Article No. mb982221 *J. Mol. Biol.* (1998) **284**, 1201–1210

JMB

Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods

Jong Park[1], Kevin Karplus[2], Christian Barrett[3], Richard Hughey[2] David Haussler[3,4], Tim Hubbard[5] and Cyrus Chothia[1]

[1]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[2]*Departments of Computer Engineering and* [3]*Computer Science, University of California, Santa Cruz CA 95064, USA*

[4]*Isaac Newton Institute for Mathematical Sciences University of Cambridge Clarkson Road, Cambridge CB3 0EH, UK*

[5]*Sanger Centre, Wellcome Trust Genome Campus Hinxton, Cambridge CB10 1SA, UK*

The sequences of related proteins can diverge beyond the point where their relationship can be recognised by pairwise sequence comparisons. In attempts to overcome this limitation, methods have been developed that use as a query, not a single sequence, but sets of related sequences or a representation of the characteristics shared by related sequences. Here we describe an assessment of three of these methods: the SAM-T98 implementation of a hidden Markov model procedure; PSI-BLAST; and the intermediate sequence search (ISS) procedure. We determined the extent to which these procedures can detect evolutionary relationships between the members of the sequence database PDBD40-J. This database, derived from the structural classification of proteins (SCOP), contains the sequences of proteins of known structure whose sequence identities with each other are 40% or less. The evolutionary relationships that exist between those that have low sequence identities were found by the examination of their structural details and, in many cases, their functional features. For nine false positive predictions out of a possible 432,680, i.e. at a false positive rate of about 1/50,000, SAM-T98 found 35% of the true homologous relationships in PDBD40-J, whilst PSI-BLAST found 30% and ISS found 25%. Overall, this is about twice the number of PDBD40-J relations that can be detected by the pairwise comparison procedures FASTA (17%) and GAP-BLAST (15%). For distantly related sequences in PDBD40-J, those pairs whose sequence identity is less than 30%, SAM-T98 and PSI-BLAST detect three times the number of relationships found by the pairwise methods.

Keywords: protein homology; intermediate sequence search; hidden Markov models; SAM-T98; PSI-BLAST, SCOP

Introduction

It is often important to determine whether new protein sequences are related to other sequences that have a known function or structure. If relationships are demonstrated, we obtain information on the probable function, structure and evolution of the new sequences. Computer programs that make pairwise comparisons of sequences, such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson, 1988), are the methods most commonly used to search for such relationships. These programs match new sequences (queries) against all the sequences in a database (targets) and report each query-target pair that represents a statistically significant match. The experience of those who first used these procedures showed that as the sequence identities of related proteins go below 30% identity, the chance of their relationship being detected by pairwise procedures becomes increasingly small. In a recent quantitative study, we found that, of the evolutionary relationships known on the basis of structure, sequence and

Present address: J. Park, Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA.

Abbreviations used: HMM, hidden Markov model; ISS, intermediate sequence search; SCOP, structural classifications of proteins; PSI, position-specific-iterated; PDBD40-J, database of sequences that have pairwise identities of 40% or less; NRP, non-redundant protein; CVE, coverage *versus* error; RFP, rate of false positives.

0022–2836/98/491201–10 $30.00/0

function in proteins with 20–30% sequence identity, only half can be detected by pairwise sequence comparisons (Brenner *et al.*, 1998) . For related proteins with less than 20% identity the proportion detected is much smaller. To try and overcome these limitations, search procedures based on the shared characteristics of sets of related sequences have been developed. Examples of such procedures are templates (Taylor, 1986; Bashford *et al.*, 1987; Tatusov *et al.*, 1994; Yi & Lander, 1994), profiles (Gribskov *et al.*, 1987; Luthy *et al.*, 1994; Thompson *et al.*, 1994), hidden Markov models (HMMs; Krogh *et al.*, 1994; Baldi *et al.*, 1994; Eddy, 1995, 1996; Eddy *et al.*, 1995; Hughey & Krogh, 1996), the position-specific-iterated (PSI) version of BLAST (Altschul *et al.*, 1997), PROBE (Neuwald *et al.*, 1997), and intermediate sequence searches (ISS; Park *et al.*, 1997).

Here we describe an assessment of three of these procedures: HMMs, PSI-BLAST, and ISS. To make these assessments, we use a database of sequences that have known distant evolutionary relationships. For the three procedures, we assess (i) their ability to detect the evolutionary relationships that are presently known to occur between the sequences in the database; (ii) how they perform relative to each other and to pairwise comparison methods; and (iii) whether or not they find the same set of evolutionary relationships. We also discuss the relation between the scoring schemes of the HMM and PSI-BLAST procedures and their observed errors.

A Database with Sequences of Low Homology and Known Evolutionary Relationships

The structural classification of proteins (SCOP) database contains a description of the evolutionary and structural relations of those proteins whose atomic structure has been determined (Murzin *et al.*, 1995). The current version is available on the World Wide Web at http://scop.mrc-lmb.cam.ac.uk/scop/. The unit of classification in the database is the protein domain. Small proteins, and most of those of medium size, have a single domain and are, therefore, treated as a whole. The domains that form large proteins are classified individually. Domains are clustered together into families if they have close evolutionary relationships. A superfamily brings together those families that have low sequence identities with each other but whose structural details and, in many cases, functional features suggest that a common evolutionary origin is very probable, e.g. the variable and constant domains of immunoglobulins. The fold classification brings together superfamilies that have the same secondary structures in the same arrangement. For most superfamilies that share a common fold, there is good evidence that they do not have evolutionary relationships. In a few cases, the situation is less clear in that current evidence weakly supports the existence of evolutionary relationships. In these cases, the super families are kept separate until the subsequent discovery of intermediate structures provides stronger support for their merger.

To test the sequence matching procedures we measured the extent to which they could find the evolutionary relationships described by the superfamilies in the SCOP database. As it is straightforward to find the relationships between those proteins that have sequence identities of 40% or more, we use here a database of sequences that have pairwise identities of 40% or less, which we call PDBD40-J. This database contains 935 sequences (it is available from http://scop.mrc-lmb.cam.ac.uk/scop/pdbd.html).

For the purposes of this experiment, a pair of these sequences is defined to be a homologous pair if they are in the same SCOP superfamily. There are 2096 homologous pairs, which is 0.48% of the possible $935 \times 934/2 = 436{,}645$ different pairs that can be formed from sequences in the database. Our test of the sequence comparison methods is to determine how well they can distinguish the 2096 homologous pairs from the 432,680 non-homologous pairs.

The accuracy of our results will, of course, depend upon the accuracy of the SCOP assignments. Errors will occur (i) when sequences are listed in SCOP as being related when they are not, and (ii) when true evolutionary relationships amongst the sequences are not listed. We expect that errors of the first kind are rare in SCOP, and in any event, they will affect all sequence comparison methods equally and so will not influence their relative performance.

In practice, the second kind of error, a failure to include true evolutionary relationships, would mean that the domains in SCOP are listed as having the same fold but different superfamilies. We noted above that the classification in SCOP tends to be somewhat conservative in that it requires what it's authors regards as good evidence, which may not be available at present, to put structures into the same superfamilies rather than just the same fold category. For this reason, pairs of sequences that have the same fold, but are not in the same superfamily, are defined as being of uncertain relationship. We label these pairs as uncertain to avoid counting it as an error when a sequence comparison procedure predicts a relationship between two sequences that share the same fold but not the same superfamily, while at the same time also not counting it as an error to fail to detect a relationship between two sequences that have a common fold, but are not in the same superfamily. Non-homologous pairs are formed by any two sequences that have different folds. For the PDBD40-J sequences there are 432,680 different non-homologous pairs and 1896 pairs of uncertain relationship. In fact, we detect only a tiny number of uncertain matches in this work, and these are discussed below.

The Sequence Comparison Procedures

In this section, we give brief descriptions of the three sequence comparison methods used in this work. For further details the reader is referred to papers cited in the descriptions. We begin by describing the simplest procedure, ISS, then PSI-BLAST, and finally an HMM procedure called SAM-T98.

Intermediate sequence searches

The essential idea of this procedure is that two homologous sequences, which have diverged beyond the point where their relationship can be recognised by a direct comparison, can be related through a third sequence if it has characteristics that are suitably intermediate between the two being matched. A high match score between the first and third sequences and between the second and the third sequences implies that the first and second sequences are related even though their own match score is low. Thus, in suitable cases, the ISS procedure finds the relationship between the two distantly related sequences by collecting, from a large sequence databank, homologues that both match with high scores (Park *et al*., 1997).

Errors can arise in the ISS procedure when two unrelated multidomain proteins match different domains of a multidomain intermediate sequence. For example, if we use A, B, C and D to indicate domains, and A′ and B′ to indicate homologues of A and B, a protein with the two domains AB can match both a protein with domains A′D and one with B′C, even though the latter two proteins are unrelated. This would not be a problem if sequence matches were limited accurately to the domain boundaries; the requirement that the two query sequences match the same region of the intermediate would eliminate such errors. In practice, however, matches extend beyond domain boundaries because a region where the match is strong can support matches made in adjacent regions where it is weak. To a large extent this problem can be avoided if the sequences of individual domains of multidomain proteins are used as separate query sequences, and the search is carried out as follows: (i) the query sequences are matched against a large non-redundant database and the matched pairs of query-intermediate sequences retained; (ii) the region of the intermediate sequence matched by the query sequence is saved and the rest removed; and (iii) the saved fragment of the intermediate sequence is then matched against a user-selected database, and significant hits are collected (Park *et al*., 1997). Details if the ISS search method are available from http://www.mrc-lmb.cam.ac.uk/genomes/jong/ISS.html.

PSI blast

This procedure, recently described by Altschul *et al*. (1997), iteratively collects sets of intermediate sequences to find homologues. The main steps of the procedure are: (i) for a given query sequence, an initial set of homologues is collected from the sequence database using GAP-BLAST, a new version of BLAST that generates gapped alignments, and a conventional score matrix (here we use BLOSUM-62); (ii) a weighted multiple alignment is made from the query sequence and the homologues whose match scores are better than a specified cut-off value (E_M); (iii) a position specific score matrix is constructed from this alignment; (iv) this matrix is then used to search the database for new homologues; (v) new homologues with a good match score are used to construct a new position-specific score matrix, which is then used in a further search for homologues; and (vi) rounds of matrix reconstruction and new searches are iterated either until no new homologues are found or until the number of iterations reaches a specified limit (j). (Altschul *et al*., 1997).

The procedure uses two parameters that can be set by the user. The first parameter, E_M, is the GAP-BLAST E-value cut-off used for selecting homologues for the alignments and the position-specific score matrix. If E_M is too low, only close homologues of the query sequence are used to make the position-specific score matrix and the sequence variation is too limited for it to find distant homologues. If E_M is too high, non-homologues of the query sequence are included amongst the sequences used to make the position-specific score matrix which, in further iterations, will collect additional non-homologues and so become too ''polluted'' to be useful.

The second parameter, j, is the number of rounds (iterations) of new sequence searches and matrix reconstruction that are carried out to find new homologues. Too few iterations may mean that distant homologues are missed, especially if E_M is too low. Too many will allow erroneous matches to be made, especially if E_M is not low enough.

To find good values for these parameters we performed tests using PDBD40-J that will be described elsewhere. We found that to achieve a low rate of false positive homology predictions in our experiments, effective parameters for detecting true sequence relationships are $E_M = 0.0005$ and $j = 20$. In the calculations described below, all PSI-BLAST calculations were carried out using these parameters. This E-value cut-off differs significantly from the default value of $E_M = 0.01$, which we observed to give considerably worse performance in our tests.

Inspection of those PSI-BLAST searches that continue to find new homologues over more than a few iterations showed that there are cases where the recognition properties of the position specific score matrix did not just grow with the number of iterations, but changed. The effect of this matrix migration is that some sequences that matched with high scores at one stage of the calculation are not matched at subsequent stages. This problem was overcome by collecting the high scoring hits

(with *E*-values less than 0.00001) made during each iteration and ensuring that they were included in all subsequent iterations.

The NCBI version of PSI-BLAST can be found at http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast.

SAM-T98 hidden Markov model method

This is a procedure somewhat similar to PSI-BLAST, but using an HMM and its associated parameter estimation procedures in place of a position specific scoring matrix. The first step is to create an initial HMM from a single given query sequence (or a query sequence along with some aligned homologues, if these are available). Then a database of potential homologues of the query sequence is constructed by searching a large protein database using WU-BLAST with a very loose score cut-off. Finally, the following steps are iterated three times: (i) select new sequences from the database of potential homologues that have good local alignment scores to the HMM; (ii) create a new HMM and a new multiple alignment for the query sequence and the selected additional sequences using sequence weighting and Dirichlet mixture priors (Sjolander *et al.*, 1996).

After these iterations, the final HMM can be used to search a user-selected database for homologues of the query sequence.

We use the UCSC SAM package to create the HMMs and to score sequences with the HMMs (Hughey & Krogh, 1996; http://www.cse.ucsc.edu/research/compbio/sam.html). To create an initial database of homologues, the procedure takes the query sequence and conducts a search of a non-redundant protein database using WU-BLASTP (Altschul & Gish, 1996; http://blast.wustl.edu/blast). Two sets of sequences are collected: (i) close homologues that match the query with *E*-values of 0.00003 or less, and (ii) a large set of sequences that match the query with *E*-values of 500 or less, and will therefore include more distant homologues. WU-BLASTP is used in this search for two reasons. Firstly, in the initial step of the procedure, an HMM estimated from one sequence would probably not perform as well as WU-BLAST. Secondly, in all steps of the procedure the HMM searches are computer-intensive, so it is more efficient to first filter the non-redundant database with WU-BLAST and thus reduce the size of subsequent searches.

In the first iteration, the use of low *E*-values by WU-BLAST, and a strict HMM scoring threshold, ensures that only close homologues are used. In the three subsequent iterations, we use gradually decreasing HMM score thresholds and search for additional matching sequences in the large set that includes distant homologues. This action with the score thresholds of HMM parameters is balanced by the use of alignment gap penalties that are more lenient for the first iterations and stricter for the last iteration. Further details can be found in publications by Karplus *et al.* (1997, 1998). The procedure is available at http://www.cse.ucsc.edu/research/compbio/HMM-apps/.

Assessment Procedure

The ISS, PSI-BLAST and SAM-T98 procedures all use intermediate sequences to find relationships between a query sequence and a target sequence. In our tests, ISS and PSI-BLAST used NRDB90 as a source for intermediate sequences. This database contains 152,228 sequences assembled from various other databases and has been made non-redundant by removing sequences that have greater than 90% sequence identity (Holm & Sander, 1998). For its initial WU-BLASTP search for homologues, SAM-T98 used the NRP (non-redundant protein) Database. (This is distributed on the Internet *via* anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute's Frederick Biomedical Supercomputing Center.) The performance differences using different non-redundant database versions were negligible for the ISS search method (data not shown), and would be expected to be negligible for the other methods as well.

To test the ISS procedure, we performed an all-against-all search of the PDBD40-J database. As described above, each of the 935 sequencess was used to search NRDB90 for intermediate sequences and these intermediate sequences were then used to search PDBD40-J for homologues of the original sequence. Since each of the PDBD40-J sequences is used both as a query and a target sequence, every pair (seq1, seq2) of sequences in PDBD40-J will have two raw scores, one when seq1 is the query and one when seq2 is the query. The final score for this pair is taken to be the better of these two scores; it should be noted that the two scores are usually very close, if not identical.

This all-against-all search was repeated using PSI-BLAST and SAM-T98. For these procedures, each query sequence was used to build either a position-specific scoring matrix or an HMM with the help of intermediate sequences from NRDB90 or NRP, and this position-specific scoring matrix or HMM was used to score all sequences in the PDBD40-J database. In the case of PSI-BLAST, PDBD40-J was simply added to NRDB90 before the PSI-BLAST iterations were begun. With SAM-T98, PDBD40-J was not included in this initial phase, but only scored at the end, after the HMM was built from NRP sequences. The PSI-BLAST final score for the pair (seq1, seq2) was the better of the two raw scores (as for ISS), but, for historical reasons, the SAM-T98 final score was the sum of the two raw scores. For all three methods, taking the sum of the raw scores instead of the better of the raw scores gave no significant difference in performance.

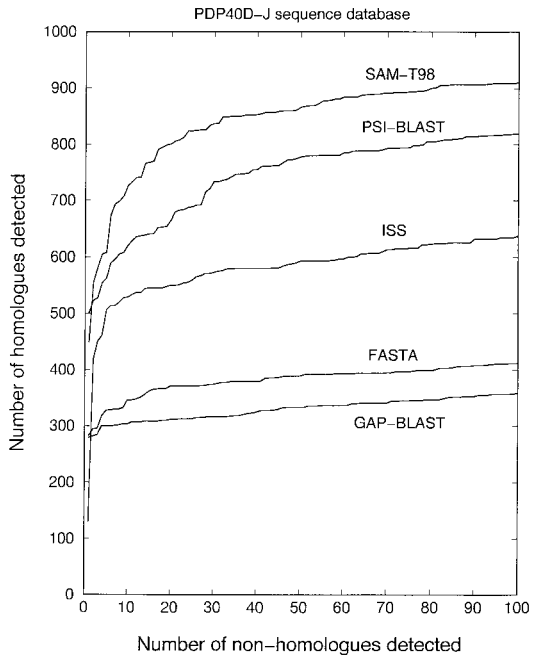


Figure 1. Coverage *versus* error plot. To construct this plot, we calculate, for each method, all matches between a query and a target PDBD40-J sequence. These matches include both homologous pairs (true matches) and non-homologous pairs (false matches). Pairs with uncertain relationship are excluded. The matches are ranked by score from best to worst. Starting from the best score, we sweep down, calculating the number of true matches and the number of false matches at each possible score threshold.

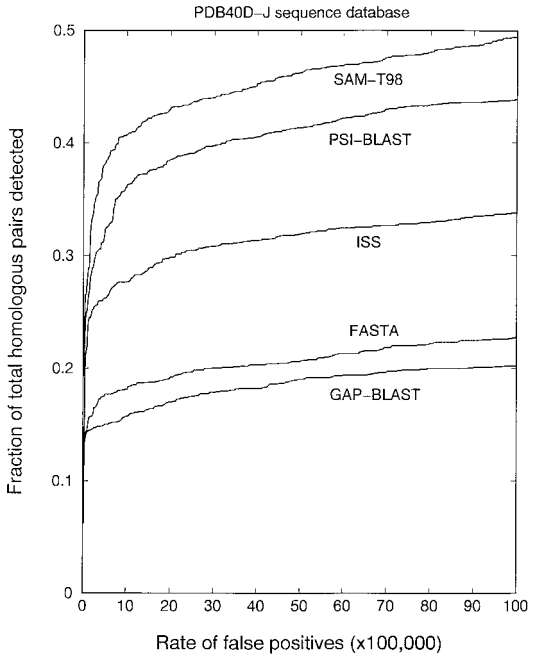


Figure 2. Plot of the proportion of the total number of true matches *versus* the rate of false positives. This plot is constructed similarly to Figure 1 except that the *X*-axis is the rate of false positives (RFP), where RFP = (number of false matches)/(935 × 934/2 − 3965; see the text.) On the *Y*-axis we show the proportion of the total number of true matches at a given RFP (see the text).

In the PSI-BLAST calculation the number of iterations allowed to be made by the query sequence, *j*, was 20. Of the 935 sequences, 61% finished their search after 2–4 iterations, 18% after 5–10 iterations and 11% after 11–20 iterations.

To display and compare the results of these tests we use the ''coverage *versus* error (CVE)'' plots described by Brenner *et al*. (1998). To construct these plots, we first collect all matches between a query PDBD40-J sequence and target PDBD40-J sequences. This collection includes both the pair of homologous sequences (which we call true matches) and pairs of non-homologous sequences (which we call false matches or false positive predictions). Pairs that form uncertain relationships are excluded. The entries in this collection of matches are ranked by their scores from best to worst score. Ideally, all true matches would be ranked before any false matches. In practice, of course, this does not occur. At the top of the list almost all matches are true; at the bottom almost all matches are false, and in between they are mixed. By moving down the list and tabulating the number of true and false matches, we can calculate the number of true match pairs for different numbers of false match pairs. By plotting the number of true matches *versus* false matches, as in Figure 1, one can graphically depict the performance of the methods.

The Comparative Performance and Overall Detection Rate of ISS, PSI-BLAST and SAM-T98

For each procedure, we can calculate the proportion of the total number of true matches found at a given error rate. The proportion of the total number of true matches is the number of true matches (homologous pairs) divided by 2096 (the number of homologous sequence pairs in PDBD40-J, see above). For reasons described in the section on scoring schemes, the error rate is best given as the observed rate of false positive (RFP) predictions: if for a given score threshold the procedure makes four false positive predictions out of the possible 432,680, then the observed rate of false positives is approximately 1/100,000. Rates of 1/10,000 and 1/1000 correspond to 43 and 433 false matches, respectively. In Figure 2, for the three procedures, the proportion of the total number of the possible true matches made is plotted against the RFP. Table 1A shows, for RFP values in the range of 1/100,000 to 1/1000, the actual number of true homologues detected and the percentage of the total number of the possible true matches that this number represents.

The data in Table 1A and Figures 1 and 2 show that at a low rate of false positives, 1/100,000, SAM-T98 detects 29% of the homologues relation-

Table 1. The number of homologous matches made between PDBD40-J sequences at different rates of false positives

Errors		Homologous matches found at different error rates (actual number and its percentage of total in PDBD40-J)				
Rate of false positives	Number of errors	SAM-T98	PSI-BLAST	ISS	FASTA ktup = 1	GAP-BLAST
A. *All homologous pairs found in PDBD40-J*						
1/100,000	4	608; 29	562; 27	506; 24	328; 16	300; 14
1/50,000	9	726; 35	620; 30	529; 25	346; 17	304; 15
1/10,000	43	857; 41	763; 36	580; 28	386; 18	329; 16
1/5000	87	907; 43	814; 39	626; 30	407; 19	353; 17
1/1000	432	1037; 50	920; 44	708; 34	477; 23	396; 19
B. *Matches found for homologous pairs in PDBD40-J that have sequence identities of less than 30%*						
1/100,000	4	414; 22	411; 22	318; 17	153; 8	134; 7
1/50,000	9	522; 28	455; 25	340; 18	165; 9	137; 7
1/10,000	43	649; 35	574; 31	383; 22	199; 11	154; 8
1/5000	87	697; 38	621; 34	423; 23	214; 12	176; 10
1/1000	432	822; 45	722; 39	502; 27	276; 15	213; 12

ships in PDBD40-J, PSI-BLAST detects 27% and ISS detects 24%. At a high rate of false positives, 1/1000, the three methods detect 50%, 44% and 34%, respectively.

The Performance Relative to that of Pairwise Comparison Methods

To see how results found for the multiple sequence procedures compare with those produced by pairwise comparison procedures, FASTA (ktup = 1) and GAP-BLAST were used to match the PDBD40-J sequences in a similar all-against-all test. The sequence matches made by FASTA and GAP-BLAST were processed in a manner described above for the multiple sequence methods and the results are given in Figures 1 and 2 and Table 1A. Inspection of these data shows that SAM-T98 and PSI-BLAST find about twice as many homologous matches as the pairwise procedures FASTA and GAP-BLAST (Table 1A). Thus, the proportion of the PDBD40-J relationships that are found at a rate of false positive of 1/50,000 by the various multiple sequence procedures is about one-third in contrast to the one-sixth found by the two pairwise procedures.

We would expect the other pairwise methods to give similar results. In a previous assessment of pairwise sequence methods using a more recent PDB40D sequence database we showed that the performances of SSEARCH and WU-BLAST are only slightly better than that of FASTA (ktup = 1), whilst BLAST is a little worse (Brenner *et al.*, 1998).

PDBD40-J contains homologues that have sequence identities of up to 40%. It is found generally that relationships between homologous sequences that have greater than 30% identity can be found by pairwise comparison methods. This view is supported by calculations on PDBD40 which showed that, except for short sequences, pairwise methods detect essentially all the relationships between homologous pairs that have sequence identities of over 30% (Brenner *et al.*, 1998). Serious failures by pairwise methods are in the region where homologues have sequence identities below 30%. PDBD40-J has 222 homologous pairs whose sequence identities are 30% or greater (as measued by the sequence identity in the match region given by a FASTA alignment.) To make comparisons of the results for the region below 30% identity, we removed from the matches made by each method all homologous pairs that have sequence identities of 30% or over. The number of matches that remain are given in Table 1B. Inspection of this Table shows that, for homologous pairs with sequence identities of less than 30%, multiple sequence comparison methods detect three times as many homologues as pairwise methods.

The Relation of the Scoring Schemes of PSI-BLAST and SAM-T98 to the Observed Error Rate

The matches found by PSI-BLAST are given a raw log odds score S, which is then converted into a normalised score S', expressed in bits, by the equation:

$$S' = (\lambda g S - \ln Kg)/\ln 2$$

where $\lambda_g = 0.251$ and $K_g = 0.031$ are experimentally estimated constants, and ln denotes the natural logarithm (Altschul *et al.*, 1997). The expected number of (false) matches with score S' or better that would be found searching a database of random protein sequences that has N possible matches should be approximated by the E-value:

$$E = N2^{-S'}$$

Thus:

$$\log_2 E = \log_2 N - S' = -aS + b,$$

where $a = \lambda_g/\ln 2$ and $b = \ln K_g/\ln 2 + \log_2 N$. Thus, the theory predicts that the logarithm of the expected number of false matches is linearly related to the log odds score S. However, when using gapped alignments, the precise theory of the log odds scores has been difficult to work out, so exact relationship between the expected number of

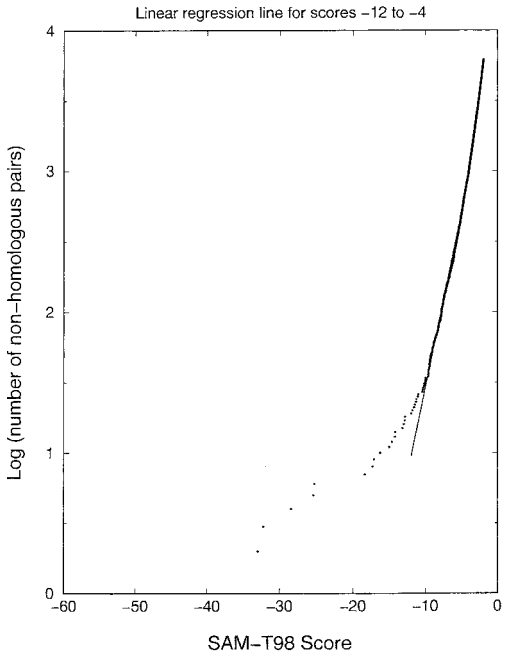


Figure 3. SAM-T98, sequence match scores and the observed error rate. Using the same data as in Figure 1, on the *X*-axis we plot the SAM-T98 score and on the *Y*-axis the logarithm base ten of the number of false positive matches observed with that score or better. One box is plotted for each point in the ranked list of matches. Also shown is a linear regression computed for the data with scores between −12 and −4 (917 data points). The slope of this regression line is 0.248 and the *Y*-intercept is 3.95.

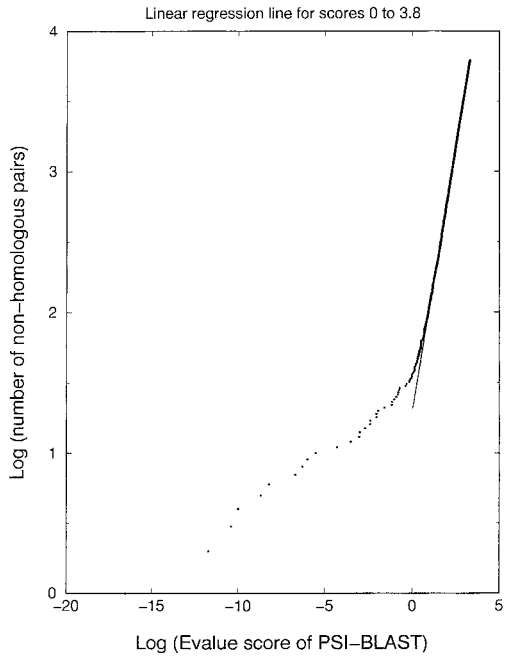


Figure 4. PSI-BLAST, sequence match scores and the observed error rate. This is the same as Figure 3, except that here we used the ranked list of matches for PSI-BLAST, and plot the logarithm base ten of the PSI-BLAST *E*-value on the *X*-axis. We show a linear regression for the data with the logarithm of the *E*-value between 0 and 3.8 (6162 data points). The slope of this regression line is 0.76 and the *Y*-intercept is 1.29.

false matches and the log odds score is to some extent empirical.

The same is true for the log odds scores used by SAM-T98. In this method, the score S for a given sequence X and HMM M is defined to be the log likelihood ratio:

$$S = \ln(P(\mathrm{X}|\mathrm{M})P(\mathrm{X}^{\mathrm{R}}|\mathrm{M}))$$

where X^{R} denotes the reverse of the sequence X and $P(\mathrm{X}|\mathrm{M})$ is the probability of the sequence X under the model defined by the HMM M. Here a local match model is used. This means that the modelling and probability calculation is done in such a way that it is allowed (and expected) that only a part of the sequence X will match to a part of the model M; see Karplus *et al*. (1998) for further discussion.

Recall that in our all-against-all searches, the final score for (seq1, seq2) is the best of the two possible scores for PSI-BLAST and the sum of the two scores for SAM-T98. However, in each case, for significant scores, this should not, in theory, greatly affect the predicted linear relationship between the score and the logarithm of the number of false positives. When p is the probability of getting a false match from one of the two raw scores, then (assuming independence) the probability of getting a false match from taking the better of the two scores is $2p - p^2$, which is larger than p by a nearly constant factor of 2 for all small p, preserving the linear relationship. Thus, for small p we can discount the event of probability p^2 in which both raw scores are significant purely by chance. Similar reasoning shows that the linear relationship should also be approximately preserved for small p when the log odds scores are added.

We examined the extent to which the scores from PSI-BLAST and SAM-T98 are linearly related to the logarithm of the observed number of false positive matches. The results are shown in Figures 3 and 4, which plot the score *versus* the logarithm of the number of observed false positives with that score or better for SAM-T98 and for PSI-BLAST, respectively, as well as a linear regression to these data. For PSI-BLAST, we tried plotting both the normalised bit score *versus* the logarithm of the number of false positives, and the logarithm of the *E*-value *versus* the logarithm of the number of false positives. Because the number N of possible matches depends to a certain extent on the length of the query sequence, the latter gives a smoother plot when data from different length queries are combined, so that is the plot we show for PSI-BLAST.

The curves do seem to be linear for most of the range of scores, but not for the extremely good scores. Here we seem to be getting more false positives than expected. It is possible that some of these false positives that score extremely well vio-

late the assumptions implicit in the null models used in the score theory, and hence have unusual scores. When estimating the regression lines, these data points from unusually good scores were considered outliers and removed from the data set. Data from very poor scores were also removed. The usual statistics for the fit of the regression line to the remaining data (*t*-values, etc.) show this fit to be quite significant, but because the data points are generated by sweeping through one ranked list, they are highly dependent, and hence these statistics are not appropriate here.

One can extrapolate the linear regression lines for SAM-T98 and PSI-BLAST scores to obtain empirical predictions of the rate of false positives for extremely good scores. For example, a SAM-T98 score of −15 would, by this extrapolation, seem to give a *Y*-value of zero, representing one false positive out of about 430,000. This is more optimistic than the estimate of the rate of false positives for a score of −15 that is obtained from the data, which is 10(±6.3) out of 432,680, where 10 is the observed number of false positives and the ±6.3 represents about two standard deviations, or roughly a 95% confidence interval. The relatively high variance of this latter estimate, and the fact that these ten data points appear as outliers, leads us to expect that the estimation of the rate of false positives obtained by the linear extrapolation would prove to be the more accurate one were we to do this experiment on a much larger data set. On the other hand, because of the dependence in the data mentioned above, we do not have good knowledge of the variance in the slope of the regression line, so this extrapolation must be interpreted with caution as well. In particular, we cannot explain the fact that the slope of the regression line for PSI-BLAST was 0.76, when the theory would predict a slope of 1.0.

Finally, we note that, for observed rates of false positives of 1/100,000, 1/50,000, 1/10,000, and 1/1000, the SAM-T98 scores are −25.3, −15.0, −9.4 and −5.1, respectively, and the PSI-BLAST normalised bit score S' are 59.8 ($E = 6 \times 10^{-9}$), 45.7 ($E = 5 \times 10^{-5}$), 31.4 ($E = 2.0$) and 26.2 ($E = 64$), respectively. These values may provide some guidance to practitioners who use these methods, though it should be noted that the searches described here use the sequences of the protein domains in SCOP. Small proteins, and most medium sized proteins, are built of one such domain and searches with their sequences would be expected to behave in a manner similar to those used here. Large proteins built of multiple domains would probably give a somewhat higher false positive rate. Further calculations and experiments are needed to obtain a general quantitative theory of the expected number of false matches that would be directly applicable to more general settings of protein homology search and match the empirical data well.

Similarities and Differences in the True Matches and Errors made by the Three Methods

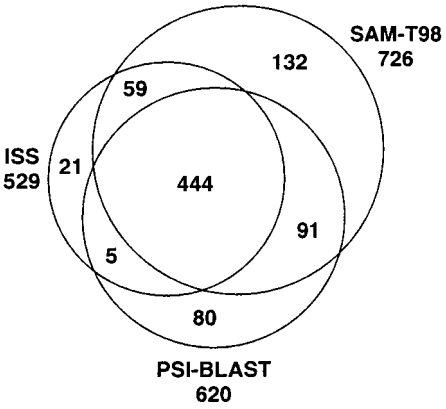
Although all three of the procedures discussed here use multiple sequences, the ways in which they do so are different. Thus, we might expect some procedures to be more appropriate for particular families than others. We determined the extent to which each of the procedures found the same pairs of homologues as one or both of the other two procedures, and the extent to which they made unique matches. The results are given in Figure 5(a) for all matches and in Figure 5(b) for matches between distantly related homologues (those whose sequence identities are less than 30%). At a rate of false positives of 1/50,000, the three procedures found altogether 830 homologous pairs. Of these, 444 were found by all three procedures, and 535 by both SAM-T98 and PSI-BLAST. The number of unique matches made by SAM-T98, PSI-BLAST and ISS are 132, 80 and 21, respectively. A total of 624 distantly related homologous pairs were found and their distribution between the different procedues is shown in Figures 5(b).

In Figure 5(c), we show the extent to which the top 10, 20 and 30 false positive predictions for each of the three procedures are found by more than one procedure. This Figure shows that there are only a few instances where two or more of these procedures make the same false positive predictions with very good scores. Thus, unlike the discovery of true matches, the production of errors is strongly dependant on the exact nature of the procedures.

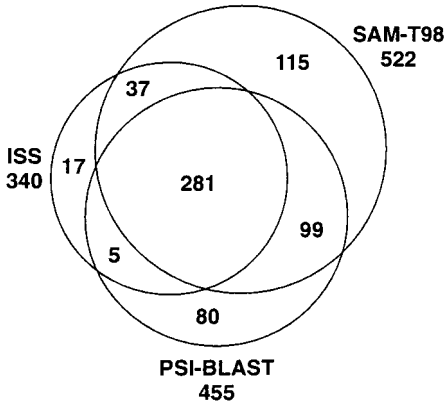
Relationships between Different Superfamilies in the Same Fold Category

At the beginning of this paper we described how the fold classification brings together superfamilies that have the same secondary structures in the same arrangement. Most superfamilies in this category are strongly believed not to have evolutionary relationships. For some superfamilies, however, the situation is less clear in that the current evidence makes the relationships possible but weak. In these cases, the superfamilies are kept separate until the subsequent discovery of intermediate structures provides strong support for their merger (see Murzin (1998) for a description of such cases). Given this ambiguity, pairs of matched sequences that have the same fold, but are not in the same superfamily, are defined here as being uncertain relationships and not counted as true homologues or errors (see above).

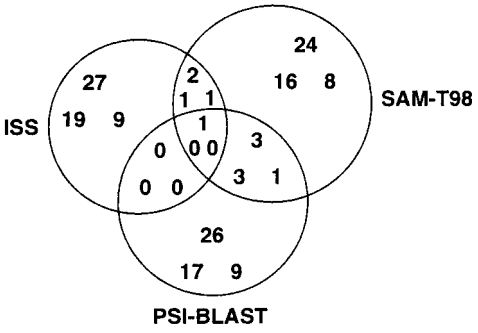
We examined the matched sequences to determine the number of uncertain relationships found with good match scores. For a low rate of false positives of 1/50,000 or less, only a small number are found: one by SAM-T98, one by PSI-BLAST



(a) All matches between Homologues



(b) Matches between Homologues with sequence identies less than 30%



(c) Matches between Non-Homologues

Figure 5. (a and b) Venn diagrams showing, for a RFP of 1/50,000 or less, the extent to which matched sequences are found by one or more of the three sequence comparison methods. The numbers in (a) are for all true matches in PDBD40-J. The numbers in (b) are for the true matches between those pairs of sequences that have less than 30% sequence identity. (c) Venn diagram showing the extent to which the same non-homologous pairs are found by one or more of the three sequence comparison methods. The numbers here are for the top 30, 20 and 10 errors found by each procedure

and none by ISS. Going to a higher rate of 1/10,000 or less increases the number of matches only a little: three for SAM-T98, two by PSI-BLAST and one by ISS. The folds in which these matches are found are those of the NAD binding domains; the barrel-sandwich hybrid; the α/β barrels; and the knottins. For the first three of these four, evolutionary relationships between the particular superfamilies are plausible. Proteins in the different knottins superfamilies are not thought to be related, however. These proteins are small and have a high proportion of cysteine residues and false matches between such proteins are common.

Discussion

We have described an assessment of three sequence comparison methods that use multiple sequence information to find distant evolutionary relatives. Using additional sequences increases the sensitivity of search methods by a factor of 3 in the region where the sequence identities of homologous pairs are less than 30%. Because the number of sequences in databases continues to increase dramatically from the output of genome sequencing projects, we will reach a stage in the near future where many unknown sequences can be matched instantly with multiple intermediate sequences and the models and profiles derived from them.

All methods tested contain parameters that must be adjusted, based on how remote a homologue one is interested in finding. These include substitution matrices, gap costs, sequence weighting methods, number of iterations, and threshold for inclusion on each iteration. For all methods, parameters were adjusted to try and find those that would be most effective for assessment described here, although for SAM-T98 most of the adjustments were made on just part of the dataset (Karplus *et al.*, 1998).

In the procedures described here the SAM-T98 HMMs are created by four iterative searches for homologous sequences. PSI-BLAST was allowed to perform up to 20 iterations and for 39% of queries found homologues after the fourth round (see above). ISS only makes one search for homologues. The performance of SAM-T98 and ISS would be expected to be further improved by increasing their number of iterative searches.

Scoring is another area where these methods may be improved. The scoring system for PSI-BLAST is based on a solid statistical theory, but some of the assumptions made in this theory are occasionally violated in practice, and the theory is not easily applied in cases were many insertions and deletions are made. On the other hand, no significant theoretical work has been done yet on the SAM-T98 scoring method used in this study, which used the reversed sequence as a null model. Further analysis may lead to improvements in this scoring function.

Even though these methods are still at a relatively early stage in their development, this study shows that ISS is about twice as effective as FASTA and GAP-BLAST (and by implication other pairwise methods) at finding the distant evolutionary relations in PDBD40-J, and that SAM-T98 and PSI-BLAST are three times as effective.

Acknowledgements

We are grateful for the advice from Dr Alex Bateman and comments from the referees of the paper. J.P. is grateful for the support of Dr George Church and Harvard Medical School for his scientific research. K.K., C.B., R.H., and D.H. were supported by NSF grant DBI-9408579, DOE grant DE-FG03-95ER62112, and a gift from the Digital Equipment Corporation. D.H. gratefully acknowledges the support of the Isaac Newton Institute for Mathematical Sciences and the Sanger Centre.

References

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* **266**, 460–480.

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

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. *Nucl. Acid Res.* **25**, 3389–3402.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059–1063.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold: unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

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

Eddy, S. R. (1995). *ISMB 95, Intelligent Systems in Molecular Biology* (Conference), *Multiple Alignment Using Hidden Markov Models*.

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

Eddy, S. R., Mitchison, G. & Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* **2**, 9–23.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355–4358.

Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics,* **14**, 423–429.

Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS,* **12**, 95–107.

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

Karplus, K., Barrett, C. & Hughey, R. (1998). *Fold Recognition by Remote Homology Detection Using Hidden Markov Models,* Technical Report UCSC-CRL-98–06, Department of Computer Science, University of California at Santa Cruz, CA 95064. Obtainable at ftp://ftp.cse.ucsc.edu/pub/tr/ucsc-crl-98-06.ps.Z.

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

Luthy, R., Xenarios, I. & Bucher, P. (1994). Improving the sensitivity of the sequence of the sequence profile method. *Protein Sci.* **3**, 139–146.

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380–387.

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**, 536–540.

Neuwald, A., Liu, J., Lipman, D. & Lawrence, C. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* **25**, 1665–1667.

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

Pearson, W. R. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, S. & Haussler, D. (1996). Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *CABIOS,* **12**, 327–345.

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. USA,* **91**, 12091–12095.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS,* **10**, 19–29.

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

Edited by J. Thornton

(Received 18 May 1998; received in revised form 3 September 1998; accepted 14 September 1998)