Clustering Protein Sequences — Structure Prediction by transitive homology

Eva Bolten[1,2], Alexander Schliep[2], Sebastian Schneckener[3],

Dietmar Schomburg[1], Rainer Schrader[2]

[1] Institut für Biochemie, Universität zu Köln

[2] ZAIK/ZPR, Universität zu Köln

Weyertal 80, D-50937 Köln, Germany

Tel. +49-221-470-6040 Fax- +49-221-470-5160

schliep@zpr.uni-koeln.de

[3] Science Factory, Köln, Germany

Keywords: Structure prediction, Proteins, Remote Homologues, Transitivity, Clustering

Abstract

Motivation: It is widely believed that for two proteins A and B a sequence identity above some threshold implies structural similarity due to a common evolutionary ancestor. Since this is only a sufficient, but not a necessary condition for structural similarity, the question remains what other criteria can be used to identify remote homologues.

Transitivity refers to the concept of deducing a structural similarity between proteins A and C from the existence of a third protein B, such that A and B as well as B and C are homologues, as ascertained if the sequence identity between A and B as well as that between B and C is above the aforementioned threshold. It is not fully understood, if transitivity always holds and whether transitivity can be extended ad infinitum.

Results: We developed a graph-based clustering approach, where transitivity plays a crucial role. We determined all pair-wise similarities for the sequences in the SwissProt database using the Smith-Waterman local alignment algorithm. This data was transformed into a directed graph, where protein sequences constitute vertices. A directed edge was drawn from vertex A to vertex B if the sequences A and B showed similarity, scaled with respect to the self-similarity of A, above a fixed threshold. Transitivity was important in the clustering process, as intermediate sequences were used, limited though by the requirement of having directed paths in both directions between proteins linked over such sequences. The length dependency — implied by the self-similarity — of the scaling of the alignment scores appears to be an effective criterion to avoid clustering errors due to multi-domain proteins.

To deal with the resulting large graphs we have developed an efficient library. Methods include the novel graph-based clustering algorithm capable of handling multi-domain proteins and cluster comparison algorithms. SCOP was used as an evaluation data set for our method, yielding a 24 percent improvement over pair-wise comparisons in terms of detecting remote homologues.

Availability: The software is available to academic users on request from the authors.

Contact: schliep@zpr.uni-koeln.de

Supplementary Information: http://www.zaik.uni-koeln.de/~schliep/ProtClust.html

Introduction

Finding the three-dimensional structure of proteins is one of the fundamental problems in molecular biology today. The improvements in throughput of classical methods for determining the structure — e.g., using x-ray diffraction analysis or NMR — cannot keep up with the ever-increasing speed at which proteins are sequenced. It is thus desirable to have methods allowing structure prediction solely from sequence data, either ab initio, modeling the molecular folding process, or homology based, using protein sequences with known structures as a template.

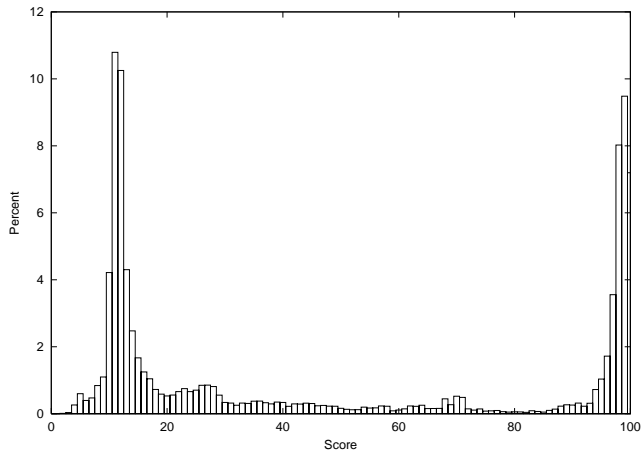The main idea in the latter case is based on the fact

Figure 1: Histogram of pair-wise alignment scores for all pairs from the same super-family in the SCOP1 data set. Note, the large proportion well below the 30% mark which shows that is is impossible to distinguish super-family and unrelated pairs only by their pair-wise alignment score.

that sequence similarity allows to detect homology, i.e., the existence of a common evolutionary predecessor, and thus to infer similar structure and even function virtue of this shared history (Pearson, 1997; Yona *et al.*, 1998). Note, that the same structure or function does not imply a common ancestor; likewise, a common ancestor does not imply a common function, but probably a shared fold.

The relation of sequence similarity — as obtained by pair-wise alignments — to structural or functional properties has been the goal of a number of publications (Brenner *et al.*, 1998; Pearson, 1997; Pearson, 1995). An widely accepted rule-of-thumb is that 30% identity over aligned regions (Chothia & Lesk, 1986) suffices. More recent studies (Sander & Schneider, 1991; Rost, 1999) qualified this rule. We will call a sequence similarity above this threshold *significant.*

There are lots of examples — e.g., pairs of SCOP super-family sequences with low similarity scores, cf. Fig. 1 — of homologue proteins with a sequence similarity below any reasonable threshold. Detecting those *distant homologues*, bringing light into the so-called *twilight zone* of low similarity, has been investigated with a number of different approaches (Abagyan & Batalov, 1997; Park *et al.*, 1997; Pearson, 1997; Gerstein, 1998; Krause & Vingron, 1998; Salamov *et al.*, 1999; Arvestad *et al.*, 2000). Fundamental in several of those approaches was the concept of *transitivity* of homology.

Transitivity is a property of (mathematical) relations. In the context of homology, the relation between proteins is defined as having a shared ancestor, which can
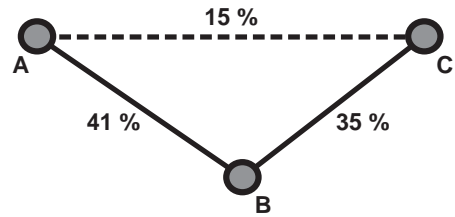

Figure 2: An example of transitivity of homology. Note the low similarity score of A and C . The link between A and C via B is called *indirect* compared to the *direct* link between A and B, respectively B and C.

be interpreted as follows. If for three given proteins A, B and C, A and B as well as B and C have a common ancestor, then also A and C have a common ancestor. This can be used to detect remote homologues, when the sequence similarity between A and C is too low to infer homology with a sufficient degree of confidence. If the level of sequence similarity between A and B as well as that between B and C allows to infer homology directly, we can then use transitivity to infer homology between A and C indirectly, using B as an *intermediate sequence* (cf. Fig. 2). The question remains, if transitivity extends to arbitrary numbers of intermediate sequences, and whether it holds on data sets as large as SwissProt (Bairoch & Boeckman, 1992).

We will show later that multi-domain proteins constitute a problem in the use of transitivity for inferring remote homologues. We have developed a method capable of dealing with this problem, while still employing transitivity to a large degree. In the following sections, we will give a detailed account on this novel graph based clustering algorithm, discuss the choices in the implementation and the data-sets used, present the results of our method and conclude with a discussion and an outlook on further extensions to our method.

Algorithm

We developed a novel graph-based clustering algorithm for structure prediction by transitive homology. Graphs are a natural model for objects and relations between those objects and have been successfully used for clustering in other contexts. See e.g. (Roberts, 1984) for an introduction to graph theory and (Jain & Dubes, 1988) for graph based clustering. We identified protein sequences with nodes of our cluster graph and will from now on use the terms interchangeably.
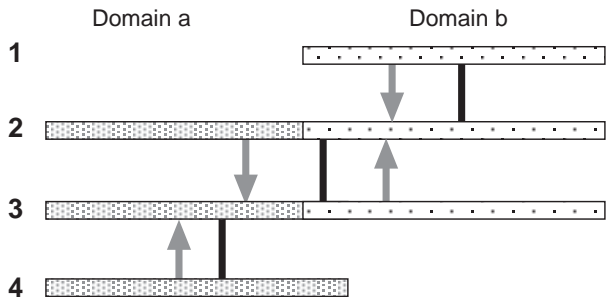
Figure 3: The problem arising from multi-domain proteins is illustrated. If we use an undirected graph, the solid black edges provide a path from protein #1 to protein #4. In the directed case, the grey edges avoid this possible problem.

Computing pairwise similarities

At first, a complete undirected graph G was computed, where each edge between proteins P and Q was weighted with their raw Smith-Waterman (Smith & Waterman, 1981) local alignment score[1], denoted by $raw(P,Q)$. Note, that an arbitrary similarity measure can be used as input for the clustering.

One concern in clustering protein-sequences are multi-domain proteins which form unwanted "bridges", connecting clearly unrelated proteins. As Fig. 3 shows, this is caused by the relation being symmetric, i.e., by not distinguishing between (A, B) and (B, A). If protein A is more or less a domain of protein B, instead of the whole of protein A being similar to the whole of protein B, false positives will result during clustering (Park et al., 1997).

A computationally inexpensive method for reliable prediction of domains solely from sequence information would be highly desirable to accurately establish such a domain relation between protein sequences. Unfortunately, to our knowledge, no such method exists.

Directing the edges

We developed the following simple criterion to approximate the knowledge we would obtain from a domain-prediction method. Noting that there has to be a difference in length between sequences if multi-domain proteins cause a problem, we decided to *direct* the edges in the graph. Each undirected edge was replaced by two *directed* edges, where the weight of the edge from P to Q, (P, Q), was computed as

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

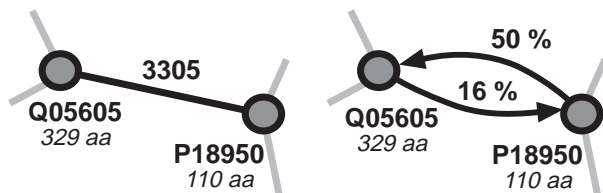[1]See the next section for implementation details.

Figure 4: A single edge (left) is replaced by two *directed* edges in opposite directions (right). Note the distinct similarity scores on the edges.

and similarly for $w(Q, P)$, resulting in a similarity score between zero and one-hundred percent scaled by the raw score of an alignment of P with itself. Note, that the raw self similarity score $raw(P, P)$ is approximately proportional to the length of P. In the case of two sequences P and Q of distinct length this results in distinct similarity scores for the edges (P, Q) and (Q, P); see Fig. 4. The resulting graph is denoted by G_d.

Clustering in a Threshold Graph

The next step in the clustering procedure is to proceed to a threshold graph. That is, all edges from G_d with a similarity score of less or equal than some fixed threshold τ are removed, resulting in the graph $G_d(\tau)$. All similarity values below this threshold are assumed to be produced by chance and not to be an indicator of true structural homology. Since using a single-link clustering algorithm would only solve part of the problem with multi-domain proteins, we made use of a standard concept in graph theory, the so called *strongly connected component*.

Definition 1 *In a directed graph G, a strongly connected component (SCC) is a maximal set C of nodes of G, such that for every pair of nodes p and q in C there is one directed path in G from p to q and one from q to p.*

The nodes of a directed graph are partitioned into SCCs, all of which can be computed efficiently with complexity $O(n + e)$ (Sedgewick, 1990) for a graph with n nodes and e edges. Some care has to be taken in an efficient implementation of the algorithm and the underlying graph data structures to allow interactive handling of graphs of close to 100,000 nodes and 20,000,000 edges.

As at least two distinct paths, one in each direction, have to exist, using SCCs as clusters is comparable to using a more conservative algorithm such as average linking. Note, that in Fig. 3 only proteins number two and three are in a SCC. Thus, using a SCC as a cluster does not make use of a large amount of information. Nevertheless, we chose to evaluate the performance of our
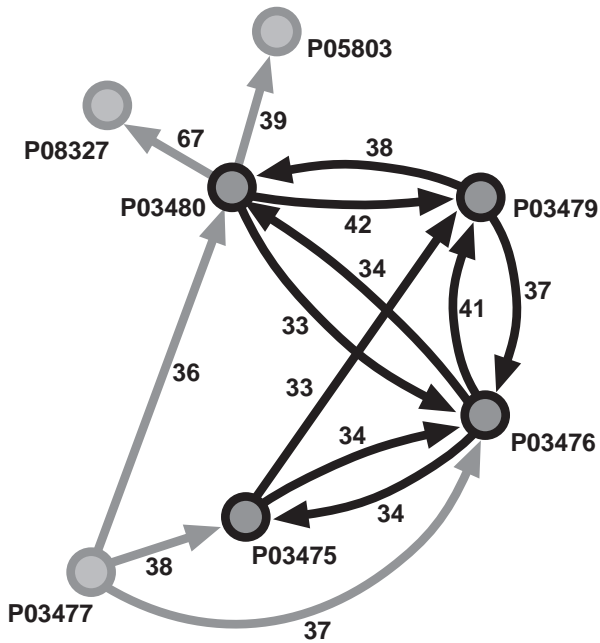
Figure 5: An example of a SCC in SwissProt. The grey nodes are not part of the SCC, but are clearly related. Note the sub-threshold scores (no edge present) between nodes P03480 and P03475. The threshold used in this example was 32 percent.

algorithm on the basis of the SCCs alone to establish the validity of our approach.

Implementation and Evaluation

The algorithms presented here have been implemented in the C++ package ProtClust. They have 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, as well as on a Compaq ES40 running Tru64 Unix V5.1, employing Compaq's cxx compiler, version 6.20.

We used our own implementation of the Smith-Waterman local alignment algorithm (Smith & Waterman, 1981) for computing sequence similarity. The choice of the alignment algorithm was motivated by the superior sensitivity (Brenner *et al.*, 1998) for low-scoring alignments as compared with other, faster algorithms such as BLAST (Altschul *et al.*, 1990). The substitution matrix, an integerized version of BLOSUM80 (Blocks Substitution Matrix) (Henikoff & Henikoff, 1993), was chosen based on the results of the investigations of one of the authors (Schneckener, 1998).

We chose the gap opening (gop) and gap extension penalties (gep) used in the alignment algorithm to be $gop = 90$ and $gep = 9$. This was decided after trying out a wide range of choices with extensive computations on the SCOP1 data set (see below). After computing a pairwise sequence alignment for all sequences against all, we performed single-link-clustering for thresholds from 5 to 95. By assigning a SCOP super-family to each cluster by maximal intersection we implied a measure of error by the number of sequences belonging to families distinct from the assigned. The gap penalties appeared to have a limited influence on the performance.

Data

We used SwissProt (Bairoch & Boeckman, 1992) version 39 as of June 2000, excluding all sequences with less than 40 amino acids (a.a.), resulting in a set of 86,494 protein sequences after removal of identical sequences. The total running time for the pair-wise Smith-Waterman alignment was on the order of 1,400 CPU-days. We used a simple tool (Schliep, 1998) to distribute the work over about 40 computers with a total of 55 CPUs available.

For the evaluation, the Structural Classification of Proteins (SCOP) data base (Murzin *et al.*, 1995) was employed. It provides a high quality hand-crafted partition of protein sequences at different levels. For our application, the relevant levels are *family* — sets of sequences with more than 30% sequence identity and possible functional identity — *super-family* — sets of sequences likely to have a common ancestor, low sequence identity, but structural and functional similarity — and *fold* — sets of sequences having structural similarity. All data sets defined in the following were based on version 1.37 of SCOP as obtained from `http://scop.mrc-lmb.cam.ac.uk/scop/pdbd.html`.

SCOP1

This data set of 2,692 sequences contains all non-identical sequences from SCOP having at least forty amino acids excluding sequences belonging to classes 8, 9 or 10. There are 65,464 pairs of homologue sequences; i.e., pairs where both sequences are in the same super-family and 3,556,622 pairs where the sequences are in distinct super-families. On the SCOP fold level the corresponding numbers are 76,660 and 3,545,426 pairs, respectively.

SCOP1+SP

To avoid having to uniquely identify sequences from the dataset SCOP1 in the SwissProt data we used, we simply pooled the two sets of sequences, yielding a total of 85,961 sequences.

SCOP2

For comparison with (Arvestad *et al.*, 2000) we also used a subset of 609 randomly chosen sequences from SCOP. It contains sequences shorter than forty amino acids, but no sequences from classes 8, 9 or 10.

Performance measure

A natural quality measure for detecting remote homologues is counting the number of truly homologous pairs of sequences and relating this to the number of errors made. A pair is taken to be truly homologous in this context, if both sequences are for example in the same SCOP super-family. We will call correctly identified true homologue pairs *true positives* (TP), not identified true homologues *false negatives* (FN), non-homologue pairs predicted to be homologue *false positives* (FP) and non-homologue pairs correctly identified *true negatives* (TN).

Sensitivity specifies the proportion of identified homologue pairs

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

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

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

Note, that $sens = spec = 1$ would be the most highly desired performance, since it implies that neither false positive nor false negative errors are made.

Discussion

It should be noted that, for a very large proportion of pairs of sequences from the same SCOP super-family, an alignment score in the twilight zone (cf. Fig. 1) well below reasonable thresholds for a pair-wise comparison is obtained. This supports the need for methods employing intermediate linking.

At first we evaluated the method on the SCOP1 data set. As Fig. 6 shows, the algorithm has maximum specificity over a very wide range of thresholds. For a conservative threshold choice of 32, we obtain a sensitivity of 55.6 percent. One factor contributing to the relatively small proportion of homologues found due to intermediate linking, about 8.0 percent at the chosen threshold, is that SCOP1 is a small data set, providing not enough sequences for intermediate linking. Also, it cannot be expected to recover all SCOP relations, since the superfamilies contain sequences which are only *likely* to have a common ancestor. As a matter of fact, we are not aware of any method which fully recovers even SCOP families.
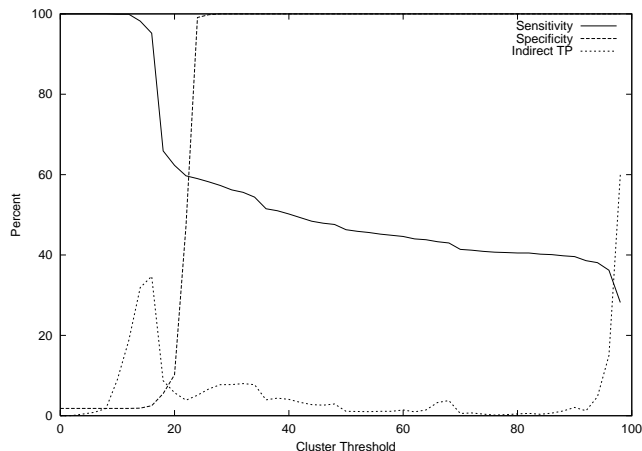

Figure 6: Sensitivity, specificity and the percentage of indirectly linked true positives vs. clustering threshold for the SCOP1 data set.

The rapid change in sensitivity and specificity at a threshold of 23 is due to lifting off the noise floor; a similar behavior is observed for pair-wise sequence comparisons.

The evaluation on the SCOP1+SP data set allows us to investigate two different aspects of our method. First, we should be able to benefit from the presence of more sequences in the clustering, as we are likely to find more indirect links. Secondly, there might be a large loss of specificity caused by unbounded transitivity. We computed the clustering for the whole data set, containing SCOP1 and SwissProt sequences. Therefore, sequences from SCOP1 might be joined by intermediate sequences from SwissProt not present in SCOP1 itself. The evaluation itself was only performed on pairs of SCOP1 sequences. Fig. 7 shows a very favorable performance of the method on this substantially larger data set. Again, we achieve a one-hundred percent specificity over a wide range of thresholds while additional sensitivity is gained. At a threshold choice of 32, a sensitivity of 57.9 percent and specificity of 99.8 percent is obtained. This is an absolute increase in sensitivity of 2.3, and a relative of 4.1 percent. The proportion of true positives found due to intermediate linking is increased by 3.6 percent to 11.6 percent. The noise floor is higher than in the previous case, as can be seen from the shift of the point of sudden decrease in specificity to the right in Fig. 7. However, a more than thirty-fold increase in the number of sequences only requires a minor adjustment of the threshold to again obtain near optimal specificity.

To investigate possible reasons for the somewhat surprising magnitude of the increase in sensitivity due to the additional sequences from SwissProt, we compared histograms of cluster sizes (cf. Table 1). The SCC-based clustering is of a very conservative nature, as can be ob-
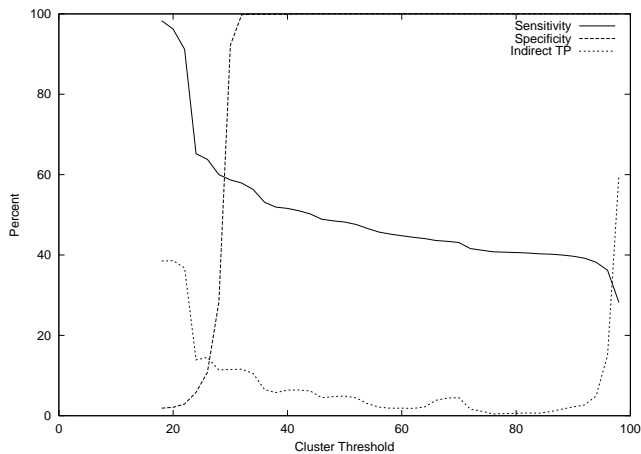
6

Figure 7: Sensitivity, specificity and the percentage of indirectly linked true positives vs. clustering threshold for the SCOP1+SP data set.

Figure 8: Sensitivity vs. specificity for the SCOP2 data set on the fold, super family and family level.

served in the number and sizes of clusters obtained. Few large and a large number of single element clusters are produced. Note, that larger clusters in general do not merge to one "super cluster" as the threshold decreases, indicating a good separation of clusters and thus the validity of our approach. This observation can be readily explained by the fact that terminal sequences, connected only with an uni-directional path, are excluded from the SCC (cf. Fig. 5).

To stress the importance of intermediate linking algorithms we compared our algorithm with (Arvestad *et al.*, 2000). While it employs only pair-wise sequence comparisons, their approach uses a more involved scoring method, optimized substitution matrices, and gap penalties, to achieve a substantial improvement over straight-forward pair-wise sequence comparisons. From Fig. 8, noting that the graphs are nearly parallel to the x-axis and that a high sensitivity is obtained at nearly maximum specificity, we can deduce the quality and the conservative nature of our method on the SCOP2 test set. The SCCs based on the Smith-Waterman score scaled by self-similarity thus realize a 24 percent better sensitivity at virtually equal specificity.

The handling of the problem with multi-domain proteins was investigated by manual inspection of pair-wise alignments along paths connecting false positives (not shown) present at low clustering thresholds. The analysis of the alignments and a comparison with scores not scaled by self-similarity showed the expected behaviour. Unfortunately, a comprehensive analysis was not feasible due to the limited amount of domain data available.

The clustering algorithm, queries on the clusters etc. are low both in theoretical complexity and practical running-time. Thus, they are even suitable for inter-
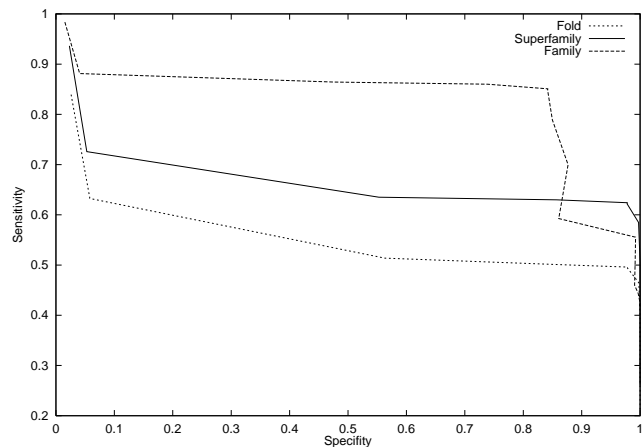
active work. As our experience shows, even the computation of the alignment matrix can be performed without need for specialized hardware or high-performance computers. Nevertheless, using substantially larger databases such as TrEMBL (Bairoch & Apweiler, 1999) might not be feasible without additional computing hardware.

We have demonstrated that the SCCs are conservative and highly specific clusters of protein sequences at the super-family level. They can be used as *seeds* for further advances in the clustering, ultimately obtaining an automated, biologically meaningful partition of protein space.

In the next stage, the information from nodes connected to but not part of a SCC will be taken into account. This should greatly improve the sensitivity of the method. As a further step, statistical significance, e.g., classical length dependent extremal value distribution fitting (Levitt & Gerstein, 1998), E-values, can be used to improve the quality of the threshold graph and thus the quality of the resulting clusters. Also, length dependent threshold functions (Arvestad *et al.*, 2000) should be investigated.

Acknowledgments

We would like to thank Oliver Leven, Sebastian Meller, Frank Nübel, Peter Pipenbacher, Barthel Steckemetz and Olav Zimmermann for helpful discussions and their support.

References

threshold	total	1	2–5	6–10	11–20	21–50	51–100	> 100
25	688	348	237	50	37	12	3	1
30	743	388	256	53	30	12	3	1
35	814	450	271	46	31	13	2	1
40	867	485	289	51	29	10	2	1
45	932	535	305	58	22	9	2	1
50	989	587	315	54	24	6	2	1
25	17,964	11,419	4,566	915	560	384	83	37
30	23,395	15,079	5,886	1,123	692	433	120	62
35	27,154	17,933	6,685	1,231	709	432	109	55
40	30,748	20,791	7,465	1,266	710	379	95	42
45	34,323	23,814	8,148	1,263	654	330	79	35
50	38,033	27,138	8,669	1,262	585	298	51	30
-	384	153	126	43	32	24	5	1

Table 1: Total number of SCC clusters, and SCC clusters of sizes 1, 2–5, 6–10 etc. for varying thresholds from 25 to 50 percent are displayed for the SCOP1 dataset (top) and the SCOP1+SP data set (bottom). The bottom-most row are the corresponding numbers for SCOP super-families.

Abagyan, R. A. & Batalov, S. (1997) Do aligned sequences share the same fold? *J Mol Biol,* **273** (1), 355–68.

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990) Basic local alignment search tool. *J Mol Biol,* **215** (3), 403–10.

Arvestad, L., Ivansson, L., Lagergren, J. & Elofsson, A. 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. & Apweiler, R. (1999) The swiss-prot protein sequence data bank and its supplement trembl in 1999. *Nucleic Acids Res,* **27** (1), 49–54.

Bairoch, A. & Boeckman, B. (1992) The swissprot protein sequence data bank. *Nucleic Acid Research,* **20**, 2019–2022.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A,* **95** (11), 6073–8.

Chothia, C. & Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J,* **5** (4), 823–6.

Gerstein, M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics,* **14** (8), 707–14.

Henikoff, S. & Henikoff, J. (1993) Performance evaluation of amino acid substitution matrices. *Proteins,* **17** (1), 49–61.

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

Krause, A. & Vingron, M. (1998) A set-theoretic approach to database searching and clustering. *Bioinformatics,* **14** (5), 430–8.

Levitt, M. & Gerstein, M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A,* **95** (11), 5913–20.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol,* **247** (4), 536–40.

Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997) Intermediate sequences increase the detection of homology between sequences. *J Mol Biol,* **273** (1), 349–54.

Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci,* **4** (6), 1145–60.

Pearson, W. R. (1997) Identifying distantly related protein sequences. *Comput Appl Biosci,* **13** (4), 325–32.

Roberts, F. S. (1984) *Applied Combinatorics.* Prentice-Hall, Inc.

Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng,* **12** (2), 85–94.

Salamov, A. A., Suwa, M., Orengo, C. A. & Swindells, M. B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues [in process citation]. *Protein Eng,* **12** (2), 95–100.

Sander, C. & Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins,* **9** (1), 56–68.

Schliep, A. PBQ – the python batch queue. Available from `http://www.zpr.uni-koeln.de/~schliep/PBQ`.

Schneckener, S. (1998). *Positionsgenaues Alignment von Proteinsequenzen.* PhD thesis, Universität zu Köln.

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

Smith, T. F. & Waterman, M. S. (1981) Identification of common molecular subsequences. *J Mol Biol,* **147** (1), 195–7.

Yona, G., Linial, N., Tishby, N. & Linial, M. (1998) A map of the protein space–an automatic hierarchical classification of all protein sequences. *Ismb,* **6**, 212–21.