Measurement of the Effectiveness of Transitive Sequence Comparison, through a Third "Intermediate" Sequence

Mark Gerstein

Department of Molecular Biophysics & Biochemistry
266 Whitney Avenue, Yale University
PO Box 208114, New Haven, CT 06520
(203) 432-6105, FAX (203) 432-5175
Mark.Gerstein@yale.edu

(Version w703 for *Bioinformatics*)

Abstract

Motivation: Transitive sequence matching expands the scope of sequence comparison by re-running the results of a given query against the databank as a new query. This sometimes results in the initial query sequence (Q) being related to a final match (M) indirectly, through a third, "intermediate" sequence (Q -> I -> M). This approach has often been suggested as providing greater sensitivity in sequence comparison; however, it has not yet been possible to precisely gauge its improvement.

Results: Here this improvement is comprehensively measured by seeing what fraction of the known structural relationships transitive sequence matching can uncover beyond that found by normal pairwise comparison (i.e. direct linkage). The structural relationships are taken from a well-characterized test set, the scop classification of protein structure. Specifically, 2055 known structural similarities (called "pairs") between distantly related proteins constitute the basic test set. To make the measurement of transitive matching properly, special data sets, called "baseline sets," are derived from this. They consist of pairs of sequences that have a clear structural relationship that *cannot* be found by normal sequence comparison (i.e. they cannot be directly linked). Specifically, using standard sequence comparison protocols (FASTA with an e-value cutoff of .001), it is found that the baseline set consists of 1742 pairs. A third intermediate sequence can link 86 of these indirectly (5%), where this third sequence is drawn from the entire, current universe of protein sequences. The number of false positives is minimal. Furthermore, when one considers only the relationships within the test set that correspond to a close structural alignment, the coverage increases considerably. In particular, 862 of the baseline set pairs fit to better than 2.6 A RMS, and transitive matching can find 62 of these (9%).

Availability: All the test data, including precise similarity values calculated from structural alignment, are available in tabular format over the web from http://bioinfo.mbb.yale.edu/align.

Introduction

Transitive sequence matching is an approach taken toward improving sequence comparison. It entails taking the matches found after running a sequence comparison and then re-running them as new queries against the databank. The resulting matches consist of many of the previous matches plus (hopefully) some new ones. These new matches are, in turn, related back to the initial query only indirectly through an intermediate sequence (see TIL in figure 1). This whole process can be repeated again, iteratively, with these new matches.

The idea of transitive matching has been previously suggested and implemented. In particular, it has been used to improve the sensitivity of single sequence comparison and to refine templates and Hidden Markov Models (Yi & Lander, 1994, 1996; Tatusov et al., 1994; Sonhammer et al., 1997; Eddy, 1996; Wolf et al., 1997; Gribskov et al., 1990; Pearson, 1997; Altschul et al., 1997; Park et al., 1997; Abagyan & Batalov, 1997). In the latter application, one forms a template from a small "seed" alignment, which is then used to find homologues. These are added to the template, and the process is repeated. While this technique has been demonstrated to be effective to varying degrees on specific protein families, it can lead to incorrect assignments.

The objective here is to assess the effectiveness of a simple form of transitive matching, in a comprehensive fashion. For this assessment, the structural classification of proteins (scop) (Murzin et al., 1995), which arranges *all* the known structures in the protein databank into a few hundred domain-level, fold families, provides a "gold-standard" reference dataset. Not all the structures in a given fold family are highly similar to each other in terms of sequence, so one can assess the usefulness of a given sequence comparison method by seeing how many of the structural similarities between only marginally similar sequences the method is able to detect.

Brenner et al. (1995, 1996, 1998) used this approach to assess the effectiveness of the popular FASTA and BLASTP programs and their probabilistic scoring schemes (i.e. the e-value) (Pearson & Lipman, 1988; Pearson, 1996; Altschul et al., 1990, 1994; Karlin & Altschul, 1993). They found that the FASTA e-value closely tracked the number of false positives, i.e. the error rate, and that at a conservative e-value cutoff of .001, the FASTA program could detect nearly all the relationships that a full Smith-Waterman

comparison would (Smith & Waterman, 1981). Specifically, they found that FASTA with a .001 threshold would find 16% more of the structural relationships in scop than would be found by standard sequence comparison with a 40% identity threshold. Here a similar approach is taken to assess the effectiveness of transitive matching.

Results

The Test Data: Sets of Structural Relationships

The analysis begins with the 8330 domains in the PDB indexed by the current version of scop. These are clustered into 905 representative sequences at a 40% identity level in the publicly distributed PDB40D dataset (see methods). About 400,000 pairs can be formed from these representative sequences (i.e. 408156 = 970 x 969 / 2).

Test Set 1. By definition, all these pairs have a distant ("twilight-zone") level of sequence similarity. However, according to the scop classification (Murzin et al., 1995), 2055 (~0.5%) of them have a significant structural relationship, in being joined to one of 171 structural superfamilies (see methods). These 2055 form the first set of "scop pairs." They are not distributed equally amongst the scop superfamilies, with one superfamily (the Rossman fold) containing 231 pairs and 70 others with just a single pair (Table 2). Furthermore, not all the structural relationships in these pairs are of equal weight, so it is worthwhile to consider two further "selections" based on somewhat closer structural relationships.

Test Set 2. Because determining the structural similarity of short sequences is particularly problematic, one can exclude sequences of less than 60 amino acids. This gives 783 representative sequences, 305371 possible pairs and 1801 structural relationships, which constitute a selection of 2055 original scop pairs. Gerstein & Levitt (1996, 1998) constructed structural alignments for all the preceding 1801 structural similarities of full-length sequences. These alignments allow one to determine the precise degree of structural similarity by calculating an RMS value from fitting the aligned atoms. There are 862 pairs that align with a scaled RMS of less than 2.6 Å (see methods), and these form a second test set of scop pairs. They are (roughly) the more structurally similar half of the scop pairs.

Direct Linkage: The Baseline for Comparison

To measure the effectiveness of transitive matching, one can look at how many of the scop pairs in each of the three sets indirect linkage can find, relative to the number of false positives. However, before doing this one has to determine how many pairs normal sequence comparison can find, in order to establish a baseline upon which transitive matching can improve. Here pairs found by normal sequence comparison are called direct linkages since they involve no intermediate sequences (TDL, figure 1).

This is essentially what Brenner et al. (1995, 1996, 1998) did in asking how many of these real structural relationships could be found using the FASTA and BLASTP programs with their probabilistic scoring schemes. Here these results are reproduced for the three specific sets of scop pairs discussed above. As Brenner et al. found, at a practical e-value threshold of .001, FASTA can find "direct linkages" for about 15% of the relationships in the first set of the scop pairs with only a few false positives.

Notice that when one considers just the 862 pairs with a close structural alignment (test set 2), it is possible to find a higher fraction of the pairs (25%). This last result is reasonable, given the established relationship between divergence in structure and sequence (Chothia & Lesk, 1986, 1987; Chothia & Gerstein, 1997). That is, the pairs with greater structural similarity are expected to have more sequence similarity.

As shown in tables 2 and 3, the fraction of pairs found varies considerably amongst the scop superfamilies, with a larger fraction of pairs found in the smaller superfamilies.

Work on direct linkage provides a necessary background against which to examine indirect linkage. Here the idea is to find out how many *additional* structurally related pairs can be found by considering a third, intermediate sequence linked to both. These indirect linkages, both true and false, are illustrated schematically in figure 1. To find them, it was necessary to construct more sets of scop pairs, the same as those described previously but now with all the pairs found by direct matching removed. These are called baseline sets:

Baseline Set 1-3. This consists of 1742 pairs taken from the 2055 pairs in test set 1 with direct matches removed. It involves 697 sequences in total. It is based on a FASTA e-value cutoff of 10e-3. If this cutoff is changed, obviously the number of pairs will

change, so one also has baseline sets 1-4 and 1-5 for cutoffs of 10e-4 and 10e-5, and so forth. (See Table 1.)

Baseline Set 2-3. This consists of 643 pairs taken from the 862 closely aligned pairs in test set 2 with direct matches removed. It involves 491 sequences in total. It is also derived with a cutoff of 10e-3, and baseline sets 2-4 and 2-5 can be defined in similar fashion.

Indirect Linkage: The Improvement Over the Baseline

By definition, each of the sequences in the baseline sets has no sequence similarity to any other sequence within the same set. Consequently, it is now readily possible to gauge the improvement provided by transitive matching: any new pairs found constitute the improvement. A transitive match could in principle be through the sequences within the baseline sets -- i.e. if pair AB and pair BC exist in set 1-3, but not pair AC, there would be an indirect link between A and C. As shown in Table 1, this occurs, but not that frequently. For instance, for baseline set 1-3, one can find 23 of the 1742 pairs.

One can find more transitive matches by considering the entire population of protein sequences as candidate “intermediate sequences.” Specifically, one can run the sequences in each of the baseline sets against the OWL composite databank (which contains all currently known protein sequences) and determine whether any of the homologues found in OWL linked a scop pair in the baseline sets. Used in this way, the sequences in the baseline sets are better thought of as cluster representatives for whole families than individual sequences. Each indirect link made between them is effectively between all the members of two distant families.

The results are summarized in table 1. For an e-value threshold of .001 (baseline set 1-3), one can find 86 of the 1742 baseline pairs through indirect linkage (5%), with 13 false positives. This means that using both direct and indirect linkage, sequence comparison with FASTA can find about a fifth of the scop pairs (399 of 2055 in test set 1) with 16 total false positives, about one for every 25 true positives.

On the 862 closely aligned scop pairs (test set 2), the coverage improves significantly. In particular, transitive matching can find 74 of the 643 pairs in baseline set 2-3 (12%).

As shown in table 2, the fraction of extra pairs found by transitive matching varies somewhat among the 171 scop superfamilies, with the larger families having a greater degree of improvement relative to the smaller ones. This is perhaps because direct matching was more successful with the smaller superfamilies and because the larger superfamilies are potentially associated with a larger and more diversified collection of intermediate sequences. For instance, indirect plus direct linkage can find 30% of the 120 pairs in the “FAD/NAD(P)-binding domain” superfamily (representative identifier d2tpra2), whereas direct matching can only find 10% (Table 3). Likewise, for the globin superfamily (d3sdha_), the comparable statistics are 63% of 91 pairs found, improving on 40%. In contrast, for the 70 scop superfamilies containing only a single pair, indirect plus direct linkage finds 47% of pairs, only a small improvement over the 41% found by direct matching alone.

Qualifications

The specific "improvement" values quoted here for the effect of transitive sequence matching are intended to be representative of the performance of the method on a comprehensive data set using reasonable parameters. They are, nevertheless, contingent upon the selection of proteins in the test set (the scop classification), the particular comparison baselines established (i.e. an e-value cutoff of .001 for the direct linkage baseline), and the precise criteria for overlap (as discussed in the methods section). These parameters have been selected in a reasonable fashion to exclude highly similar sequences and give a sense of how indirect sequence matching performs near the margin, in the "twilight zone." The scop data set (Murzin et al., 1995), in particular, is a popular and well-documented set of similarities. It has been validated by both automatic and manual methods (Gerstein & Levitt, 1997) and should give the most comprehensive possible indication of how indirect matching performs on the whole range of known similarities, rather than just on specific families.

Moreover, while the exact improvement values quoted here may change somewhat with different choices for test data, baselines, and overlap criteria, with any reasonable choices, transitive matching will be able to find additional real pairs without generating many false positives. That is, while the absolute values may change the relative improvement will remain. This is shown to some degree in table 1, where the

improvement statistics for a variety of baselines are collected, e.g. .0001 and .00001. (One could develop this table further to build up a complete analysis of coverage vs error rate.) Furthermore, the entire test set and related data files (including all the precise similarity values from structural alignment) are available over the web, thus enabling the analysis to be easily repeated with any parameters of one's choice.

Conclusion

The results reported here show that transitive sequence matching is an effective technique for improving the sensitivity of standard sequence comparison methods in searching for structural similarities. Specifically, one is able to find considerably more of the known structural similarities in a "gold-standard" set of test data (scop) by combining indirect linkage via an intermediate sequence with direct linkage than by direct matching alone. Moreover, there are few false positives. One can intuitively rationalize the success of transitive sequence matching as this approach makes use of the (presumably) more diversified outliers of a cluster in a search, instead of searching with the centroid (as is the case, for instance, for profiles).[*]

The measurement of the effectiveness of transitive sequence matching was done here for the FASTA program, but a similar analysis could easily be done for the other popular sequence comparison approaches, such as profiles and HMMs (Bowie et al., 1991; Johnson et al., 1993; Eddy et al., 1994; Krogh et al., 1994). In fact, such analyses have recently been performed successfully by other groups (Chothia & Park, pers. communication). It is expected that careful measurement of the effectiveness of sequence comparison methods for detecting structural similarities will allow these methods to be used as a baseline for assessing more elaborate fold recognition methods such as threading (Jones et al., 1992; Jones & Thornton, 1996; Bryant & Lawrence, 1993).

Details of the Methods

Data

Sequences with known structure were taken from the Protein Databank (Bernstein

[*] In this comparison one visualizes each sequence in a multiple alignment as occupying an integral grid point in a high dimensional sequence space which has an axis for each position in the alignment (Maynard Smith, 1970; Vingron & Sibbald, 1994). As discussed by Vingron & Sibbald, the profile occupies a potentially off grid position midway between all the sequences.

et al., 1977). Fold definitions were taken from scop, version 1.32 (May 1996) (Murzin et al., 1996; Brenner et al., 1996; Hubbard et al., 1997). Only the superfamily, as opposed to fold pairs, were used as these have a much clearer structural relationship. It is the intention of the creators of scop that the superfamily pairs represent an evolutionary relationship between proteins with no appreciable sequence similarity -- i.e. link proteins that are true homologues (Murzin et al., 1995; Hubbard, 1997). However, this is necessarily speculative, and all one can know for certain is that these pairs have a close structural relationship. Furthermore, the scop pairs have been extensively checked by both manual and automatic methods (Gerstein & Levitt, 1998) and are believed not to contain any false positives.

Based on the scop pairs, Brenner et al. (1995, 1998) clustered the PDB into 905 representative sequences at 40% identity (domains split between different chains are omitted from this count), making a list denoted pdb40d, which is distributed through the scop website (http://scop.mrc-lmb.cam.ac.uk/scop). The clustering employed a single-linkage approach similar to that in Hobohm et al. (1992, 1994), i.e. "select until done." For the indirect sequence matching, pdb40d was compared against the 142737 total sequences in the OWL composite databank (version 27.1) (Bleasby et al., 1994). Low complexity sequences were filtered out of OWL using the SEG program (Wooton & Federhen, 1993).

The overall analysis was greatly expedited by using a simple relational database implemented using DBM and perl5 (Wall et al., 1996). A number of detailed tables relevant to this paper will be made available over the Internet at http://bioinfo.mbb.yale.edu/align .

Sequence Comparison

All sequence matching was done with the FASTA program (version 2.0) (Lipman & Pearson, 1985; Pearson, 1996, 1998; Pearson & Lipman, 1988; Pearson et al., 1997) with a k-tup value of 1. This program was chosen for a number of reasons:

(i) FASTA is commonly used in the comparison of sequences corresponding to structures -- for instance, it was used for the original definition of superfolds by Orengo et al. (1994) and is the sequence comparison method used by the PDB browser (Stampf et al., 1995). Consequently, FASTA forms an established standard on which to base this work.

(ii) FASTA was used in the work of Brenner et al. (1998) which forms a necessary background for this work. Brenner et al. showed, furthermore, that FASTA performs better than BLAST and essentially the same as Smith-Waterman for the detection of distant similarities between structurally similar proteins, obviating the need to consider these other approaches here.

(iii) Assessing the improvement in indirect linkage is much more straightforward for single-sequence comparison methods, such as FASTA, than for multiple-sequence methods, such as HMMs or PSI BLAST (Altschul et al., 1997; Krogh et al., 1994), where the performance of the method varies depending on the number of members in the family.

Structure Comparison

Structure matching was done by the iterative dynamic programming method from Gerstein & Levitt (1996, 1998). This method aligns protein sequences on the basis of direct comparison of the corresponding three-dimensional structures. Two numbers characterize the alignment: the number of residues aligned (N) and the RMS deviation in Cα positions after these atoms are fit onto each other (RMS). Since an alignment with a higher RMS value can be more significant than one with a lower RMS if there are more residues included in the first alignment, Gerstein & Levitt (1998) define a scaled RMS: RMS' = 225 RMS / (N + 135). For an approximately average match of 90 residues, the scaled RMS is nearly the same as RMS (both quantities agree to within 10% for N between 70 and 110 residues). The distribution of scaled RMS values has a median value of 2.65 Å (with a mean of 2.68 Å and a standard deviation of 0.87 Å), so 2.6 Å marks the approximate halfway point in the range of values and a reasonable division point. Levitt & Gerstein (1998), furthermore, show that this scaled RMS threshold corresponds approximately to a structural similarity P-value of .01.

Overlap on Intermediate Sequence

In the transitive matching procedure, two sequences corresponding to structures (denoted Q for the query and M for the match) are linked through an intermediate sequence I. One has to take care that the region of match of Q on I overlaps with that of I on M. The criteria for overlap used here was quite conservative: the overlap region of Q

on I must share at least 60 residues with that of I on M. Other less stringent criteria were tried. These tend to increase the number of matches, both true and false, but not to affect the results greatly, as long as they were reasonable.

As a practical matter one can deal with the overlap when doing a search as follows. One runs the query against the whole databank, finding a number of direct matches I_i. Then the precise matching regions of each I_i is "cut out" and this is re-run against the databank again, producing matches $M_{i,j}$, which are then indirectly linked back to the original query Q. While simple and elegant, this procedure has the effect of changing the scores linking I_i and $M_{i,j}$ relative to those found in an all-vs-all of the databank since the length of the intermediate sequence I_i is different when it is matched by Q or used as the query to find $M_{i,j}$.

Acknowledgements

The authors of scop (Alexey G. Murzin, Bartlett G. Ailey, Steven E. Brenner, Tim J.P. Hubbard, and Cyrus Chothia) are acknowledged for supplying the pdb40d dataset (via website) and suggestions on the manuscript. Further thanks to C Chothia and J Park for communicating similar work to that done here prior to publication; S Brenner for providing a copy of his thesis on CD-ROM; H Hegyi and R Rambo for carefully reading the manuscript; and M Levitt for suggestions on overlap criteria. The NSF is thanked for support (Grant DBI-9723182).

References

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

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. [Review]. *Nature Genetics* **6**, 119-29.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

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

Bleasby, A. J. & Wootton, J. C. (1990). Construction of validated, non-redundant composite protein sequence databases. *Protein Eng* **3**, 153-9.

Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL -- a non-redundant composite protein sequence database. *Nuc. Acid. Res.* **22**, 3574-3577.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science (Washington D C)* **253**, 164-170.

Brenner, S. E. (1996). *Molecular Propinquity: Evolutionary and Structural Relationships of Proteins.* PhD Thesis, Cambridge University.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications [In Process Citation]. *Curr Opin Struct Biol* **7**, 369-76.

Brenner, S., Chothia, C. & Hubbard, T. (1998). Assessing Sequence Comparison Methods. *Proc. Natl. Acad. Sci. USA* (in press).

Brenner, S., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene Duplication in H. Influenzae. *Nature* **378**, 140.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct Funct Genet* **16**, 92-112.

Chothia, C. & Gerstein, M. (1997). How far can sequences diverge? *Nature* **385**, 579-581.

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

Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **LII**, 399-405.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struc. Biol.* **6**, 361-365.

Eddy, S. R., Mitchison, G. & Durbin, R. (1994). Maximum Discrimination Hidden Markov Models of Sequence Consensus. *J. Comp. Bio.* **9**, 9-23.

Gerstein (1997). A Structural Census of Genomes: Comparing Eukaryotic, Bacterial and Archaeal Genomes in terms of Protein Structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. & Levitt, M. (1996). Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures. In *Proc. Fourth Int. Conf. on Intell. Sys. Mol. Biol.,* pp. 59-67, AAAI Press, Menlo Park, CA.

Gerstein, M. & Levitt, M. (1998). Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the Scop Classification of Proteins. *Protein Science* **7**, 445-456.

Gribskov, M., Lüthy, R. & Eisenberg, D. (1990). Profile Analysis. *Meth. Enz.* **183**, 146-159.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Science* **3**, 522.

Hobohm, W., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Prot. Sci.* **1**, 409-417.

Hubbard, T. J. (1997). New horizons in sequence analysis. *Curr Opin Struct Biol* **7**, 190-3.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Res* **25**, 236-9.

Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993). Alignment and searching for common protein folds using a databank of structural templates. *J. Mol. Biol.* **231**, 735-752.

Jones, D. T. & Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struc. Biol.* **6**, 210-216.

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

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 5873-7.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modelling. *J. Mol. Biol.* **235**, 1501-1531.

Levitt, M. & Gerstein, M. (1998). A Unified Statistical Framework for Sequence Comparison and Structure Comparison. *Proceedings of the National Academy of Sciences USA* **95**, 5913-5920.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441.

Maynard Smith, J. (1970). Natural Selection and the concept of a protein space. *Nature* **225**, 563- 564.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: A Structural Classification of Proteins for the Investigation of Sequences and Structures. *J. Mol. Biol.* **247**, 536-540.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631-634.

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

Pearson, W. R. & Lipman, D. J. (1988). Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448.

Pearson, W. R. (1996). Effective Protein Sequence Comparison. *Meth. Enz.* **266**, 227-259.

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

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71-84.

Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. (1997). Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24-36.

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

Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a Comprehensive Database of Protein Domain Families Based on Seed Alignments. *Proteins* (in press).

Stampf, D. R., Felder, C. E. & Sussman, J. L. (1995). PDBbrowse--a graphics interface to the Brookhaven Protein Data Bank. *Nature* **374**, 572-4.

Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A* **91**, 12091-5.

Vingron, M. & Sibbald, P. R. (1993). Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. USA* **90**, 8777-8781.

Wall, L., Christiansen, D. & Schwartz, R. (1996). *Programming Perl.* O'Reilly and Associates, Sebastapol, CA.

Wolf, E., Kim, P. S. & Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils [In Process Citation]. *Protein Sci* **6**, 1179-89.

Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry* **17**, 149-163.

Yi, T. & Lander, E. S. (1996). Iterative Template Refinement: Protein-Fold Prediction Using Iterative Search and Hybrid Sequence/Structure Templates. *Meth. Enz.* **266**, 322-339.

Yi, T. M. & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (itr). *Protein Science* **3**, 1315-1328.

Table 1 Overall Statistics for Sequence Matching

FASTA e-value cutoff	Type of Pair	Pairs in Total: Test Sets	Linked Directly	Pairs Remaining: Baseline Sets	Linked Indirectly within dataset	Linked Indirectly via OWL sequence
1.0E-05	TPs	2055	220 11%	1835	19 1.0%	67 3.7%
	low-RMS TPs	862	162 19%	700	15 2.1%	62 8.9%
	FPs		0		0	12
1.0E-04	TPs	2055	271 13%	1784	28 1.6%	73 4.1%
	low-RMS TPs	862	198 23%	664	17 2.6%	64 9.6%
	FPs		1		0	13
1.0E-03	TPs	2055	313 15%	1742	23 1.3%	86 4.9%
	low-RMS TPs	862	219 25%	643	18 2.8%	74 12%
	FPs		3		1	13

The table shows how many of the scop pairs can be found by direct linkage (i.e. normal sequence comparison) and indirect linkage (i.e. transitive matching through an intermediate sequence). The rows show the number of true and false positive linkages (TPs and FPs) for various FASTA e-value thresholds. These linkages are computed for two test sets: test set 1, which has 2055 scop pairs, and test set 2, which has 862 scop pairs, corresponding to more closely aligned structures that have a structural alignment with atoms fitting to better than 2.6 Å RMS. The true positives for the latter test set are denoted by "low-RMS TPs." There were no false positives for test set 2 data (so there are no "low-RMS FPs" rows). The first column ("test sets") shows the total number of pairs that one starts with. The next column ("linked directly") shows the number of these pairs that can be found by direct linkage. These are subtracted away to give the baseline sets shown in the third column ("baseline sets"). The final two columns give the number of pairs from the baseline sets than can be found by indirect linkage. The first of these ("indirectly within dataset") lists the transitive matches that can be found purely within a given baseline set. The next column ("indirectly via OWL") lists the larger number of transitive matches that can be found if one allows any sequence in the large OWL database to function as an intermediate sequence. Note that it is possible to have a pair linked directly but not indirectly if no suitable intermediate sequence exists. Also, a number of the false positive linkages were between scop class 8 sequences ("peptides"). (In particular, the pairs d1tiv__-d1tvs__ and d1bba__-d1ppt__.) These were excluded from the statistics (but they are still listed, for completeness, in the web presentation).

Table 2 Matching Statistics for the Various Scop Superfamilies, Divided by Family Size

Pairs per Super-family	Num. of Super-families	Total Num. Scop Pairs	Num. Linked Directly	Frac. Linked Directly	Num. Indirect or Dir. Links	Frac. Linked Ind. or Dir.
231	1	231	21	9%	26	11%
171	1	171	5	3%	9	5%
153	1	153	3	2%	3	2%
120	2	240	22	9%	47	20%
91	1	91	36	40%	57	63%
78	1	78	4	5%	4	5%
55	3	165	12	7%	12	7%
36	3	108	22	20%	28	26%
28	8	224	21	9%	24	11%
21	6	126	45	36%	53	42%
15	4	60	7	12%	8	13%
10	11	110	15	14%	16	15%
6	17	102	29	28%	33	32%
3	42	126	42	33%	46	37%
1	70	70	29	41%	33	47%
Total	171	2055	313	15%	399	19%

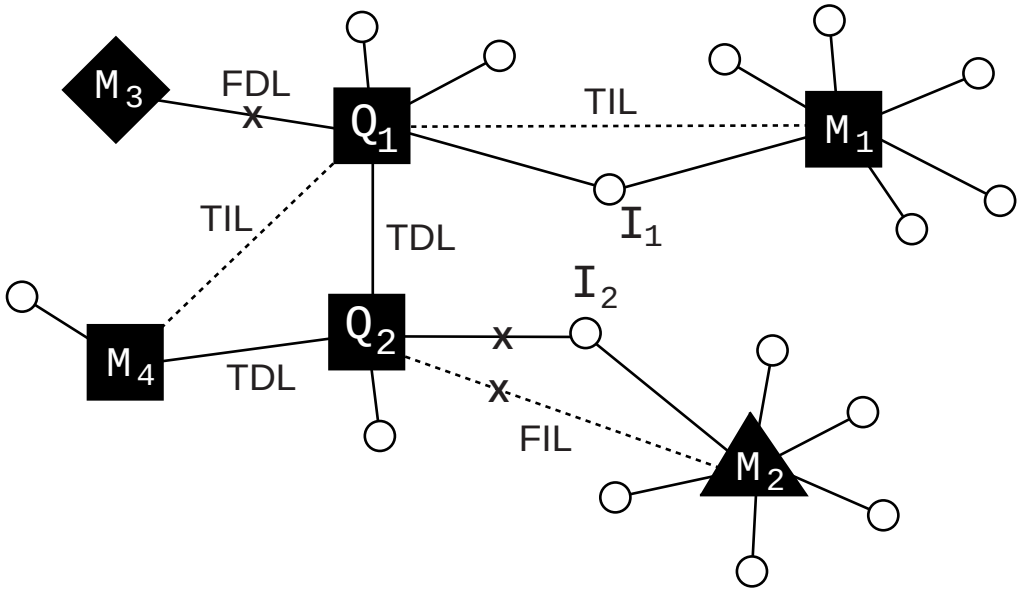
This table shows the statistics for direct and indirect linkage for the 171 scop superfamilies. The statistics are broken down by the size of the superfamily. Details on each column follow: (1) The number of pairs P in a superfamily, i.e. its size (using PDB40D in scop 1.32 as described in the methods). (2) The number N of superfamilies of this size in scop. (3) The total number of pairs then follows by multiplication, T = NP; (4) The number of pairs D that can be directly linked by sequence comparison with FASTA and an e-value cutoff of .001. (5) The fraction of the total number of pairs that the number of directly linked pairs comprises, F = D/T. (6) The number of pairs I that can be linked by either indirect or direct linkage. (7) The fraction that I is of the total (I/T).

Table 3 Detailed Matching Statistics for the Largest Scop Superfamilies

Super-family ID	Num. Scop Pairs	Num. Direct Links	Frac. Linked Directly	Num. Indirect or Dir. Links	Frac. Linked Ind. or Dir.
d2pgd_2	231	21	9%	26	11%
d3dpva_	171	5	3%	9	5%
d3cd4_2	153	3	2%	3	2%
d2tpra2	120	12	10%	36	30%
d2ebn__	120	10	8%	11	9%
d3sdha_	91	36	40%	57	63%
d5p21__	78	4	5%	4	5%
d1yrnb_	55	8	15%	8	15%
d1pmy__	55	4	7%	4	7%
d3inkc_	55	0	0%	0	0%
d5znf__	36	10	28%	10	28%
d4icb__	36	8	22%	14	39%
d2trxa_	36	4	11%	4	11%
d2tgf__	28	13	46%	13	46%
d1r69__	28	3	11%	3	11%
d5cytr_	28	2	7%	3	11%
d3hhrb2	28	1	4%	1	4%
d3tgl__	28	1	4%	2	7%
d2olba_	28	1	4%	1	4%
d1tssa1	28	0	0%	0	0%
d2yhx_2	28	0	0%	1	4%

This table shows the statistics for direct and indirect linkage for the 21 largest scop superfamilies, those with at least 28 pairs. Details on each column follow, many of them being very similar to those in table 2: (1) The first column gives the identifier for the superfamily. Here this is a scop identifier for a representative domain in this superfamily. Scop identifiers have the following syntax: d1pdbcN, where “1pdb” is a PDB id, “c” is a chain identifier, and “N” describes if this is the first, second, or only domain in the chain. Thus, d1ggta1 is the first domain in the A chain of 1GGT. (2) The number of pairs P in the superfamily, i.e. its size (using PDB40D in scop 1.32 as described in the methods). (3) The number of pairs D that can be directly linked by sequence comparison with FASTA and an e-value cutoff of .001. (4) The fraction of the total number of pairs that the number of directly linked pairs comprises, F = D/P. (5) The number of pairs I that can be linked by either indirect or direct linkage. (6) The fraction that I is of the total (I/P).

Figure 1, Schematic Illustrating Transitive Matching via Indirect Linkage



This schematic illustrates direct and indirect linkage and how either linkage can result in a true or false positive. The black shapes (square, triangle, diamond) indicate sequences with a known structure (i.e. PDB sequences). For the purposes of the discussion here one imagines that each of these sequences is a representative sequence drawn from scop-pair test sets and so it stands for a whole sequence family. The four black squares indicate sequences that share a common fold, the “square fold,” while the folds corresponding to the black triangle and diamond are supposed to be different. The line marked TDL (“true direct linkage”) indicates a “true” sequence similarity linking two of these sequences (Q_1 and Q_2). In contrast, the line marked FDL indicates a "false" direct linkage between Q_1 and M_3 i.e. a linkage that is not a scop pair. False linkages, indicated by “x” in the figure, give rise to the false positives in table 1. It is possible for two direct linkages to join two PDB sequences that are not linked directly. This is “indirect linkage” and is indicated by the dotted line between Q_1 and M_4 in the figure. Here Q_2 is functioning as the “intermediate sequence.” One can expand the possible intermediate sequences by considering sequences that do not correspond directly to PDB structures -- i.e. sequences from OWL homologous to PDB sequences (the OWL sequences). These are indicated by the small white circles linked to the black shapes. The OWL sequences can function as intermediate sequences (denoted I) linking two PDB sequences (the query Q and the match M) via indirect linkages that are either true or false (i.e. TIL or FIL, “true or false, indirect linkage”).