP instead of SAM-PSIBLAST, see figure 3. The spec-sens curves, figure 2 and 3, show that for some specificities the performance is better using the PSI-BLAST multiple alignment and for others the HSSP one. One notable feature is that using SAM-PSIBLAST it is not possible to reach high specificities either at family or superfamily levels. We have not been

able to trace the reason why SAM-T98 has this problem but, somehow, HMMER solves the problem using exactly the same multiple sequence alignments. It is obvious that using a HMM efficiently is not that trivial, since it is necessary to choose the best combination of program, database and alignment algorithms to reach the best possible performance.

THREADER is much better at fold recognition

Even more difficult than detecting proteins that belong to the same superfamily is detecting proteins not belonging to the same superfamily but only sharing a common fold. This can be seen as no method reliably detects more than a few proteins related on the fold level, see figure 4.

In figure 4 and table V it can be seen that for the fold recognition targets the best performance is obtained by THREADER, Jones *et al.* (1992). This method detects more than twice as many targets in first rank, 15%, as well as among the top five, 38%, than any other method. Further the sensitivity is much better for THREADER than for all other methods. The performance of the other methods are impossible to separate from the ranking but from figure 4 it can be seen that BLAST-LINK and ssHMM, perform slightly better than the others.

It was a bit surprising to us that the performance of THREADER was so outstanding compared to all other methods, as for instance Karplus *et al.* (1997) obtained good results using SAM-T98 in CASP2 and CASP3. It is our belief that this is partly due to the mix of superfamily and fold level targets in CASP and to the poor performance of THREADER at identifying superfamily, or family, targets. There are many other fold recognition meth-

ods not evaluated in this study therefore we do not know if the superb performance of THREADER is generally true for fold recognition methods or only for THREADER. Certainly THREADER performs much better than ssHMM, however it should be noted that ssHMM is developed from version HMMER-1.8, which performs significantly worse than HMMER-2.1. This makes us believe that it might be possible to increase the performance of ssHMM, and other prediction based methods, by using the improvements from HMMER-2.1. Of the sequence based methods it is clear that the linking method performed significantly better than the other methods, see figure 4.

Finally it should be noted that no method detects any fold related targets reliably, while THREADER detects almost one out of six targets in the first rank, and more than one out of three among the five first ranked targets. This shows that to do well in fold recognition it is as important to have a method that filter out the false matches as a method that detects correct ones.

One reviewer noted that THREADER contained thousands of parameters and since it is not unlikely that some of the proteins used in our testset were used to calculate these parameters, the results from the all against all comparisons could be biased by this. This issue can be adressed by additionally studying a set of proteins submitted to PDB after January 1st, 1998, and thus not used for parametrizing THREADER. Also for this set the performance of THREADER, detecting 42% of the proteins in first rank, is far superior to the other methods, none of which detects more than 10%, table **??** . This supports and strengthens the results of the all against all comparison.

Combining methods is difficult

From the results above it is very tempting trying to use a combination of methods to improve the performance; however we have only obtained limited success with such approaches. The only significant improvement that we obtained was by combining PSI-BLAST and BLAST-LINK, taking the average of a scaled score. With this the performance increase on the superfamily level was significant, detecting 36% of the pairs in first rank and 43% among the top five, however the combined method did not improve the sensitivity, instead it dropped to zero before a specificity of one was obtained.

Conclusions

Detecting related proteins is of extreme importance as the genome projects proceed, as this is the best method to assign structural, evolutionary and functional knowledge to a gene. For many years the standard method for detecting relationships between two proteins was to use a pairwise sequence alignment method. It was generally assumed that using multiple sequence alignments helps to find more proteins, but only limited large scale benchmarking was done until recently. A few years ago things changed as it became easier to create a benchmark thanks to the excellent manual classification of protein relationships in Scop. These benchmark gave quite a conclusive picture of the best method to detect related proteins, including using an expectation value scoring-scheme and multiple sequence information. These studies failed to use the full potential of the benchmark as they did not carefully separate the different types of relationships described in Scop, i.e. it is not at all certain that the method best at detecting proteins related on family level is also best at finding superfamily relationships. To gain a better understanding on how to detect related proteins we have extended earlier studies by (i) completely separating relationship levels, (ii) including all categories of sequence recognition methods and (iii) including fold recognition (threading) methods.

By the separation of relationship levels we show that a benchmark developed to compare algorithms that detect related proteins has to include several relationship levels. If this is not done carefully it is not possible to obtain a full understanding of how different methods perform, and wrong conclusions might be drawn. Further our study shows that: (a) Multiple sequence information helps to detect proteins related on the family and superfamily levels.

(b) For lower specificities the best way of including multiple sequence information is to use a linking method, while at higher specificities better performance is obtained by PSI-BLAST, Altschul *et al.* (1997), and some hidden Markov models (Krogh *et al.*, 1994; Karplus *et al.*, 1997; Eddy, 1998; Karplus *et al.*, 1998). (c) The HMMER, Eddy (1998), HMM program performs better than SAM-T98 (Krogh *et al.*, 1994; Karplus *et al.*, 1998) at family level, but worse on superfamily levels. (d) The exact method to create multiple sequence alignments is of extreme importance to the HMM. (e) The fold recognition program THREADER, Jones *et al.* (1992), performs significantly better than any other method on fold level targets, but worse on all other levels.

Methods

Benchmark database

In order to assess the performance of protein recognition algorithms it is important to use a large and broad set of related and unrelated protein domains with few errors. We created our benchmark by starting from the pdb40d set of Scop version 1.37. This database consists of a Scop subset where no two proteins have more than 40% sequence identity, Brenner *et al.* (1998). Since some of the algorithms needed the secondary structure and multiple sequence alignments, we used the definition in the latest release of HSSP (Sander & Schneider, 1991). Unfortunately the sequences don't match completely, since (i) HSSP is created from another subset of pdb and (ii) proteins in Scop are divided into domains, which isn't the case in HSSP. To overcome this we matched each sequence in pdb40 to the HSSP database and replaced it with the HSSP sequence if the match had a significance better than $1 \cdot 10^{-5}$ using FASTA, (Pearson & Lipman, 1988; Pearson, 1995), and an alignment length equal to the original sequence. This procedure also removed all "non-protein" entries in Scop, leaving 1130 out of the original 1272 sequences from pdb40. Some of the structures had missing atoms or sidechains, or the secondary structure elements were too random to create templates for THREADER, Jones *et al.* (1992). Discarding these left us with 976 sequences for which multiple sequence alignments and secondary structures were read from HSSP. For the benchmark runs needing predicted secondary structure input we used the prediction from PhD, Rost & Sander (1993). The complete benchmark, including all multiple sequence alignments, secondary structure predictions and evaluation scripts is available from

http://www.biokemi.su.se/~arne/protein-id/.

Comparison & assessment

Scop being a hierarchical database, relationships can be studied on different levels. In our case we wanted to perform the comparison throughout the hierarchy so the Scop fold, superfamily and family of each sequence were recorded and the benchmark run by matching every protein against all others in the set. Since all methods are much better at identifying closely related protein domains, we consistently chose to ignore hits from lower levels in Scop. Otherwise the score on e.g. fold level would be a sum of fold, superfamily and family, dominated by the easier family level. Further, to avoid being biased by possible misassignments in Scop all proteins belonging to the same fold where ignored when annotating as false matches, as in Park *et al.* (1998). Since there are almost 1 million pairs with different folds, this will only remove a very small fraction of the possible false hits. The size of the benchmark is described in table I, where it can be seen that on all levels there are between 300 and 600 possible true hits to be found, and 944, 162 false ones.

We have used two different criteria to analyze the performance of a particular method on our benchmark. First we simply examined the fraction of true hits in first and top five ranks, respectively. This is a very intuitive measure, but it tells nothing about the reliability of the match, i.e. a match could be the top rank but still have a very low score as long as all other hits have even lower scores. To overcome this limitation we have used spec-sens plots (Rice & Eisenberg, 1997; Arvestad *et al.*, 1999; Hargbo & Elofsson, 1999) as a complementary measure, describing the fraction possible correct hits found as a function of the fraction

found hits being correct. The main advantage of this is that it measures the ability of a method to reliably find all pairwise matches in the database. The fraction possible correct hits found, sensitivity, is defined as:

$$\mathrm{SENS}(score) = \frac{\mathrm{TP}(score)}{\mathrm{TP}(score) + \mathrm{FN}(score)} \tag{1}$$

where TP(*score*) is the number of correct hits having a score above *score*, and FN(*score*) being the number of correct hits with a score less than *score*. The specificity measures the probability that a pair of sequences with a score greater than a certain threshold really is a true hit, defined as:

$$\mathrm{SPEC}(score) = \frac{\mathrm{TP}(score)}{\mathrm{TP}(score) + \mathrm{FP}(score)} \tag{2}$$

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 corresponding to a certain score. This measure is similar but not identical to the plots in, Park *et al.* (1997) and Park *et al.* (1998) where sensitivity, referred to as "Fraction of homologous pairs detected", was plotted against "Rate of false positives".

Additional testset

One reviewer noted that THREADER (Jones *et al.*, 1992) contains many parameters and it is not unlikely that some of the proteins might have been used in our testset were used to calculate these parameters, the results from the all against all comparisons could be influenced by this. Therefore we have in addition to the all against all comparison used another set of proteins that are matched against the 976 target folds. These proteins were selected from all structures deposited in PDB (Bernstein *et al.*, 1977) after January first 1998. From these proteins only the ones that were classified in scop-1.41 to belong to the same fold as one target protein domain was selected. Finally all proteins with ore than 40% identity to any of the 976 targets proteins were ignored. After this procedure 119 proteins remained. We have only studied the ranking of these studies as there are too few data points to create reliable spec-sens curves. This testset will be referred to as the new protein domains below.

Matching methods

All algorithms used in this study are summarized in table II, classified into three categories: methods using - only single sequence information (BLAST2, SSEARCH) - multiple sequence alignments (PSI-BLAST, HMMER-HSSP, HMMER-PSIBLAST, SAM-HSSP, SAM-PSIBLAST & BLAST-LINK) and - structural information (ssHMM & THREADER).

BLAST2, Altschul *et al.* (1997), and SSEARCH (Pearson & Lipman, 1988; Pearson, 1995) were used with default parameters, taking the expectation values for score and matching a query sequence against the benchmark database. We refrained from including

the results from FASTA (Pearson & Lipman, 1988; Pearson, 1995) as it employs the same scoring scheme as SSEARCH, but with a heuristic algorithm, delivering at best performance equal to SSEARCH.

PSI-BLAST iteratively collects sets of related sequences to find more proteins. We have used the method with default parameters, allowing for a maximum of 25 iterations. During the iterations we used a larger dataset consisting of the complete SWISSPROT-35 and TREMBL-5 databases (Bairoch & Apweiler, 1996) together with the benchmark sequences. Finally only sequences that belonged to our benchmark were recorded.

An alternative use of multiple sequence information is to first create a multiple sequence alignment of a family of proteins and then use this multiple sequence information for a search. This would thus be similar to a single round of an iterative method. The main advantage with this method is that it is much faster, allowing for computationally more demanding algorithms to be used in the multiple sequence alignment and search algorithms. We have created multiple sequence alignments from two sources, either the HSSP database or a maximum of three rounds of PSI-BLAST. Other alignments we tried, such as more rounds of PSI-BLAST, produced worse results. The resulting alignments were used with the SAM-T98 and HMMER-2.1 hidden Markov model programs, using default parameters including the use of a substitution matrix.

In linking (or intermediate search) methods (Holm & Sander, 1997; Abagyan & Batalov, 1997; Park *et al.*, 1997; Arvestad *et al.*, 1999; Salamov *et al.*, 1999a) two proteins are defined as related either if the score is above a certain cutoff or if there is a third protein related to both above a cutoff. Our approach, BLAST-LINK, differs from Park *et al.* (1997) since

we have based our study on BLAST2 instead of on FASTA, the advantage with BLAST2 being that the expectation values do not depend on the size of the database searched. This makes it possible to use a single cutoff instead of two as in, Park *et al.* (1997). Arvestad *et al.* (1999) used a similar method, but with non-heuristic algorithms for the alignments and different scoring. The main difference between this and our approach is that we have used a larger databases in the intermediate search, while they used the same database as for the final matching step. Salamov *et al.* (1999a) used a method very similar to ours, the main difference being that they recalculated the expectation values from BLAST2 or FASTA, while we used the default ones.

We have studied two fold recognition methods, THREADER, Jones *et al.* (1992), and ssHMM, (Hargbo & Elofsson, 1999). THREADER can be used both with and without predicted secondary structure constraints and also perform a shuffling step to help detecting false matches, but since we didn't notice any significant difference with any of these we chose not to include the results from the shuffling or constrained searches. ssHMM was used with default parameters, using the multiple sequence alignments from HSSP and predicted secondary structures from PhD.

Acknowledgment

This work was supported by grants form the Swedish Natural Sciences Research Council and the Swedish Research Council for Engineering Sciences to AE. We thank Jeanette Hargbo, Björn Larsson, Erik Wallin and Gunnar von Heijne for valuable discussions and help.

References

Abagyan, R. A. and Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355–368.

Altschul, S. F. and Madden, T. L. and Schaffer, A. A. and Zhang, J. and Zhang, Z. and Miller, W. and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Arvestad, L. and Ivansson, L. and Lagergren, L. and Elofsson, A. (1999). What is the best method to determine if two proteins are related ? a study on the structural and evolutionary significance of pairwise protein sequence alignmente. *submitted.*

Bairoch, A. and Apweiler, R. (1996). The swiss-prot protein sequence data bank and its new supplement trembl. *Nucleic Acids Res.* **24**, 17–21.

F. C. Bernstein and T. F. Koetzle and G. J. B. Williams and E. F. Meyer and Jr. and M. D. Brice and J. R. Rodgers and O. Kennard and T. Shimanouchi and M. Tasumi. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol* **112**, 535–542.

Bowie, J. U. and Lüthy, R. and Eisenberg, D. (1991). A method to identify protein sequence that fold into a known three-dimensional structure. *Science* **253**, 164–170.

Brenner, S. E. and Chothia. C. and Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073–6078.

Di Francesco, V. and Geetha, V. and Garnier, J. and Munson, P. J. (1997). Fold recognition using predicted secondary structure sequences and hidden Markov models of proteins folds. *Proteins: Struct. Funct. Genet., Suppl* **1**, 123–128.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics* **14**, 755–763.

Elofsson, A. and Fischer, D. and Rice, D. W. and Le Grand, S. M. and Eisenberg D. (1996). A study of combined structure/sequence profiles. *Fold Des* **1**, 451–461.

Fischer, D. and Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.

Flöckner, H. and Domingues, F. and Sippl, M. J. (1997). Proteins folds from pair interactions: A blind test in fold recogition. *Proteins: Struct. Funct. Genet., Suppl* **1**, 129–133.

Hargbo, J. and Elofsson, A. (1999). A study of hidden markov models that use predicted secondary structures for fold recognition. *Proteins: Struct. Funct. Genet.* **36**, 68–87.

Holm, L. and Sander, C. (1997). An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72–82.

Jones, D. T. and Taylor, W. R. and Thornton, J. M. (1992). A new appoach to protein fold recognition. *Nature* **358**, 86–89.

Karplus, K. and Sjölander, K. and Barrett, C. and Cline, M. and Haussler, D. and Hughey, R. and Holm, L. and Sander, C. (1997). Predicting structures using hidden Markov models. *Proteins: Struct. Funct. Genet., Suppl* **1**, 134–139.

Karplus, K. and Barrett, C. and Hughey, R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856.

Krogh, A. and Brown, M. and Mian, I. S. and Sjölander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Moult, J. and Hubbard, T. and Bryant, S. H. and Fidelis, K. and Pedersen, J. T. (1997). Critical assesment of methods of proteins structure predictions (CASP): Round II. *Proteins: Struct. Funct. Genet., Suppl* **1**, 2–6.

Murzin, A. G. and Bateman, A. (1997). Distant homology recognition using structural classification of proteins. *Proteins: Struct. Funct. Genet., Suppl* **1**, 105–112.

Murzin, A. G. and Brenner, S. E. and Hubbard, T. and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.

Orengo, C. A. and Michi, A. D. and Jones, S. and Jones, D. T. and Swindels, M. B. and Thornton, J. M. (1997). Cath - a hierarchical classification of protein domain structures. *Structure* **5**, 1093–1108.

Park, J. and Teichmann, S. A. and Hubbard, T. and Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 249–254.

Park, J. and Karplus, K. and Barrett, C. and Hughey, R. and Haussler, D. and Hubbard,

T. and Chothia C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**, 1201–1210.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444–2448.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1145–1160.

Rice, D. and Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026–1038.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. and Schneider, R. and Sander, C. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480.

Salamov, A. A. and , Suwa, M. and Orengo, C. A. and Swindells, M. B. (1999a). Genome analysis: Assigning protein coding regions ato three-dimensional structures. *Protein Science* **8**, 771–777.

Salamov, A. A. and Suwa, M. and Orengo, C. A. and Swindells M. B. (1999b). Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng* **12**, 95–100.

Sander, C. and Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Sheridan, R. P. and Dixon, J. S. and Venkataraghavan, R. (1985). Gnerating plausible protein folds by secondary structure similarity. *Int. J. Pept. Protein Res.* **25**, 132–143.

Sonnhammer, E. L. and Eddy, S. R. and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins, Structure function and genetics* **28**, 405–420.

Table I:

Description	Number
Number of protein domains	976
Number of different families	597
Number of different superfamilies	443
Number of different folds	330
Number of possible correct hits on family level	555
Number of possible correct hits on superfamily level	434
Number of possible correct hits on fold level	321
Number of possible false hits	944, 162

Table II:

Method	Multiple sequence information	Parameters	Description
BLAST2 Altschul *et al.* (1997)	-	-e 99	Scoring was done using the expectation values.
SSEARCH Pearson & Lipman (1988)	-	-E 99	Scoring was done using the expectation values.
PSI-BLAST Altschul *et al.* (1997)	SWISSPROT+ TREMBL+ benchmark	blastpgp -j 3 -m 4 -e 1e12	Scoring was done using the expectation values.
HMMER-HSSP Eddy (1998)	HSSP Sander & Schneider (1991)	hmmbuild –pam blosum62.dat	For each sequence the multiple sequence alignment from HSSP was used to create a hidden Markov model for HMMER-2.1.
SAM-HSSP Karplus *et al.* (1998)	HSSP	buildmodel -priorlibrary recode1.20comp	The same as HMMER-HSSP but using SAM-T98.
HMMER-PSIBLAST	SWISSPROT+ TREMBL+ benchmark	–pam blosum62.dat	PSI-BLAST was used to create the multiple sequence alignment, allowing max three iterations.
SAM-PSIBLAST	SWISSPROT+ TREMBL+ benchmark	-priorlibrary recode1.20comp	The same as HMMER-PSIBLAST but using SAM-T98
BLAST-LINK	SWISSPROT+ TREMBL+ benchmark	-	BLAST2 is used for intermediate searches, the score is calculated as the max of the two expectation values.
ssHMM Hargbo & Elofsson (1999)	HSSP	PhD Rost & Sander (1993)	Multiple sequence alignment from HSSP and predicted secondary structure from PhD
THREADER Jones *et al.* (1992)	-	-	For each domain a template for THREADER was created, then the query sequences was threaded against all templates.

Table III:

Method	Rank 1	Rank 5
BLAST2	366 (66%)	389 (70%)
SSEARCH	381 (69%)	**420 (76%)**
PSI-BLAST	**395 (71%)**	401 (72%)
HMMER-HSSP	382 (69%)	413 (74%)
SAM-HSSP	380 (68%)	406 (73%)
HMMER-PSIBLAST	376 (68%)	408 (74%)
SAM-PSIBLAST	389 (70%)	413 (74%)
BLAST-LINK	**414 (75%)**	**438 (79%)**
ssHMM	350 (63%)	398 (72%)
THREADER	273 (49%)	327 (59%)

Table IV:

Method	Rank 1	Rank 5
BLAST2	81 (19%)	129 (30%)
SSEARCH	90 (21%)	141 (32%)
PSI-BLAST	**119 (27%)**	121 (28%)
HMMER-HSSP	70 (16%)	114 (26%)
SAM-HSSP	109 (25%)	**164 (38%)**
HMMER-PSIBLAST	90 (21%)	136 (31%)
SAM-PSIBLAST	**123 (28%)**	**169 (39%)**
BLAST-LINK	**127 (29%)**	**176 (41%)**
ssHMM	80 (18%)	137 (32%)
THREADER	47 (11%)	107 (25%)

Table V:

Method	Rank 1	Rank 5
BLAST2	18 (6%)	42 (13%)
SSEARCH	18 (6%)	50 (16%)
PSI-BLAST	13 (4%)	15 (5%)
HMMER-HSSP	17 (5%)	38 (12%)
SAM-HSSP	15 (5%)	47 (15%)
HMMER-PSIBLAST	14 (4%)	47 (15%)
SAM-PSIBLAST	11 (3%)	60 (19%)
BLAST-LINK	22 (7%)	53 (17%)
ssHMM	22 (7%)	77 (24%)
THREADER	**47 (15%)**	**121 (38%)**

Table VI:

Method	Rank 1	Rank 5
BLAST2	0 (0%)	9 (17%)
SSEARCH	0 (0%)	3 (6%)
PSI-BLAST	0 (0%)	0 (0%)
HMMER-HSSP	0 (0%)	8 (15%)
SAM-HSSP	0 (0%)	8 (15%)
HMMER-PSIBLAST	5 (10%)	15 (29%)
SAM-PSIBLAST	3 (6%)	8 (15%)
BLAST-LINK	5 (10%)	10 (19%)
ssHMM	5 (10%)	16 (31%)
THREADER	**22 (42%)**	**34 (65%)**

Figure 1: Different ways of using multiple sequence information. Each circle represents a family member and the distance between two circles represents some arbitrary scoring distance. The black circle is the query sequence that is used as a start for the database search. *(i) A pairwise alignment algorithm*: would find all pairs within a certain distance in scoring space from the query sequence. *(ii) A two step method:* would then use all these sequences to create a multiple sequence alignment and then perform a search using the profile. This search would find all proteins within a certain distance from the center of the sequences included in the profile. *(iii) In an iterative method:* the search starts from a query protein to find the closest relatives and then creates the profile from this to do another search, etc. This approach would find all related proteins as long as it is never too far away from the center of the proteins found. *(iv) In a linking method:* all proteins related to the query protein will be clustered together, i.e. it does not matter where the center of related proteins is placed, as long as there is never too large a distance to its closest neighbor.

Figure 2: Spec-sens curve at family level. In the top panel the two threading methods, THREADER and ssHMM, are shown together with PSI-BLAST. In the middle panel the hidden Markov models are shown, and at the bottom the single sequence methods and BLAST-LINK. All curves have been smoothed by a running average of the raw data to make them easier to analyze.

Figure 3: Spec-sens curve at at superfamily level. In the top panel the two threading methods, THREADER and ssHMM, are shown together with PSI-BLAST. In the middle panel the hidden Markov models are shown, and at the bottom the single sequence methods and BLAST-LINK. All curves are created by using a running average of the raw data.

Figure 4: Spec-sens curve at fold level. In the top panel the two threading methods, THREADER and ssHMM, are shown together with PSI-BLAST. In the middle panel the hidden Markov models are shown, and at the bottom the single sequence methods and BLAST-LINK. All curves are crated from a running average of the raw data.