Protein Science (1992), *1*, 1691-1698. Cambridge University Press. Printed in the USA.

A database of protein structure families with common folding motifs

LIISA HOLM, CHRISTOS OUZOUNIS, CHRIS SANDER, GEORG TUPAREV, AND GERT VRIEND
European Molecular Biology Laboratory, Heidelberg, Germany

(RECEIVED June 9, 1992; REVISED MANUSCRIPT RECEIVED August 27, 1992)

Abstract

The availability of fast and robust algorithms for protein structure comparison provides an opportunity to produce a database of three-dimensional comparisons, called families of structurally similar proteins (FSSP). The database currently contains an extended structural family for each of 154 representative (below 30% sequence identity) protein chains. Each data set contains: the search structure; all its relatives with 70–30% sequence identity, aligned structurally; and all other proteins from the representative set that contain substructures significantly similar to the search structure. Very close relatives (above 70% sequence identity) rarely have significant structural differences and are excluded. The alignments of remote relatives are the result of pairwise all-against-all structural comparisons in the set of 154 representative protein chains. The comparisons were carried out with each of three novel automatic algorithms that cover different aspects of protein structure similarity. The user of the database has the choice between strict rigid-body comparisons and comparisons that take into account interdomain motion or geometrical distortions; and, between comparisons that require strictly sequential ordering of segments and comparisons, which allow altered topology of loop connections or chain reversals. The data sets report the structurally equivalent residues in the form of a multiple alignment and as a list of matching fragments to facilitate inspection by three-dimensional graphics. If substructures are ignored, the result is a database of structure alignments of full-length proteins, including those in the twilight zone of sequence similarity. The database makes explicitly visible architectural similarities in the known part of the universe of protein folds and may be useful for understanding protein folding and for extracting structural modules for protein design. The data sets are available via Internet.

Keywords: computer algorithm; modular protein design; protein folding; protein structure alignment; protein structure database

Recent years have seen an explosion both in the number of newly determined protein sequences and in the number of structures solved by X-ray diffraction or NMR. However, the gap between available sequence and structure information is widening rapidly. As long as the protein folding problem remains unsolved, empirical observations of regularities gleaned from the database of known structures can be very useful. The analysis of known structures has already led to the discovery of more and more similarities at all hierarchical levels of protein architecture: secondary structure; folding units, e.g., the $(\beta\alpha)_3$ or $(\beta\alpha)_4$ substructures in many nucleotide-binding proteins; domains, e.g., the globin fold found in the membrane-insertion domain of colicin A (Holm & Sander, 1992b); and entire proteins, e.g., the similarity between ubiquitin and ferredoxin (Vriend & Sander, 1991). Both general structural similarities and direct evolutionary connections are successfully exploited in model building by homology, currently the most powerful method of predicting protein tertiary structure.

Reprint requests to: Protein Design Group, EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, Germany.

Abbreviations: rmsd, root mean square deviation of atomic positions of C^{α} atoms after optimal superposition; PDB, Protein Data Bank; DSSP, dictionary of secondary structure of proteins; HSSP, homology derived structures of proteins; 3D, three-dimensional; CPU, central processing unit.

Protein structure comparison

The aim of structure comparison is to identify residues occupying spatially equivalent positions in a pair of proteins. The resulting set of equivalenced residue pairs is similar to the classical notion of sequence alignment, but

more general: sequential ordering of residues need not be preserved. Once a list of equivalent atoms is known, several algorithms (e.g., Kabsch, 1978) can be used to transform the coordinates of one protein into the framework of the other.

Previous solutions

The recognition of similarities in 3D objects, such as protein folds, is often obvious to the eye, but presents a difficult computational problem. Nevertheless, it has been solved, partially or completely, several times. The diversity of the solutions arises from the combinatorial complexity of the problem and from the multitude of perspectives on what it means to be similar (Rao & Rossmann, 1973; Rossmann & Argos, 1976; Lesk, 1979; Remington & Matthews, 1978, 1980; Levine et al., 1984; Zuker & Samorjai, 1989; Abagyan & Maiorov, 1988; Mitchell et al., 1989; Taylor & Orengo, 1989; Sali & Blundell, 1990; Subbarao & Haneef, 1991; Alexandrov et al., 1992; Fischer et al., 1992). Each of these methods has one or more of the following limitations: (1) large insertions and deletions are difficult to recognize; (2) only sequential alignments can be detected, i.e., spatial similarity in spite of different loop connections is missed; (3) the occurrence of multiple copies of a motif is not detectable; (4) gradual distortions or interdomain motions between two structures cannot be detected; (5) manual initial alignment is required; (6) massive CPU or memory resources are needed.

Three new algorithms

Three new algorithms for pairwise protein structure comparison have been developed to address these difficulties. The quick fragment pair clustering algorithm called Suppos (Vriend & Sander, 1991) finds a largest common 3D substructure that fits within given cutoffs of positional deviations in rigid-body superposition. The fast dynamic programming algorithm called Comp3D (Tuparev & Sander, unpubl.) determines optimal sequential alignments by rigid-body superposition of a trailing trace and can detect interdomain motion. Both topological reconnections and geometrical distortions are detected by the algorithm called Dali (Holm & Sander, 1992a). Dali uses Monte Carlo optimization to compare distance matrices, exploiting the fact that the set of all intramolecular distances is a rotation and translation invariant representation of 3D protein structure.

The availability of fast, fully automatic, and robust algorithms for structure comparison provides an opportunity to produce and maintain a database of structural alignments. The database explicitly shows architectural relationships in the known part of the universe of protein structures.

Results

Automatically derived structural alignments

The database contains structural alignments corresponding to medium and weak ("twilight" zone) sequence similarities (Fig. 1), by each of the three algorithms. Strong to medium sequence similarity implies evolutionary conservation of protein structures. Similarities of protein structures or substructures in the twilight zone of sequence similarity may result from either convergent or divergent evolution. To capture these similarities, a representative set of 154 search structures was selected from the Protein Data Bank (Bernstein et al., 1977; release 58) such that no pair in the set has higher than 30% sequence identity (Hobohm et al., 1992), and an all-against-all structure comparison was performed, a total of 11,781 pair comparisons. Proteins with medium similarity (30–70% sequence identity) to a search structure were identified from the HSSP database of sequence alignments (Sander & Schneider, 1991) and then aligned in three dimensions. These pairs retain the shape of the structural core, whereas the length and conformation of loops may vary drastically. Strong similarities (70–100% sequence identity) are not included: two proteins with more than 70% identical residues normally have very little structural variation and their sequence-derived and structure-derived alignments are in almost all cases identical.

Format of the database

The results of the structural comparisons are stored in data sets named PID_method.FSSP, where PID stands for the PDB data set and chain identifier (e.g., PID = 2PABA for prealbumin chain A), "method" is one of Suppos, Comp3D, or Dali, and FSSP stands for family of structurally similar proteins. The file format (Table 1) consists of four sections: (1) a general header, (2) a one-line summary for each aligned protein, (3) a multiple alignment showing the aligned sequence segments and their secondary structure, and (4) a straight list of the matching fragments with their residue numbers. The list of aligned proteins is sorted by the number of equivalenced residues.

Exclusions from the database

Subjective empirical cutoffs were applied to select the most interesting similarities and to exclude trivial short alignments of pieces of secondary structure. At present, only the best alignment per protein pair is included. For Suppos, we report all alignments with at least four fragments or 45 residues. For Comp3D, we include all alignments where the number of equivalences is at least 35 residues or 50% of the length of the shorter protein. For Dali, we select all alignments with similarity scores more than one standard deviation above the mean value of all

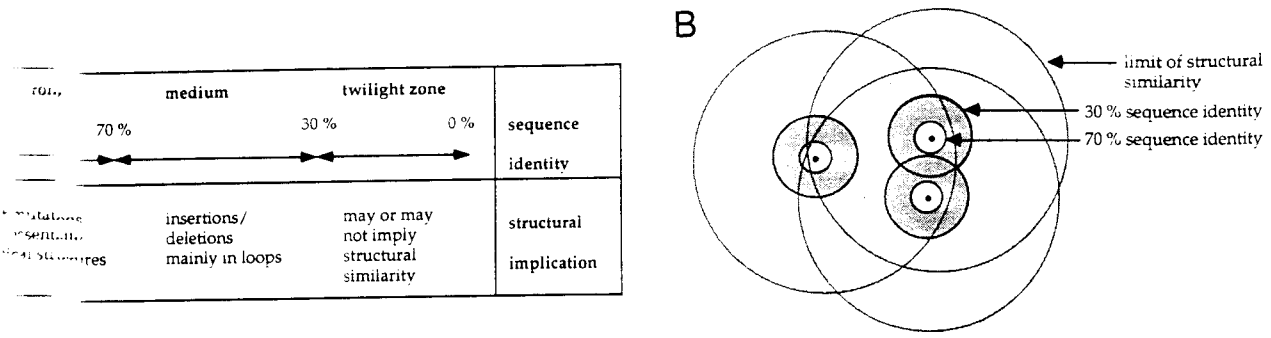

Fig. 1. Ranges of similarity between proteins. **A:** The structural implication of sequence similarity. Very little structural variation is observed in the 70–100% range of sequence identity. **B:** Protein families and relatives included in the database. There is one data set for each of 154 representative families. The central member of each family, i.e., the search structure, is shown as a dot in the center of concentric circles corresponding to strong sequence similarity (white), medium sequence similarity (shaded), and twilight zone (lightly shaded) with structural similarity but without significant sequence similarity. The central member of each family has less than 30% sequence identity with any other central member. For reasons of economy, the database of three-dimensional (3D) alignments only includes relatives in the medium and twilight zone range, but not in the strong range. In the medium range, the database contains all available 3D structures; in the twilight zone range, only representatives of each protein family. By definition, families can overlap, especially in the twilight zone; only a few proteins in the representative set share structures in the medium zone (pair on the right). The large radius of convergence of the structure comparison search precludes reconstruction of nonoverlapping families but is of great advantage in discovering remote structural similarities.

alignments with the given search structure. The cutoffs used are given in the header of the files and may be readjusted in the future, as a result of user experience.

Below, we show examples of structural alignments generated with each of the three methods.

Detection of common substructures: β-sandwiches

The immunoglobulin fold is a sandwich of two sheets formed by seven or nine antiparallel β-strands. Domains belonging to the immunoglobulin superfamily are widely distributed in nature, particularly in cell-surface and secreted proteins, e.g., immunoglobulins, T-cell receptors, and fibronectin type III repeats of extracellular modular proteins (Bork, 1992). The very close structural similarity of papD, a bacterial chaperonin protein, to the immunoglobulin fold was taken to suggest horizontal gene transfer (Holmgren & Brändén, 1989). Figure 2 and Kinemage 1 show the alignment of the CD4 T-cell receptor structure with papD, generated with the Suppos option of WHAT IF (Vriend, 1990).

Recognition of folding motifs: $(\alpha\beta)_n$ domains

Cores built of alternating α-helices and β-strands, sometimes called α/β or $(\alpha\beta)_n$ domains, are found in a large number of proteins. The $(\alpha\beta)_n$ folds differ in the number of β-strands, their orientations, and the way they are connected. For example, parallel sheets are present in dehydrogenases with strand order 3-2-1-4-5-6; in GTPase domains such as p21 ras with strand order 2-1-3-4-5-6; in flavodoxin and cheY with strand order 2-1-3-4-5. Figure 3 and Kinemage 2 show a subset of alignments obtained between arabinose binding protein (1ABP) and the rest of

Fig. 2. Structure alignment of papD and CD4. Structural alignment and optimal superposition of papD protein (3DPA) and CD4 T-cell receptor (1CD4). The C^{α}-rmsd after optimal superposition is 1.5 Å for 41 aligned residues. PDB residue numbers of the aligned fragments follow. For 3DPA: 16–38, 85–94, 105–112. For 1CD4: 111–133, 154–163, 166–173. The alignment was generated using the Suppos algorithm.

Table 1. *Example of an FSSP file*[a]

A. Header block

```
FSSP       FAMILIES OF STRUCTURALLY SIMILAR PROTEINS, VERSION 0.1 1992
PDBID      51dh
//
HEADER     OXIDOREDUCTASE, CHOH DONOR, NAD ACCEPTR 29-OCT-80   5LDH
COMPND     LACTATE DEHYDROGENASE H=4= AND S-$LAC-/NAD$==+== COMPLEX
SOURCE     PIG (SUS SCROFA) HEART
AUTHOR     U.M.GRAU,M.G.ROSSMANN
SEQLENGTH   333
NALIGN       35
//
```

B. Summary

```
## PROTEINS : PDB/chain identifiers and structural alignment statistics
  NR. STRID1 STRID2 RMSD LALI LSEQ2 %IDE REVERS PERMUT NFRAG TOPO  PROTEIN
   1: 51dh   21dx    2.0  312   331   62      0      0     8 S     APO-LACTATE DEHYDROGENASE (E.C.1.1.1.27), ISOENZYME C=4=
   2: 51dh   111c    1.8  311   320   32      0      0    11 S     L-*LACTATE DEHYDROGENASE (E.C.1.1.1.27) COMPLEX WITH
   3: 51dh   21db    2.1  297   301   35      0      0     8 S     L-*LACTATE DEHYDROGENASE (E.C.1.1.1.27) COMPLEX WITH
   4: 51dh   4mdh-A  2.7  296   333   19      0      0    21 S     CYTOPLASMIC MALATE DEHYDROGENASE (E.C.1.1.1.37)
   5: 51dh   1gdi-O  8.7  164   334   16      0      0    16 S     $HOLO-*D-*GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
   6: 51dh   1s01    4.1  137   275    9      0      0    17 S     SUBTILISIN /BPN$(PRIME) 8350 (E.C.3.4.21.14) (MUTANT WITH
   7: 51dh   2aat    4.8  126   396    6      0      0    14 S     ASPARTATE AMINOTRANSFERASE (E.C.2.6.1.1) MUTANT K258A
   8: 51dh   1etu    3.5  122   177    9      0      0    12 S     ELONGATION FACTOR TU (DOMAIN I) - *GUANOSINE DIPHOSPHATE
   9: 51dh   1fxi    3.5  122   147    9      0      0    11 S     FLAVODOXIN
  10: 51dh   21bp    7.2  117   346   18      0      2    12 N     LEUCINE-BINDING PROTEIN (/LBP$)
  11: 51dh   3pgm    3.7  117   230    9      1      3    12 N     PHOSPHOGLYCERATE MUTASE (E.C.2.7.5.3) DE-PHOSPHO ENZYME
  12: 51dh   4fxn    3.2  116   138    9      0      0     9 S     FLAVODOXIN (SEMIQUINONE FORM)
//
```

C. Multiple structure alignment

```
## ALIGNMENTS     1 -   25
 SeqNo PDBNo AA STRUCTURE BP1 BP2  ACC NOCC  .  .  .  .  :  .  .  .  .  :  .  .  .  .  :  .  .  .  .  :  .  .  .  .  :
//
   23    24   I  E      -ab  48  91A  19   34  Ie Ve Ve Ve Ve Ve Ve Ae Ve L_ M_ Ve Fe Ve Le Ve Ce Le .  Le Ie Ve Ye Me
   24    25   T  E      -ab  49  92A   7   34  Te Ie Ve Le Ge Ae We Ge Le Ge Ve Ke Ye Le Ae Le Ae Ae Ae .  Ye Pe Le Le Ve
   25    26   V  E      -ab  50  93A  17   34  Ve Le Ve Ve Ie Ve V_ Te Ie Ve Le Ie Ge Ve Ie De Me Ve Ie .  Le Ye G_ Ve De
   26    27   V  S    S+     0   0   19   32  Ve Ve Ie Te Ne Ie S_ Ie Ve Ee V_ Ve Ve Ke Ie Ge Ve Fe N_ .  Ke Ie M_ Ie Ve
   27    28   G        +     0   0    5   25  G_ G_ G_ Ge G_ De Ne .  Y_ .  .  Ye .  Q_ G_ .  .  G_ L_ .  Re E_ Sh G_ V_
   28    29   V        +     0   0    3   21  V_ Dt A_ At F_ Ss Pe Kh .  .  .  We .  Pt A_ .  .  L_ Ve .  Eg .  Eh G_ At
   29    30   G  S  > S-    0   0   22   25  Ge Gt Ge Ge Ge G_ .  Th Eh Gh Gt Eh .  Et Ge .  G_ Ge .  .  Dg Vh .  Ge Nt
   30    31   Q  H  > S+    0   0  111   26  Dh Ah Fh Qh Rh Se .  Th Yh Ih Nt Kh .  .  Ph Ge Rh Gh .  .  Lg Dh .  Sh Hb
   31    32   V  H  > S+    0   0   25   26  Vh Vh Vh Ih Ih Hh S_ Lh Th Mh Sh Mh Ah .  Sh Kh Th Vh .  .  .  Vh .  Gh .
   32    33   G  H  > S+    0   0    0   27  Gh Gh Gh Ah Gh Gh W_ Th Ah Eh Lh Ah Ih .  Gh Gh Vh Gh .  .  .  Th .  Gh .
   33    34   M  H  X S+    0   0   28   28  Mh Sh Ah Yh Rh Th Pt Ah Eh Fh Rh Eh Rh Ee Lh Th Eh Lh .  .  .  Ph .  Lh .
   34    35   A  H  X S+    0   0    6   28  Ah Sh Sh Sh Nh Hh Nt Ah Th Mh Gh Lh Sh Ph Lh Vh Sh Sh .  .  .  Ah .  Ah .
   35    36   C  H  X S+    0   0   21   28  Ch Yh Yh Lh Vh Vh Ht Ih Ih Gh Lh Ih Vh Wh Lh Sh Lh Vh .  .  .  Mh .  Sh .
   36    37   A  H  X S+    0   0    7   28  Ah Ah Vh Lh Fh Ah Kt Th Ah Ah Vh Ah Vh Fh Gh Sh Th Ih .  .  .  Lh .  Ah .
//
```

D. Aligned fragments

```
## FRAGMENTS: ranges of superimposed residues
  NR. STRID1 STRID2                          RANGE1 <--> RANGE2
//
   2: 51dh   111c    16 THR (  16) -  80 ASN (  81) <-->   4 THR (  16) -  68 ALA (  81)
   2: 51dh   111c    82 ASP (  84) -  94 VAL (  96) <-->  69 GLU (  84) -  81 ILE (  96)
   2: 51dh   111c    95 THR (  97) - 101 GLN ( 103) <-->  83 ALA (  98) -  89 PRO ( 105)
   2: 51dh   111c   104 GLU ( 107) - 109 LEU ( 112) <-->  90 GLY ( 106) -  95 ASP ( 111)
   2: 51dh   111c   110 VAL ( 113) - 206 VAL ( 208) <-->  97 VAL ( 113) - 193 ILE ( 208)
   2: 51dh   111c   208 GLY (209B) - 213 GLN ( 211) <--> 194 GLY (209A) - 199 ALA (210B)
   2: 51dh   111c   214 LEU ( 212) - 225 GLU ( 223) <--> 201 TRP ( 212) - 212 LYS ( 223)
   2: 51dh   111c   227 TRP ( 225) - 240 GLU ( 238) <--> 213 LEU ( 225) - 226 GLU ( 238)
   2: 51dh   111c   241 VAL ( 239) - 246 GLY ( 244) <--> 228 ILE ( 240) - 233 ALA ( 245)
   2: 51dh   111c   248 THR ( 246) - 284 GLU ( 282) <--> 234 THR ( 246) - 270 ASN ( 282)
   2: 51dh   111c   286 GLU ( 284) - 333 LEU ( 331) <--> 271 ASP ( 283) - 318 LYS ( 331)
//
   4: 51dh   4mdh-A  19 PRO (  19) -  28 VAL (  29) <-->   2 GLU (   2) -  11 ALA (  11)
   4: 51dh   4mdh-A  29 GLY (  30) -  42 LYS (  43) <-->  13 GLY (  13) -  26 GLY (  26)
   4: 51dh   4mdh-A  44 LEU (  45) -  53 VAL (  54) <-->  33 GLN (  33) -  42 ILE (  42)
   4: 51dh   4mdh-A  54 LEU (  55) -  67 HIS (  68) <-->  45 MET (  45) -  58 ASP (  58)
   4: 51dh   4mdh-A  70 LEU (  71) -  75 PRO (  76) <-->  59 CYS (  59) -  64 LEU (  64)
   4: 51dh   4mdh-A  76 LYS (  77) -  86 THR (  88) <-->  66 ASP (  66) -  76 ALA (  76)
   4: 51dh   4mdh-A  87 ALA (  89) - 101 GLN ( 103) <-->  78 LYS (  78) -  92 ARG (  92)
   4: 51dh   4mdh-A 108 ASN ( 111) - 130 ASN (132B) <-->  99 ASP (  99) - 121 LYS ( 121)
   4: 51dh   4mdh-A 131 CYS ( 133) - 152 GLY ( 154) <--> 123 VAL ( 123) - 144 PRO ( 144)
   4: 51dh   4mdh-A 153 LEU ( 155) - 162 GLY ( 164) <--> 146 ILE ( 146) - 155 THR ( 155)
   4: 51dh   4mdh-A 164 ASN ( 166) - 186 HIS ( 188) <--> 156 ARG ( 156) - 178 LYS ( 178)
   4: 51dh   4mdh-A 187 GLY ( 189) - 208 GLY (209B) <--> 180 VAL ( 180) - 201 LEU ( 201)
   4: 51dh   4mdh-A 209 VAL (209C) - 214 LEU ( 212) <--> 205 GLU ( 205) - 210 GLU ( 210)
   4: 51dh   4mdh-A 215 ASN ( 213) - 220 THR ( 218) <--> 212 VAL ( 212) - 217 TRP ( 217)
   4: 51dh   4mdh-A 227 TRP ( 225) - 240 GLU ( 238) <--> 219 LYS ( 219) - 232 ALA ( 232)
   4: 51dh   4mdh-A 241 VAL ( 239) - 268 SER ( 266) <--> 234 ILE ( 234) - 261 PRO ( 261)
   4: 51dh   4mdh-A 270 ILE ( 268) - 277 VAL ( 275) <--> 265 PHE ( 265) - 272 SER ( 272)
   4: 51dh   4mdh-A 278 GLN ( 276) - 297 ALA ( 295) <--> 274 GLY ( 274) - 293 ASP ( 293)
   4: 51dh   4mdh-A 299 GLY ( 297) - 305 ASN ( 304) <--> 294 LYS ( 294) - 300 GLU ( 300)
   4: 51dh   4mdh-A 306 GLN ( 305) - 324 TRP ( 323) <--> 302 LEU ( 302) - 320 ALA ( 320)
   4: 51dh   4mdh-A 326 ILE ( 325) - 333 LEU ( 331) <--> 321 GLU ( 321) - 328 GLU ( 328)
//
```

[a] The listing is truncated; deleted parts are marked by //. The full file (5ldh_dali.fssp) is given on the Diskette Appendix.

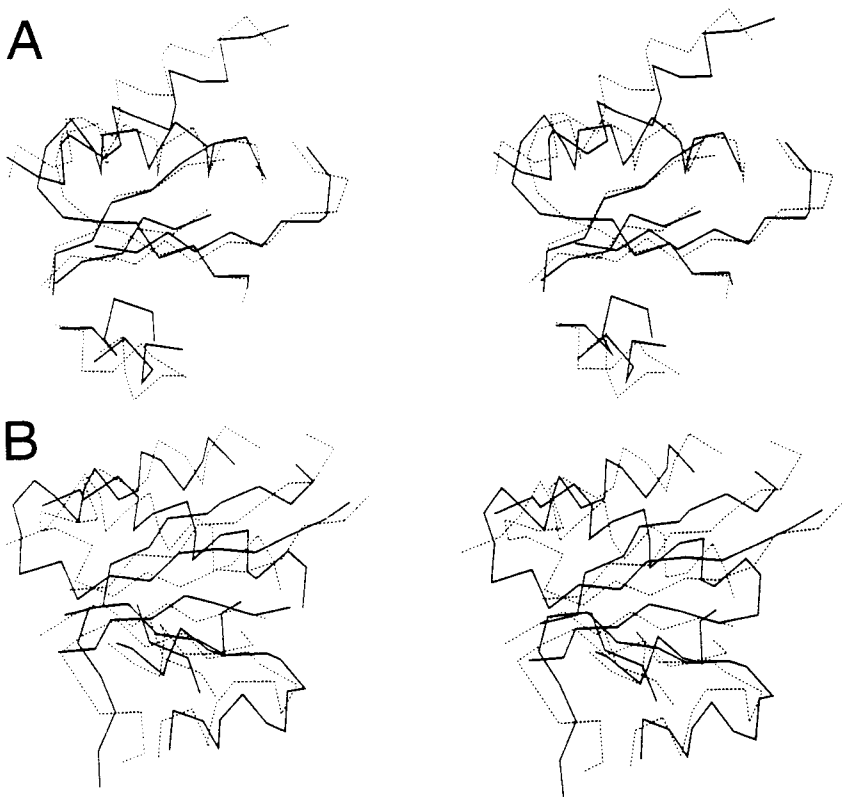

Fig. 3. Structure alignment of arabinose binding protein (ABP) with flavodoxin (FXN) and malate dehydrogenase (MDH). **A:** Structural alignment and optimal superposition of 1ABP and 4FXN. The C^{α}-rmsd after optimal superposition is 2.7 Å for 78 aligned residues. PDB residue numbers of the aligned fragments follow. For 1ABP: 5–15, 17–40, 49–54, 59–64, 73–79, 83–89, 253–269. For 4FXN: 1–35, 39–44, 47–52, 72–78, 80–86, 122–138. **B:** Structural alignment and optimal superposition of 1ABP and 4MDH. The C^{α}-rmsd after optimal superposition is 3.4 Å for 97 aligned residues. Residue numbers of the aligned fragments follow. For 1ABP: 2–27, 31–41, 45–53, 59–66, 73–94, 103–106, 254–270. For 4MDH (chain A): 1–26, 31–41, 72–88, 113–130, 135–138, 151–154, 240–256. The alignments were generated using the Comp3D algorithm.

the database, using the Comp3D algorithm. The two domains of 1ABP are similar to GTPases. The substructure most frequently identified in the database scan is a parallel β-sheet with four or five strands sandwiched between two helices each on either side of the sheet.

Identification of common structural cores: $(\alpha\beta)_8$ barrels

To date, 20 structures belonging to the $(\alpha\beta)_8$ fold family have been determined by X-ray crystallography. The structural principles of this apparently very robust folding motif are partly understood (Lasters et al., 1988; Lesk et al., 1989). The optimal 3D rigid-body superpositions of a selection of five $(\alpha\beta)_8$ barrel proteins with widely different sequences in Figure 4 and Kinemage 3 correspond to alignments generated by the program Dali using the α-subunit of tryptophan synthase as the guide structure. All eight β-strands and α-helices that form the barrel are identified as structurally equivalent in spite of considerable geometrical distortions of the basic barrel shape.

Discussion

Comparison of algorithms

The three methods used for structural alignment differ in basic concept, algorithm, and speed. For full-length protein domains clearly related by evolution, all three algorithms and similarity measures give very similar alignments. As can be expected, the alignments differ in detail for more remote cases. Nonsequential alignments are identifiable by the Suppos and Dali algorithms, but not by the Comp3D algorithm. The Suppos and Comp3D algorithms use positional rmsd criteria calculated from rigid-body superpositions, whereas the Dali algorithm optimizes the agreement of intramolecular pair relationships. Interdomain motion is detectable by the Comp3D and Dali algorithms, but not by the Suppos algorithm. Note that the rmsd values for the rigid-body superposition corresponding to the full alignment can be quite high for methods that allow flexibility between different parts of the molecules. The programs Suppos and Comp3D create the full database in 1 and 3 days of CPU time, respectively. The program Dali is presently an order of magnitude slower, but it implements the most general method that is both sensitive and accurate.

Updates and supplementary information

The FSSP database provides a rich source of comparative information on protein anatomy. Incremental updates to the present database are planned with new PDB releases. It is often useful to complement the compilation of structures with sequence and variability information by direct reference to the latest version of the HSSP database of carefully sequence-aligned protein families, par-

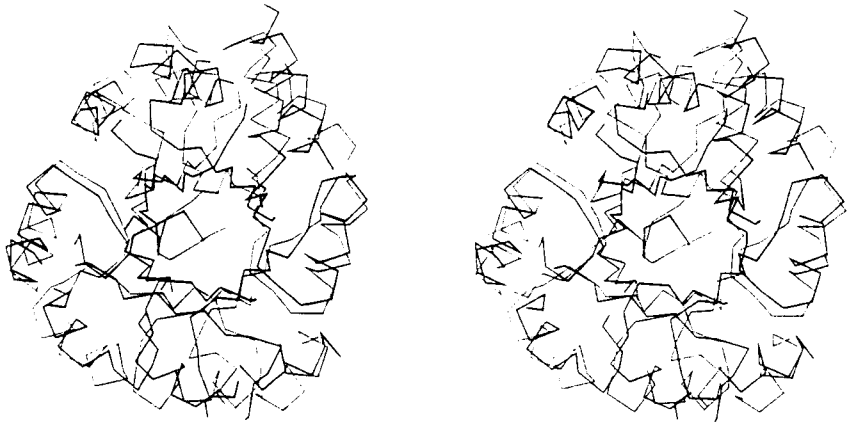

Fig. 4. Structure alignment of $(\alpha\beta)_8$ barrels. Structural alignment and optimal superposition of tryptophan synthase (1WSY) and flavocytochrome b2 (1FCB). The C^{α}-rmsd after optimal superposition is 3.1 Å for 198 aligned residues. PDB residue numbers of the aligned fragments follow. For 1WSY (chain A): 3–8, 17–28, 29–43, 44–53, 84–92, 93–105, 108–130, 131–144, 148–159, 162–177, 192–201, 202–243, 250–265. For 1FCB (chain A): 182–187, 190–201, 205–219, 221–230, 234–242, 245–257, 259–281, 330–343, 344–355, 356–371, 387–396, 400–441, 442–457. The alignment was generated using the Dali algorithm.

ticularly in the 70–100% range of sequence identity where structural and sequence-based alignments are usually identical (Sander & Schneider, 1991; data sets PID.hssp, where PID is the PDB identifier). For example, many additional globin sequences, aligned to any particular globin search structure, can be extracted from the HSSP database. Users interested in detailed local structural properties of each protein, such as solvent accessibilities, may refer to the DSSP database of secondary structures, derived from PDB files (Kabsch & Sander, 1983; data sets PID.dssp). For statistical investigations, FSSP, HSSP, and DSSP data for a set of representative protein chains, rather than for all known structures, may be used, avoiding double counting in the partially redundant PDB. A representative set is chosen such that no two protein chains have sequence identity exceeding a user-defined threshold, e.g., 25% (Hobohm et al., 1992; data set pdb_select.NN, where NN is the number of the PDB release, e.g., NN = 58).

Automatic versus manual alignments

Pascarella and Argos (1992) have described a collection of structure superpositions gathered in part from the literature and cross referenced with sequence-based alignments. Our approach in generating the FSSP database has been to use automatic procedures (as done for seven protein families in Overington et al., 1990) for completeness, consistency, and ease of maintenance; and, to include only structure-based alignments. Users interested in sequence-aligned families are referred to the new version of the HSSP database, which is based on an improved method of profile alignments (Sander & Schneider, unpubl.).

Recommendations for use of the database

Understanding the structural role of residues in remote relatives requires reliable structural alignments. For example, the FSSP data sets can be used to derive generalized sequence patterns specific for a given family of divergently related proteins, such as actin, hexokinase, and the heat shock protein hsp70 (Bork et al., 1992); to study the structural principles of one particular type of fold, such as that of $(\alpha\beta)_8$ barrels; to define structural cores for modular construction of novel proteins or for model building by homology; or, to discover new folding units, i.e., compact substructures that are part of otherwise different folds. The user can select the appropriate subset of aligned structures or substructures depending on the scientific question under consideration. In the database, the number of fragments and the number of residues equivalenced between the guide structure and the aligned protein as well as the topological type (sequential or nonsequential) and sequence identity of the alignment are given as a guide.

We encourage the use of 3D graphics to evaluate and study the structural alignments, especially those of remotely related pairs. A graphical interface for interactively "dialing through" all superposed members of a structural family is provided in the program package WHAT IF (Vriend, 1990; available on license agreement from G.V.). An X-windows-based protein query and 3D inspection system, ProtQuiz (Sander & Scharf, unpubl.; test version available via anonymous ftp from ftp.embl-heidelberg.de), can be used for interactive evaluation of pairwise alignments.

Availability of the database

The FSSP data sets are available via Internet, either by electronic mail from the EMBL file server or by anonymous ftp (file transfer protocol). For example, to request a file with structures aligned to flavodoxin data set 4FXN using the algorithm Suppos, send the electronic mail message help proteindata and send proteindata:4fxn_suppos.fssp to netserv@embl-heidelberg.de. Alternatively, start ftp, connect to ftp.embl-heidelberg.de and log in as anonymous, give your email address as password, and type get /pub/databases/protein_extras/fssp/4fxn_suppos.fssp to transfer the file directly. FSSP data sets may

not be incorporated into other databases and may not be used commercially without written permission of the authors.

Methods

Fragment pair cluster algorithm (Suppos)

This quick search algorithm (Vriend & Sander, 1991) uses a clustering approach to join pairs of fragments with similar backbone conformation into a set of globally best 3D matches. The clustering step allows different topological connections and reversal of chain direction. The best match is defined as the cluster that has the most residues and fits within given distance cutoffs in the optimal rigid-body superposition of equivalenced C^{α} positions. The cutoffs used were 2.5 Å maximal rmsd of C^{α} positions and 4.5 Å maximal absolute deviation of any single pair of C^{α} positions. The algorithm has three steps. (1) Find all fragment pairs that superpose well. Distance geometry criteria are used to compare all nongapped fragments of length 11 residues or longer in one protein to those in the other. The optimal 3D superposition for each fragment pair is calculated using the fast algorithm by Kabsch (1978). (2) Cluster pairs of fragments into larger structural units. The criteria for extending a cluster are based on comparing the rotation/translation matrices for optimal superposition. The idea is that all fragments that are part of a multifragment 3D substructure must obey similar rotation–translation transformations. (3) Fine tune the set of equivalenced residues. The largest cluster identified in the previous step is kept. Optimal 3D superpositions are calculated explicitly and residues are reassigned until a self-consistent set of equivalences is obtained. The implementation of this algorithm in WHAT IF (Vriend, 1990) is fast enough to allow for a pairwise alignment of all 154 representative structures in 1 day of CPU time on a workstation.

Trailing trace superposition algorithm (Comp3D)

The second method (Sander & Tuparev, unpubl.) makes use of the dynamic programming techniques used in sequence alignment (Smith & Waterman, 1981; Gotoh, 1982). Optimal alignment is achieved by finding the best sequential path through a matrix of local residue-residue similarities scores. The best path is the one with the highest score, as summed for all residue pairs in the path. Paths may include insertions and deletions. The balance between extending an alignment and inserting gaps is governed by empirical gap penalties. Gaps are allowed in either protein or in both. The crucial difference of the Comp3D algorithm compared to sequence alignment is that the definition of the local similarity score is based on 3D structure rather than on one-dimensional sequence. This is done by exploiting a special property of this type of algorithm. The best path through the similarity matrix is given by the simple induction rule that a forward extension of the optimal path at a given point (i, j) in the similarity matrix must include the optimal path up to that point. As a result, the entire "traceback" up to the point (i, j) is available at the time the local choice at (i, j) is made. Comp3D optimally (Kabsch, 1978) superposes all equivalenced residues in the alignment trace up to (i, j) and then measures the structural similarity of residues $(i + 1, j + 1)$ in the coordinate frame of that superposition, based on the distance between the C^{α} atoms of residues $i + 1$ and $j + 1$. Alternatively, the local similarity can be calculated from the difference in rmsd for the superpositions with and without the new pair.

This "trailing trace" superposition provides the necessary cooperativity in 3D. As the alignments get longer, the superpositions gradually change. Interdomain motion can be detected, if the memory of the trailing alignment used for superposition is limited to about half the size of a typical domain, say 40–50 residues (folding unit size). In the present implementation, the program Comp3D can do an all-against-all comparison of the set of 154 proteins in 3 days CPU time on a workstation.

Distance matrix alignment algorithm (Dali)

The third method (Holm & Sander, 1992a) evaluates the match of a pair of structures by comparing the agreement of intramolecular distances rather than positional deviations after 3D superposition. Similarity between two aligned substructures is defined as a sum over similarities of *all* equivalent intramolecular C^{α}–C^{α} distances. In this way, 3D cooperativity is built in. Because of the pairwise dependencies the additivity condition of dynamic programming is not valid and a Monte Carlo procedure is used for maximizing the similarity score.

The similarity score is defined so that smaller distance deviations correspond to higher similarity. The method is made tolerant to spatially extended geometrical distortions by using relative rather than absolute distance deviations, preventing dominance of the longest intramolecular distances; and, to large interdomain rotations by using a damping factor as a function of intramolecular distance, which limits the range of the score to the radius of a typical domain (20 Å).

The algorithm has two steps. (1) Identify similar distance patterns in the two distance matrices. Elementary distance patterns of a given size, e.g., hexapeptide–hexapeptide submatrices, are systematically compared to identify pairs of fragments in one protein that have similar (tertiary) interactions as a pair of fragments in the other protein. (2) Monte Carlo optimization of the similarity score. Alignments are initiated by a pair of similar distance patterns, i.e., two hexapeptide pairs, one pair in each protein. The basic moves are addition, replacement, and removal of residue equivalence assignments. Coherence in three di-

mensions is assured by requiring that a distance pattern can be added only if it shares at least one equivalenced fragment with a distance pattern already in the alignment. Several alignments are built up in parallel to cover different regions of alignment space. The algorithm does not require sequential ordering of fragments, so that changes in the topology of loop connections on an otherwise similar core (permutations) can be detected, as well as reversals of chain direction, e.g., equivalence between an antiparallel and a parallel strand pair. The present implementation compares one structure against the representative database in an overnight run on a workstation.

Acknowledgments

We thank Philippe Cronet, Uwe Hobohm, Michael Scharf, Reinhard Schneider, Jolanta Stouten, Alfonso Valencia, and Anna Tramontano for discussions and help; members of the EMBL Computer Group for systems support; and, the Bridge program of the Commission of the European Communities and the Human Frontiers Science Program for financial support and EMBO for a fellowship (L.H.).

References

Abagyan, R.A. & Maiorov, V.N. (1988). A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. *J. Biomol. Struct. Dyn. 5*, 1267–1279.

Alexandrov, N.N., Takahashi, K., & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol. 225*, 5–9.

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

Bork, P. (1992). Mobile modules and motifs. *Curr. Opin. Struct. Biol. 2*, 413–421.

Bork, P., Sander, C., & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl. Acad. Sci. USA 89*, 7290–7294.

Fischer, D., Bachar, O., Nussinov, R., & Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J. Biomol. Struct. Dyn. 9*, 769–789.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol. 162*, 705–708.

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

Holm, L. & Sander, C. (1992a). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, in press.

Holm, L. & Sander, C. (1992b). Globin fold in a bacterial toxin. *Nature*, in press.

Holmgren, A. & Brändén, C.-I. (1989). Crystal structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature 342*, 248–251.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A34*, 827–828.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*, 2577–2637.

Lasters, I., Wodak, S., Alard, P., & van Cutsem, E. (1988). Structural principles of parallel β-barrels in proteins. *Proc. Natl. Acad. Sci. USA 85*, 3338–3342.

Lesk, A.M. (1979). Detection of 3-D patterns of atoms in chemical structures. *Commun. ACM (Assoc. Comput. Machinery) 22*, 219–224.

Lesk, A.M., Brändén, C.-I., & Chothia, C. (1989). Structural principles of α/β barrel proteins: The packing of the interior of the sheet. *Proteins Struct. Funct. Genet. 5*, 139–148.

Levine, M., Stuart, D., & Williams, J. (1984). A method for the systematic comparison of the three-dimensional structures of proteins and some results. *Acta Crystallogr. A40*, 600–610.

Mitchell, E.M., Artymiuk, P.J., Rice, D.W., & Willett, P. (1989). Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol. 212*, 154–166.

Overington, J., Johnson, M.S., Sali, A., & Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity. *Proc. R. Soc. Lond. B 241*, 132–145.

Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng. 5*, 121–137.

Rao, S.T. & Rossmann, M.G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol. 76*, 241–256.

Remington, S.J. & Matthews, B.W. (1978). A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. *Proc. Natl. Acad. Sci. USA 75*, 2180–2184.

Remington, S.J. & Matthews, B.W. (1980). A systematic approach to the comparison of protein structures. *J. Mol. Biol. 140*, 77–99.

Rossmann, M.G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol. 105*, 75–95.

Sali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol. 212*, 403–428.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet. 9*, 56–68.

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

Subbarao, N. & Haneef, I. (1991). Defining topological equivalences in macromolecules. *Protein Eng. 4*, 877–884.

Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol. 208*, 1–22.

Vriend, G. (1990). WHAT IF: A molecular modelling and drug design program. *J. Mol. Graph. 8*, 52–56.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins Struct. Funct. Genet. 11*, 52–58.

Zuker, M. & Samorjai, R.L. (1989). The alignment of protein structures in three dimensions. *Bull. Math. Biol. 51*, 55–78.