BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S182–S191

ProClust: improved clustering of protein sequences with an extended graph-based approach

P. Pipenbacher[1, 3], A. Schliep[†], S. Schneckener[2], A. Schönhuth[1], D. Schomburg[3] and R. Schrader[1]

[1]ZAIK/ZPR, Universität zu Köln, Weyertal 80, Köln, 50937, DE, [2] Science Factory, Unter Käster, Köln, 50937, DE and [3]Institut für Biochemie, Universität zu Köln, Zülpicher Straße, Köln, 50937, DE

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Motivation: The problem of finding remote homologues of a given protein sequence via alignment methods is not fully solved. In fact, the task seems to become more difficult with more data. As the size of the database increases, so does the noise level; the highest alignment scores due to random similarities increase and can be higher than the alignment score between true homologues. Comparing two sequences with an arbitrary alignment method yields a similarity value which may indicate an evolutionary relationship between them. A threshold value is usually chosen to distinguish between true homologue relationships and random similarities. To compensate for the higher probability of spurious hits in larger databases, this threshold is increased. Increasing specificity however leads to decreased sensitivity as a matter of principle.

Sensitivity can be recovered by utilizing refined protocols. A number of approaches to this challenge have made use of the fact that proteins are often members of some larger protein family. This can be exploited by using position-specific substitution matrices or profiles, or by making use of *transitivity* of homology. Transitivity refers to the concept of concluding homology between proteins A and C based on homology between A and a third protein B and between B and C. It has been demonstrated that transitivity can lead to substantial improvement in recognition of remote homologues particularly in cases where the alignment score of A and C is below the noise level.

A natural limit to the use of transitivity is imposed by domains. Domains, compact independent sub-units of proteins, are often shared between otherwise distinct proteins, and can cause substantial problems by incorrectly linking otherwise unrelated proteins.

Results: We extend a graph-based clustering algorithm which uses an asymmetric distance measure, scaling similarity values based on the length of the protein sequences compared. Additionally, the significance of alignment scores is taken into account and used for a filtering step in the algorithm. Post-processing, to merge further clusters based on profile HMMs is proposed. SCOP sequences and their super-family level classification are used as a test set for a clustering computed with our method for the joint data set containing both SCOP and SWISS-PROT. Note, the joint data set includes all multi-domain proteins, which contain the SCOP domains that are a potential source of incorrect links. Our method compares at high specificities very favorably with PSI-Blast, which is probably the most widely-used tool for finding remote homologues.

We demonstrate that using transitivity with as many as twelve intermediate sequences is crucial to achieving this level of performance. Moreover, from analysis of false positives we conclude that our method seems to correctly bound the degree of transitivity used. This analysis also yields *explicit* guidance in choosing parameters.

The heuristics of the asymmetric distance measure used neither solve the multi-domain problem from a theoretical point of view, nor do they avoid all types of problems we have observed in real data. Nevertheless, they do provide a substantial improvement over existing approaches.

Availability: The complete software source is freely available to all users under the GNU General Public License (GPL) from http://www.bioinformatik.uni-koeln.de/~proclust/download/

Contact: proclust@www.bioinformatik.uni-koeln.de, schliep@zpr.uni-koeln.de

Supplementary Information: A web interface to the software allowing to run query sequences against the set of clusters is available at http://www.bioinformatik.uni-koeln.de/~proclust.

[†] Current address: Max Planck Institute for Molecular Genetics, Berlin, 14195, DE

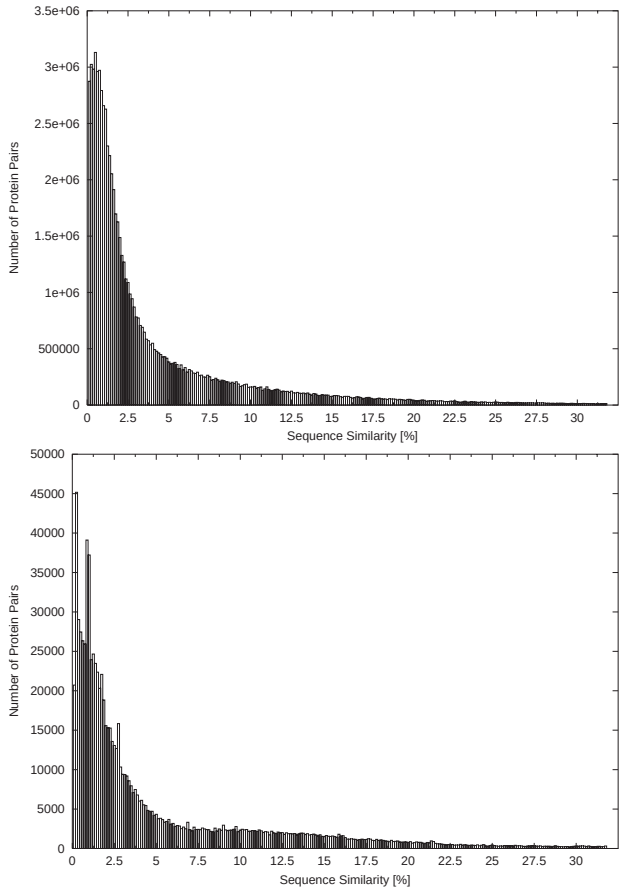

Fig. 1. For pairs of domain sequences from SCOP 1.53 we show histograms of alignment scores: the sequences are either members from distinct SCOP super-families (top) or the same SCOP super-family (bottom). Only pairs with sequence similarities of up to 30% are shown. Note the extensive overlap; separating the two classes of pairs by alignment score is virtually impossible. There are even more true homologues with very low sequence similarity compared with SCOP 1.37 (not shown).

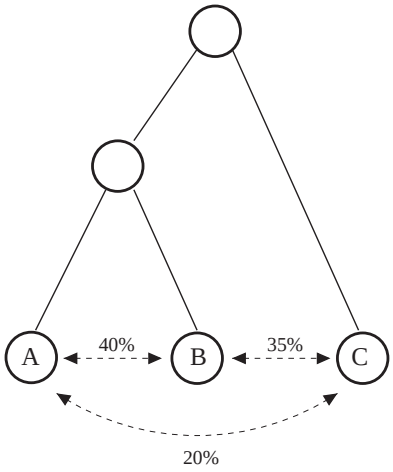

Fig. 2. This partial evolutionary tree demonstrates the biological mechanism which allows the use of transitivity. Proteins *A* and *C* have diverged too far to establish homology based on their low similarity value of 20%. However, as the existence of an (unknown) common ancestor of *B* and *C* as well as *A* and *C* can be established due to the reasonably high similarity values of 35% respectively. 40% of their sequences, protein *B* might serve as the missing link.

INTRODUCTION

The advances in experimentally determining or verifying the three-dimensional structure of proteins do not keep up with the ever-increasing sequencing capacities. A standard method for alleviating this problem is using homology between a target sequence of unknown structure and a protein of known structure to predict the structure of the target. Homology, the existence of a common ancestor, can be detected by a pair-wise comparison if the sequence similarity is 'significant' (Chothia and Lesk, 1986; Sander and Schneider, 1991; Rost, 1999). This allows to infer structural or even functional similarity (Brenner *et al.*, 1998; Pearson, 1997, 1995).

It is well known that a large proportion of true homologues are hidden in the so-called *twilight zone*; their sequence similarity is too low to separate them from pairs of sequences with equal or even higher sequence similarity due to chance, which is drastically apparent in Figure 1. This problem becomes more acute as the increase of the size of sequence databases (Spang and Vingron, 2001) leads to increased noise level; the twilight zone is ever increasing.

Proteins are often part of protein families. This kinship can be used to find a related known structure, which is hidden in the twilight zone, by using other, more closely related family members as *intermediate* sequences. This concept is called *transitivity* and refers to the following property of mathematical relations: If *A* and *B* are related as well as *B* and *C*, then *A* and *C* are also related. Transitivity, as it applies to the problem of finding remote homologues, is depicted in Figure 2; it has been examined in a number of approaches (Abagyan and Batalov, 1997; Park *et al.*, 1997; Pearson, 1997; Gerstein, 1998; Salamov *et al.*, 1999; Arvestad *et al.*, 2000; Bolten *et al.*, 2001). They establish that transitivity does work in this context, but unfortunately only to a limited extent. Note, the relation we are considering and which we are trying to detect through sequence similarity is structural homology, which is not truly a transitive relation in the mathematical sense.

There are a number of factors contributing to this limitations. On one hand we know from the theory of random graphs (Spencer, 2001) that large enough random

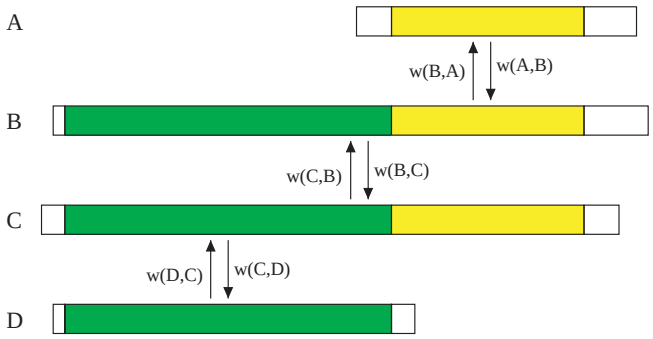

Fig. 3. This figure motivates our desire for an asymmetric distance measure. Here A and D are proteins consisting of distinct domains depicted in the different shades of gray. B and C are two multi-domain proteins each containing both domains. If a *symmetric* distance measure is used, i.e. $w(A, B) = w(B, A)$, then an incorrect link from A to D is established as long as the edges are present in the threshold graph. In the asymmetric case the length-dependent scaling will result in $w(A, B) < w(B, A)$ and possibly removal of the edge (A, B) when going over to a threshold graph. Thus, the links from B to A and from C to D will be lost and hence A and D will not longer be linked.

similarities will produce so-called super-clusters, very large clusters connecting large parts of the sequence space. This can be dealt with by using more stringent criteria for significance.

Multi-domain proteins pose the more acute problem. Domains are compact, semi-independent structural units of proteins, which often appear highly conserved in a number of multi-domain proteins; i.e. proteins containing two or more domains, see Figure 3 for a schematic view. The edges in Figure 3 represent significant sequence similarity, and, considered individually, are correct. However, a *symmetric* similarity relation does not distinguish between two proteins being globally similar and one protein being similar to an individual domain of a multi-domain protein. This leads to incorrect links via intermediate sequences between distinct single-domain proteins (cf. protein A and D in Figure 3). An *asymmetric* similarity relation or distance measure can be employed to distinguish between the two distinct flavors of similarity mentioned above. Since obtaining domain annotation is neither possible in general nor computationally feasible, a simple heuristic (Bolten *et al.*, 2001) was proposed which we extend in this manuscript to deal with the aforementioned problems.

A number of related approaches have used the concept of transitivity for large scale analysis of protein sequences.

Systers (Krause and Vingron, 1998) uses an iterated BLAST or FASTA search for computing clusters. The iteration proceeds by picking the protein most distantly related to the query subject to some consistency and termination conditions. Clusters computed for all proteins are subsequently merged and processed further. Potential multi-domain problems are not explicitly dealt with in this approach.

Protomap (Yona *et al.*, 1999) also uses a graph-based approach where edges represent sequence comparisons and the corresponding edge weights result from a scoring scheme combining BLAST, FASTA and Smith–Waterman E-values. A hierarchy of clusters is obtained by iteratively lowering thresholds in a threshold graph and computing strongly connected components at each threshold level. Presence of ambiguous proteins, which potentially are multi-domain proteins, results in cluster splitting.

Enright and Ouzounis (2000) employ a routine all against all BLAST search and subsequently ignore hits below a specified E-value threshold, yielding a $(0, 1)$-similarity matrix. They disregard all differences in similarity for hits above the threshold. Extensive post-processing requiring additional Smith–Waterman is performed to symmetrize the matrix and to deal with multi-domain proteins, assuring that transitivity holds row-wise in the similarity matrix. Subsequently, rows are clustered using single links. An evaluation of performance is provided by inspection of some examples.

Tatusov *et al.* (1997) build clusters of orthologous groups (COG's) starting with proteins from seven different species. At first significant hits across species are detected and so-called 'triangle relationships' used as seeds for clusters. An iterative merging process is performed, which tries to account for the multi-domain problem in the merging step. Novel protein sequences can be compared to the existing clusters to provide structure and function prediction.

Our method is designed to provide a clustering as an aid in finding remote homologues; the multi-domain problem is directly addressed although we do not pretend to fully solve it. However, the asymmetric distance employed results in very high sensitivity while keeping error rates at a minimum, as the large-scale evaluation shows. In the following sections, we give a detailed account on the extensions to the graph-based clustering algorithm we have developed, describe the data sets used, present and discuss our results with an emphasis on the extend of transitivity used and problems with multi-domain proteins. The evaluation process also provides *explicit* guidance for choosing parameters. We conclude with an outlook on further developments.

ALGORITHM

The algorithm is an extension of the graph-based clustering proposed in Bolten *et al.* (2001), which we will summarize very briefly in the following. An introduction to graph-based clustering can be found in Jain and Dubes (1988).

- Compute a complete undirected graph G where vertices are identified with protein sequences and each edge represents a Smith–Waterman local alignment (Smith and Waterman, 1981) of the two incident sequences P and Q, weighted with the raw Smith–Waterman score, denoted by raw(P, Q). Note, an arbitrary distance measure can be used as the weight instead of the Smith–Waterman score.
- Replace each undirected edge $\{P, Q\}$ with two directed edges (P, Q) and (Q, P) modifying the weights such that

$$w(P, Q) = \frac{\text{raw}(P, Q) * 100}{\text{raw}(P, P)}$$

  and similarly for $w(Q, P)$. Dividing by the self-similarity raw(P, P) corrects for amino-acid composition and, as the self-similarity is proportional to the number of amino-acids of P, scales the similarity value by the length of P. Hence, $w(P, Q)$ and $w(Q, P)$ will generally be distinct; the distance measure we defined between protein sequences is *asymmetric*. The resulting graph is denoted by G_d.
- Proceed to the threshold graph $G_d(\tau)$ by removing all edges of weight—i.e. a similarity percentage value of—less than τ.
- Compute all strongly connected components (SCCs) (Sedgewick, 1990) in $G_d(\tau)$. The strongly connected components, maximal sets of vertices such that directed paths exists from P to Q and from Q to P for all vertices P, Q in a SCC, are output as the resulting clusters.

This algorithm and the results presented in (Bolten *et al.*, 2001) raised a number of questions and opened several possible avenues for improving the performance. To avoid over-fitting and to concentrate on the highest performance pay-off extensions, we chose to restrict ourselves to graph pruning based on score significance and a Profile-HMM based post-processing step presented in the following. Other extensions were implemented and evaluated, and either matched or, in combination, insignificantly exceeded the performance presented in this paper (not shown).

Filtering by score significance

Preliminary analysis (not shown) suggested that especially for short sequences an improvement in performance might be gained by pruning the graph further based on the statistical significance of the score. We employed the standard extremal value distribution (Karlin and Altschul, 1990) to estimate maximal scores observable with the Smith–Waterman algorithm for random sequences (Waterman and Vingron, 1994) of given lengths. The parameters of the extremal value distribution, $\gamma = 0.04469$ and $p = 0.971029$, were estimated by computing alignments of random sequences using our Smith–Waterman implementation with the parameters listed below. The pruning consisted of removing edges (P, Q) from the graph G if the significance of the score $w(P, Q)$ was below the chosen significance threshold t_σ. Various values for t_σ were tested.

Post-processing: merging clusters

As was noted before (Bolten *et al.*, 2001), the clustering procedure seems to be rather conservative and likely to produce clusters which partition SCOP super-families. It was a natural extension to investigate whether a criterion could be found to merge those clusters without introducing false positives. Given the high quality of the clusters, we propose to use Profile-HMMs for that task. The protocol providing the greatest gain was the following:

- Clusters containing at least twenty sequences were selected.
- A multiple alignment was built for each set of sequences with the ClustalW (Thompson *et al.*, 1994) software version 1.7 using the default parameters.
- With the HMMER package (Eddy, 1998), version 2.1.1 from http://hmmer.wustl.edu/, profiles were built with the hmmbuild and hmmcalibrate programs. Again, default parameters were used.
- For each such cluster profile all sequences not contained in the cluster were scored using the profile and the E-value was recorded.
- Clusters C and D were merged, if, using the profile for cluster C, the average E-value of sequences from D was below some threshold t.

Complexity and running time

The dominating term is the computation of the pair-wise sequence comparisons, which is quadratic in the number of sequences. However, it only has to be performed once, it is trivial to distribute to a large cluster of CPU's, and additions or changes to the computed data set can be made incrementally. The resulting graphs G and G_d are large but can be easily dealt with in real-time. The computation of the SCC's is linear in the number of vertices plus the number of edges (Sedgewick, 1990). The clustering as well as subsequent filtering operations on the graphs benefit greatly from the fact that the threshold graphs $G_d(\tau)$ are typically very sparse. We observed an average vertex degree of 17.6.

For the data set ALL (see below for details) the Smith–Waterman computations needed 70 CPU days, the

Table 1. Descriptive statistics of the datasets used

	SCOP	SPROT	ALL
Number of sequences	9.403	47.160	56.563
Average length	176	381	346
Number of families	1.264	/	/
Number of super-families	807	/	/
Number of folds	534	/	/
Proteins per super-family	11,7	/	/
Homologous pairs	608.578	/	/
Non-homologous pairs	43.594.925	/	/

clustering needs about 30 seconds. For the cluster merging using HMMs about 21 CPU days were needed.

IMPLEMENTATION AND EVALUATION

The method has been implemented in a C++ software which has been published under the GNU General Public License (GPL). It has been developed and tested on a Compaq ES40 running Tru64 Unix V5.1, using Compaq's cxx compiler, version 6.20. In addition, it has been tested and used on various Sun Ultra computers (Ultra 5 up to Sun Enterprise 10000), running Solaris 7 and earlier versions, using the GNU g++ compiler version 2.9x and above.

In the Smith–Waterman algorithm (our own implementation is included in software), the following parameters (Bolten *et al.*, 2001) were used: an integerized version of the BLOSUM80 substitution matrix, gap opening penalty 90 (about 1.5 times the average a.a. identity score), and gap extension penalty 9. The substitution matrix was chosen based on experiments of one of the authors (Schneckener, 1998). Guidance for choosing the gap penalties was provided by experimentation with single-link clustering on a subset of SCOP; cf. (Bolten *et al.*, 2001). The choice of gap penalties proved not to be critical (not shown).

Data sets

We used the following datasets, which are available from http://www.bioinformatik.uni-koeln.de/∼proclust/download/ for easier reference.

SCOP. We used SCOP (Hubbard *et al.*, 1999) version 1.53 from http://astral.stanford.edu/scopseq-1.53.html. The domain sequences and classification were obtained from http://astral.stanford.edu/seq.cgi?get=scopdom-seqres-all;ver=1.53. This file does not contain any sequences from SCOP classes 8–9. After removing all sequences with less than 40 a.a., the sequences were filtered for low complexity regions by using the software seg (Wootton and Federhen, 1993) with parameters '12 1.8 2.0 -x'. Sequences containing masked a.a. as well as duplicate sequences were removed.

SPROT. SWISS-PROT (Bairoch and Apweiler, 2000) release 39 from ftp://ftp.ebi.ac.uk/pub/databases/swissprot was processed analogously to SCOP: short sequences of less than 40 a.a. as well as sequences containing a.a. masked due to low complexity were removed. To speed up computations the CD-HI (Cluster Database at High Identity) software (Li *et al.*, 2001) was used to remove redundant sequences at the 80% or higher sequence identity level. A cutoff at this high identity level is unlikely to influence the results and greatly facilitates re-computations should they become necessary.

ALL. This dataset was created by merging SCOP and SPROT.

Evaluating performance

We evaluated the validity of our hypothesis, that the asymmetric sequence length-dependent distance measure improves recognition of remote homologues while avoiding false positives due to problems with multi-domain proteins, by using SCOP as test set. The annotation given by the SCOP super-family classification of a (domain) sequence was taken as the 'truth' to which we compared the clustering we computed. Note, the clustering for the analysis was performed on the combined data set ALL containing the domain sequences from SCOP as well as virtually all non-redundant SWISS-PROT sequences. In particular, ALL included the *complete* sequences which contain the domain sequences from SCOP. A failure of the method would be clearly detectable by incorrectly joining pairs of sequences from distinct SCOP super-families by virtue of a multi-domain SWISS-PROT sequence containing them both.

For the further analysis we will refer to a (unordered) pair of sequences from the same SCOP super-family as *true* homologues, and to a pair of sequences in the same computed cluster as *predicted* homologues. We will call a predicted true homologue pair *true positive* (TP), a true homologue which has been not predicted *false negative* (FN), a true non-homologue pair predicted to be homologue *false positive* (FP) and a true non-homologue pair not predicted *true negative* (TN). The following derived quantities allow to summarize the performance: *Sensitivity* specifies the proportion of homologue pairs detected

$$sens = \frac{\#TP}{\#TP + \#FN}$$

and *specificity* the proportion of correct predictions among the pairs predicted to be homologues

$$spec = \frac{\#TP}{\#FP + \#TP}.$$

A perfect method would have $sens = spec = 1$, which implies that neither false positive nor false negative errors are made.

Table 2. Histogram of super-family sizes in SCOP versus cluster sizes

Size	Proportion of super-families	Proportion of clusters
1	22.7%	65.7%
2–5	37.8%	24.5%
6–10	16.0%	4.4%
11–20	11.0%	3.0%
21–50	8.3%	1.9%
51–100	3.0%	0.4%
>100	1.2%	0.1%

For comparison with PSI-Blast (Altschul *et al.*, 1997) we used PSI-Blast version 2.1.2 from ftp://ncbi.nlm.nih.gov/blast/executables with the following parameters '-h E-Value -e E-Value -j 20 -M BLOSUM80 -b 0 -F T'. PSI-Blast is *not* symmetric in the sense that it does not necessarily find sequence P starting from a query sequence Q, even if the reverse search, using sequence P as the query, does find Q. To compensate for that, we considered *ordered* pairs of sequences from SCOP in the comparison. That is, for the two sequences P and Q we considered both pairs (P, Q) and (Q, P), running two separate PSI-Blast searches with P and Q as query sequences. Given a query sequence P, we defined (P, Q) to be a homologue predicted by PSI-Blast, if Q was among those sequences found and vice versa for queries from Q. Since the SCOP classification is identical for both (P, Q) and (Q, P) whereas predictions by PSI-Blast might differ, it can occur that (P, Q) and (Q, P) are different with respect to their status of true/false positives respectively negatives, when evaluating PSI-Blast.

This way of counting predictions is in favor of PSI-Blast. It results in a higher sensitivity of PSI-Blast, as the many cases where asymmetric, i.e. only one pair of (P, Q) or (Q, P) was predicted, PSI-Blast search results were observed (not shown) gave at least partial credit. All searches were performed on the ALL dataset.

DISCUSSION

We chose parameters as to achieve maximal sensitivity at a specificity of 99%. This rate of 1% false positives has been chosen in a number of publications (Park *et al.*, 1997; Brenner *et al.*, 1998; Park *et al.*, 2000; Enright and Ouzounis, 2000) as a reasonable compromise which substantially improves sensitivity compared with requiring perfect specificity.

The clustering computed appears to have a tendency to partition SCOP super-families. This can be deduced from the high specificity and the histogram of cluster sizes in Table 2. On a positive note, super-components due to random similarities do not emerge. Roughly half of the resulting clusters are 'non-trivial'. That is, they contain pairs of sequences which are connected via a number of intermediate sequences, pair-wise similarities vary over a wide range with a large proportion of pairs having score below the clustering threshold τ (not shown)[†].

Observe the rather 'flat' shape of the sensitivity versus specificity curve in Figure 4 and the sudden rise in

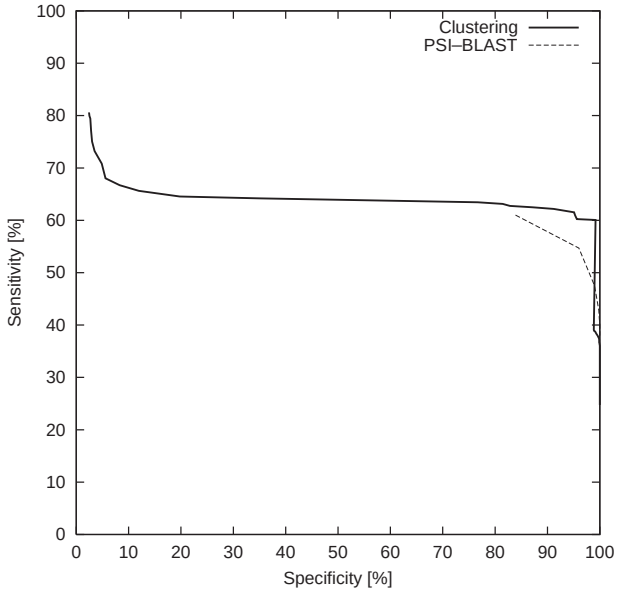

Fig. 4. This comparison with PSI-BLAST shows sensitivity versus specificity for both methods: the clustering has been computed on the data set ALL, evaluation is done on the data set SCOP. As already the partial curve indicates, a greater flexibility with respect to choosing an appropriate specificity versus sensitivity tradeoff is provided by PSI-Blast.

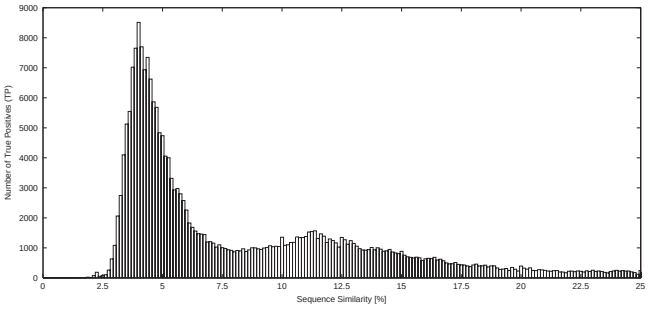

Fig. 5. The histogram of scores for true positives in a clustering of data set ALL with threshold $\tau = 13.1\%$ shows a larger number of pairs with very low sequence similarity. Scores were taken from the complete directed graph G_d.

[†] Length-, distance- and score-histograms per cluster are available for download from http://www.bioinformatik.uni-koeln.de/~proclust/download/

Table 3. Results obtained for our clustering (top) as well as PSI-Blast (bottom). Searches have been performed on the ALL dataset, the evaluation on sequences from SCOP. The number of TN is 87 084 524 minus the number of FP

τ	$1 - t_\sigma$	#TP	#FP	#FN	*sens.*	*spec.*
16.3%	$2.3\,10^{-6}$	230 240	2 083	374 949	38.0%	99.1%
13.1%	$8.3\,10^{-7}$	364 458	3 096	240 731	60.2%	99.2%
13.1%	$3.1\,10^{-7}$	363 545	3 028	241 644	60.1%	99.2%

E-value	#TP	#FP	#FN	*sens.*	*spec.*
10^{+0}	662 076	27 467	547 612	54.7%	96.0%
10^{-1}	**574 887**	**5 934**	**635 491**	**47.5%**	**99.0%**
10^{-2}	517 851	1 187	692 527	42.8%	99.8%
10^{-4}	485 437	188	724 939	40.1%	>99.9%

sensitivity and loss of specificity, once the threshold is lowered below $\tau = 4\%$, at which level spurious similarities due to random similarities appear. This does not allow to improve the recognition of homologues, even at the expense of a lower specificity, by varying the threshold. In contrast PSI-Blast can easily be re-run, after the initial profiles have been generated, with a larger *E*-values threshold to find a larger proportion of homologues, while sacrificing specificity.

Using transitivity

We estimated the degree of transitivity used by computing distances between true positive and false positive SCOP sequences. In graphs a distance between two vertices is naturally given by the length—i.e. the number of edges—of a shortest path connecting them. Note, in directed graphs the distance from P to Q is not necessarily equal to the distance from Q to P; in a SCC paths from P to Q and vice versa exist by definition. As Figure 6 shows, a substantial proportion of true homologues have distance two or larger, with a significant drop-off at distance five. That is, one up to four intermediate sequences are needed for about 50% of the super-family pairs. However, still a sizable proportion has larger distance up to a maximum of 13.

False positives are rare (note the different scale for the *y*-axis) and have an average distance of about 4.6, which is substantially larger than the 2.1 we observe for the true positives. However, there is a wide variation of distances as well as a substantial overlap of the two histograms for the two different classes of positives. Hence, true positives cannot be separated from false positives by their distance. If high distances were an indicator for false positives, this would show an overuse of transitivity. The opposite seems true, errors are rather due to high

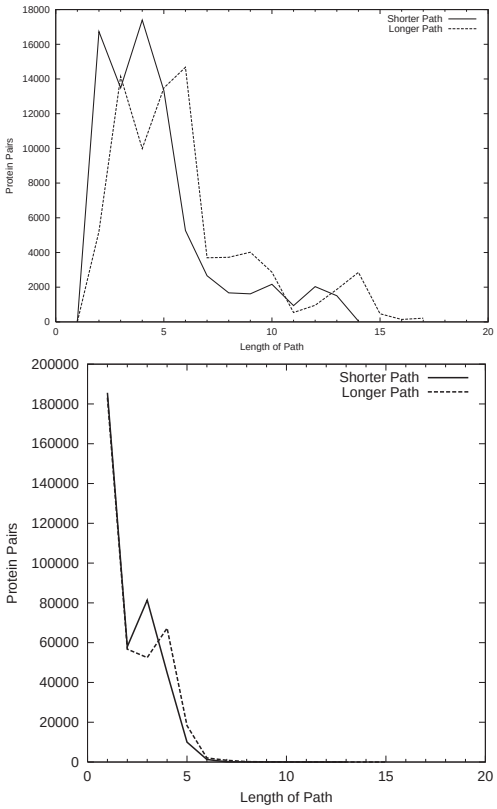

Fig. 6. Since $G_d(\tau)$ is directed, and TP as well as FP are in the same SCC, directed paths going in both directions exist by definition. We show histograms of the shorter respectively longer path between false positives (top) and true positives (bottom). Observe the abundance of the latter for distance 3-5 and the existence of true positives at even larger distances up to a maximal distance of 13.

sequence similarity by chance which supports the claim that our method limits the degree of transitivity inherently and correctly. Figure 6 also indirectly demonstrates that clusters are inhomogeneous with respect to distances between members.

Dealing with multi-domain proteins

The evidence supporting the success of our method in dealing with multi-domain proteins (their abundance is depicted in Table 4) is indirect and relies on the presence of the multi-domain protein sequences from SWISS-PROT in the ALL dataset. As we have demonstrated, intermediate sequences are used to a large extend to link SCOP domain sequences but nevertheless few of those links are incorrect, as indicated by a specificity of 99.2%.

We analyzed the false positive errors and observed the following causes of errors. In general such errors are the result of 'unwanted' edges, e.g. (A, B) or (D, C)

Table 4. The abundance of multi-domain proteins and the number of domains is tabulated for the 33 409 sequences from ALL for which information about the domain composition could be derived from Pfam 6.2 (Bateman *et al.*, 2000)

No. of domains	No. of proteins
1	25.738
2	4.902
3	1.261
4	580
5	261
6	181
7	137
≥8	349
total	33.409

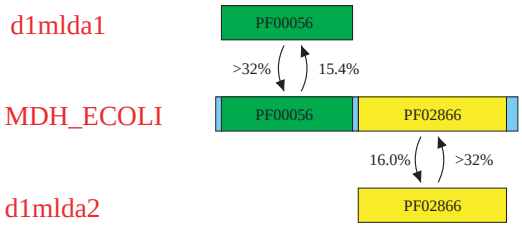

Fig. 7. This multi-domain problem is present in cluster # 1779 for data set ALL using a threshold of $\tau = 13.1\%$ and a significance threshold $t_\sigma = 1 - 3.1 \cdot 10^{-7}$. MDH_ECOLI is a Malat-Dehydrogenase from *E. coli*, *d1mlda1* is the NAD(P)-binding N-terminal (PF00056) and *d1mlda2* the C-terminal domain (PF02866) of the Malat-Dehydrogenase from *Sus scrofa*. The multi-domain problem depicted in this picture disappears if the threshold is raised above 15.4%, since the edge linking MDH_ECOLI to *d1mlda1* vanishes. There are further examples of these two domains causing problems at thresholds as high as $\tau = 18.8\%$. There are no edges between *d1mlda1* and *d1mlda2*. Domain annotation obtained from Pfam 6.2 (Bateman *et al.*, 2000).

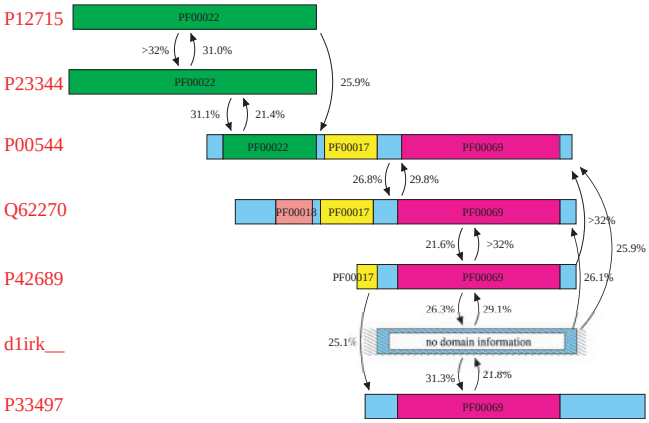

Fig. 8. A larger multi-domain problem in cluster # 1517: Clustering of ALL, using a a threshold of $\tau = 21.3\%$ and a significance threshold $t_\sigma = 1 - 3.1 \cdot 10^{-7}$. There are no edges between P12715 and P33497 in the graph. Domain annotation obtained from Pfam 6.2 (Bateman *et al.*, 2000).

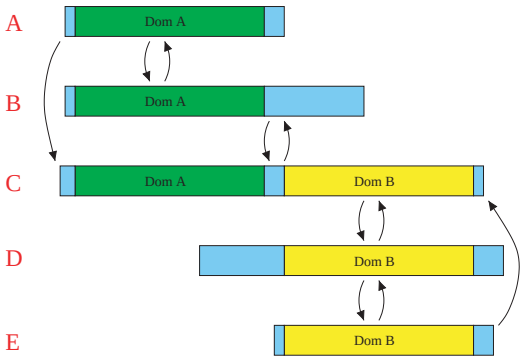

Fig. 9. This schematic picture shows a case where our simple heuristic fails. Due to the 'ladder' of proteins with just the right increase in length, none of unwanted edges are removed when going over to the threshold graph. Such cases have been observed in the analysis of false positive appearing for lower thresholds τ (not shown).

in Figure 3, not being removed when going over to the threshold graph $G_d(\tau)$. In the examples in Figure 7 and Figure 8, the length-dependent scaling heuristic we employ fails because the shared domains are too well conserved for the length ratio of the proteins involved.

There are also some systematic errors associated with the heuristic. The most common one is caused by edges from a single-domain protein to a multi-domain protein (cf. proteins *A* and *B* respectively in Figure 3 or Figure 9) having weights above the threshold τ since the differences in length are not large enough to have enough of a scaling effect. Typically this will appear, see Figure 9, as a 'ladder' of proteins of increasing length. Each step of this ladder is a valid edge in itself.

Another problem is posed by multi-domain proteins of similar length, sharing exactly one well conserved domain. Besides incorrectly linking those two proteins, this can also lead to incorrect links between distinct single-domain proteins analogously to Figure 3. These and other possible problems appear however to be rare as indicated by the very high specificity of our method.

Merging clusters

The use of profile HMMs to merge clusters with and assign singletons, or one-element clusters, to those large enough to allow proper training of HMMs showed only a very modest improvement of 3.3% in sensitivity with a small loss of 0.14% in specificity, cf. Table 5.

Table 5. Changes in true positives, ΔTP, and false positives, ΔFP, using the HMM-based cluster merging for varying E-value are shown. Our choice of 10^2 is displayed in bold

E-value	ΔTP	ΔFP	Sensitivity	Specificity
10^{+0}	+9.087	+113	61.6%	99.2%
10^{+1}	+12.633	+228	62.2%	99.1%
10^{+2}	**+20.959**	**+745**	**63.5%**	**99.0%**
$5 \cdot 10^{+2}$	+30.084	+9.064	65.0%	97.0%

We also investigated the following graph theoretical approach. Compute the average number of edges connecting cluster C to cluster D. Note, all edges between clusters must have the same direction by definition of a SCC. If that average is above some threshold t_m, merge C and D. However, this resulted only in very marginal improvements (not shown) which is indicative of SCC-clusters having few and probably spurious links to the rest of the sequence space. The relative scarcity of unidirectional edges between clusters appears somewhat surprising as a cluster of single domain proteins should have many edges towards a cluster of multi-domain proteins sharing that particular domain.

CONCLUSION AND OUTLOOK

We were able to significantly improve the detection of remote homologues using a graph based approach, where vertices represent protein sequences and each edge corresponds to a Smith–Waterman local alignment. Scaling the raw alignment scores essentially based on the length of the proteins results in an asymmetric distance measure. Clustering was performed by computing strongly connected components in the resulting directed graph after an additional edge pruning based on score significance. Clusters were subsequently merged using profile HMM's.

False positives due to problems with multi-domain proteins are largely avoided. Transitivity, or intermediate sequences are used for recognition of about 50% of the true positives. Altogether, the method achieves a sensitivity of 63.5% at 99.0% specificity improving about 34% upon PSI-Blast's performance of 47.5% sensitivity also at 99.0% specificity. This improvement is gained at the expense of a much larger computational effort, which can be leveraged through the use of CPU clusters. An improved version of the method using PSI-Blast as the input for the clustering is under development and will be described elsewhere.

An efficient and accurate method to predict shared domains based on a sequence alignment, without resorting to databases of known domains, would be highly desirable. Potentially, performance of our method can already be improved by taking the length and position of conserved regions of the alignments into account.

ACKNOWLEDGMENTS

The second author would like to thank Stephen Altschul, Kevin Karplus and Alejandro Schaffer for helpful discussions.

REFERENCES

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

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., 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., Ivansson,L., Lagergren,J. and Elofsson,A. (2000) What is the best method to determine if two proteins are related? A study on the structural and evolutionary significance of pairwise protein sequence alignments. Submitted for publication.

Bairoch,A. and Apweiler,R. (2000) The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Bolten,E., Schliep,A., Schneckener,S., Schomburg,D. and Schrader,R. (2001) Clustering protein sequences-structure prediction by transitive homology. *Bioinformatics*, **10**, 935–942.

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

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

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

Enright,A.J. and Ouzounis,C.A. (2000) Generage: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Gerstein,M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, **14**, 707–714.

Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) Scop: a structural classification of proteins database. *Nucleic Acids Res.*, **27**, 254–256.

Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Krause,A. and Vingron,M. (1998) A set-theoretic approach to database searching and clustering. *Bioinformatics*, **14**, 430–438.

Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.

Park,J., Holm,L., Heger,A. and Chothia,C. (2000) Rsdb: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.

Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.

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

Pearson,W.R. (1997) Identifying distantly related protein sequences. *Comput. Appl. Biosci.*, **13**, 325–332.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Salamov,A.A., Suwa,M., Orengo,C.A. and Swindells,M.B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues [in process citation]). *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*, **9**, 56–68.

Schneckener,S. (1998) Positionsgenaues Alignment von Proteinsequenzen, Ph.D. thesis, Universität zu Köln.

Sedgewick,R. (1990) *Algorithms in C*. Addison Wesley, Reading, MA.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Spang,R. and Vingron,M. (2001) Limits of homology detection by pairwise sequence comparison. *Bioinformatics*, **17**, 338–342.

Spencer,J. (2001) *The Strange Logic of Random Graphs*. Springer, Berlin.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Waterman,M.S. and Vingron,M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci. USA*, **91**, 4625–4628.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163. Software available at http://blast.wustl.edu/pub/seg

Yona,G., Linial,N. and Linial,M. (1999) Protomap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins*, **37**, 360–378.