**DDBASE2.0: updated domain database with improved identification of structural domains**

A. Vinayagam¹,†, J. Shi²,§, G. Pugalenthi¹, B. Meenakshi¹, T.L. Blundell² and R. Sowdhamini¹,∗

¹National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK campus, Bellary Road, Bangalore, Karnataka 560 065, India and
²Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB21GA, UK

Received on October 27, 2002; revised on March 3, 2003; accepted on March 28, 2003

**ABSTRACT**

**Motivation:** Although many methods are available for the identification of structural domains from protein three-dimensional structures, accurate definition of protein domains and the curation of such data for a large number of proteins are often possible only after manual intervention. The availability of domain definitions for protein structural entries is useful for the sequence analysis of aligned domains, structure comparison, fold recognition procedures and understanding protein folding, domain stability and flexibility.

**Results:** We have improved our method of domain identification starting from the concept of clustering secondary structural elements, but with an intention of reducing the number of discontinuous segments in identified domains. The results of our modified and automatic approach have been compared with the domain definitions from other databases. On a test data set of 55 proteins, this method acquires high agreement (88%) in the number of domains with the crystallographers' definition and resources such as SCOP, CATH, DALI, 3Dee and PDP databases. This method also obtains 98% overlap score with the other resources in the definition of domain boundaries of the 55 proteins. We have examined the domain arrangements of 4592 non-redundant protein chains using the improved method to include 5409 domains leading to an update of the structural domain database.

**Availability:** The latest version of the domain database and online domain identification methods are available from http://www.ncbs.res.in/~faculty/mini/ddbase/ddbase.html

*To whom correspondence should be addressed.
†Present address: Department of Molecular Biophysics, German Cancer Research Centre (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.
§Present address: Celltech R&D, Inc., 1631 220th Street SE, Bothell, WA 98021, USA.

**INTRODUCTION**

Many proteins involved in cell regulation and signalling contain multiple modules that perform their functions, sometimes independently and sometimes synergistically, often involving inter-domain interactions, and domain rearrangements induced due to ligand binding. To understand these, accurate domain boundaries and reliable domain identification procedures are required. The compilation and maintenance of protein domain-fold libraries or databases also make important contributions to knowledge-based fold recognition methods. However, several public domain resources that curate domain boundaries of protein structural entries often require careful manual intervention.

Domain identification methods have, in the past, been achieved by measuring distances: for example, successive inter-residue distances (Schultz, 1977) or by clustering segments (Crippen, 1978) or by plane dissection (Rose, 1979) or from hydrophobic clusters (Swindells, 1995) or simple harmonic approximation (Wernisch et al., 1999) or graph heuristics of residue interactions (Holm and Sander, 1994; Xu et al., 2000). We have re-examined our previous approach, DIAL (Sowdhamini and Blundell, 1995), to the identification of structural domains that depends on clustering of secondary structures (Sowdhamini and Blundell, 1995). After incorporating features such as the inclusion of loop regions as elements of structure, this approach (DIAL4) gives rise to domain boundaries that are less discontinuous. Domain definitions obtained by the modified method compare quite well with crystallographers’ definition and databases like SCOP (Murzin et al., 1995), 3Dee (Siddiqui and Barton, 1995; Siddiqui et al., 2001), CATH ( Orengo, 1994; Orengo et al., 1997), DALI (Holm and Sander, 1994) and Protein Domain Parser
(Xu et al., 2000). This new method has been employed to obtain an update of the domain database.

SYSTEMS AND METHODS

Features of DIAL and limitations

The original version of DIAL (DIAL0; Sowdhamini and Blundell, 1995) clusters secondary structures of a protein on the basis of average C\(^\alpha\)-distances using protein three-dimensional coordinates. Multiple clusters indicate the possibility of multiple-domain organization. However, the compactness of the individual clusters are also examined, so that a domain organization that provides maximum domain compactness, as measured by disjoint factor, is considered.

Domain boundaries suggested by DIAL0 correspond to structural domains and the agreement with other domain database definitions was only about 45%. Moreover, there are a number of concerns in providing DIAL-derived domain definitions as inputs to create profiles for fold recognition:

1. The algorithm can allow compact subdomains to be treated as a separate domain with high disjoint factors. This can sometimes lead to overprediction of the number of domains.
2. Since the method considers a protein as a string of secondary structures and clusters the elements on the basis of distances, several small discontinuous segments are often considered part of a domain.
3. Long secondary structures in proteins are treated as single elements in clustering, but may actually have interactions with more than one domain.
4. Long insertions of supersecondary structures or domain insertions often pose problems in the accurate delineation of domain boundaries due to local interactions. For example, extra helices in some TIM barrels form local clusters with the component helices leading to incorrect domain boundaries.
5. Interactions involving loop regions were ignored in DIAL since the method only considers \(\alpha\)-helices and \(\beta\)-strands. Some proteins, like the viral coat proteins and some oxidoreductases, have elaborate loop regions where the interactions could be substantial.
6. In others like ricin and muconelactate dehydrogenase, due to local interactions, a small part of one domain gets associated with another domain.

Features of the new version of DIAL

The DIAL program has been modified to achieve better domain definition by employing six different schemes (see Table 1 for details). We have assessed the performance of the individual schemes on 40 proteins. 32 of which had caused problems using the earlier method (DIAL0; Sowdhamini and Blundell, 1995) and eight of which caused no problems but we introduced as controls (see legend to Fig. 1 and Supplementary information for details on the 40 proteins). The problematic entries are those protein structures where DIAL0 did not agree with the manual/visual crystallographers’ definition of domain boundaries even in the number of domains. Six different schemes for the domain identification were developed that are minor variants of the original DIAL0. Table 1 provides a description of the new features in different schemes. Figure 1 shows the number of proteins whose domain definitions remain problematic after applying different schemes. DIAL3, DIAL4 and DIAL6 perform quite well for problematic single-domain and multi-domain proteins, whereas the domain definitions of non-problematic control proteins are not severely affected. DIAL4 performs better in proteins that have long secondary structures and the domain boundaries suggested by this method are also satisfactory when compared to other methods. Features of DIAL4 are:

- Loop regions have been included in the clustering and domain definition.
- Loops and secondary structures were split into fixed length segments (loops, helices and strands longer than 9, 15 and 10 residues, respectively).
- N and C terminal loops were also considered for domain definition when longer than nine residues.

### Table 1. Different schemes employed for the improvement of domain definitions

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIAL0</td>
<td>Original DIAL algorithm (Sowdhamini and Blundell, 1995) that considers only (\alpha)-helices and (\beta)-strands for clustering.</td>
</tr>
<tr>
<td>DIAL1</td>
<td>Loop regions also considered for clustering</td>
</tr>
<tr>
<td>DIAL2</td>
<td>Loop regions also considered for clustering; if loop region crosses 9 residues then they are split into 9-residue and smaller segments</td>
</tr>
<tr>
<td>DIAL3</td>
<td>Loop regions also considered for clustering; long loops and secondary structures are split into 9-residue segments</td>
</tr>
<tr>
<td>DIAL4</td>
<td>Loop regions also considered for clustering; helix, strand, loops are split if longer than 15, 10 and 9 residue segments respectively. Extra residues, if more than two, in such long segments are considered as separate elements for clustering. N- and C-terminal loops (if &gt;9 residues) are included</td>
</tr>
<tr>
<td>DIAL5</td>
<td>Loop regions also considered for clustering; helix, strands, loops that exceed 15, 10 and 9 residues, respectively, are alone split into equal residue segments (where the resulting segment length is less than given cut-off)</td>
</tr>
<tr>
<td>DIAL6</td>
<td>Same as DIAL5 but N- and C-terminal loops (if &gt;9 residues) are included</td>
</tr>
</tbody>
</table>

Downloaded from bioinformatics.oxfordjournals.org at University of Portland on May 24, 2011
Fig. 1. Comparison of different DIAL schemes with a data set of 40 proteins. Please refer to Table 1 for the definition of schemes. PDB (Berman et al., 2000) codes for single domain problematic entries considered are 1ccr-, 1ab-, 1babB, 1bgeB, 1bcfA, 1fha-, 1aaj-, 1cauA, 4tvh1, 1add-, 1ads-, 1tph1, 4xiaA, 153l-, 1ayaA, 1hjrA, 7rsa-, 5fdl-, 1bet-, 1ccf-, 1dta and 1colA; multi-domain problematic entries include 1chrA, 4gcr-, 4gp-, 4cc-, 8abp-, 8acn 8atcA, 8catA and 9rubA; non-problematic entries include 8rxnA, 9rnt-, 7znf-, 7pti-, 1yaaA, 8fabB and 8tlnE. Please see Supplementary information for further details about the PDB entries. The problematic entries refer to those protein chains where the number of domains identified by the current scheme does not agree with a graphical inspection or with the crystallographers’ definition. A general reduction in the number of problematic entries with no loss in accuracy for the controls reflects improvement in the algorithm for this test data set.

- Small clusters \( \leq 25 \) residues that were ignored in the first version (Sowdhamini and Blundell, 1995) and small discontinuous segments \( \leq 15 \) residues were merged into the domain that is closest in the amino acid sequence.

RESULTS

Comparison with other databases

In order to verify that DIAL4 is the most robust scheme for the update of the domain database, we have further compared the domain definitions proposed by this scheme with other definitions as used in SCOP (Murzin et al., 1995), CATH (Orengo, 1994; Orengo et al., 1997), DALI (Holm and Sander, 1993, 1994), 3Dee (Dengler et al., 2001; Siddiqui et al., 2001) databases and DOMAINPARSER (Xu et al., 2000). This comparison has been performed using the 55-protein data set, employed earlier by Thornton’s group (Jones et al., 1998) and Xu’s group (Xu et al., 2000). Here, agreement score is hundred times the fraction of the number of proteins that attain common domain definition by two resources to the number of proteins that are common between the two resources. Domain definition is the number of domains suggested for a protein chain in a particular resource. Figure 2 compares the performance of the various databases with each other by the projection of pairwise agreement scores. The agreement scores between DIAL4 and the other databases is quite high (see Supplementary information for comparison of the number of domains identified by DIAL4 against those recorded in other resources for the 55-protein data set). In this data set, we find that 29 out of 30 single-domain folds have been correctly predicted by DIAL4. The definitions compare very well with SCOP for rhodanese and ferredoxin reductase (see Supplementary information for these illustrative examples). However, for \( p \)-hydroxybenzoate hydroxylase, DIAL can propose local clusters of secondary structures as individual domains when they are probably accurately defined as subdomains. The splitting of long secondary structures in DIAL4 can sometimes lead to an over-prediction of the number of domains; for example, the coiled-coil fold of 1vsg is defined as a single-domain fold but DIAL4 suggests that it is a two-domain fold. However, the agreement score observed between DIAL4 and the other resources in the 55-protein data set is 88% where the highest agreement score (88.7%) is between CATH and other databases.

The domain boundary definitions suggested by DIAL4 are also comparable to those recorded in SCOP and other databases (Jones et al., 1998; Xu et al., 2000) and the mean overlap score between DIAL4 and the other resources is 98%. Overlap score is the same as described earlier (Jones et al., 1998) and can be defined using a 17-residue peptide where two domains have been assigned (numbered as 1 and 2) by two independent methods of domain identification or resources A and B:

**OVERLAP SCORE CALCULATION**

<table>
<thead>
<tr>
<th>Residue No.</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Overlap table:**

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>0</td>
</tr>
<tr>
<td>B2</td>
<td>2</td>
</tr>
</tbody>
</table>

Overlap score = \( \frac{(8 + 7)}{17} \) * 100 = 88.23%
Table 2. Domain boundary overlap score* for 9 proteins from the 55 protein data set (Jones et al., 1998) that obtain the same number of domains as defined in different resources

<table>
<thead>
<tr>
<th>PDB code</th>
<th>DIAL4</th>
<th>SCOP</th>
<th>3Dee</th>
<th>CATH</th>
<th>CRYST^b</th>
<th>DALI</th>
<th>PDP^c</th>
</tr>
</thead>
<tbody>
<tr>
<td>8atcA</td>
<td>100</td>
<td>92.90</td>
<td>92.58</td>
<td>93.54</td>
<td>88.38</td>
<td>93.22</td>
<td>89.03</td>
</tr>
<tr>
<td>1fnr-</td>
<td>100</td>
<td>98.99</td>
<td>98.99</td>
<td>100</td>
<td>96.62</td>
<td>99.66</td>
<td>97.62</td>
</tr>
<tr>
<td>1lap-</td>
<td>100</td>
<td>98.96</td>
<td>99.79</td>
<td>99.79</td>
<td>95.86</td>
<td>72.52</td>
<td>98.55</td>
</tr>
<tr>
<td>1rhd-</td>
<td>100</td>
<td>97.95</td>
<td>99.66</td>
<td>99.66</td>
<td>98.97</td>
<td>98.98</td>
<td>99.31</td>
</tr>
<tr>
<td>8atcB</td>
<td>100</td>
<td>99.44</td>
<td>97.20</td>
<td>99.43</td>
<td>99.43</td>
<td>99.43</td>
<td>98.87</td>
</tr>
<tr>
<td>3cd4-</td>
<td>100</td>
<td>96.97</td>
<td>98.39</td>
<td>98.34</td>
<td>98.93</td>
<td>98.35</td>
<td>97.59</td>
</tr>
<tr>
<td>8adh-</td>
<td>100</td>
<td>91.97</td>
<td>91.44</td>
<td>92.69</td>
<td>90.19</td>
<td>90.01</td>
<td>96.25</td>
</tr>
</tbody>
</table>

*Overlap score is provided as percentage. Please see text for the definition of overlap score and Supplementary information for domain boundaries.

^bCrystallographers’ definition.

^cProtein Domain Parser.

eight residues assigned by method A as belonging to domain 1 is also assigned by method B to belong to domain 1; likewise, seven residues assigned by method A as belonging to domain 2 is also assigned by method B to belong to the same domain. The overlap score is 100 times the ratio of the number of residues with same domain definition and the total number of residues (as in Jones et al., 1998).

The detailed domain boundary definitions and overlap scores compared with crystallographers’ definition, SCOP, CATH, DALI, 3Dee and PDP resources for all the protein chains in the 55-protein data set and the mean overlap scores are provided in the Supplementary information. Table 2 provides the overlap scores between all the seven resources for nine proteins in the 55-protein data set where an identical domain definition was obtained.

DISCUSSION

Statistics of DDBASE2.0

DIAL4 has been applied to a culled data set (Hobohm et al., 1993; Dunbrack, R.L. and Muquit, M., unpublished results) consisting of 4592 protein entries corresponding to the August 2002 PDB release (Berman et al., 2000), where no protein chains are more than 60% identical in sequence with each other. After removing small protein chains, DDBASE2.0 contains information on about 4159 non-redundant protein entries involving 5409 domains. Figure 3 shows the distribution of domains in different structural classes. A majority of the domains identified contain both helices and strands. Nearly 50% of the proteins are single domain folds by DIAL4 definitions. Figure 4 shows the classification of protein entries in DDBASE2.0 on the basis of the number of domains identified in individual entries. In general, as observed in various test data sets and by other groups (Jones et al., 1998), accurate identification of single domain folds is usually possible.

Web interface

A user-friendly web interface has been created to access DDBASE Version 2.0 that has the following features as in the previous version: individual entries in the database accessible through keyword search, PDB code search or...
from the full entries list. The individual PDB entries contain essential information on secondary structure-based clustering and domain arrangements that includes general information about the protein, sequence information and secondary structure definition, proximity Index matrix of segments, dendrogram of segment clustering (Rough dendrogram, PS file and GIF files), table of possible domain definitions (with DI values, link to detailed domain boundaries for each possibility) best domain definition (automatically identified) with inter domain-distance matrix, links to view the individual domains in Rasmol (Sayle and Milner-White, 1995), sequence files for individual domains are available in both PIR and FASTA format (sequence file for full protein chain), classification of domain at class level and to obtain PDB files for individual domains. In addition, this version contains the following new features:

- Option for running BLAST and PSI-BLAST search against DDBASE version 2.0 at domain level (E-value option ranges from 0.0001 to 10).
- Option for running DIAL online using the user-supplied PDB file—provision for running any one of DIAL versions (DIAL0, DIAL1, DIAL2, DIAL3, DIAL4, DIAL5 or DIAL6).

**SUMMARY**

The inclusion of loop regions, the N- and C-terminal overhangs and the breakdown of long secondary structures in DIAL4 has significantly improved the structure of the segments giving rise to better clusters of secondary structures and domain definitions. The domain definitions proposed by the improved method are comparable to CA TH (98%). A new version of the domain database, with this improved method for the identification of structural domains, is now available. We will be constantly updating the database and extending it to include domain information for all protein structural entries. Proteins with multiple chain domain organization will be specifically examined for inter-chain domain connections to record domain swapping events (Bennet et al., 1995; Liu and Eisenberg, 2002).

**ACKNOWLEDGEMENTS**

R.S. is a recipient of Senior Research Fellowship awarded by the Wellcome Trust, UK. A.V. and, G.P. are supported by the Wellcome Trust. We also thank NCBS (TIFR) for infrastructural support.

**REFERENCES**


