SAD—a normalized structural alignment database: improving sequence–structure alignments

Brian Marsden* and Ruben Abagyan

The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA

Received on August 24, 2003; revised on February 17, 2004; accepted on March 24, 2004
Advance Access publication April 15, 2004

ABSTRACT

Motivation: We present a structural alignment database that is specifically targeted for use in derivation and optimization of sequence–structure alignment algorithms for homology modeling. We have paid attention to ensure that fold-space is properly sampled, that the structures involved in alignments are of significant resolution (better than 2.5 Å) and the alignments are accurate and reliable.

Results: Alignments have been taken from the HOMSTRAD, BAliBASE and SCOP-based Gerstein databases along with alignments generated by a global structural alignment method described here. In order to discriminate between equivalent alignments from these different sources, we have developed a novel scoring function, Contact Alignment Quality score, which evaluates trial alignments by their statistical significance combined with their ability to reproduce conserved three-dimensional residue contacts. The resulting non-redundant, unbiased database contains 1927 alignments from across fold-space with high-resolution structures and a wide range of sequence identities.

Availability: The database can be interactively queried either over the web at http://abagyan.scripps.edu/lab/web/sad/show.cgi or by using MySQL, and is also available to download over the web.

Contact: marsden@scripps.edu

INTRODUCTION

There is a continuing expansion of the chasm between the quantity of known protein sequences and solved three-dimensional (3D) structures of these sequences. This reinforced the need for other predictive methods, such as homology modeling, to bridge this gap. Homology modeling critically depends upon the availability of one or more structural homologs and accurate sequence–structure alignments. While the number and breadth of available structures limit the former, the latter is dependent upon the alignment methods used. The recent CASP4 experiment has highlighted this aspect of homology modeling as an area of 'concern' (Tramontano et al., 2001) and 'THE bottleneck to improving the quality of the model' (Venclovas et al., 2001).

Sequence–structure alignments of homologous proteins are often generated using an iterative use of methods such as BLAST or PSI-BLAST (Altschul et al., 1997) to create profile alignments that span target and template (Peitsch, 1996; Venclovas, 2001). Threading algorithms and double-dynamic programming methods that use the structure and may or may not include information about sequence and secondary structure of the target and template(s) are also used (Taylor, 1997; Bates et al., 2001; Sali and Blundell, 1993; Orengo and Taylor, 1996). All these methods intrinsically depend upon some sequence-adjacent method that in turn uses a scoring matrix. This scoring matrix is often derived from other alignments without any regard to structural information. Even if such information is included at this stage or later, there is often little attempt to derive optimal weighting schemes for the sequence and structural components. Venclovas (2003) pointed out that sequence–structure alignments 'remain a significant hindrance'.

In order to derive such schemes, databases of 'standard-of-truth' structure–structure alignments are used to train sequence–structure alignment methods. There are currently a number of such databases available that might be used for this purpose (de Bakker et al., 2001; Gerstein and Levitt, 1998; Holm and Sander, 1993, 1999; Mallika et al., 2002; Marti-Renom et al., 2001; Mizuguchi et al., 1998; Shindyalov and Bourne, 1998, 2001; Thompson et al., 1999, 2001). Indeed, databases such as HOMSTRAD (de Bakker et al., 2001; Mizuguchi et al., 1998), PASS2 (Mallika et al., 2002) and that derived by combinatorial extension (CE) (Shindyalov and Bourne, 2001) have been derived specifically for generic method-derivation purposes. Although these databases cover fold-space well, they do not focus upon the quality of the structures included in the alignments and also work at the whole-chain rather than the domain level. Other databases, such as BAliBASE (Thompson et al., 1999, 2001) are limited...
in their coverage of fold-space. Clearly, a structural alignment of two proteins with poor resolution is likely to add noise to any method derivation. Similarly, the use of multidomain chains runs the risk of misalignment, especially if one or more of the domains in the chain are related, but not identical (in terms of domain structure classification), to one or more of the remaining domains in the chain.

A traditional structural alignment of two rigid structures is frequently wrong because structures are deformable. It is worth considering, for a moment, what we want from a reference alignment. Clearly we desire an alignment that as far as possible matches, in terms of spatial positions, equivalent residues between the two structures. For pairs of structures that are relatively compact and that remain in a very similar conformation, a structural alignment will most likely result in such a useful alignment. However, for structures that undergo plastic deformation (e.g. due to conformational changes occurring during binding or catalytic processes) such an alignment will most probably be wrong. Calmodulin is an extreme example of this where the binding of calcium causes a gross change in conformation. An example at the other end of the scale is that of tyrosine/serine/threonine kinases which, when bound to ATP, undergo much smaller changes in conformation (plastic deformations). At either end of this scale, such differences will most probably lead to errors in an alignment produced by a purely structural alignment method.

It is important to note that sequence alignments are necessarily immune to this phenomenon since they include no direct structural information in their methods whatsoever. Therefore, it might make sense to include the best of both worlds—the direct structural information from a pure structural alignment, and the local conformation-unbiased information of a pure sequence alignment.

How do we monitor the acceptability of any structural alignment? Analyses of the RMSD between the two structures, based upon a structural alignment, are often made during the derivation of structural alignment databases (SADs). RMSD calculations are necessarily dominated by contributions from large local structural differences between the two homologous structures, such as loop deformations or even more significant plastic deformations of subunits of the structures. This sensitivity leads to an over-emphasis upon plastic deformations over local structural conservations—i.e. conformational differences to the deficit of the conserved structural regions. This might lead to the acceptance of structural alignments that are sub-optimal.

A measure that is not as sensitive to such large differences is contact area difference (CAD) (Abagyan and Totrov, 1997). This method quantitatively evaluates the changes in residue–residue contacts between structurally equivalent residue pairs in both structures. While regions of structural differences contribute significantly to the RMSD value as the square of distances between the two structures, they do not over-contribute to the CAD value, since once a residue–residue contact has been broken, no matter how much farther the two residues are moved apart from each other the CAD contribution remains the same. Although it is currently impossible to provide structural alignments that are exact (Feng and Sippl, 1996; Godzik, 1996), the use of CAD, rather than any form of RMSD calculation, is more likely to be able to provide an independent estimate of the quality of the structural alignment of a pair of structures (Abagyan and Totrov, 1997).

In generating a database of structural alignments for methods derivation, a number of criteria should be met. The database must

1. Be non-redundant. No sequence should be represented in the database more than once.
2. Contain representatives from as many different folds as possible so that it may reflect fold-space as it is currently known. However, since the content of the Protein Data Bank (PDB) (Berman et al., 2002) is currently dominated by a number of folds such as IgG domains as well as globins and proteases, the database should be normalized so that no one fold-type is over-represented.
3. Contain a sufficient number of alignments to be statistically viable. Quality is important, but a small number of alignments in the database (i.e. low quantity) is unlikely to provide a good basis for derivation of novel alignment methods, for example.
4. Contain alignments from pairs of structures that have good resolution.
5. Contain alignments that are structurally significant. Alignments with only a small number of aligned pairs are not likely to be structurally significant.
6. Provide alignments with a good range of sequence identities. This will allow subsets of the database to be used to analyze the effect of sequence identity upon the efficacy of alignment methods.

In this paper, we describe the derivation of a new SAD, which satisfies the above criteria. SAD is a normalized database that contains high-quality structural alignments whose sequences are present in the PDB and can be used for optimization of sequence–structure alignment algorithms. We also describe a method for the estimation of CAD for homologous structures that we use to ensure that the alignments generated are of the best quality. In addition, we also introduce a new structure–structure alignment method that considers local sequence information as well as structural information during the derivation of structural alignments, thereby removing some error introduced by plastic deformations.

**MATERIALS AND METHODS**

**Component databases**

Structural alignments were taken from a number of publicly available databases, HOMSTRAD (de Bakker et al., 2001;
Structural alignment database for methods derivation

Mizuguchi et al., 1998) (June 1, 2001 version), BALiBASE (Thompson et al., 1999) and the Yale alignment database derived from SCOP (Gerstein and Levitt, 1998) (hereby known as SCOP-Gerstein). The BALiBASE database contains alignments that are made under different conditions such as the use of large insertions or deletions, low sequence identity or inclusion of orphan sequences. In addition, many of the sequences are not represented by structures in the PDB. Therefore, pairwise alignments of sequences with known structure were extracted from ‘Reference 1’ multiple sequence alignments of the database (no large insertions or deletions, sequences are of similar length).

In cases where the length of one or both sequences in an alignment from HOMSTRAD or BALiBASE were significantly longer than those in equivalent alignments in the SCOP-Gerstein database, a sub-alignment was extracted from the longer alignment(s) such that the sequences of this sub-alignment were identical to those found in the SCOP-Gerstein alignment. Entries with unusual amino acids or groups were also removed. Alignments were further only accepted if the resolutions of the associated structures were better than 2.5 Å.

The resulting alignments from the three databases were used as seeds for the generation of further structural alignments using the same pair of sequences. Two sources of information are used to derive the optimum global structural alignment (ICM global alignment) between two structures:

- The structural similarity in a 15-residue window between two fragments surrounding residues $i$ and $j$. This similarity is calculated as the local RMSD of the $\alpha$ atoms.
- The sequence similarity. The average local sequence alignment score using the ZEGA alignment method (Abagyan and Batalov, 1997) in a 15-residue window is calculated for $i, j$-centered pairs of fragments. ZEGA uses the Needleman and Wunsch algorithm (Needleman and Wunsch, 1970) that employs zero-end gap penalties.

The sequence similarity score for each window is added to the window-RMSD using a weight of 0.5 to produce a matrix. A dynamic programming method is then used upon this matrix to find the optimum alignment.

In addition to the resolution criteria, the probability of structural insignificance ($pP$) (Abagyan and Batalov, 1997) can be used to provide a measure of the probability that an alignment of two sequences is structurally insignificant. This function was derived via a statistical analysis of 1.3 million sequence alignments of structurally unrelated domains and has been shown to exhibit better discrimination between alignments compared with a simple alignment identity or similarity score. Alignments from any source were discarded if the $pP$ value was <6.0. At this level of alignment significance, the alignments may have up to 15–30% errors (Abagyan and Batalov, 1997) and therefore can be considered to be highly unreliable.

**Structural alignment quality**

Derivation of an estimate of the contact area difference between two non-homologous structures

The best way to evaluate the quality of a trial sequence–structure alignment is to look at the errors in the 3D model derived from it. The quality of this alignment can be evaluated directly in the 3D space by comparing the preservation of inter-residue contacts of imaginary homology model generated via the trial alignment with the true structure corresponding to the source sequence.

In order to estimate the difference in contact area of a residue position in an alignment, we assume that the contact area estimate, $\text{CAE}_{ij}$, between two residues, $i$ and $j$, is directly related to the distance, $d_{ij}$, between the two residues such that

$$\text{CAE}_{ij} = a_{ij}d_{ij} + b_{ij}. \quad (1)$$

where $I$ and $J$ are residue types (glycine, non-glycine) and $a$ and $b$ are to be parameterized. Here, we take $d_{ij}$ to be a projected point in space along the $\alpha-C\beta$ vector such that

$$d_{ij} = [(\overrightarrow{OA}_{i} + 1.5\overrightarrow{AB}_{i}) - (\overrightarrow{OA}_{j} + 1.5\overrightarrow{AB}_{j})], \quad (2)$$

where $\overrightarrow{AB}$ is the vector between the $\alpha$ and $C\beta$ atoms of a residue and $\overrightarrow{OA}$ is the vector from the origin to the $\alpha$ atom of the same residue.

For glycine residues, we use the $\alpha$ atoms as the measuring point rather than a projected point since glycine does not contain $C\beta$ atoms. Therefore, for glycine/non-glycine contacts where residue $i$ is a glycine

$$d_{ij} = |\overrightarrow{OA}_{i} - (\overrightarrow{OA}_{j} + 1.5\overrightarrow{AB}_{j})| \quad (3)$$

and for glycine/glycine contacts

$$d_{ij} = |\overrightarrow{OA}_{i} - \overrightarrow{OA}_{j}|. \quad (4)$$

The contact area of a pair of residues as a function of their distance, $d_{ij}$, was calculated for all residue pairs in all PDB structures whose resolution was better than 2.5 Å. Equation (1) was then parameterized for each type of pair of residues using linear regression. The mean of all parameters $a$ and $b$ [Equation (1)] for all non-glycine/non-glycine pairs and the mean of all glycine-containing pairs was then used for the ensuing work.

Given a pairwise sequence alignment that provides the mapping of equivalent residues between two non-homologous structures, the CAD estimate ($\text{CADE}$) of the two structures is then calculated in the same spirit as reported previously for CAD (Abagyan and Totrov, 1997) such that

$$\text{CADE} = \frac{100}{2} \sum_{i,j} \frac{|\text{CAE}_{ij} - \text{CAE}_{ij}^M|}{\overline{\text{CAE}}_{ij} + \text{CAE}_{ij}^M}, \quad (5)$$

where $\text{CAE}_{ij}^R$ is the estimated contact area of residues $i$ and $j$ for the reference structure and $\text{CAE}_{ij}^M$ for the trial structure.
The structural alignment algorithm, implemented in ICM by S. Batalov, was less likely to be structurally insignificant. The dependency of CADE upon the length of the alignment concerned was investigated by taking 455 random alignments from the SCOP-Gerstein whose structures had resolutions better than 2.5 Å and whose alignments showed pP values better than 6.0. The average CADE value of a window of varying alignment positions along the length of each alignment was calculated. The window sizes were taken as multiples of 20 starting from 20 to the nearest multiple of 20 of the full length of the alignment. To enable fair comparison of results between alignments, the values calculated in each window within an alignment were normalized by the overall CADE of that alignment.

**Derivation of structural alignment scoring function** The estimation of CAD via an alignment (CADE) provides information about the differences in packing between two structures, given an alignment that provides the information about equivalent residues between the two structures. Therefore, this provides a measure of how accurately the alignment equivalences the residues of the reference structure to the trial model. If a residue is incorrectly mapped, it will have significantly different packing in one structure than the other, leading to an increase in the CADE. Although this is a good measure of the alignment’s structural quality, it does not penalize for sparse alignments, where there may be few paired residues and many large gaps. Such alignments should only occur when the two structures are very dissimilar. Therefore, along with CADE, we include the absolute value of the logarithm of pP (Abagyan and Batalov, 1997) of the alignment in our final quality score, the Contact Alignment Quality (CAQ) score:

\[
CAQ = c \times CADE - \log(pP),
\]

where \( c \) is a weighting factor. Low values of \( pP \) indicate poor alignments while high values indicate that the alignments are less likely to be structurally insignificant.

**ICM structural alignment algorithm** The structural alignment algorithm, implemented in ICM by S. Batalov, was described previously (Abagyan and Totrov, 2002) and is similar to the SSAP algorithm (Orengo and Taylor, 1996). Here, we give a short description of the algorithm. It finds the residue alignment (or residue-to-residue correspondence) for two arbitrary molecules having superimposable backbone fragments. The structural alignment identification and optimal superposition is based on the Cα-atom coordinates, and the sequence information is added with an equal weight. The structural alignment algorithm follows these steps:

- Calculate a pairwise residue matrix of local RMSD values in 14-residue windows around a given pair of residues and convert these distances into scores.
- Add amino acid type-based substitution scores to the previous structural score [we used the Gonnet matrix (Gonnet et al., 1992)] with an equal weight.
- Apply the ZEGA (zero-end gap alignment) dynamic programming procedure in which substitution scores for each \( i,j \)-pair of residues contain the two terms described above.
- Clean up the alignment by iterative superpositions in which the Alexandrov and Go function (Alexandrov and Go, 1994), which describes a balance between RMSD and the alignment length, is optimized.

**Derivation of SAD, an optimal structural alignment database**

For each pair of structures, there may be three or more alignments from the sources detailed above. Each alignment was scored using Equation (6). The alignment with the lowest CAQ-score was taken to be the optimal alignment. In cases where two or more alignment achieved the same CAQ-score, it was always found that the alignments were identical.

In order to determine which alignment method was the most accurate, each alignment for a given pair of structures was ranked according to the alignment’s CAQ-score with the alignment with the lowest CAQ-score being ranked first. Each alignment was then further scored as \( 1/n \), where \( n \) is the rank position of that score. The total rank-scores of all alignments from a given source were then summed.

The resulting database of alignments was partially normalized to avoid over-representation of certain folds that are ubiquitous in the PDB (Table 2). The fold of the two structures represented in each alignment was classified using SCOP version 1.55 (Murzin et al., 1995) and the total number of alignments in each SCOP family was tallied. As anticipated a number of folds were over-represented. One approach for normalizing such folds would be to pick the 50 alignments with the best average structural resolution (i.e. the mean of the resolution of the two structures concerned) within the fold. However, we found that this left representatives that were dominated by one or two very-high-resolution structures, potentially biasing the fold towards these structures. Since, we wish to include as much structural diversity for each fold we used a different approach. For folds with more than 50 representative alignments, a simple protocol of selecting 50 alignments was adopted.

**RESULTS**

**Derivation of Contact Alignment Quality score**

A good pairwise structure–structure alignment should maximize the number of aligned pairs of residues while minimizing some measure of local 3D deviation between the two structures. Note that it is important that this is a local, not a global, measure; otherwise we will be incorporating error derived from plastic deformations as discussed above. For
large proteins with local structural plasticity (i.e. structural deviations that cannot necessarily be accounted for at the sequence level), this problem becomes increasingly challenging. RMSD is commonly used to measure structural deviation between two structures. However, RMSD is very sensitive to any rearrangement of parts of a structure that are wrong or unimportant (e.g. long side-chains that are exposed) (Abagyan and Totrov, 1997).

Previously, we introduced the CAD measure (Abagyan and Totrov, 1997) that addresses this problem by measuring the significant differences in packing between two homologous structures. We cannot simply use this measure in the case of two non-homologous structures, such as those involved in pairwise structure alignments, since the possibility of residue substitution renders the absolute difference of two equivalent contacts meaningless. For example, a glycine–leucine contact in one structure compared with an alanine–leucine contact in another will contribute a large value to the total CAD but yet may not indicate a significant local structural deviation. Hence, we have introduced an estimate of CAD that is related to the distance between residues. The use of projected points along the Cα–Cβ vectors of each pair of residues as the termini for the inter-residue distance measurement provides information not only about the packing of the backbone, but also about the orientation of the side-chains—a factor that other distance measures often neglect (Abagyan and Totrov, 1997).

In order to parameterize our new measure, we considered all residue–residue pairs in structures whose resolution is better than 2.5 Å and calculated the contact surface area (CSA) against inter-residue distance for each type of pair. Figure 1a shows an example of the results of the analysis for alanine–leucine pairs of residues. There are distinct values of inter-residue distance and CSA which are more likely to be populated than others, but in general there is a trend of increasing inter-residue distance with decreasing CSA, as would be expected. Almost all other non-glycine-containing residue pairs show the same general properties with the same regions of inter-residue distance and CSA being populated.

**Fig. 1.** Examples of population density of residue–residue contact surface area versus inter-residue distance for all PDB protein structures with resolution better than 2.5 Å. (a) Alanine–leucine; (b) tryptophan–arginine; (c) glycine–glycine; and (d) glycine–arginine. Pseudo-Cb refers to the projection along the Cα–Cβ vector of a residue by 1.5 times its length.
An exception is that of tryptophan–arginine Figure 1b). Here, a region of large CSA (∼95 Å²), not seen for any other residue pair, along with a relatively small inter-residue distance (6.7 Å) is observed. This is due to arginine side-chains often packing across the planar surface of the tryptophan, leading to an unusually high degree of contact between the two side-chains (data not shown). Figure 1c and d show the results for glycine–glycine interactions and glycine–arginine interactions, respectively. They are typical of all glycine-containing pairs, the highest density being shifted to smaller inter-residue distances due to the use of Cα atoms as the termini of the distance measure, rather than the use of projected points.

When comparing best-fit lines of CSA against inter-Cα distance, non-glycine-containing pairs show similar trends to each other, apart from the tryptophan–arginine pair, as discussed previously (data not shown). Glycine-containing pairs also show similar trends to each other with smaller inter-residue distances than non-glycine-containing pairs for the reasons discussed above. Since we are interested in deriving an estimate of contact area, these results indicate that it is viable to take the mean of the gradient and intercept of the two sets of pairs (non-glycine- and glycine-containing) and use these as estimators of the CSA based upon the inter-residue distance. For non-glycine-containing pairs, the estimator is

\[ \text{CAE} = 84.6 - 6.96d, \]

where \( d \) is the inter-residue distance, and for glycine-containing pairs

\[ \text{CAE} = 90.0 - 9.41d. \]

We used the above results in order to calculate the CADE, for non-homologous proteins related by a sequence alignment. It is important that the CADE value is not dependent upon the length of the alignment. Analysis of the alignment length compared with the normalized CADE shows that this is indeed so for this measure, as long as the alignments are longer than about 75 residues (data not shown). This is acceptable, since the residue-lengths of the majority of single domains are greater than this figure.

Figure 2 shows the relationship between CADE and RMSD for homologous pairs of structures. For low values of CADE, RMSD is very well defined since there are few differences in packing at such low levels of structural difference. As CADE increases, the range of RMSDs increases as more gross packing differences occur (e.g. insertions, deletions or plastic deformations). This result indicates that at low levels of structural differences, CADE is as accurate as RMSD, vindicating our choice of inter-residue distance and contact area estimation. At higher levels of structural difference, CADE is more sensitive to significant packing differences than RMSD.

**Derivation of alignment database**

For the derivation of an accurate database of structure–structure alignments, we took a number of freely available SADs; the HOMSTRAD database (Mizuguchi et al., 1998), BAliBASE database (Thompson et al., 1999) and Yale-Gerstein (SCOP-Gerstein) database derived from the SCOP database (Gerstein and Levitt, 1998). In addition, using the pairs of sequences from these databases, we generated a database of global pairwise structural alignments using the ICM program (Abagyan and Totrov, 2002). The use of the SCOP-derived database helped to ensure that the alignments to be considered cover as much fold-space as possible, as defined by SCOP (Murzin et al., 1995). Table 1, column 2 ‘number of alignments’, indicates the contribution of each source of alignments after rejection of alignments that were.

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of alignments(^a)</th>
<th>Number before normalization(^b,d)</th>
<th>Number of alignments in SAD(^c,d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICM</td>
<td>4845</td>
<td>3370 (70%)</td>
<td>1298 (27%)</td>
</tr>
<tr>
<td>HOMSTRAD</td>
<td>2029</td>
<td>822 (40%)</td>
<td>468 (23%)</td>
</tr>
<tr>
<td>SCOP-Gerstein</td>
<td>3305</td>
<td>1206 (37%)</td>
<td>346 (10%)</td>
</tr>
<tr>
<td>BAliBASE</td>
<td>39</td>
<td>9 (25%)</td>
<td>6 (17%)</td>
</tr>
<tr>
<td>Total</td>
<td>4845</td>
<td>4845</td>
<td>1927</td>
</tr>
</tbody>
</table>

\(^a\) Contributions from each alignment source.
\(^b\) Contributions from each alignment source after removal of poor-quality alignments and structures.
\(^c\) Contributions of alignment sources to the final SAD.
\(^d\) Values in parentheses are the percentage of original set of alignments remaining for the source.

Note that these figures sum up to more than the total number of alignments due to the fact that identical alignments are sometimes provided by two or more sources.

\(^e\) Total number of actual alignments at each stage.
found to be sparse, had poor structural resolution or included post-translational modifications. The SCOP-Gerstein database contributes most to the raw database with 3305 alignments, followed by HOMSTRAD (2029) and BAliBASE (39). The ICM structural alignment method provides 4845 alignments based upon the structure pairs from the HOMSTRAD, BAliBASE and SCOP-Gerstein databases.

In order to remove degeneracy within this set of alignments we employed two measures of alignment quality. First, we used information from the superposition of the two structures as defined by the alignment itself. The quality of the fit of the two structures due to the alignment was measured using our estimate of differences in contact area between equivalent residues in the two structures (CADE), as defined by the alignment. The smaller the difference in CADE, the more accurate the structure-structure alignment is likely to be. However, it is possible that the CADE can become smaller as the number of aligned pairs within the alignment decreases leading to a sparse alignment. This problem can manifest itself if only a few core and conserved residues are aligned leaving large unpaired gaps in the alignment. To counteract this detrimental effect, we included a measure of the likelihood that the alignment is structurally insignificant (pP) (Abagyan and Batalov, 1997). This is directly related to the number of aligned pairs in the alignment and also the sequence-alignment quality.

In order to parameterize the CAQ-scoring function between CADE and pP [Equation (6)], we performed visual inspections of a wide variety of structure-structure sequence alignments (i.e. different folds, different sequence identities, etc.) and the resulting superposition of the homologous structures based upon these alignments. By comparing the redundant alignments for a given homologous pair we determined that the optimum value of \( c \) in Equation (6) is 1.0. This value gives, visually, the best superposition of the structures and reduces the number of false positives due to alignments that are sparser than the others in the set of redundant alignments. A value of 1.0 is reasonable since it scales both CADE and pP values to the same order of magnitude—CADE can range from 0 to about 60 while pP ranges from 0 to about 40. Therefore, none of the measure dominates the other leading to a good balance between structural and sequence alignment information. We tried values of 0.5 and 1.5 (instead of 1.0) for \( c \) and found that the quality of the alignments was clearly worse, both in terms of the visual quality of the superposition of pairs of structures and in terms of the resulting sequence alignment.

The application of the CAQ-score to the redundant set of alignments led to a set of 4845 unique alignments made up from the redundant set of alignments as described in column 3 (‘before normalization’) of Table 1. The ICM global structural alignment method produced the largest number of high-quality alignments, as determined by Equation (6), followed by the contents of the SCOP-Gerstein database, the HOMSTRAD database and the BAliBASE database. However, since each source of structural alignments does not provide the same number of alignments, it is more instructive to infer the relative success of each alignment method in terms of a normalized measure, such as the percentage of alignments of each method kept during the redundancy removal procedure (values in parentheses, column 3, Table 1). Once again, the ICM structural alignment method fared the best (70%), followed by HOMSTRAD (40%), SCOP-Gerstein (37%) and BAliBASE (25%).

What is the rank of quality for the different alignment databases? Figure 3 shows that the ICM global alignment method produces structural alignments that are of the highest quality, more often relative to the other structural alignment sources. BAliBASE, HOMSTRAD and SCOP-Gerstein are ranked second most often. None of the method ranks significantly in third or fourth place.

Why do the various databases of alignments rank with respect to each other in this way? Figure 4 shows a multiple structural alignment for the pair of structures 1mrj and 1abr, chain A, and shows typical characteristics of each source or method of structural alignment. It can be seen that the core regions are almost completely conserved with the majority of differences between the pairwise alignments occurring within the loop regions of the structures. These differences are often subtle, usually consisting of small shifts of local sequence by one or two residues. However, some general features of each method can be visually detected from this example, and other structural alignments.

In addition to the ICM global structural alignment method’s ability to incorporate sequence as well as structural information, it is also able to introduce consecutive insertions in both
sequences, i.e. gaps that follow consecutively between the sequences (arrows 1, 2 and 10, Fig. 4). This allows consecutive regions of both structures to be left unaligned, which is particularly useful for loop regions where the topologies of equivalent loops in the two structures are very different and therefore should not be structurally aligned. This approach is not used by the methods, which result in alignments that are consistently over-aligned compared with alignments generated by the ICM global structural alignment method (percentage of residues aligned: ICM-align, 95.3%; HOMSTRAD, 96.6%; BALiBASE, 97.7%; and SCOP-Gerstein, 98.6%). Although this reduction in paired residues necessarily reduces the pP value, the structural fit improves, leading to an overall reduction in CAQ-score. This leads to a higher quality structural alignment. HOMSTRAD alignments are usually very similar to ICM global structural alignments with few local sequence shifts. However, it is not uncommon to observe a reduction in the number of indels (e.g. arrows 3 and 4, Fig. 4). BALiBASE and SCOP-Gerstein alignments also show this reduction in the number of indels but also include, to varying degrees, local sequence shifts (arrows 5–9, Fig. 4). Critically, sometimes these shifts occur within secondary structural elements (arrows 7 and 9, Fig. 4), which cause significant increases in the CAQ-score. These shifts and indels led to an improvement in sequence-alignment (and hence improved pP values) to a small degree, but also result in a significant reduction in structural overlap that manifests itself in low CADE values.

An analysis of the SCOP database (Murzin et al., 1995) indicates that there are a significant number of structural families that are highly populated in the PDB (Table 2). Indeed up to this point of the database derivation, SAD reflects this (Table 3) with IgG domains, eukaryotic proteases and globins, e.g. being over-represented. This over-representation will bias any future method that uses this database. Therefore, the database was normalized by removing alignments in those SCOP families that had more than 50 representatives. This number was picked based upon the distribution of family populations (Fig. 5), the goal being to reduce the most highly populated families but not to adversely affect the fold-space sampling. The procedure reduced the content of the database to 1927 alignments (Table 1, column 4). The disproportionately large reduction in the contribution from SCOP-Gerstein is due to the databases’ domination by IgG-related alignments.
Table 2. Distribution of the 10 most populated structural families as defined in SCOP v1.55

<table>
<thead>
<tr>
<th>SCOP description</th>
<th>Number of structures (SCOP v1.55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1 set domains (antibody constant domain-like)</td>
<td>992</td>
</tr>
<tr>
<td>V set domains (antibody variable domain-like)</td>
<td>953</td>
</tr>
<tr>
<td>Globins</td>
<td>657</td>
</tr>
<tr>
<td>Eukaryotic proteases</td>
<td>549</td>
</tr>
<tr>
<td>C-type lysozyme</td>
<td>377</td>
</tr>
<tr>
<td>Phage T4 lysozyme</td>
<td>359</td>
</tr>
<tr>
<td>Bacterial AB5 toxins, B-subunits</td>
<td>281</td>
</tr>
<tr>
<td>Retroviral protease (retropepsin)</td>
<td>278</td>
</tr>
<tr>
<td>Animal virus proteins</td>
<td>251</td>
</tr>
<tr>
<td>Legume lectins</td>
<td>243</td>
</tr>
</tbody>
</table>

Table 3. SCOP classes with more than 50 alignments in the un-normalized SAD

<table>
<thead>
<tr>
<th>SCOP class</th>
<th>Description</th>
<th>Un-normalized SAD content</th>
</tr>
</thead>
<tbody>
<tr>
<td>b.1.1.1</td>
<td>V set domains (antibody variable domain-like)</td>
<td>1869</td>
</tr>
<tr>
<td>a.1.1.2</td>
<td>Globins</td>
<td>523</td>
</tr>
<tr>
<td>b.47.1.2</td>
<td>Eukaryotic proteases</td>
<td>287</td>
</tr>
<tr>
<td>b.1.1.2</td>
<td>C1 set domains (antibody constant domain-like)</td>
<td>146</td>
</tr>
<tr>
<td>a.133.1.2</td>
<td>Vertebrate phospholipase A2</td>
<td>142</td>
</tr>
<tr>
<td>d.2.1.2</td>
<td>C-type lysozyme</td>
<td>123</td>
</tr>
<tr>
<td>b.34.2.1</td>
<td>SH3-domain</td>
<td>73</td>
</tr>
<tr>
<td>d.3.1.1</td>
<td>Papain-like</td>
<td>73</td>
</tr>
<tr>
<td>a.3.1.1</td>
<td>Monodomain cytochrome c</td>
<td>66</td>
</tr>
<tr>
<td>b.29.1.1</td>
<td>Legume lectins</td>
<td>57</td>
</tr>
<tr>
<td>a.45.1.1</td>
<td>Glutathione S-transferases, C-terminal domain</td>
<td>55</td>
</tr>
<tr>
<td>c.41.1.1</td>
<td>Subtilases</td>
<td>52</td>
</tr>
</tbody>
</table>

Fig. 5. Distribution of the number of alignments in SCOP families in the database before normalization. The majority of SCOP families are populated by less than 20 structural alignments with only 12 families containing more than 50 structural alignments.

Table 4. Distribution of structural families and alignments in SAD in terms of fold-type

<table>
<thead>
<tr>
<th>Fold-type (SCOP class)</th>
<th>Number of families represented</th>
<th>Number of alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha (a)</td>
<td>48</td>
<td>427</td>
</tr>
<tr>
<td>All beta (b)</td>
<td>66</td>
<td>583</td>
</tr>
<tr>
<td>Alpha/beta (c)</td>
<td>67</td>
<td>428</td>
</tr>
<tr>
<td>Alpha+beta (d)</td>
<td>56</td>
<td>394</td>
</tr>
<tr>
<td>Multi-domain (e)</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Membrane and cell surface proteins and peptides (f)</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Small proteins (g)</td>
<td>16</td>
<td>71</td>
</tr>
<tr>
<td>Total</td>
<td>261</td>
<td>1926</td>
</tr>
</tbody>
</table>

DISCUSSION AND CONCLUSIONS

The normalization of the database resulted in a set of alignments that are not biased towards any particular fold-type (Table 4), with the same order of magnitude of alignments in all four major fold-types, as defined by SCOP (alpha, beta, alpha/beta and alpha+beta). In addition, there are alignments representing a small number of membrane and cell surface proteins as well as a significant number of small proteins.

Figure 6 indicates that the distribution of alignment sequence identities within the normalized database is towards the twilight zone, but there are still a significant number of alignments at higher identities.

The mean structural resolution of each alignment was calculated and the distribution of these values is shown in Figure 7. The predominant mean resolution is ~2 Å with a number of alignments with mean resolutions below 1.5 Å.
Fig. 6. Distribution of sequence identities of alignments in the final SAD database. Despite the majority of SAD structural alignments having sequence identities in the ‘twilight zone’ (below ∼30%), SAD still contains significant amounts of alignments that cover higher ranges of sequence identities.

Fig. 7. Distribution of mean resolutions of structures of all alignments in the final SAD database. The average resolution of structures involved in SAD structural alignments is ∼2 Å.

Bourne, 1998, 2001; Thompson et al., 1999); whilst it is desirable to have as many data-points as possible, it is not necessary to include structural alignments for almost every conceivable pair of structurally related sequences.

A further departure from many methods of SAD generation is the use of estimation of CAD rather than Cartesian RMSD as part of the structural-alignment quality score used to select better alignments. This was done, in part, to remove the deleterious influence of plastic deformation that plagues the Cartesian RMSD-based residue-equivalencing methods. It is important to use contacts in structural alignment of plastic protein structures, e.g. the DALI method for finding structural similarity is based upon the use of a residue–residue distance matrix (Holm and Sander, 1999). However, the DALI method focuses upon Cα-atoms only, leaving out information about the orientation of side-chains as a consequence. We have derived a method that includes such side-chain information based around the ideas first developed for CAD (Abagyan and Totrov, 2002). Since it is not possible to use CAD directly due to the non-conservation of residues in homologous structures, we have derived a method that estimates the contact area of each residue (CAE) with respect to the surrounding structure using projected vectors that are used to measure the distance between the side-chains. The relationship between this measure and the actual contact area is remarkably similar between different non-glycine residue pairs with only a few exceptions. This prompted the use of a linear approximation of the contact area in the final CADE method.

In non-alignable areas, CAD has the beneficial feature of not being sensitive to the magnitude of irrelevant differences and allowing plastic deformations in the aligned areas, which are completely prohibited by the RMSD measures. An example of such a difference is a loop that diverges very differently which would provide a significant contribution to any heavy-atom RMSD calculation. This is clearly shown in Figure 2—the relationship between CADE with heavy-atom RMSD is linear until RMSD becomes greater than about 7 Å—pairs of structures with larger RMSDs usually have significant structural differences often due to loops. Therefore, CADE is more sensitive to differences in packing in regions of the structures that are more structurally conserved and therefore more structurally significant.

Since it is possible to generate structural alignments that show good values of CADE but are sparse (i.e. highly underaligned), we have included a measure of the structural insignificance of the alignment (pP) in our structural alignment scoring function which we call CAQ-score. The parameterization of CADE with respect to pP can only be performed effectively by eye, but we have found that a simple linear combination of the two performs satisfactorily. We believe that the CAQ-score is a powerful alternative to simply testing the quality of a structural alignment by RMSD means alone.

We have derived a database of structural alignments using alignments from a number of available databases (de Bakker et al., 2001; Gerstein and Levitt, 1998; Mizuguchi et al., 1998; Thompson et al., 1999, 2001) and an ab initio method, ICM global structural alignment. The ICM global structural alignment method not only takes into consideration contributions from the structural quality of the resulting fit of the two structures, but also information from the compatibility of the sequences via the alignment, in spirit with previously described methods such as SSAP (Orengo and Taylor, 1996;
Taylor and Orengo, 1989). We used a new CAQ-score function of alignment quality to discriminate between alignments from different databases. The ICM global alignment method with accessibility and secondary structure terms provided the largest number of highest quality alignments (Fig. 3 and Table 1) with 61% of the final alignments derived from this method. The ability of the method to assign continuous gaps across both sequences to portions of the alignment that might lead to significant structural deviation (e.g. arrows 1, 2 and 10 in Fig. 4) is the key to this method’s success. This ability to allow two local sequences to remain completely unaligned is critical for the success of this method in generating the alignments for this database and may go some way to resolve the problem of lack-of-convergence for structure–structure superposition (Feng and Sippl, 1996; Godzik, 1996)

The resulting SAD of 1927 alignments efficiently covers known fold-space in an unbiased manner (Table 4), contains alignments with a wide range of sequence identities (Fig. 6) particularly in the <40% identity range where many current methods are challenged, and is made up of structures whose resolutions are, on average, ∼2.0 Å (Fig. 7). In short, this database fulfils the objectives set out above and provides an excellent test set for derivation of new and improved approaches such as sequence–structure alignment models.

SAD is available in interactive form over the web at http://abagyan.scripps.edu/lab/web/sad/show.cgi and is also available in the form of MySQL tables upon request.

ACKNOWLEDGEMENTS

We thank Wen Hwa Lee and Lars Brive for critically reading this manuscript and Sergei Batalov for implementing the structural alignment algorithm in ICM. B.M. was supported by a Wellcome Trust Prize Traveling Research Fellowship.

REFERENCES


Venclovas C. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins*, **53** (Suppl. 6), 380–388.