Large-Scale Comparison of Protein Sequence Alignment Algorithms With Structure Alignments

J. Michael Sauder, Jonathan W. Arthur, and Roland L. Dunbrack, Jr.*
Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania

ABSTRACT **Sequence alignment programs such as BLAST and PSI-BLAST are used routinely in pairwise, profile-based, or intermediate-sequence-search (ISS) methods to detect remote homologies for the purposes of fold assignment and comparative modeling. Yet, the sequence alignment quality of these methods at low sequence identity is not known. We have used the CE structure alignment program (Shindyalov and Bourne, Prot Eng 1998;11: 739) to derive sequence alignments for all superfamily and family-level related proteins in the SCOP domain database. CE aligns structures and their sequences based on distances within each protein, rather than on interprotein distances. We compared BLAST, PSI-BLAST, CLUSTALW, and ISS alignments with the CE structural alignments. We found that global alignments with CLUSTALW were very poor at low sequence identity (<25%), as judged by the CE alignments. We used PSI-BLAST to search the nonredundant sequence database (*nr*) with every sequence in SCOP using up to four iterations. The resulting matrix was used to search a database of SCOP sequences. PSI-BLAST is only slightly better than BLAST in alignment accuracy on a per-residue basis, but PSI-BLAST matrix alignments are much longer than BLAST's, and so align correctly a larger fraction of the total number of aligned residues in the structure alignments. Any two SCOP sequences in the same superfamily that shared a hit or hits in the *nr* PSI-BLAST searches were identified as linked by the shared intermediate sequence. We examined the quality of the longest SCOP-query/SCOP-hit alignment via an intermediate sequence, and found that ISS produced longer alignments than PSI-BLAST searches alone, of nearly comparable per-residue quality. At 10–15% sequence identity, BLAST correctly aligns 28%, PSI-BLAST 40%, and ISS 46% of residues according to the structure alignments. We also compared CE structure alignments with FSSP structure alignments generated by the DALI program. In contrast to the sequence methods, CE and structure alignments from the FSSP database identically align 75% of residue pairs at the 10–15% level of sequence identity, indicating that there is substantial room for improvement in these sequence alignment methods. BLAST produced alignments for 8% of the 10,665 nonimmunoglobulin SCOP superfamily sequence pairs (nearly all <25% sequence identity), PSI-BLAST matched 17% and the double-PSI-BLAST ISS method aligned 38% with E-values <10.0. The results indicate that intermediate sequences may be useful not only in fold assignment but also in achieving more complete sequence alignments for comparative modeling. Proteins 2000;40:6–22.**

Key words: **homology modeling; PSI-BLAST; intermediate sequence; SCOP; alignment benchmark**

INTRODUCTION

With the experimental determination of many new protein structures in recent years and the development of more sensitive remote homologue detection methods that exploit rapidly growing sequence databases, it has become increasingly likely that a protein of biological interest but unknown three-dimensional structure will have a homologue of known structure. From a homologue of known structure, it is possible to build a model of the target sequence of unknown structure using methods developed by many research groups over the last 30 years.[1–3] Comparative modeling helps to bridge the gap between primary and tertiary structure by allowing the construction of models that may be used to identify critical residues involved in catalysis, binding, or structural stability; examine protein–protein or protein–ligand interactions; and correlate genotypic and phenotypic mutation data. Comparative modeling, or homology modeling, is usually based on a number of steps[4, 5]: (1) identifying a known structure (the parent structure) homologous to the sequence of unknown structure (the target sequence); (2) aligning the target sequence to the parent sequence and structure; (3) building the backbone for the model from the alignment and the parent structure, possibly rebuilding loop regions

Abbreviations: CE, combinatorial extension structure alignment program; SCOP, structural classification of proteins; ISS, intermediate sequence search procedure; PDB, protein data bank; BLAST, basic local alignment search tool; PSI-BLAST, position specific iterated-BLAST; FSSP, fold classification by structure-structure alignment of proteins.

Grant Sponsor: National Institutes of Health; Grant numbers: CA06927 and CA09035; Grant sponsors: American Cancer Society; Commonwealth of Pennsylvania; and National Cancer Institute.

J.M.S. and J.W.A. contributed equally to this work.

*Correspondence to: Roland L. Dunbrack, Jr., Institute for Cancer Research, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111. E-mail: RL_Dunbrack@fccc.edu

Received 5 October 1999; Accepted 6 January 2000

containing insertions or deletions; and (4) placing sidechains onto the backbone model. If more than one homologous structure is found, then model building can be based on information derived from multiple structures.[6] In practice, the first two modeling steps are accomplished together, because the identification is based on the statistical significance of a proposed sequence alignment. Therefore, the second step sometimes involves manual adjustment of the original alignment. Once the identification is made, another computer program may be used to realign the parent structure and target sequence. The fold identification step in comparative modeling can be accomplished by sequence comparison[7,8] or by fold recognition methods based on sequence–structure compatibility.[9–13] Sequence comparison methods can either be global, such that two sequences are aligned along their full lengths,[7] or local, such that only segments of the two sequences with some similarity are aligned.[8] For fold assignment by sequence comparison, local alignment methods are usually used because two proteins may share homology over only a portion of their full lengths. This can be accomplished by a pairwise search of a database of sequences of known structure, and fold assignments made when the alignment score between query and hit attains a certain significance level. If no hit is found in the Protein Data Bank (PDB) sequence database, then the target protein is not related to a protein of known structure or the homology to such a protein is more remote than can be detected through a pairwise search. More remote homologies can sometimes be detected by "sequence hopping." In this strategy, a pairwise search of a large protein sequence database such as GenBank or SwissProt (i.e., including proteins of unknown structure) is performed to create a list of homologues of the target sequence. Any of these homologues can be used as a query to the PDB sequence database and, if a significant hit is found, homology between the target and a sequence of known structure is established through the intermediate sequence. This method is therefore sometimes referred to as an "intermediate sequence search" (ISS).[14–17] To extend the evolutionary distance that can be traversed, a number of hops can be made before the PDB sequence database is searched.

In addition to pairwise comparison, local alignment methods can also be used to build a profile of the target sequence family from sequences in GenBank or SwissProt.[18] Such a profile consists of the proportions of the 20 amino acids aligned at each position of the target sequence. New sequences can be added to the profile by an iterative search of the sequence database for proteins with significant similarity to the profile. At any step of the iterative search, the profile can be used to search the PDB sequence database for a significant match. We call such an identification method a *profilewise* fold assignment and sequence alignment.

To assess their general utility for homology modeling of proteins, we need to assess these sequence comparison methods for their ability to detect remote homologues of known structure and for their ability to achieve the correct alignments of the target and PDB sequences. Recently, several sequence comparison methods have been tested for their ability to identify remote homologues. Brenner et al.[19] tested the pairwise local sequence alignment methods FASTA,[20] BLAST,[21] WU-BLAST2,[22] and SSEARCH,[8] comparing their ability to identify known homology relationships in the SCOP protein structure classification.[23] They found that only 10–18% of the known homologies with sequence identity less than 40% were found by these pairwise methods. Park et al.[14] used the SCOP database as a test set to study intermediate sequence search methods. They tested two procedures: (1) searching the nonredundant OWL sequence database with each sequence in the SCOP PDB40D sequence database using FASTA,[20] and establishing an identification between two SCOP sequences if they shared the same hit in OWL with each E-value less than 5 and an overlap along the OWL sequence of at least 30 residues; and (2) searching the nonredundant OWL sequence database with each SCOP sequence, and then taking the fragments of OWL sequences aligned to the query to search the PDB40D database. The latter procedure guards against false positives that sometimes occur when weakly aligned regions exist next to strongly aligned regions. At an error-per-query rate of 1%, their "double-search" procedure found 70% more remote homologues than pairwise methods and found a total of 26% of 2143 homologous sequence pairs with percent identity of 40% or less. In a subsequent paper, Park et al.[24] compared ISS to PSI-BLAST and a hidden Markov model method, SAM-T98,[25] and found that at a false positive rate of 1/10,000, SAM-T98, PSI-BLAST, ISS, FASTA, and BLAST found 28, 25, 18, 9, and 7%, respectively, of homologous pairs with less than 30% identity.

Gerstein[15] also used the SCOP sequence database to test an ISS method based on FASTA. His procedure was similar to Park's double search procedure, except that instead of searching PDB40D in the second step, the full OWL sequence database was searched with the intermediate sequence fragment. Gerstein used a more stringent E-value cutoff of 10^{-4} and overlap lengths of sixty amino acids to make identifications. He found that ISS identified 33–38% more relationships than direct pairwise comparisons, but that only 19% of all homologous relationships in PDB40D (all at 40% sequence identity or less) could be identified by direct or intermediate sequence searches. Salamov et al. tested a multiple-intermediate search procedure.[16] A pairwise search program (either Gapped-BLAST or FASTA) was used to find hits in the nonredundant OWL protein sequence database. One of the more distant hits was randomly chosen and used to search OWL again. This process was repeated until the sought-after PDB sequence was found in the list of hits (because the PDB is contained within OWL). With a single intermediate sequence, they found that ISS returned 38% more homologous pairs in their test set, derived from the CATH database,[26] than a pairwise search. Multiple intermediate sequence searches (ironically abbreviated "MISS") found an additional 12%. The combined ISS methods found 1.5 times as many pairs as the pairwise search. They did not break down their results by sequence identity.

Although sequence alignment methods have been tested for their ability to identify remote homologues, there have been only a few systematic studies to assess the quality of the resulting sequence alignments.[27,28] The evaluation requires a standard for the "true" alignment. This is usually derived from alignments of experimentally determined protein structures.[29] There are two important aspects to determining the "structural alignment" between two protein chains. The first is to use the positions of the atoms to determine a rotation and translation matrix. This matrix can be applied to the three-dimensional coordinates of one protein to superimpose, or structurally align, it with another protein. The second step involves determining which residues on the two chains are aligned with each other given the structural alignment, thereby producing a sequence alignment based on the structure alignment. In the literature, the term "structural alignment" is often used to describe either or both of the above steps.

Structural alignment programs include MINAREA,[30] CE,[31] LOCK,[32] DALI,[33] VAST (www.ncbi.nlm.nih.gov/Structure/VAST/), STAMP,[34] and SSAP[35] among many others. The various techniques fall into three broad categories. The most common methods are based on minimizing the distances between the α-carbon atoms in the two protein chains to determine the best alignment.[30,34] Alternatively, some methods compare the internal distances between various atoms on one protein chain with internal distances on the second chain.[31,33,35] Other programs attempt to analyze the location and orientation of the major elements of secondary structure and, finally, some methods use a combination of these techniques.[32]

Although the superposition produced by two structure alignment programs may be quite similar, the sequence alignments produced may differ substantially when the sequence identity between them is low.[36,37] Nevertheless, it is necessary to define a standard to assess the quality of the alignments produced by the sequence alignment methods. We would like a structural alignment algorithm that produces the most useful sequence alignment in terms of building homology models. Such an algorithm must have specific features. It must align major elements of secondary structure correctly even if they are translated or rotated relative to one another, as well as correctly identify gaps and insertions. Further research in the area of structural alignment will undoubtedly produce better sequence alignments for the purposes of homology modeling. Currently, we believe the best techniques are those using an algorithm based on the calculation and comparison of intramolecular distances and angles such as CE and DALI. The current work is based on CE, because this program was publicly available at the time this work was started. DALI has only recently become available (L. Holm, personal communication).

There have been a small number of head-to-head comparisons of sequence alignment quality as compared to structural alignments. Gotoh[27] compared several global multiple sequence alignment programs against structure alignments in the Joy3.2 database.[38] Sequences related to the structures were identified in SwissProt using the ungapped BLAST program. Sequences were grouped with their sequence relatives (from SwissProt) and their structure relatives (from Joy3.2) into families. The families were aligned with four different multiple sequence alignment methods: (1) a progressive alignment method, CLUSTALW; (2) a randomized iterative method that optimizes a sum-of-pairs score; (3) a randomized iterative method that optimizes a weighted sum-of-pairs score; and (4) a doubly nested iterative method optimizing weighted scores. Except for CLUSTALW, the other three methods are from Gotoh.[39–41] The percentage of correctly aligned residues as a function of sequence identity was not given, but in general, Gotoh found that the iterative weighted sum-of-pairs score performed better than the sum-of-pairs score or the progressive alignments. It should also be noted that N- and C-terminal extensions in the structural alignments were deleted from the sequence alignment benchmarks.

Thompson et al.[28] recently compared several local and global multiple sequence alignment methods against the manually constructed BAliBASE benchmark of structure-derived sequence alignments. Each aligned set of sequences consists primarily of sequences of known structure, rather than including a larger number of sequences from the full sequence database. In cases with no N- or C-terminal extensions or large insertion regions, they found that global alignment methods performed better than local methods. But in the presence of large insertions or N- and C-terminal extensions, the local methods performed better. The number of test cases was relatively small. For instance, for alignments with less than 25% sequence identity, there were 23 alignments tested. The only profile-based method that was tested was a hidden Markov method[42] that performed rather poorly in terms of alignment quality.

In contrast to multiple sequence alignment methods, the sequence-based methods commonly used for fold assignment, BLAST, PSI-BLAST, and ISS methods, have not been tested for the quality of their sequence alignments. Because the significance of the detection is based on the score of the sequence alignment, it is of interest to assess the alignment quality of these methods. In this work, we compare these methods for their alignment accuracy. We have used the CE program of Shindyalov and Bourne[31] to align structures in each SCOP family and superfamily. The resulting sequence alignments were used as a standard for judging the sequence alignments used in fold assignment methods. We also compared the DALI-generated alignments available in the FSSP database[43] with those produced by CE. A goal in this work is to examine various sequence alignment techniques using a test set that is extremely large in comparison to those often used in assessing algorithms in computational biology.

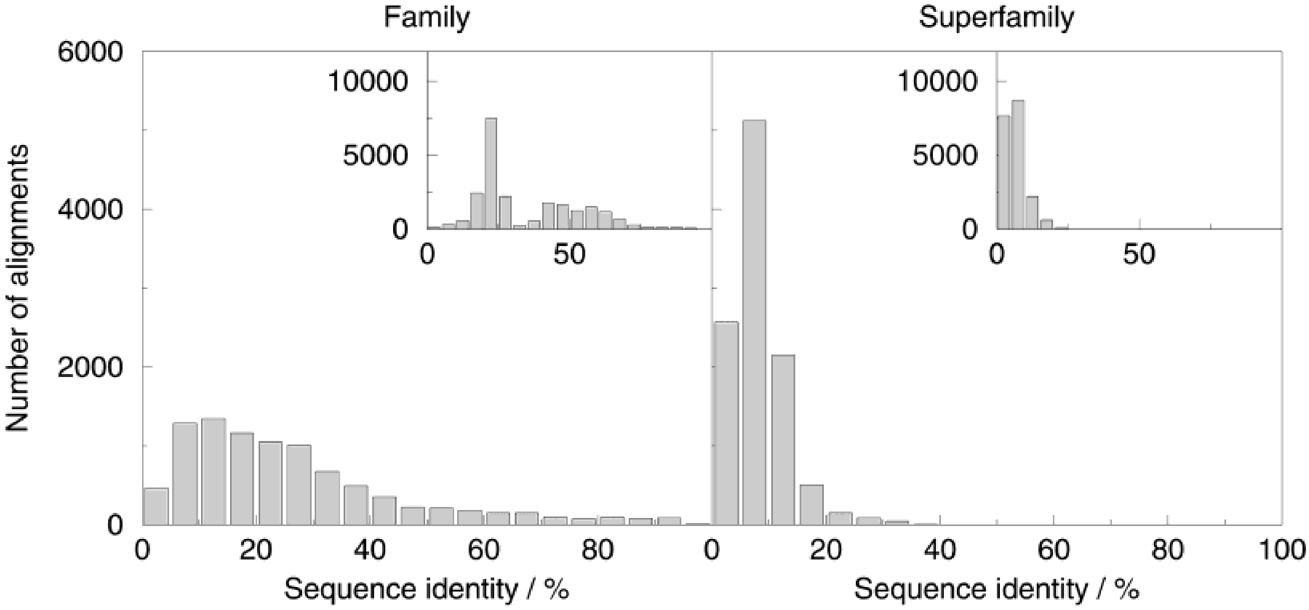

Fig. 1. Histograms of the number of CE structural alignments as a function of sequence identity. The left panel shows the results for alignments where the two SCOP domains are related at the family level. The right panel shows the results for alignments at the superfamily level (i.e., excluding pairs with the same family designation). The insets show the results for the alignments involving an immunoglobulin variable domain. The main panel shows all the other alignments. The sequence identities were binned into intervals of width 5%.

RESULTS

Structure Alignments

We used the CE program of Shindyalov and Bourne to produce structure alignments of all superfamily and family-level pairs of proteins in the SCOP 1.38 database. Details of the CE structural alignment program can be found in the original paper.[31] In brief, the program aligns two proteins by breaking both protein chains into a number of fragments. Each pair of fragments, one from each protein, forms an "aligned fragment pair." Using a series of rules and scoring functions, the program gradually attempts to extend the alignment of the two proteins combinatorially, starting from each aligned fragment pair to find the longest and best alignment between the two protein chains. Once the structural alignment is completed, a sequence alignment corresponding to that structural alignment can be obtained by examining the aligned fragment pairs included in the final structural alignment.

Structural classification of proteins (SCOP) is a database of protein structure and homology relationships.[23,44] The top level of the SCOP hierarchy divides proteins of known structure into all-α, all-β, α/β, $\alpha + \beta$ domains, as well as classes of membrane proteins, peptides, etc. Each class is then broken down into separate fold categories, such as the globin fold (in the all-α class) or the immunoglobulin fold (in the all-β class). Each fold is then divided into "superfamilies," each of which appears to be an evolutionarily distinct lineage. Proteins in different superfamilies seemingly have unrelated functions and may share the same structure only through convergent evolution. The superfamilies are further divided into "families" of more closely related proteins. Proteins in a single family are usually orthologous, that is, having the same function in different organisms. Sequences in different families are usually paralogues, related by descent from a duplicated common ancestor but having different functions. We modified the SCOP 1.38 database, PDB95D, by removing multichain domains, domains consisting of separated segments from the same chain, small proteins and peptides, and membrane proteins. This resulted in 2622 sequences of known structure. We derived sequences for these structures from the SEQRES records of the PDB entries and the SCOP domain descriptions. Pairs of sequences where both sequences are in the same SCOP family are defined as related at the family level. We define sequences related at the superfamily level as pairs of sequences in the same superfamily but in different families.

In Figure 1, we show the sequence identities arising from the CE alignments for proteins related at the family and superfamily level. The 2,622 structures and the SCOP family-superfamily definitions yield a total of approximately 32,000 family level pairs and 30,000 superfamily level pairs. Of these, 22,791 family level pairs and 19,260 superfamily level pairs were alignments consisting of at least one immunoglobulin variable domain. The next largest family, the globins, contains 51 members and accounts for only 1,806 family and 344 superfamily alignments. Because the SCOP family containing immunoglobulin variable domains has almost 5 times more entries than the next largest family, the number of alignments was dominated by comparisons involving these immunoglobulin domains. To avoid overrepresentation of the immunoglobulins, the data for these variable domains have been plotted separately in the inset. Also, in analyzing the

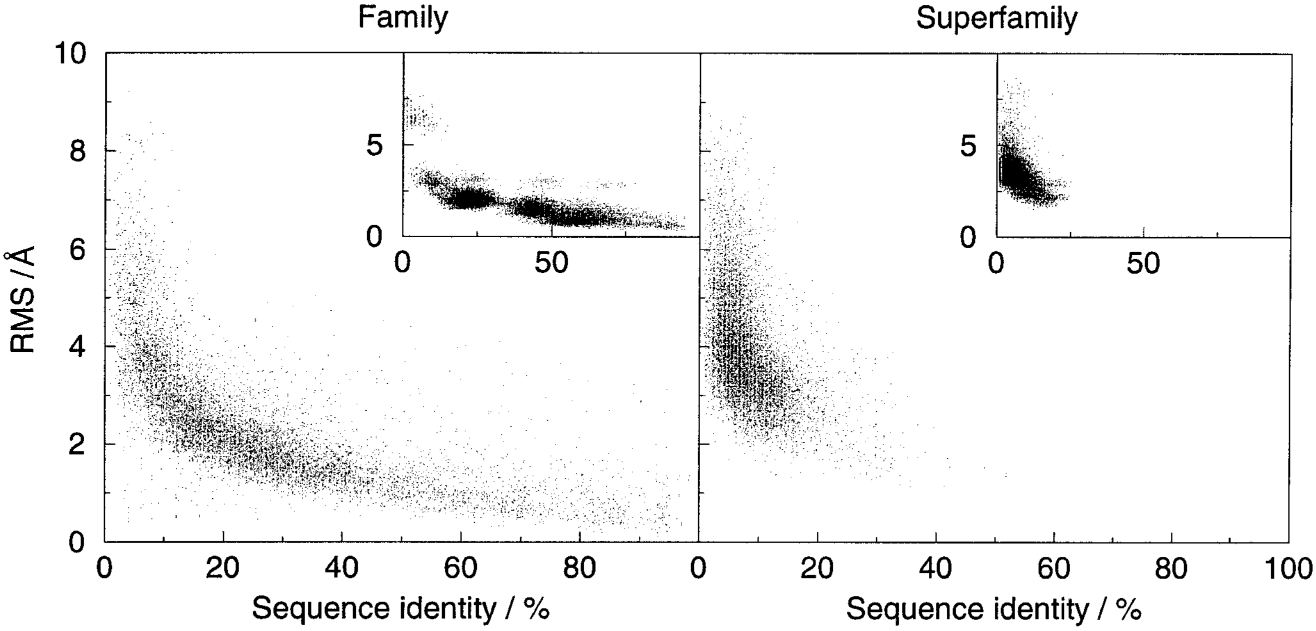

Fig. 2. Variation of the RMS deviation of the positions of the α-carbons of the two proteins as a function of sequence identity. The left panel shows the family level alignments, and the right panel shows the superfamily level alignments. The alignments involving an immunoglobulin variable domain are shown in the inset whereas all others are in the main panels.

sequence alignment algorithms, we have normalized the scores for each family or superfamily, so that all families and superfamilies are counted equally, regardless of size.

Figure 1 demonstrates that the family level sequence alignments represent a good test set for sequence alignment methods. The largest proportion of alignments have a sequence identity around 20% (68.4% of the alignments had sequence identity of less than 30%), which is near the detection threshold for most sequence alignment methods. In addition, the alignments at the family level have protein pairs that span the whole range of sequence identity values. This allows us to evaluate the various sequence search techniques over a broad range of sequence identity. The superfamily level contains more remote relationships with 99% of the alignments having a sequence identity below 30%.

Before analyzing the various sequence alignment methods, it is important to evaluate whether our structural alignment program will serve as a reasonable standard for comparing the various sequence alignment methods. It is worthwhile noting whether, and how often, the structural alignment program produces a poor alignment between the two proteins. CE defines the quality of the alignment in terms of a Z-score. There is less than one chance in a thousand of making the alignment by chance if the alignment has a Z-score higher than 3.5.[31] The quality of the alignment increases as the Z-score increases. The distribution of the Z-scores for the total set of structural alignments, at both the family and the superfamily level, is roughly normal. The family level alignments produce a mean Z-score of 5.7 ± 0.8. (The error is quoted to one standard deviation.) At the superfamily level the mean Z-score decreases to 3.8 ± 0.7. Alignments at the family level are much more statistically significant than at the superfamily level. The mean score is well above the threshold value of 3.5 and 93% of these alignments are above this threshold. At the superfamily level, the Z-score is much closer to the threshold value and only 74% of the alignments have a Z-score above this cutoff. This reflects the larger evolutionary distance in superfamily alignments compared to family alignments.

Another common method for assessing the quality of a structural alignment is the root mean square (RMS) deviation of the positions of the α-carbon atoms in the protein chain. The RMS deviation is correlated with both the sequence identity between the two domains being compared and the Z-score (or the equivalent internal scoring function used by a particular structural alignment program). The structural alignment should have the lowest RMS deviation consistent with the evolutionary distance between the two proteins. This ensures the two structures are properly superimposed in space, keeping related residues close together. Figure 2 shows the correlation between the RMS deviation and the sequence identity. As previously, the insets show the results for the alignments involving immunoglobulin variable domains whereas the main frames show the results for the remainder of the alignments. The data show an inverse asymptotic relationship between the RMS deviation and the sequence identity, as shown long ago by Chothia and Lesk.[45] The RMS deviation between the positions of the α-carbons decreases as the sequence identity increases. At the family level, where the alignments have a spread of values at various sequence identities, this trend can be seen clearly. The majority of the alignments (76%) have an RMS deviation of less than 3 Å. The majority of the remaining 24% of

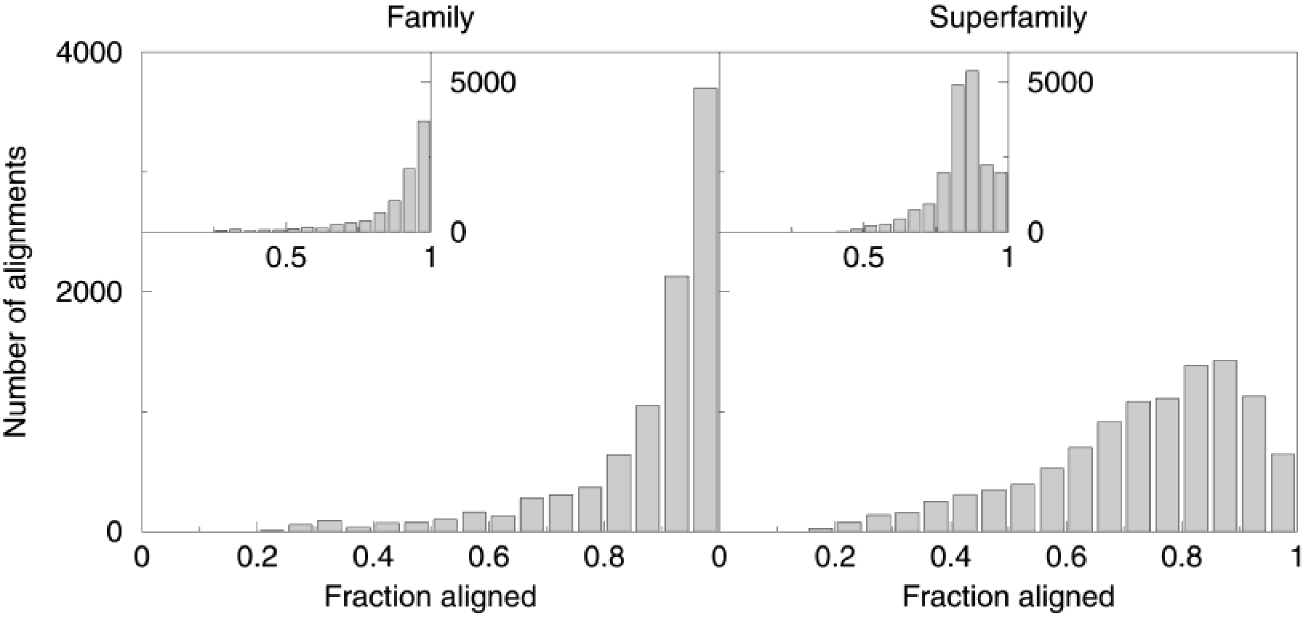

Fig. 3. Histogram of the number of structural alignments having a particular fraction of the shorter of the two proteins aligned. The left panel shows the family level alignments, and the right panel shows the superfamily level alignments. The alignments involving an immunoglobulin variable domain are shown in the inset whereas all others are in the main panels. The alignments were binned in intervals of width 0.05.

alignments involve pairs of proteins with sequence identity less than 20% indicating that the two proteins are more distantly related. The inset shows the immunoglobulin variable domains concentrated in a few regions with RMS deviations of 1–2 Å. At the superfamily level, the structures are more diverse with only 20% of the alignments having an RMS deviation of less than 3 Å. The alignments cluster at 5–10% sequence identity with an RMS deviation of approximately 3–4 Å. The immunoglobulin variable domains aligned with other immunoglobulin domains are also clustered in this region.

An important aspect of sequence alignments derived from structure alignments is whether the sequence alignment covers most or all of the topology common to the two proteins. In some cases, if secondary structure units have shifted significantly in position, the structure alignment program will fail to align these in the resulting sequence alignment. We have used the SCOP database in part because the entries are divided by fold and domain, such that there should not be significant N- and C-terminal extensions in any of the structural alignments. Figure 3 shows a histogram of the number of alignments having a certain fraction of the two proteins aligned. The fraction of residues aligned measures the ability of the structural alignment program to align the complete length of the shorter of the two proteins. CE gradually extends a series of alignments starting at various different positions to determine the best and longest alignments. Sometimes, the best alignment does not cover the entire length of the shorter of the two proteins. The figure shows that, at the family level, 68% of the alignments have greater than 80% of the shorter protein in each pair aligned, and 93% of the alignments have greater than 50% of the shorter protein aligned. The alignments involving an immunoglobulin variable domain display similar trends to the other alignments. At the superfamily level, the immunoglobulin domains have a high peak where 80–90% of the two chains are aligned.

In Figure 4, we show the fraction of secondary structures in each protein aligned with the same type of secondary structure in the structure alignments (helix aligned to helix, sheet to sheet, coil to coil). At the family level, the majority of proteins (67%) have over 80% of secondary structure aligned to equivalent secondary structures. A similar trend is evident at the superfamily level, although the average fraction of secondary structure alignment is a little lower. This indicates that the alignments are preserving at least one determinant of the topology.

It is worth considering the nature of the worst structure alignments produced by CE. This includes alignments with a very low Z-score, a very high RMS deviation, or, most critically, where the fraction of aligned residues is very low. If the structure-based sequence alignment is shorter than the alignment produced by CLUSTALW or PSI-BLAST, for example, then we have no way of evaluating whether or not paired residues in the extended region are correctly aligned. We examined a number of structural alignments of poor quality and found various reasons for the poor structural alignments, although a few general trends did emerge. In some alignments, an abnormally large insertion in one of the chains appeared to be the cause of the poor alignment. One example of this is the alignment between SCOP domains 1waj (residues 376–901) and 1dpi (residues 519–928) where the former chain becomes radically different from the latter at the N-terminus of the alignment. In the alignment of 1n2c (G

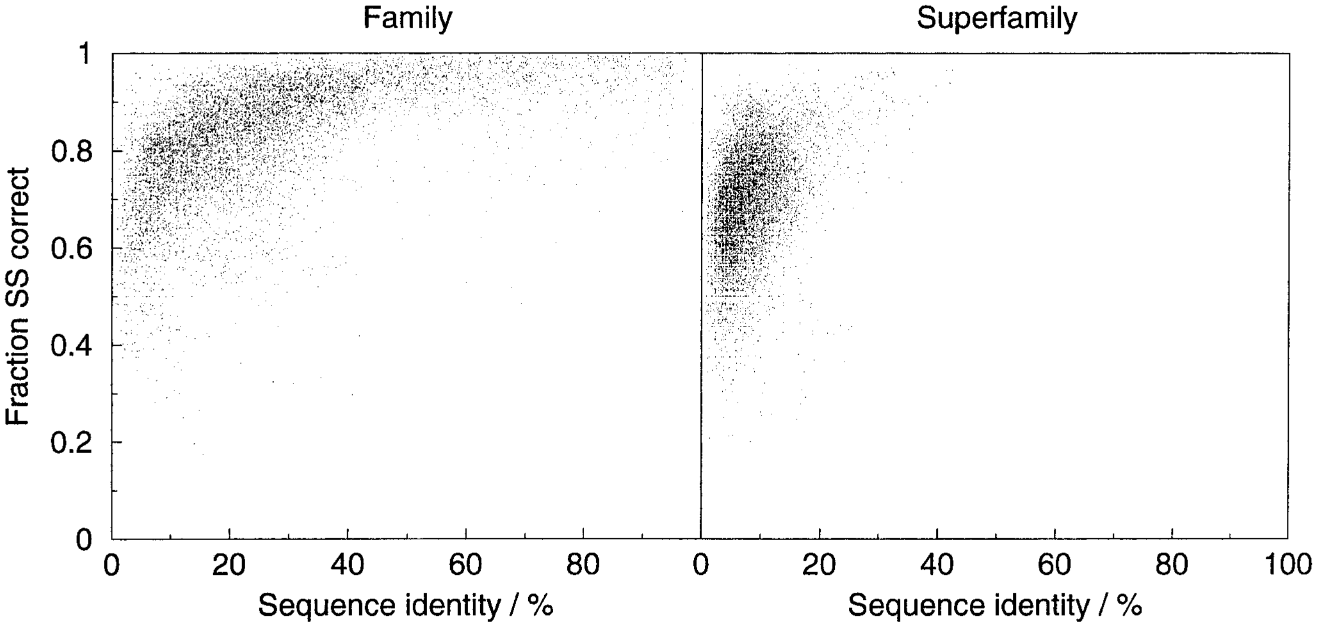

Fig. 4. Variation in the fraction of secondary structure correctly aligned in the structural alignments as a function of the sequence identity. The left panel shows the family level alignments, and the right panel shows the superfamily level alignments.

chain) with 1ade (B chain) the two protein chains have an extreme departure in the paths of the chains through three-dimensional space. Incorrectly aligned elements of secondary structure also seem to be involved in a number of cases, along with large regions of coiled protein or areas of missing density in the PDB file. Finally, in a small number of cases, the difficulty appears to lie in the specification of the SCOP domain itself. An example of this is chain A of 2eze. This protein is in the HMG box superfamily, which includes three helices in an irregular array. PDB entry 2eze contains only a 25 amino acid fragment with an irregular secondary structure. Many of the other members in this SCOP superfamily contain the full protein of 70–80 amino acids. This accounts for the poor alignments of this protein with other members of the same family or superfamily.

Fortunately, as discussed previously, the number of these poor alignments is small, particularly at the family level. Because the sequence alignment methods are tested against the same structural alignment, they all will be equally affected by the quality of that structural alignment. Nevertheless, future structural alignment programs that satisfactorily handle the various difficult cases examined here, and improvements in SCOP domain definitions would provide more adequate tools for comparison of protein sequence alignment methods.

As a final check on the alignment quality, we compared the CE alignments with those available from FSSP.[46] This necessitated forming transitive alignments between SCOP sequences and their representative structures in FSSP, and between the alignments of representative structures themselves (see Methods). Only a subset of the CE alignments we performed on SCOP structures were present in FSSP. In some cases, this was because some SCOP sequences were missing in FSSP because of formatting problems in the PDB record and maximum length cutoffs (L. Holm, personal communication). In others, the DALI program did not apparently detect sufficient similarity in the structures (a Z-value >2.0) to list the alignment in the FSSP file. In fact, FSSP contained only 5,629 of 10,665 pairs (53%) of the SCOP superfamily-level alignments and 5,269 of 9,248 (57%) of the family-level alignments.

In Figure 5, we show the fraction of aligned residue pairs in each CE alignment that are present in each FSSP alignment for each SCOP pair as a function of sequence identity (according to CE's alignment). At low sequence identity (below 10%), there is significant disagreement between the structure alignments. But above this value, the alignments are quite similar. This is in contrast to Gotoh's comparison of Joy3.2[38] with 3D_ALI[47] structure alignments, where the cutoff for agreement was close to 20%.[27] As we will show, the agreement of CE with FSSP is substantially higher than the agreement of any pure sequence alignment algorithm with the CE structure alignments.

Homologue Identification

Several research groups have tested BLAST, PSI-BLAST, and ISS methods for their ability to detect remote homologues within the SCOP database.[14,19,48] The reported results vary somewhat because the program parameters and methodology can significantly affect the results, especially in the case of ISS. We used our modified version of SCOP to test the detection abilities and alignment quality of BLAST, PSI-BLAST, and an ISS method based on PSI-BLAST. For comparison in alignment quality, we

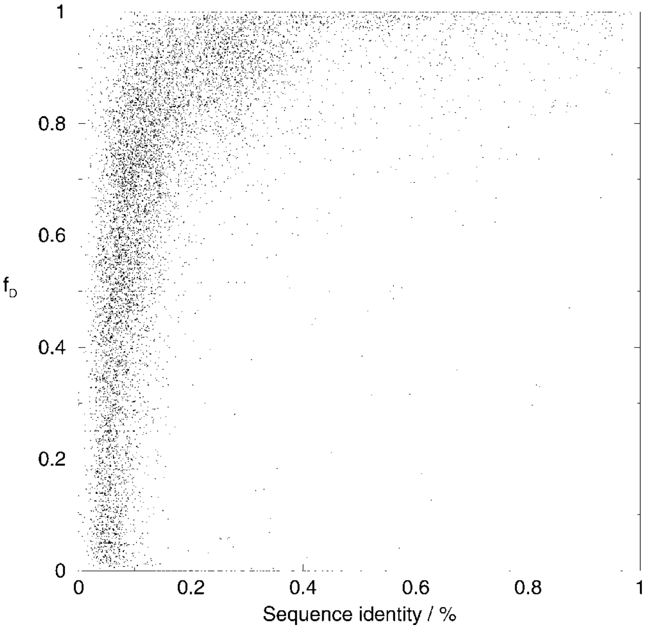

Fig. 5. Comparison of CE and FSSP structure alignments. The fraction of residue pairs in each CE alignment that are present in each FSSP alignment for each SCOP pair is plotted as a function of sequence identity (according to CE's structure alignment).

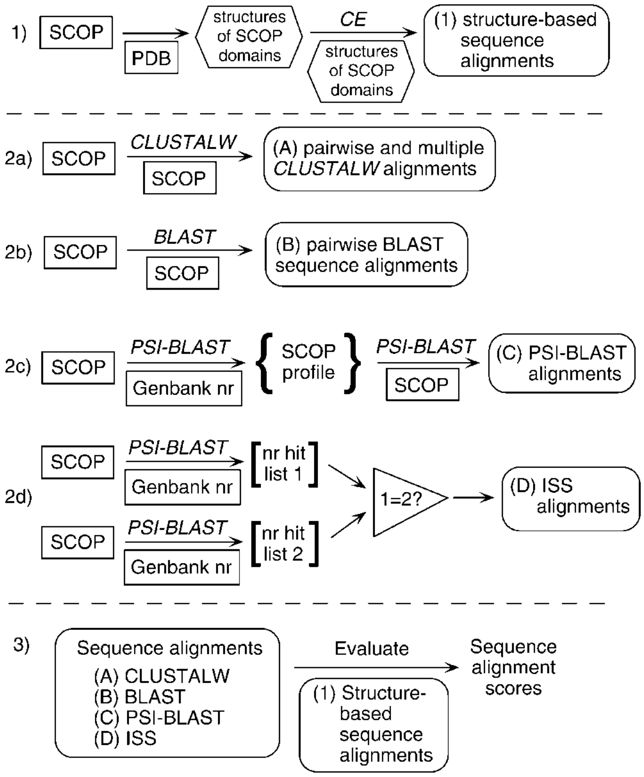

Fig. 6. Schematic representation of procedures used for fold assignment and comparing sequence alignment methods against structure alignments. The sequence databases "GenBank nr" and SCOP are pictured in rectangles. The SCOP structure database is shown as a hexagon. Sets of sequence alignments are shown in rounded ovals. Line 1 shows the procedure for deriving the structure-based sequence alignments using the CE program. Lines 2a–2d illustrate the usage of sequence search programs for fold assignment and sequence alignment. Line 3 describes the comparison of the structure and sequence alignments.

also tested the global alignment program CLUSTALW.[49,50] In Figure 6, the procedures we used for deriving sequence alignments with these methods are shown schematically.

In Table I, we list the number of homologous pairs identified with each method at the family and superfamily level in 5% intervals of sequence identity. The total number of pairs in each row is given in the pairwise CLUSTALW column ("CPAIR," line 2a in Fig. 6), because the global alignment program will align any two sequences at least partially. The CMULT column contains the number of pairs aligned by grouping the sequences in each family or superfamily and aligning the group using CLUSTALW to form a multiple alignment. These numbers differ from CPAIR because, in a handful of cases, this resulted in sequence alignments with no residues actually aligned between a given pair. The set of superfamily relationships is a good test of remote homology detection, because 97% of the sequence pairs (10,364 out of 10,665) have ≤20% sequence identity. Gapped BLAST (line 2b in Fig. 6; the column labeled "BLAST" in Table I) was able to identify only 8.0% of the superfamily pairs (850 out of 10,665) with E-values of 10.0 or better. To test PSI-BLAST, we used each SCOP sequence as a query to the nonredundant protein sequence database (*nr*) available from NCBI for a total of four iterations each. The final position-specific similarity matrix was saved, and used to search our SCOP 1.38 sequence database (line 2c in Fig. 6). Hits in the same family or superfamily were identified with E-values better than 10.0. In some cases, the profile for sequence "A" would identify the related sequence "B," but the profile for sequence "B" would not identify sequence "A." The numbers in the PSIBL column are the total number of relationships identified in both directions divided by 2 (and rounded up if necessary). PSI-BLAST is able to detect over 2 times as many superfamily pairs (17%) as BLAST at the same E-value cutoff.

We tested an ISS method based on PSI-BLAST, shown in line 2d in Figure 6. The results of the PSI-BLAST searches against *nr* for each SCOP sequence were saved. Any *nr* sequence found in each of two members of the same family or superfamily was used to identify a match between the sequences, as long as the resulting alignment of the SCOP sequences via the intermediate was at least 30 amino acids in length. Again, a cutoff E-value of 10.0 was used, but in this case, the E-value is based on the size of the *nr* database and not the SCOP database, on which the BLAST and PSI-BLAST E-values were based. This ISS procedure was able to find 38.2% of the superfamily pairs. ISS identifies 98.7% of all SCOP family level alignments and 95.3% of all superfamily level alignments having sequence identity greater than 20%. In addition, this method found 63.9% of the family pairs with less than 20% sequence identity (2,724 out of 4,264). PSI-BLAST finds

TABLE I. The Number of Comparisons at the Family and Superfamily Level for Each Sequence Alignment Method†

	Family					Superfamily				
Identity	CPAIR	CMULT	BLAST	PSIBL	ISS	CPAIR	CMULT	BLAST	PSIBL	ISS
0–4%	464	463	17	29	78	2570	2570	47	141	485
5–9%	1288	1288	63	220	517	5128	5128	110	482	1581
10–14%	1324	1324	270	732	1027	2135	2135	221	595	1294
15–19%	1188	1188	730	1031	1102	531	531	211	317	427
20–24%	1051	1050	941	1002	1016	151	151	112	127	137
25–29%	995	994	976	966	984	89	89	88	87	89
30–34%	688	688	685	677	676	45	45	45	43	45
35–39%	502	502	502	500	502	9	9	9	9	9
40–44%	354	354	354	353	354	4	4	4	4	4
45–49%	225	225	225	225	225	1	1	1	1	1
50–54%	213	213	213	213	213	2	2	2	2	2
55–59%	179	179	179	179	179	0	0	0	0	0
60–64%	155	155	155	155	155	0	0	0	0	0
65–69%	147	147	147	147	145	0	0	0	0	0
70–74%	109	109	109	109	107	0	0	0	0	0
75–79%	79	79	79	79	79	0	0	0	0	0
80–84%	101	101	101	101	101	0	0	0	0	0
85–89%	85	85	85	85	84	0	0	0	0	0
90–94%	86	86	86	86	86	0	0	0	0	0
95–99%	14	14	14	14	14	0	0	0	0	0
100%	1	1	1	1	1	0	0	0	0	0
Total:	9248	9245	5932	7241	7645	10665	10665	850	2462	4074

†Pairwise CLUSTALW (CPAIR), multiple alignment CLUSTALW (CMULT), BLAST, PSI-BLAST (PSIBL), and ISS. The number of CE structure alignments is identical to the number of pairwise CLUSTALW alignments. Only the numbers of comparisons not involving an immunoglobulin variable domain are shown. The total number of CE and CLUSTALW comparisons including immunoglobulin variable domains is 32,039 at the family level and 29,925 at the superfamily level. The small number of alignments in the 95–100% sequence identity range result from the use of SCOP95, which only includes sequences that have less than 95% pairwise identity.

47.2% of the family pairs, and BLAST finds 25.3% below 20% identity.

It should be noted that since our aim was to test alignment quality and not detection ability, we did not evaluate the false positive rates for any of the methods. In many cases, despite poor E-values, functional information may be sufficient to indicate a correct match, even in the presence of false positives with better E-values. Another complication is that a match between sequences in different SCOP fold classifications is often taken as a false positive.[14,19,48] This assumption, however, is not always valid. This is especially true of several of the Rossmann type folds, some of which are in fact quite likely to be homologous. They exist in separate fold categories because the central β-sheet with strand order 32145 is surrounded by varying numbers of α helices on each side and in some cases additional strands are added to the sheet. Nevertheless, our results suggest that this implementation of ISS is at least as sensitive as other ISS implementations.[14,15,24]

Comparing Sequence Alignments With Structural Alignments

To test the quality of sequence alignments from BLAST, PSI-BLAST, ISS, and CLUSTALW, we compared the aligned residue pairs in these alignments with those from the structure alignments from CE. We defined a residue pair as correctly aligned in the sequence alignment if the same pair is also aligned in the corresponding structure alignment. We assessed the quality of the sequence alignments in two different ways. The first we refer to as the "modeler's viewpoint" and the second we refer to as the "developer's viewpoint." From the modeler's viewpoint, the quality of the sequence alignment (from BLAST, PSI-BLAST, etc.) is measured by the score, f_M, which is the number of amino acids correctly aligned in the sequence alignment divided by the total number of aligned residues in the sequence alignment. As an example, we might have the following alignments from CE and BLAST:

```
CE alignment   BLAST alignment
IEYFGPVEEV      IEYF-GP-VEEV
VEFFSPALQG      VE-FFSPALQG-
                **    **
```

In this case, the sequence alignment has eight pairs aligned, but only four of these pairs are also aligned in the structure alignment (marked with asterisks). f_M is therefore 50%, representing the fraction of the residues in a model built from the sequence alignment that can be considered "correctly" modeled. This score assesses the quality of the model as given by the alignment, and therefore is the number a modeler using the sequence alignment would want to know, in the absence of knowing the actual structure alignment. From the developer's viewpoint, a similar score, f_D, is calculated as the ratio of the number of residues correctly aligned in the sequence

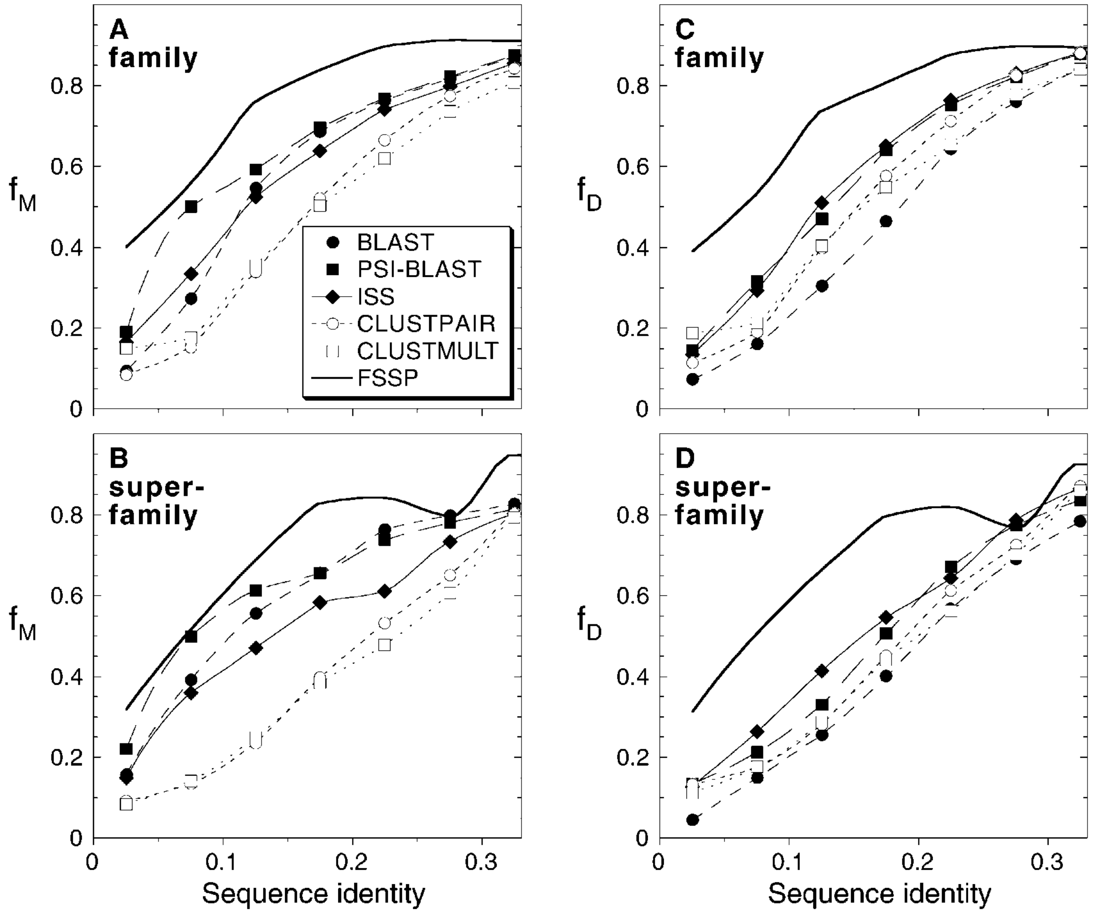

Fig. 7. Fraction of correctly aligned residues as a function of sequence identity for pairwise and multiple alignment CLUSTALW, pairwise Gapped BLAST, PSI-BLAST, ISS, and FSSP. The results are shown at the family (**A**,**C**) and superfamily (**B**,**D**) level as defined by SCOP. The identities were binned every 5% and the fractions were normalized according to the number of family/superfamily alignments. Panels A and B show the fraction, f_M, from the modeler's viewpoint—the fraction of correctly aligned residues. Panels C and D show the fraction, f_D, from the developer's viewpoint—the fraction of the alignment correct as judged by the structural alignment.

alignment, divided by the total number of aligned residues in the structure alignment. In the example given, f_D is 40% because only four of the ten pairs in the CE alignment are aligned in the BLAST alignment. This is the developer's view, because it entails determining what percentage of the best possible model that can be built could be built from the sequence alignment. In other words, how similar is a model built from the sequence alignment compared to a model built using the structure-based sequence alignment? As developers of a homology modeling strategy, we want the model to be both as accurate as possible and include the maximum number of residues from the original sequence. We are therefore very interested in high values of f_D. The two scores, f_M and f_D, can differ significantly for the same protein pair. For example, this happens when the sequence alignment algorithm being tested (such as BLAST) provides only a short (but accurate) alignment between the two proteins. In this case, f_M might be near 1.0, whereas f_D might be well below 0.5.

The average modeler and developer scores for each method are plotted in Figure 7A and B and 7C and D, respectively, as a function of sequence identity. The results are shown at both the family (Fig. 7A and C) and superfamily (B and D) levels. Because the number of sequences and aligned pairs varies substantially among SCOP families (and superfamilies), we calculated average f_M and f_D scores for each family/superfamily in each sequence identity range, and then averaged these scores over the families/superfamilies represented in the sequence identity interval. For example, if $f_M(A, I, f, p)$ is the f_M score for sequence pair p in family f using alignment method A with a structure-derived sequence identity (from CE) in the range $I \pm 2.5\%$, then $f_M(A, I)$, the weighted average score for alignment method A in sequence identity interval I, is given by the expression:

$$f_M(A, I) = \frac{1}{N_{\text{families}}(A, I)} \sum_{f=1}^{N_{\text{families}}(A,I)} \frac{\sum_{p=1}^{N_{\text{pairs}}(A,I,f)} f_M(A, I, f, p)}{N_{\text{pairs}}(A, I, f)} \quad (1)$$

$f_M(A, I)$ can range from 0 to 1, where a value of 1 means complete agreement of the sequence and structure alignments.

We note first that the FSSP alignments are substantially more similar to the CE alignments at almost all levels of sequence identity below 30%, when compared to BLAST, PSI-BLAST, ISS, and CLUSTALW. Above 30%, both FSSP and the sequence alignment methods are in agreement with the CE alignments. Even at 10–15% sequence identity, f_D scores for FSSP are in the range of 0.7–0.8, compared to sequence alignment f_D scores in the range of 0.2–0.5. This indicates that for these sequence alignment methods, there is substantial room for improvement. These results are in marked contrast to those of Gotoh, where two different sets of structure alignments showed more variability than the comparison of sequence and structure alignments.[27] We also note that the dip in the superfamily FSSP curves in Figure 7C and D is because of the small number of superfamilies in the 25–30% (five superfamilies) and 30–35% ranges (three superfamilies). This occurs because the superfamily relationships are very distant and there are few superfamily pairs at these levels of identity.

Among the sequence alignment methods in Figure 7A and B, the largest distinction is between global and local alignment methods. CLUSTALW consistently performs worse than BLAST and PSI-BLAST in the "twilight zone" of low sequence identity. At these low levels of similarity, the global alignment algorithm is forced to align regions that may not share structural homology. A local alignment algorithm, by contrast, will choose not to align this region. Multiple alignments of every SCOP sequence within a given superfamily fail to improve the performance of CLUSTALW by a significant degree.

PSI-BLAST and ISS, the two profile-based local alignment methods, have the largest fraction of correct residues aligned. At 10–15% identity, PSI-BLAST correctly assigns almost 25% more residues than pairwise CLUSTALW at the family level and 38% more residues at the superfamily level. From the modeler's viewpoint, PSI-BLAST performs better than ISS at almost every level of sequence similarity. BLAST does not do nearly as well as PSI-BLAST below 10% sequence identity, but above 20% identity there is no distinction between BLAST and PSI-BLAST regarding their ability to align residues correctly. There is a difference, however, with regard to the length of the alignments that each method produces, and this becomes very apparent from the developer's viewpoint (f_D). Not only is it essential to align target-parent residues correctly during homology modeling, but it is also important to achieve the longest alignments possible. The fraction f_D represents the portion of the target model that is correct compared to the best possible model that can theoretically be built based on the parent–target structure alignment. A model based on a short alignment may have several regions of secondary structure that are correct (good f_M), but if it represents less than half of the protein (f_D <0.5), this would be an unsatisfactory model. Figure 7C and D compare each alignment method from the developer's viewpoint. The distinction is no longer "global" vs. "local" as it was from the modeler's perspective, but it is now pairwise vs. profilewise. The most obvious difference is that pairwise BLAST performs the worst; its f_D values are consistently lower than CLUSTALW. Although BLAST scores very well from the modeler's perspective (its alignments are mostly correct), it seems to truncate its alignments prematurely (BLAST alignments are always shorter than pairwise CLUSTALW).

ISS alignments have higher f_D scores at very low sequence identity than PSI-BLAST alignments, even though their f_M scores are lower. The reason is that the ISS alignments are significantly longer than the PSI-BLAST alignments. In any ISS method, we are faced with the problem of which sequence to choose as the intermediate among the set of *nr* sequences that the proteins to be aligned have in common. These can be chosen either by the best E-values between the intermediate and the query and target sequences, or by choosing the intermediate that produces the longest alignment between the query and target sequences. In our ISS method, we chose the intermediate sequence from the *nr* sequence hits that produced the longest alignment of the SCOP sequences. The choice of intermediate sequence has not been addressed significantly before in descriptions of ISS methods, although it is important in determining the final sequence alignment.

An interesting consequence of the results in Figure 7 is that the superfamily alignments are much more difficult than the family alignments, at comparable sequence identities. Because superfamily relatives (i.e., pairs in different families) are usually proteins with different functions or substrates, their sequences contain more insertions and deletions for the same level of sequence similarity than do family relatives.

In Figure 7, the sequence alignment methods are compared using all alignments that each method was able to detect. All related sequence pairs were supplied to CLUSTALW, totaling 20,000 alignments (excluding the immunoglobulin variable domains). But the methods based on BLAST and PSI-BLAST are only able to identify a fraction of the total number of alignments, so Figure 7 may be biased by which sequence pairs each method finds and the total number of pairs in each sequence identity range (given in Table I). In Table II,[10] we compare the alignment lengths and alignment quality for each possible pair of alignment methods over the sequence pairs they have in common. These numbers are calculated from the average log ratios of alignment length, f_M, and f_D. For example, if we define $N(A_1, A_2)$ as the number of homologue pairs two sequence alignment methods, A_1 and A_2, have jointly identified, then the average ratio of their f_M values for these sequence pairs (indexed by p) is given by Eq. 2:

$$r_M(A_1, A_2) = \exp\left(\frac{1}{N(A_1, A_2)} \sum_{p=1}^{N(A_1,A_2)} \ln\left(\frac{f_M(A_1, p)}{f_M(A_2, p)}\right)\right) \quad (2)$$

The averaging of logs is necessary, since ratios of the f_Ms would not produce a normal distribution, whereas the distribution of the log-ratios is approximately normal. In Table II, both the mean and median of the log-ratios are listed for each pair of methods (in both directions to facilitate comparison of one method with another). We also

TABLE IIA. Comparison of Sequence Alignment Methods at the SCOP Family Level (for Alignments With ≤30% Identity)[†]

			Alignment length			f_M			f_D		
Method 1	Method 2	Number	Mean	Med.	Prop.	Mean	Med.	Prop.	Mean	Med.	Prop.
ClustPair	ClustMult	2,164	1.03	1.01	0.87	1.06	1.00	0.50	1.09	1.00	0.52
	BLAST	2,875	1.36	1.19	1.00	0.87	0.95	0.30	1.19	1.10	0.83
	PSIBLAST	3,767	1.18	1.05	0.99	0.79	0.89	0.18	0.93	0.97	0.39
	ISS	2,032	1.11	1.02	0.91	0.77	0.91	0.23	0.85	0.95	0.31
	FSSP	1,332	1.18	1.12	1.00	0.51	0.66	0.03	0.60	0.76	0.12
ClustMult	ClustPair	2,164	0.98	0.99	0.13	0.94	1.00	0.50	0.92	1.00	0.48
	BLAST	1,340	1.32	1.16	0.96	0.83	0.94	0.33	1.09	1.09	0.70
	PSIBLAST	3,683	1.14	1.03	0.88	0.77	0.93	0.25	0.88	0.98	0.40
	ISS	3,260	1.06	1.01	0.66	0.77	0.95	0.33	0.82	0.96	0.37
	FSSP	2,827	1.16	1.11	0.95	0.43	0.68	0.04	0.49	0.78	0.14
BLAST	ClustPair	2,875	0.73	0.84	0.00	1.15	1.06	0.70	0.84	0.91	0.17
	ClustMult	1,340	0.76	0.86	0.04	1.20	1.06	0.67	0.91	0.92	0.29
	PSIBLAST	2,818	0.83	0.93	0.06	0.97	0.99	0.41	0.80	0.89	0.11
	ISS	1,313	0.77	0.87	0.01	1.03	1.01	0.54	0.79	0.87	0.13
	FSSP	779	0.80	0.92	0.27	0.83	0.86	0.15	0.67	0.74	0.07
PSIBLAST	ClustPair	3,767	0.85	0.95	0.01	1.27	1.12	0.82	1.07	1.03	0.61
	ClustMult	3,683	0.87	0.97	0.12	1.31	1.08	0.75	1.14	1.03	0.60
	BLAST	2,818	1.21	1.08	0.94	1.03	1.01	0.59	1.24	1.12	0.89
	ISS	3,683	0.89	0.97	0.01	1.08	1.02	0.68	0.96	0.99	0.42
	FSSP	2,145	0.92	1.01	0.58	0.83	0.87	0.14	0.77	0.88	0.19
ISS	ClustPair	2,032	0.90	0.98	0.09	1.31	1.10	0.77	1.18	1.06	0.69
	ClustMult	3,260	0.94	0.99	0.34	1.30	1.06	0.67	1.22	1.04	0.63
	BLAST	1,313	1.30	1.15	0.99	0.97	0.99	0.46	1.26	1.15	0.87
	PSIBLAST	3,683	1.13	1.03	0.99	0.93	0.98	0.32	1.04	1.01	0.58
	FSSP	2,043	1.03	1.06	0.81	0.75	0.80	0.07	0.77	0.86	0.17
FSSP	ClustPair	1,332	0.85	0.89	0.00	1.95	1.51	0.97	1.66	1.32	0.88
	ClustMult	2,827	0.86	0.90	0.05	2.35	1.47	0.96	2.03	1.29	0.86
	BLAST	779	1.24	1.09	0.73	1.21	1.16	0.85	1.50	1.35	0.93
	PSIBLAST	2,145	1.08	0.99	0.42	1.20	1.15	0.86	1.30	1.14	0.81
	ISS	2,043	0.98	0.95	0.19	1.34	1.24	0.93	1.31	1.16	0.83

[†]These values were calculated using Eq. 2 and 3 for the mean, median, and proportion of fractions greater than 1 of the log-ratios of the alignment length, f_M and f_D (see Results). ClustPair refers to CLUSTALW pairwise alignments, ClustMult to CLUSTALW multiple sequence alignments, and BLAST to pairwise Gapped BLAST. PSIBLAST refers to alignments created by searching the SCOP database with a profile generated through multiple iterations against Genbank *nr*. ISS, intermediate sequence search, uses the PSI-BLAST results to link two SCOP sequences through an intermediate *nr* sequence. Alignments from the FSSP structure alignment database were analyzed in the same manner as the sequence alignment methods. The number of alignments in each comparison is indicated.

give the proportion of ratios greater than 1 for each pair of methods. This can be calculated from the expression:

$$\frac{\sum_{p=1}^{N(A_1, A_2)} \delta(f_M(A_1, p) > f_M(A_2, p))}{\sum_{p=1}^{N(A_1,A_2)} \delta(f_M(A_1, p) \neq f_M(A_2, p))} \qquad (3)$$

where the δ function is 1 if the expression in parentheses is true and 0 if it is false. That is, of the pairs they have in common for which their prediction scores are different ($f_M(A_1, p) \neq f_M(A_2, p)$), for what fraction does method A_1 perform better than method A_2? A value of 0.5 means the methods behave equally well. A value greater than 0.5 means method A_1 performs better than A_2 more often than A_2 performs better than A_1.

The values for average ratios of alignment lengths, f_Ms, and f_Ds in Table II confirm and extend the results in Figure 7. For instance, the average f_M ratio for BLAST and PSI-BLAST is equal to 0.97 in the family alignments and 1.02 in the superfamily alignments (only pairs with sequence identity below 30% are considered in both cases). Their mean alignment length ratios are only 0.83 and 0.77 whereas their $f_D$ratios are 0.80 and 0.79 respectively (see line 13 in Table IIA and IIB). This indicates that the profile aspect of PSI-BLAST does not improve alignments on a per-residue basis of the alignments (f_M), but instead provides longer alignments at approximately the same accuracy as BLAST. And although PSI-BLAST alignments are 8% more accurate than ISS alignments at the family level, the length of PSI-BLAST alignments are on average 89% of the ISS lengths, resulting in an average f_D ratio of 0.96. In the more difficult superfamily case, the numbers

TABLE IIB. Comparison of Sequence Alignment Methods at the SCOP Superfamily Level (for Alignments With ≤30% Identity)†

			Alignment length			f_M			f_D		
Method 1	Method 2	Number	Mean	Med.	Prop.	Mean	Med.	Prop.	Mean	Med.	Prop.
ClustPair	ClustMult	1,421	1.07	1.04	0.91	1.13	1.02	0.52	1.21	1.07	0.57
	BLAST	493	2.15	1.70	1.00	0.58	0.69	0.15	1.24	1.14	0.81
	PSIBLAST	975	2.13	1.66	1.00	0.46	0.53	0.08	0.98	1.00	0.50
	ISS	1,093	1.35	1.21	0.99	0.56	0.60	0.17	0.76	0.79	0.31
	FSSP	1,154	1.34	1.30	1.00	0.28	0.29	0.05	0.38	0.39	0.10
ClustMult	ClustPair	1,421	0.93	0.97	0.09	0.88	0.98	0.48	0.82	0.93	0.43
	BLAST	210	2.12	1.60	0.99	0.55	0.67	0.21	1.16	1.18	0.72
	PSIBLAST	860	2.06	1.63	0.97	0.47	0.52	0.10	0.96	1.00	0.50
	ISS	939	1.27	1.15	0.89	0.59	0.68	0.20	0.75	0.83	0.33
	FSSP	2,106	1.29	1.27	0.97	0.23	0.27	0.04	0.29	0.33	0.07
BLAST	ClustPair	493	0.46	0.59	0.00	1.73	1.44	0.85	0.80	0.87	0.19
	ClustMult	210	0.47	0.63	0.01	1.82	1.49	0.79	0.86	0.85	0.28
	PSIBLAST	472	0.77	0.84	0.18	1.02	1.01	0.54	0.79	0.87	0.15
	ISS	290	0.50	0.63	0.02	1.32	1.16	0.66	0.66	0.76	0.15
	FSSP	171	0.58	0.72	0.18	0.92	0.84	0.33	0.54	0.54	0.12
PSIBLAST	ClustPair	975	0.47	0.60	0.00	2.18	1.90	0.92	1.02	1.00	0.50
	ClustMult	860	0.49	0.61	0.03	2.13	1.90	0.90	1.04	1.00	0.50
	BLAST	472	1.30	1.19	0.82	0.98	0.99	0.46	1.27	1.15	0.85
	ISS	1,282	0.55	0.69	0.04	1.47	1.26	0.78	0.81	0.92	0.34
	FSSP	776	0.60	0.74	0.29	0.86	0.83	0.30	0.52	0.52	0.12
ISS	ClustPair	1,093	0.74	0.83	0.01	1.77	1.67	0.83	1.31	1.26	0.69
	ClustMult	939	0.79	0.87	0.11	1.69	1.46	0.80	1.33	1.20	0.67
	BLAST	290	1.98	1.59	0.98	0.76	0.86	0.34	1.51	1.32	0.85
	PSIBLAST	1,282	1.81	1.45	0.96	0.68	0.79	0.22	1.24	1.09	0.66
	FSSP	1,005	0.95	1.05	0.59	0.51	0.51	0.08	0.48	0.50	0.08
FSSP	ClustPair	1,154	0.74	0.77	0.00	3.57	3.39	0.95	2.66	2.54	0.90
	ClustMult	2,106	0.78	0.79	0.03	4.42	3.77	0.96	3.43	3.03	0.93
	BLAST	171	1.71	1.40	0.82	1.08	1.19	0.67	1.86	1.84	0.88
	PSIBLAST	776	1.66	1.35	0.71	1.17	1.21	0.70	1.94	1.92	0.88
	ISS	1,005	1.05	0.96	0.41	1.97	1.96	0.92	2.06	2.00	0.92

are more extreme: average f_M, alignment length, and f_D ratios are 1.47, 0.55, and 0.81, respectively (line 19).

We speculated that in regions of a structure alignment where aligned residues have the same secondary structure type (helix, sheet, or coil), the structure alignment is more likely to be correct. These regions presumably represent corresponding elements of the protein topology and should show higher sequence conservation than other regions. To examine this, we calculated the f_D scores for only those residue pairs in the structure alignments that were of the same secondary structure type. These values are plotted in Figure 8 for each sequence alignment method and each secondary structure type. The plots resemble the general f_D plots in Figure 7C and D. The alignment of sheet residues is slightly worse than the alignment of helix and coil residues for all methods below 20% sequence identity.

DISCUSSION

Accurate sequence alignment remains the largest challenge to successful homology modeling efforts at low sequence identity. The BLAST programs, including the recently developed PSI-BLAST profile alignment program, are the most commonly used programs for database searches. PSI-BLAST in particular has been advocated as a powerful remote homology detection tool for the purpose of fold assignment and comparative modeling.[51] Intermediate sequence searches have also been recognized as powerful sequence-based fold assignment methods.[14–17,24,52] It is therefore of interest to gauge the sequence alignment quality of these methods as well as their detection ability. We have tested local alignment methods often used for fold assignment by sequence comparison, including Gapped-BLAST, PSI-BLAST, and an intermediate sequence search procedure. For comparison purposes, we also tested CLUSTALW in a pairwise fashion and with multiple sequences consisting of the sequences in each SCOP family or superfamily.

We have used two measures of alignment quality. The first, f_M, measures the fraction of each sequence alignment that is correct. The second, f_D, measures the fraction of each structure alignment reproduced correctly in the sequence alignment. The difference is in the denominator. In the first case, it is the number of aligned pairs in the sequence alignment from CLUSTALW, BLAST, PSI-BLAST, or ISS. In the second case, it is the number of aligned pairs in the structure alignments from CE. High

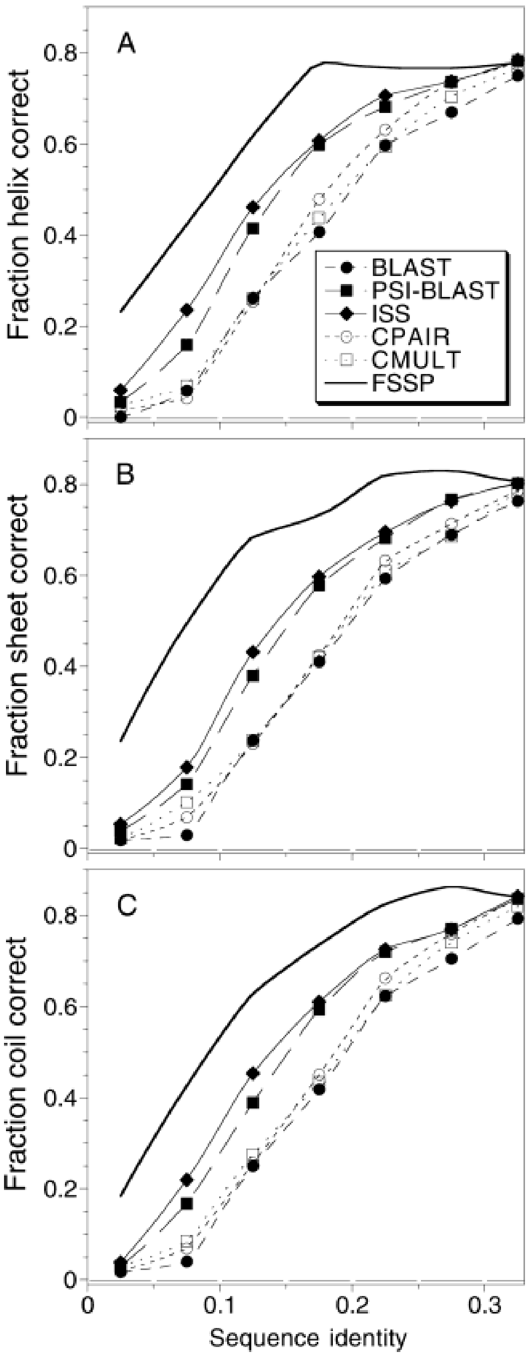

Fig. 8. Fraction of correct secondary structure aligned for each sequence alignment method. The fraction of residues correctly aligned in (**A**) helical, (**B**) sheet, and (**C**) coil regions according to each sequence alignment technique is judged based on whether those residues are matched in the structural alignment and share the same secondary structure type.

f_M but low f_D indicates underprediction; that is, only a portion of the possible number of aligned residues were actually aligned. High f_D but low f_M indicates overprediction; much of the structure alignment is reproduced in the sequence alignment but non-equivalent segments are also aligned.

Because distantly related proteins can contain many insertions as well as N- and C-terminal extensions, a key to good alignment is to align just the right amount of two sequences. Our CLUSTALW results indicate that global alignment programs perform poorly at low sequence identity because they align too much. These results are in accord with those of Thompson et al.[28] who used their BAliBASE benchmark to test global and local multiple sequence alignment methods against structure alignments. Although their test sets are significantly smaller than ours, they found that in situations where there were N- and C-terminal extensions or large insertions, local methods performed better than global methods. In other cases, they found that global alignments performed better. At very low sequence identity, however, the case of significant extensions and insertions may be more common than the contrary situation. Local alignment methods are therefore necessary. Although global methods align too much, local alignments can sometimes align too little. We found that ISS alignments were longer than PSI-BLAST profile alignments, which were longer than BLAST alignments. This is also the order of their f_D scores, indicating that manual efforts to extend alignments are generally worthwhile. This is in accord with our CASP3 results.[53] The ISS results are particularly intriguing because this method is in common use via the Entrez database[54] as well as in a number of other implementations. They indicate that even when a homologue is found reliably with BLAST or PSI-BLAST there may still be some advantage in using intermediate sequences to produce a longer alignment and a more complete model.

It might be argued that alignments that are only 50% correct (at 10–15% sequence identity) might have little practical utility. We argue that, in the context of biological analysis, they may be very useful. In most cases, the alignment produced appears uneven. Some regions are conserved with few insertions whereas other regions exhibit little similarity and large gaps. Even if the precise placement of insertions and deletions is not correct, they are usually close enough to indicate whether there is substantial change in structurally or functionally important parts of the protein. Conserved catalytic and substrate-binding residues provide evidence that the function of the remote homologue is preserved. Insertions near the active site, in contrast, indicate that the substrate and function are likely to be quite different. A great deal of biological insight can be gained from these widely available methods.

Remote homologue detection methods have been used to assign folds to sequences in completed genomes.[51,55–57] The results of these studies have two important conclusions of relevance to the field of homology modeling. First, approximately one-third of genomic sequences can be assigned a fold from the PDB, and in principle a homology model of these proteins can be produced. This presents enormous opportunities for modeling. Second, the average sequence identity between a genome sequence and its homologue of known structure is less than 25%. It is therefore important to test the sequence alignment quality of these programs at low sequence identity. In this work, we have assessed the sequence alignment accuracy of two of the more popular remote homologue detection methods, PSI-BLAST and an intermediate sequence search method, and compared these with older technologies based on pairwise local and global alignments.

The use of a large test set (over 20,000 pairwise alignments, excluding immunoglobulin variable domains) has enabled us to analyze the performance differences rigorously. This benchmark will be used to assess other sequence alignment methodologies presently available, and can also be used to assess alignment parameters (gap penalties, amino acid similarity matrices) and any future improvements in alignment algorithms.

Anyone interested in using the benchmark described in this paper to assess the alignment quality of a sequence-sequence or sequence–structure alignment method is urged to contact the authors.

MATERIALS AND METHODS

General

All calculations were performed on Silicon Graphics workstations with 195 MHz R10000 processors. The protein structures were taken from the Protein Data Bank (PDB),[58] now available at http://www.rcsb.org/pdb. Figures were generated using Xmgr 4.1.2, S-PLUS 3.4 (http://mathsoft.com/splus), and Kaleidagraph (http://synergy.com).

Protein Sequence Alignments

The sequences of proteins of known structure were chosen from version 1.38 of the SCOP (Structural Classification of Proteins) database,[23] based on the August 1998 release of the PDB. All 3,199 sequences comprise structural domains and have less than 95% sequence identity with each other (PDB95D, obtained at http://scop.mrc-lmb.cam.ac.uk/scop/). The SCOP database divides the proteins into superfamilies and families depending on the evolutionary relationship between the protein sequences. Although version 1.38 of SCOP was never fully released, we created a modified version of this database, SCOP95NEW, containing 2,622 sequences obtained from SEQRES records in PDB entries, because some SCOP 1.38 sequences were missing modified residues such as selenomethionines. SCOP95NEW has 18% fewer entries than the original PDB95D, because the following entries were removed: (1) domains involving multiple chains; (2) domains involving interrupted sequence segments; (3) entries without a structure in the PDB, including obsolete PDB entries; and (4) domains in the "small protein," "peptide," or "designed protein" SCOP classes. This modified database is available from the authors through the web site given at the end of this paper.

CLUSTALW 1.7[49] was used to perform global pairwise and multiple alignments. Default parameters were used, including the PAM250 matrix. All-against-all pairwise alignments were performed for every sequence in a SCOP family and superfamily for which there was more than one entry (9% of the SCOP sequences had no relative within the same superfamily). Multiple alignments were performed for all sequences in each of 323 different superfamilies and 588 different families (61,964 comparisons in all). The multiple alignments were repeated using CLUSTALW 1.8 with the BLOSUM62 matrix and a 25% identity cutoff for delaying the incorporation of divergent sequences into the alignment. The quality of the alignments, as judged by f_M and f_D, was almost identical to those reported here. Preliminary results using secondary structure-dependent gap penalties with pairwise CLUSTALW alignments indicate, rather surprisingly, little improvement in alignment quality.

For the local pairwise sequence alignments, each sequence was compared to SCOP95NEW using BLAST 2.0.8.[59] The PSI-BLAST comparisons were performed in the following manner. First, we downloaded a version of the non-redundant protein database (*nr*) available from NCBI (ftp.ncbi.nlm.nih.gov/blast/db/). We masked low-complexity sequence regions with the program SEG[60] using a window size of 20. This is larger than the default setting of 12. We have found that the lower setting removed many short segments contributing to the significance of matches without seriously increasing the number of false positives. We call this filtered database *nrhc*. Second, each SCOP sequence was iteratively searched against our *nrhc* database. The PSI-BLAST E-value cutoff (parameter-h) was 0.0001 for inclusion of a *nrhc* sequence into the position-specific matrix. The resulting checkpoint file, or matrix, was searched against SCOP95NEW. We limited PSI-BLAST to four iterations to prevent possible matrix migration and subsequent pollution by unrelated sequences.[24]

Sequences identified by BLAST and PSI-BLAST were considered for modeling if the E-value was less than 10 and the target (query) and parent (hit) sequences shared the same SCOP superfamily. The E-value represents the expected number of occurrences of random matches having a given score. The permissive E-value score is justified because proteins within the same superfamily are, by SCOP definition, related.

Intermediate sequence searches (ISS) were performed using the *nrhc* sequences identified by PSI-BLAST as homologues of the target and parent sequences. A library of *nrhc* accession names was created for each SCOP sequence. These libraries can be rapidly compared to identify intermediate sequences common to any two SCOP sequences being aligned. The target–intermediate and parent–intermediate alignments were used to generate a target–parent alignment for each intermediate in common. We used the longest such parent–target alignment resulting from the intermediate sequences. This implementation of ISS is guaranteed to be more sensitive than PSI-BLAST and able to identify a greater number of homologous sequences. Using multiple intermediate sequences can extend the method even further.[16]

Protein Structure Alignments

The domain definitions from version 1.38 of SCOP were used to extract the PDB coordinates comprising each SCOP domain. The structural alignments were performed using the CE program (Combinatorial Extension of the alignment path algorithm).[31] We ran the program locally on our workstations using an executable supplied by the authors. To compare both sequence and structure alignments, it was necessary to correlate the residue numbering between (1) the SCOP sequences; (2) the PDB SEQRES sequences; and (3) the PDB coordinate (ATOM) records. This was nontrivial, because there are numerous entries in the PDB files where the numbering of the residues in the coordinate (ATOM) records does not match the residue position given in the SEQRES records. We have a database available correlating the SEQRES and ATOM numbering (http://www.fccc.edu/research/labs/dunbrack/s2c).

For each family in SCOP95NEW with more than one entry, a structural alignment using the CE program was calculated for each distinct pair of protein domains. These alignments are designated as family level alignments. Each superfamily was then examined. A structural alignment was calculated for each distinct pair of proteins in a superfamily that were not also in the same family. These alignments are designated as superfamily level alignments. The CE program calculates the RMS deviation and Z-score for the alignment. It also determines a rotation and translation matrix to bring the two structures into alignment and a sequence alignment based on the structural alignment.

Comparison of FSSP With CE Alignments

The FSSP database[43] was downloaded by ftp from ftp.ebi.ac.uk/databases/fssp. Each FSSP file is created for a "representative" or parent sequence, and contains alignments with close relatives which are thereby represented by the parent sequence of the file, as well as with more distant proteins, which have their own files as well. The close relatives, or children, are not listed in other FSSP files. To compare CE to FSSP, we needed to produce FSSP alignments for each SCOP superfamily and family pair. If both SCOP sequences were parent sequences in FSSP, then the alignment was derived from the file by translating the coordinate numbers listed by FSSP into SCOP and SEQRES numbers, as described previously. If one or the other of the SCOP sequences was a child, or represented sequence, then a transitive alignment was performed. For instance, if both SCOP sequences were children, a Child1–Parent1 alignment was produced from the Parent1 FSSP file, and compared with a Parent1–Parent2 alignment, also from the Parent1 FSSP file, to produce a Child1–Parent2 alignment. This alignment was compared with a Parent2–Child2 alignment from the Parent2 FSSP file to produce a Child1–Child2 alignment. The FSSP alignments were compared with CE alignments in the same manner as alignments produced by BLAST, PSI-BLAST, etc.

For our benchmarking calculations, we calculated and stored certain quantities determined by the sequence alignment derived from the structural alignment. These include: (1) the length of each chain (n_1 and n_2); (2) the length of the portion of each chain in the aligned region. This excludes any residues on the ends of the chains that CE was unable to align; (3) the number of residue pairs aligned (n_p); (4) the number of identical residues aligned (e.g., Ile with Ile, Pro with Pro) (n_{id}); the sequence identity, SID = $100 \cdot n_{id}/n_p$; (6) the number of pairs where the secondary structure is matched or mismatched. Secondary structure was determined by the STRIDE program[61] and is classified as helix, sheet, or coil. Four quantities were calculated: the number of pairs where both residues have helical secondary structure (n_{HH}), the number of pairs where both residues have sheet secondary structure (n_{EE}), the number of pairs where both residues are in turns or random coil (n_{CC}), and the remainder where the secondary structure is mismatched (n_{mis}); (7) the alignment score (Z-score) as determined by the CE program; (8) the RMS deviation as determined by the CE program; (9)the fraction of the two chains aligned, $f_A = n_p/\min(n_1, n_2)$; and (10) the fraction of secondary structure correctly aligned, $f_{SS} = (n_{HH} + n_{EE} + n_{\mathrm{CC}})/n_p$.

Sequence Alignment Comparison

The sequence alignments, generated using CLUSTALW, BLAST, PSI-BLAST, and ISS, were assessed by calculating several quantities for each alignment, including: (1) the number of residue pairs aligned by the sequence alignment algorithm (n_A); (2) the number of residue pairs correctly aligned (n_C). A pair of residues is correctly aligned if the pair is also aligned in the structural alignment; (3) the percentage sequence identity (fraction of n_A pairs with identical residues); (4) f_M, the fraction of the sequence alignment that is correctly aligned, n_C/n_A; (5) f_D, the fraction of the structure alignment that is correctly aligned in the sequence alignment, n_C/n_p; (6) the number of correctly aligned residues where the secondary structure of both residues is either helix, sheet, or coil; (7) the fraction of the pairs of residues in the structural alignment that both have helix, sheet, or coil structure and where the sequence alignment is also correct.

Supplemental information may be found at http://www.fccc.edu/research/labs/dunbrack/.

ACKNOWLEDGMENTS

J.M.S. is supported by NIH Post-doctoral Training Grant CA09035 awarded to Fox Chase Cancer Center from the National Cancer Institute. We would like to thank Ilya Shindyalov for providing the CE program, and the Research Computing Services of Fox Chase Cancer Center for computer support.

REFERENCES

1. Blundell TL, Sibanda BL, Sternberg MJE, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 1987;326:347–352.
2. Greer J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 1990;7:317–334.
3. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
4. Browne WJ, North AC, Phillips DC. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 1969;42:65–86.
5. Greer J. Model for haptoglobin heavy chain based upon structural homology. Proc Natl Acad Sci USA 1980;77:3393–3397.
6. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based

modeling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Prot Eng 1987;5:377–384.
7. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J Mol Biol 1970;48:443–453.
8. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
9. Bowie JU, Clarke ND, Pabo CO, Sauer RT. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. Proteins 1990;7:257–264.
10. Hendlich M, Lackner P, Weitckus S, Flöckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. J Mol Biol 1990;216:167–180.
11. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
12. Bryant SH, Lawrence CE. An empirical energy function for threading protein sequence through the folding motif. Proteins 1993;16:92–112.
13. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Prot Sci 1996;5:947–955.
14. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:349–354.
15. Gerstein M. Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. Bioinformatics 1998;14:707–714.
16. Salamov AA, Suwa M, Orengo CA, Swindells MB. Combining sensitive database searched with multiple intermediates to detect distant homologues. Prot Eng 1999;12:95–100.
17. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
18. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 1987;84:4355–4358.
19. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
20. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Meth Enz 1990;183:63–98.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
22. Altschul SF, Gish W. Local alignment statistics. Meth Enz 1996;266:460–480.
23. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
24. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
25. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
26. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton J. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
27. Gotoh O. Significant improvement in accuracy of multiple sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 1996;264:823–838.
28. Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucl Acids Res 1999;27:2682–2690.
29. Barton GJ, Sternberg MJE. A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. J Mol Biol 1987;198:327–337.
30. Falicov A, Cohen FE. A surface of minimum area metric for the structural comparison of proteins. J Mol Biol 1996;258:871–892.
31. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Prot Eng 1998;11:739–747.
32. Singh AP, Brutlag DL. Hierarchical protein structure superposition using both secondary structure and atomic representations. Ismb 1997;5:284–293.
33. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
34. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323.
35. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Meth Enz 1996;266:617–635.
36. Feng Z, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Fold Des 1996;1:123–132.
37. Godzik A. The structural alignment between two proteins: is there a unique answer? Prot Sci 1996;5:1325–1338.
38. Šali A, Overington JP. Derivation of rules for comparative protein modeling from a database of protein structure alignments. Prot Sci 1994;3:1582–1596.
39. Gotoh O. Optimal alignment between groups of sequences and its application to multiple sequence alignment. Comput Appl Biosci 1993;9:361–370.
40. Gotoh O. Further improvement in methods of group-to-group sequence alignment with generalized profile operations. Comput Appl Biosci 1994;10:379–387.
41. Gotoh O. A weighting system and algorithm for aligning many phylogenetically related sequences. Comput Appl Biosci 1995;11:543–551.
42. Eddy SR. Multiple alignment using hidden Markov models. Ismb 1995;3:114–120.
43. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucl Acids Res 1998;26:316–319.
44. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucl Acids Res 1999;27:254–256.
45. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
46. Holm L, Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins. Nucl Acids Res 1996;24:206–209.
47. Pascarella S, Argos P. A data bank merging related protein structures and sequences. Prot Eng 1992;5:121–137.
48. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Prot Sci 1998;7:445–456.
49. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 1994;22:4673–4680.
50. Huggins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Meth Enz 1996;266:383–402.
51. Huynen M, Doerks T, Eisenhaber F, Orengo C, Sunyaev S, Yuan Y, Bork P. Homology-based fold predictions for *Mycoplasma genitalium* proteins. J Mol Biol 1998;280:323–326.
52. Holm L, Sander C. Protein folds and families: sequence and structure alignments. Nucl Acids Res 1999;27:244–247.
53. Dunbrack RL, Jr. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Protein (Suppl) 1999;3:81–87.
54. Marchler-Bauer A, Addess KJ, Chappey C, Geer L, Madej T, Matsuo Y, Wang Y, Bryant SH. MMDB: Entrez's 3D structure database. Nucl Acids Res 1999;27:240–243.
55. Fischer D, Eisenberg D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. Proc Natl Acad Sci USA 1997;94:11929–11934.
56. Salamov AA, Suwa M, Orengo CA, Swindells MB. Genome analysis: assigning protein coding regions to three-dimensional structures. Prot Sci 1999;8:771–777.
57. Rychlewski L, Zhang B, Godzik A. Fold and function predictions for *Mycoplasma genitalium* proteins. Fold Des 1998;3:229–238.
58. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res 2000;28:235–242.
59. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of database programs. Nucl Acids Res 1997;25:3389–3402.
60. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 1994;18:269–285.
61. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.