Protein Science (1994), *3*:1858–1870. Cambridge University Press. Printed in the USA.

Multiple protein structure alignment

WILLIAM R. TAYLOR,[1] TOMAS P. FLORES,[1] AND CHRISTINE A. ORENGO[2]

[1] Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom

[2] Biomolecular Structure and Modelling Unit, Dept. Biochemistry, University College London, Gower Street, London WC1E 6BT, United Kingdom

(RECEIVED April 6, 1994; ACCEPTED July 27, 1994)

Abstract

A method was developed to compare protein structures and to combine them into a multiple structure consensus. Previous methods of multiple structure comparison have only concatenated pairwise alignments or produced a consensus structure by averaging coordinate sets. The current method is a fusion of the fast structure comparison program SSAP and the multiple sequence alignment program MULTAL. As in MULTAL, structures are progressively combined, producing intermediate consensus structures that are compared directly to each other and all remaining single structures. This leads to a hierarchic "condensation," continually evaluated in the light of the emerging conserved core regions.

Following the SSAP approach, all interatomic vectors were retained with well-conserved regions distinguished by coherent vector bundles (the structural equivalent of a conserved sequence position). Each bundle of vectors is summarized by a resultant, whereas vector coherence is captured in an error term, which is the only distinction between conserved and variable positions. Resultant vectors are used directly in the comparison, which is weighted by their error values, giving greater importance to the matching of conserved positions. The resultant vectors and their errors can also be used directly in molecular modeling.

Applications of the method were assessed by the quality of the resulting sequence alignments, phylogenetic tree construction, and databank scanning with the consensus. Visual assessment of the structural superpositions and consensus structure for various well-characterized families confirmed that the consensus had identified a reasonable core.

Keywords: multiple alignment; protein structure comparison

The comparison of protein tertiary structures has been a rich source of insight and understanding into the nature of protein structure and the interactions that give rise to the observed forms. Systematic comparison across widely differing families has led to the identification of recurring folds and substructures that appear to constitute the fundamental building blocks of protein structure. In sequence comparison, equivalent elements (or motifs) have been found mainly through the application of automatic multiple sequence alignment methods to protein families, but the methodology for structure comparison has failed to attain the same degree of sophistication as found in multiple sequence comparison methods.

Using rigid body superposition methods, Sutcliffe et al. (1987) devised a method for the comparison of multiple protein structures. This worked well for proteins that were reasonably related, but for more diverged data the conserved core often dwindled to a small size. The application of the dynamic programming algorithm (the basic sequence alignment method both for global [Needleman & Wunsch, 1970] and local [Smith and Waterman, 1981] alignment) to rigid body superposition (Barton & Sternberg, 1988; Johnson et al., 1990b; Russell & Barton, 1992) alleviates this problem but still encounters difficulties when faced with relative internal domain movement. Methods based on the comparison of structural environments (Taylor & Orengo, 1989b; Sali & Blundell, 1990) have the capacity to overcome these problems and have been applied in a pairwise manner in which simple pair alignments were concatenated to produce a multiple alignment (Johnson et al., 1990a, 1993; Pickett et al., 1992).

The production of a multiple alignment from a matrix of pairwise comparisons, however, is equivalent to an early stage in the development of multiple sequence alignment (Taylor, 1987b) and in that field it was quickly realized that such alignments can easily become inconsistent when the sequences are remotely related. The solution to the problem was the introduction of a consensus (or average) sequence either by gradual accumulation on a

Reprint requests to: William R. Taylor, Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK; e-mail: wtaylor@uk.ac.mrc.nimr.

core (Barton & Sternberg, 1987) or by hierarchic condensation with the formation of a consensus for each subgrouping (Taylor, 1988). Despite the construction of phylogenetic trees by various structure comparison methods, no current method calculates an internal representation of multiple pairwise residue interactions.

The combination of the dynamic programming algorithm with the comparison of structural environments allows any method of multiple sequence alignment to be directly transposed to the structural problem. The current development will be relevant only to those algorithms that use the 3D structure directly (Taylor & Orengo, 1989a, 1989b; Sali & Blundell, 1990), maintaining the full pairwise interactions between residues. Of this type the method of Taylor and Orengo (1989b) is best suited for adaption because it uses only the dynamic programming algorithm, whereas the method of Sali and Blundell (1990) used a combination of dynamic programming and simulated annealing (the latter being a stochastic optimization method). Adopting the method of Taylor and Orengo (1989b), the only problem to overcome in the transposition from sequence to structure is to define the structural equivalent of a consensus sequence (or profile).

The definition of a consensus structure not only serves to reveal and quantify the conserved elements in a family of structures but also provides a structural template on which the sequence of any member of the family with unknown structure can be modeled. A true consensus structure will have advantages over a core derived from a single structure (or bits of different structures) because it will be continuous – with the core regions distinguished simply by a higher weight. This means that there will be no breaks between the core and the loops, which in standard modeling methods are generally constructed by selecting fragments from the general protein structure databank (Jones & Thirup, 1986) with little reference to the family being modeled.

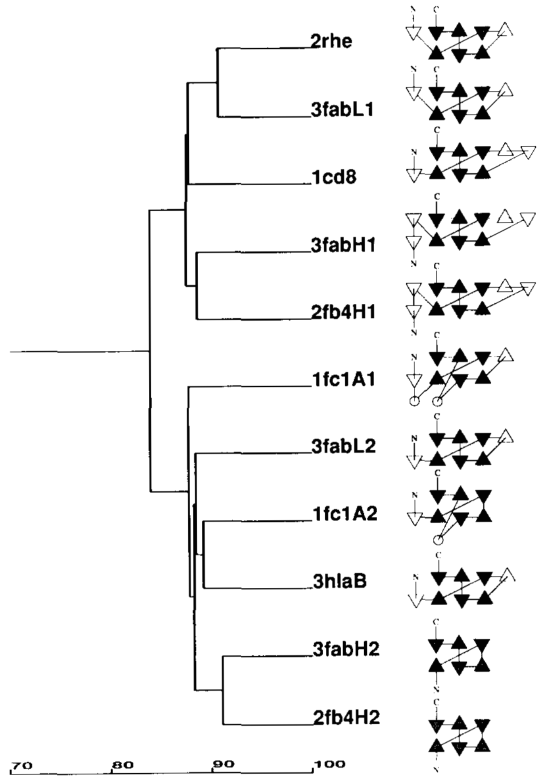

Fig. 1. Dendrogram showing structural relatedness of immunoglobulin domains. The dendrogram was generated by single-linkage cluster analysis of the SSAP pairwise score matrix (Table 1). The axis is labeled with the SSAP scores of 0 up to 100 (for identical structures). The schematic TOPS representation (Flores et al., 1994) is shown adjacent to the corresponding Brookhaven PDB code for the structure. Triangles represent strands and circles helices. Lines penetrating these symbols indicate a connection at the "front" of the secondary structure (otherwise behind).

Results

Immunoglobulin domains

The immunoglobulin domains form a tightly knit family of structures and, although the majority of sequence identities are below 30%, most of the SSAP scores were above 75 (Fig. 1; Table 1). Default values were used for the residue selection (sel_cut) and vector comparison parameters (w, g; see section on implementation details in Methods). Four cycles were applied (with successive score cutoffs 80, 75, 70, 60).

The alignments of the β-strands (A, B, C, D, E, F, G, H; see Fig. 2) agreed with that derived from an inspection of hydrogen bonding patterns (Kabsch & Sander, 1883). It can be seen that the conserved cysteines that hold the sheets (strands B, G) together and the tryptophan residue against which they pack (strand C) are among those residues having the most structurally conserved environments. A multiple superposition based on the equivalenced residues (see Fig. 3 and Kinemage 1) shows the central β-strands having the best fit, whereas edge strands and connecting loops, particularly of the variable light domains, are least well superposed.

Doubly wound domains

Table 2 shows the SSAP scores obtained by comparing pairs of the doubly wound structures. This is a more diverse set than the immunoglobulin folds (above). However, most of the structures align against 3 or more other members with significant scores above 70, suggesting a reasonably well-related group sharing the common framework of the Rossmann fold. Some structures, for example, the flavodoxins (4fxn, 2fx2, and 2fcr) are closely related, whereas there are extensive differences between others

Table 1. *Pairwise SSAP scores immunoglobulin folds*[a]

2rhe00	87.7	86.9	83.9	90.7	79.8	80.4	80.0	86.9	78.5	78.5
1cd800		84.7	76.0	87.4	80.1	79.4	80.2	87.5	78.6	78.4
3fabH1			77.0	85.7	78.2	79.8	79.2	88.6	73.0	79.3
3fabH2				74.4	85.5	84.1	86.8	77.7	91.0	84.8
3fabL1					80.6	76.5	80.0	86.9	79.1	76.6
3fabL2						86.4	88.4	80.3	86.5	86.0
1fc1A1							87.7	80.4	85.0	86.2
1fc1A2								80.9	88.2	89.1
2fb4H1									78.2	79.7
2fb4H2										84.8
3hlaB0										

[a] See section on data in the Methods for the correspondence of the PDB codes.

```
2rhe00  :     ESVLTQP PSASG  TPGQRVTISCTGSATDIG S NSVIWYQQVP  GKAPKLLI
3fabL1  :     XSVLTQP PSVSG  APGQRVTISCTGSSSNIG AGNHVKWYQQLP  GTAPKLL
1cd800  :      SQFRVSPLDRTW  NLGETVELKCQVL  LSN PTSGCSWLFQPRGAAASPTFLL
3fabH1  :     XVQLEQSG PGLV  RPSQTLSLTCTVGTSF   DDYYSTWVRQPP  GRGLEWIG
2fb4H1  :     EVQLVQSG GGVV  QPGRSLRLSCSSS GFIF SSYAMYWVRQAP  GKGLEWVA
1fclA1  :      PSVFLFPPKPKDTLMISRTPEVTCVVVDVSHEDPQVKFNWYVD       GVQVH
3fabL2  : QPKAAPSVTLFPPSSEELQA  NKATLVCLISDFY  PGAVTVAWKAD       SSPV
1fclA2  :     EPQVYTLPPSREEMTK  NQVSLTCLVKGFY  PSDIAVEWESN       GQPE
3hlaB0  :  IQRTPKIQVYSRHP AENG  KSNFLNCYVSGFH  PSDIEVDLLKN       GERI
3fabH2  :          PLAPSSKSTSG  GTAALGCLVKDYF  PEPVTVSWN         SGAL
2fb4H2  :          PLAPSSKSTSG  GTAALGCLVKDYF  PQPVTVSWN         SGAL
score   :     +**##@#+*##*#*+*::####@#@###+*#::*#*####@#***-  +**####:
        :        EE                EEEEEE           EEEEE         EE

                A                  B                C              D

2rhe00  : YYN     DLLPSGVS DRFSASKS          GTSASLAISGLESEDEADYYCAAWN
3fabL1  :            IFHNN ARFSVSKS          GSSATLAITGLQAEDEADYYCQSYD
1cd800  : YLS  QNKPKAAEGLDTQRFSGKRL          GDTFVLTLSDFRRENEGYYFCSALS
3fabH1  : YVFYHG TSDTD TPLRSRVTMLVNTS        KNQFSLRLSSVTAADTAVYYCARNL
2fb4H1  : IIWDDGSDQHYA DSVKGRFTISRNDS        KNTLFLQMDSLRPEDTGVYFCARDG
1fclA1  : N                  AKTKPR EQQY NSTYRVVSVLTVLHQNWLDGKEYKCKVSN
3fabL2  : K                  AGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTH
1fclA2  : N                   NYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMH
3hlaB0  : E                   KVEHSDLSFSKDWSFYLLYYTEFT   PTEKDEYACRVNH
3fabH2  : T                  SGVHTFPAVLQSSGLYSLSSVVTVPSSSLGT QTYICNVNH
2fb4H2  : T                  SGVHTFPAVLQSSGLYSLSSVVTVPSSSLGT QTYICNVNH
score   : *--++: :+++*:***:###*####+*#**+***######@###***#*#*##@#@####
        :                     EEE    EE     E  EEEEEE          EEEEEE

                              E                  F                G

2rhe00  : DSLD        EPGFGGGTKLTVLGQPK
3fabL1  : R  S        LRVFGGGTKLTVLR
1cd800  :   NS        IMYFSHFVPVFLPA
3fabH1  : I        AGCIDVWGQGSLVTVSS
2fb4H1  : GHGFbSSASbFGPDYWGQGTPVTVSSASTKGPSVF
1fclA1  : KAL       PA PIEK TISKAK
3fabL2  : EGS          TVEK TVAPTECS
1fclA2  : EAL       HNHYTQK SLSL
3hlaB0  : VTL       SQ PKIV KWDRDM
3fabH2  : KPS       N TKVDK KVEPKSC
2fb4H2  : KPS       N TKVDK RVEPKSC
score   : *-+-     -+--#########*+-.
        :              EE

                       H
```

Fig. 2. Structure-derived multiple sequence alignment of immunoglobulin domains. Residues in equivalent secondary structure regions, determined from the hydrogen bonding patterns, are shown bold. Residues in the β-strands common to both variable and constant domains (A, B, C, D, E, F, G, H) are all correctly aligned. The consensus SSAP score for each position is shown as a symbol ranked in the order ":-+*#" increasing with degree of conservation of the structural environment.

(Fig. 4). Six structures (4fxn, 2fcr, 2fx2, 3chy, 1etu, 5p21) are missing the edge C-strand, and in p21, the structure is further complicated by an inserted strand between the A and B strands. Proteins 1gdO1, 2fcr, and 2fx2 have additional antiparallel strands in some of the connecting loops (1gd1O1 – B/C and E/F strands, 2fcr – E/F, 2fx2 – D/E), whereas 1adh has a helical insert in the loop connecting the C/D strands. Therefore, despite the common framework, the group provides a wide range of structures with which to test the method.

Parameters for residue selection (sel_cut) and vector comparison (w, g) were set to the default values because the quality of the final alignment was not found to be sensitive to these values. Four alignment cycles were used with decreasing SSAP cutoff scores of 80, 75, 70, and 60. The alignment of the 5 β-strands (A, B, C, D, E; see Fig. 5) agreed with that derived from an inspection of the Kabsch and Sander (1983) hydrogen bonding patterns. That of the sixth edge F-strand was slightly misaligned for 4fxn, 2fx2, 2fcr, and 1etu. This secondary structure is harder to match because the geometry of the strand is distorted by a β-bulge. The correspondence of the α-helices appeared reasonable, given that their location tends to be more variable.

Multiple superposition of all the structures in the group (Fig. 6; Kinemage 2) was performed using equivalences determined by the multiple alignment. Inspection on a computer graphics workstation showed the alignment of the core strands and helices (bA, bB, bD, bE) to be very good, whereas the edge strands and helices showed greater mismatch.

A consensus structure was generated from the average vectors between structurally conserved positions (see section on construction of a consensus template in Methods). The resulting consensus fold consisted of the 4 central strands (A, B, D, E) and 2 central helices (a, b) and contained 68 residues. This minimal Rossmann fold (Mini-Ross) was scanned across a data set of 150 unique nonhomologous folds using the pairwise SSAP algorithm. All known α/β doubly wound domains were matched, both from single and multidomain proteins. TIM barrel folds, which also contain alternating α/β motifs, gave the next best hits. Comparisons between Mini-Ross and the doubly wound

Fig. 3. Multiple structure superposition of immunoglobulin domains. The domains were superposed as rigid bodies using the residue equivalences in Figure 2. Conserved elements of the average structure are drawn bold. This representation gives an impression of the consensus structure but is not an accurate guide to its true internal relationships. (Details of the method can be found in the section on multiple superpositions.)

structures used to derive it gave significant SSAP scores greater than 80 in most cases (see Table 2).

The ability of the consensus Mini-Ross fold to identify related structures was compared with the performance of a representative doubly wound structure. For this the che-Y protein (3chy) was chosen because it matched the largest number of doubly wound folds in the set with good SSAP scores. The SSAP scores observed using Mini-Ross (see Fig. 7A,B) gave a better discrimination between the Rossmann folds and TIM barrels. At least 80% of multidomain structures known to contain a Rossmann fold domain gave SSAP scores between 75 and 85, all the single-domain proteins scored above 80. By contrast, using the che-Y structure produced scores with a poorer discrimination between related and unrelated folds, with a significant proportion of multidomain folds scoring below 70. Furthermore, because the Mini-Ross domain contains fewer residues than the che-Y structure, the scan was faster.

Figure 8 shows the superposition of the structure derived from the Mini-Ross consensus (using the methods in the section on construction of a consensus template) on the equivalent domain in malate dehydrogenase (4mdhA).

TIM barrels

The level of structural similarity among the 5 TIM-barrel folds was much lower than for the immunoglobulins, with SSAP scores generally below 75 (Table 3). This was mainly due to variation in the length, packing, and orientation of the α-helices.

Table 2. *Pairwise SSAP scores for α/β doubly wound folds*[a]

cons	79.6	86.1	84.4	84.7	84.4	85.3	84.5	84.7	83.9	86.2
1grcA0		68.1	73.1	74.5	67.0	69.0	72.6	73.4	70.3	76.5
6ldh01			77.9	76.8	72.7	0.0	76.2	73.8	65.6	77.1
8adh01				78.0	68.3	68.3	67.1	73.1	64.5	77.5
1gd1O1					70.0	69.3	67.1	68.0	63.7	75.7
5p2100						80.6	70.5	74.2	69.4	74.1
1etu00							69.4	73.1	69.7	74.4
4fxn00								88.7	83.3	74.1
2fx200									86.2	76.4
2fcr00										72.5
3chy00										

[a] See section on data in the Methods for the correspondence of the PDB codes. The row cons gives the scores of the consensus structure.

Farber and Petsko (1990) suggested 4 families of TIM-barrel folds, based on a number of structural parameters such as lengths of helices and sheets and the location of additional domains and secondary structures. Most known TIM barrels are enzymes and Petsko and co-workers have suggested divergence from a common ancestor by cyclic permutation. Alternatively, Lesk et al. (1989) analyzed the different modes of residue packing within the barrels, from which they inferred convergent evolution. Using Petsko's classification, flavocytochrome *b*2 (1fcbA1) lies in one family, triose isomerase (5timA), tryptophan synthase (1wsyA), and anthranilate isomerase (1pii) in another, and aldolase (1ald) in a third. Schematic TOPS diagrams shown

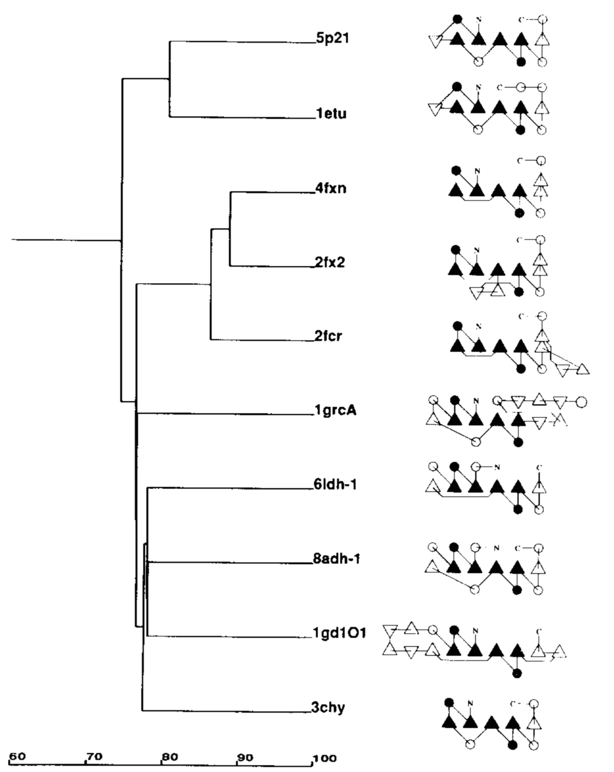

Fig. 4. Dendrogram showing structural relationships in alternating α/β doubly wound folds, generated from the SSAP pairwise score matrix (Table 2). The schematic TOPS representation is shown adjacent to the corresponding Brookhaven PDB code for the structure.

```
5p2100    :                  MTEYKLVVVGAGGV     GKSALTIQLIQNHFVDEYDPTIED
1etu00    :              FERTKPHVNVGTIGHVDH     GKTTLTAAITTVLAKTYGITINTS
4fxn00    :                       MKIVYWSGTGNTEKMAELIAKGIIESG
2fx200    :                     AKALIVYGSTTGNTEYTAETIARELADAG
2fcr00    :                      KIGIFFSTSTGNTTEVADFIGKTLG  A
1grcA0    :                     MNIVVLISG    NGSNLQAIIDACTNK
6ldh01    : ATLKDKLIGHLATSQEPRSYNKITVVGVG      AVGMACAISILMKD
8adh01    :     STGYGSAVKVAKVTQGSTCAVFGLG      GVGLSVIMGCKAAG
1gd101    :                    AVKVGINGFG      RIGRNVFRAALKNP
3chy00    :                 ADKELKFLVVDDF    S TMRRIVRNLLKELG
score     :                  ::++#@@@@@####**::#@@##@@####***-: :  ::*##
          :   HHHHHHHHHH         EEEEE      HH HHHHHHHHHHHH
                                   A

5p2100    : SYRKQVVIDGET CLLDILDTAGQEEY
1etu00    : HVEYDTPT     RHYAHVDCPGHADY
4fxn00    :             K   DVNTINVSD
2fx200    :             Y   EVDSRDAAS
2fcr00    :             K   ADAPIDVDD
1grcA0    :             IKGTVRAVFSNKADAFGLERARQAG                      I
6ldh01    :              LADEVALVDVME DKLKGEMMD            LQHGSLFLHTAKI
8adh01    :               AARIIGVDINK  DKFAKA                   KEV  GAT
1gd101    :             DI EVVAVNDLTDANTLAHLLKYDSVHGRLDAEVSVNGNNLVVNGKEI
3chy00    :              F NNVEEAED
score     : ####****    *+++*@+#@@@#**.:------..                 ..  ..-
          :                 EEEE     HHHHHHHHHHE   EE    EEHHH EHHH  EEE
                               B

5p2100    :                  SAMRDQ    YMRTGEGFLCVFAINNTK            SFE
1etu00    :                  VKNMIT    GAAQMDGAILVVAATDGP            M P
4fxn00    :     V             NI DELL     NEDILILGCS   AMG  D EVLEESEFEP
2fx200    :     V             EAGGLFE     GFDLVLLGCS   TWGDDS IELQ DDFIP
2fcr00    :     V             TDPQALK     DYDLLFLGAP   TWNTGADTERSGTSWDE
1grcA0    : ATHTLIASAFDSREAYDRELIHEIDMYA   PDVVVLAGFMR
6ldh01    : VSGKD               YSVSA     GSKLVVITAGARQQEGESRLNLVQ   RNV
8adh01    : ECVNPQD        YKKPIQEVLTEMSNGGVDFSFEVIGR
1gd101    : IVKAERD        P  ENL AWGEIG   VDIVVESTGR
3chy00    :                 GVDALNKLQAGG   YGFVISDWNMP                 N
score     : ----*.          .:****##*--+  +@@@@@@@##*: --.  . ..:.:**-:-
          : EEE         HHH H        HH       EEEE           HHHHHH   HHH
             C                                 D

5p2100    : DIH QYREQIKRVKDSDDVPMVLVGNKCDLAARTVESRQAQD    LAR     SYGI
1etu00    : QTR EHILLGRQVGV   PYIIVFLNKCDMVDDEELLELVEMEVRELLSQYDFPGDDT
4fxn00    : FIE EIST KI     SGKKVALFGSYG    W     GDGKWMRDFEERMNGYGCVVVE
2fx200    : LFD SLEETGA     QGRKVACFGCGD    SSYE  YFCGAVDAIEEKLKNLGAEIVQ
2fcr00    : FLYDKLPEVDM     KDLPVAIFGLGD    AEGYPDNFCDAIEEIHDCFAKQGAKPVG
1grcA0      ILSPAFVSHYA       GRLLNIHPSL    LA    EEHGTSVHFVTDELDGGPVILQ
6ldh01    : NIFKFIIPNIVKHS  PDCIILVVSNPV    D      VLT         YVAWKLSGL
8adh01    : L   DTMVTALSCCQEAYGVSVIVGVPP           DSQ         NLSMNPML
1gd101    : FTKREDAAKHLEAGA   KKVIISAP              AK         NEDITIVMG
3chy00    : MDGLELLKTIRADGAMSALPVLMVTAEA    KK    ENII         AAAQA
score     : #**:####**#+++. -+#@@@@@##*+###++- :- -*#*:::---::**+*++-:.
          :    HHH     H        EEE
                                 E

5p2100    :                     PYIETSAKTR        QGVEDAFYTLVREIRQH
1etu00    :                     PIVRGSALKALEGDAEWEAKILELAGFLDSYI
4fxn00    : T                    PLIVQN  EP   DEAEQDCIEFGKKIANI
2fx200    : D                    GLRIDG  DP   RAARDDIVGWAHDVRGAI
2fcr00    : FSNPDDYDYEESKSVRDGKFLGLPLDMVNDQ   IPMEKRVAGWVEAVVSETGV
1grcA0    : AKVPVFAGDSEDDIT      ARVQ         TQEHAIYPLVISWFADGRLKMHENAA
6ldh01    :          PMHRII      GSGC         NL
8adh01    :        LLSGRTWK      GAI          FGGFKSKDSVPKLVADFMAKKFA
1gd101    : VNQDKYDPK AHHVI      SNAS         CTTN
3chy00    :            GASG      YVVK         PFTAATLEEKLNKIFEKLGM
score     : .           ..::     ###*##  #:   ++-++*++++++++++-.
          :        HH     EE  EE  E                HHHHHHHHHHHHH      EE
```

Fig. 5. Structure-derived multiple sequence alignment of α/β doubly wound folds. Residues in equivalent secondary structure regions, determined from the hydrogen bonding patterns, are shown bold. Residues in the β-strands are correctly aligned with the exception of a slight displacement in the more variable C-terminal (F) strand. (See legend to Fig. 2 for details.)

in Figure 9 illustrate some of the differences between the folds. Both anthranilate isomerase and flavocytochrome $b2$ contain additional domains, which have been excluded for the purposes of this analysis. In flavocytochrome $b2$, the β-barrel is nearly circular, whereas the other barrels are more elliptical.

Because there is more diversity in this family, thresholds for residue selection were adjusted to allow more residue comparisons between structures (buried areas and angle cutoff, dtot = 200; sel_cut = 30). Similarly, the weight on errors for vector comparison was softened by reducing w in Equation 4. Five

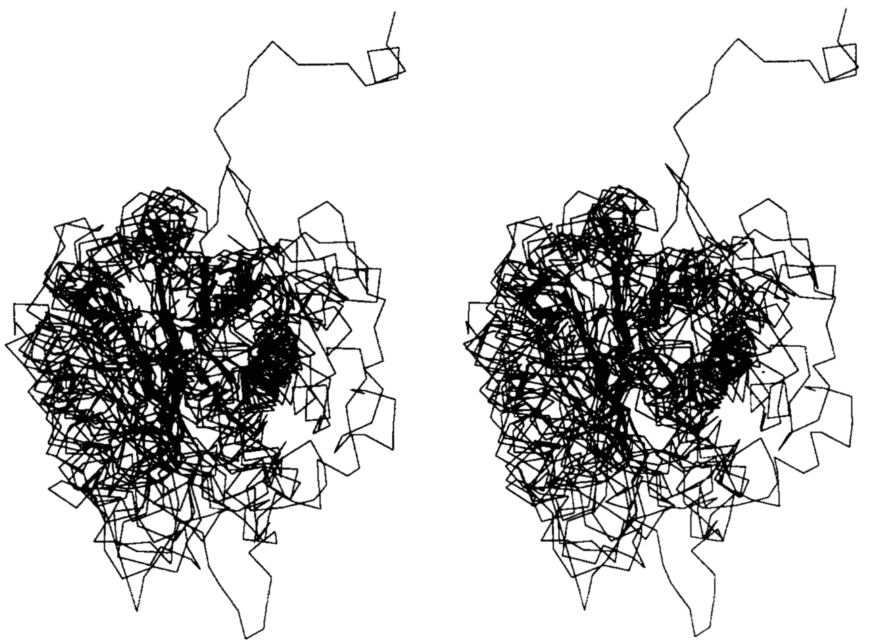

Fig. 6. Multiple structure superposition of α/β doubly wound folds. The domains were superposed as rigid bodies using the residue equivalences in Figure 5. (See the legend to Fig. 3 and the section on multiple superpositions in the Methods.)

alignment cycles were used with successive score cutoffs 80, 75, 70, 60, 50. The alignments of the 8 β-strands (Fig. 10) agreed with hydrogen bonding patterns derived from Kabsch and Sander (1983) and that of the helices was reasonable given the differences in their lengths and orientations. Figure 11 and Kinemage 3 show a multiple superposition of the structures, generated using equivalences from the multiple alignment.

A consensus TIM barrel structure was generated from the alignment consisting of 157 positions. When scanned against a data set of 150 nonhomologous folds, all the barrel structures were matched first, above the Rossmann folds. Table 4 shows all the hits giving SSAP scores above 65. All the TIM barrel folds had scores above 70 and the template matched all the structures in the set used to generate it with scores above 80 (see Table 3

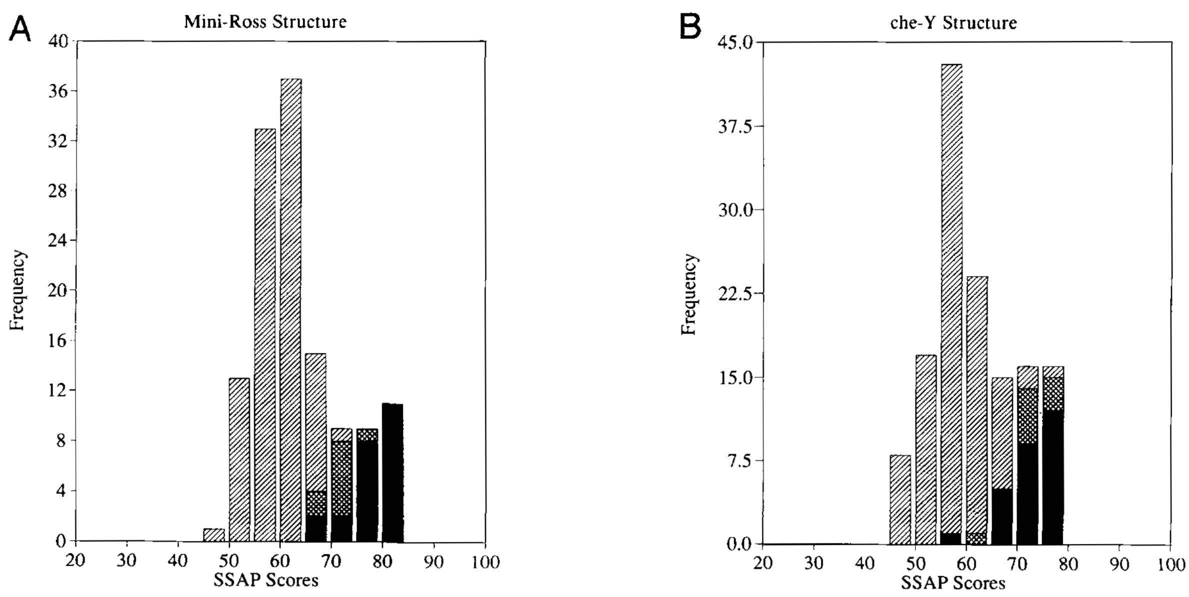

Fig. 7. Distribution of pairwise SSAP scores obtained by scanning (**A**) the consensus Rossmann fold structure Mini-Ross and (**B**) the che-Y structure, against a data set of nonhomologous folds. Solid bars represent correct hits (other alternating β/α-type proteins with a similar fold), whereas crosshatched bars represent hits on TIM-barrel proteins (alternating β/α-type proteins with a different fold). Hatched bars are unrelated folds. The Mini-Ross hits score more highly for correct folds giving better resolution.

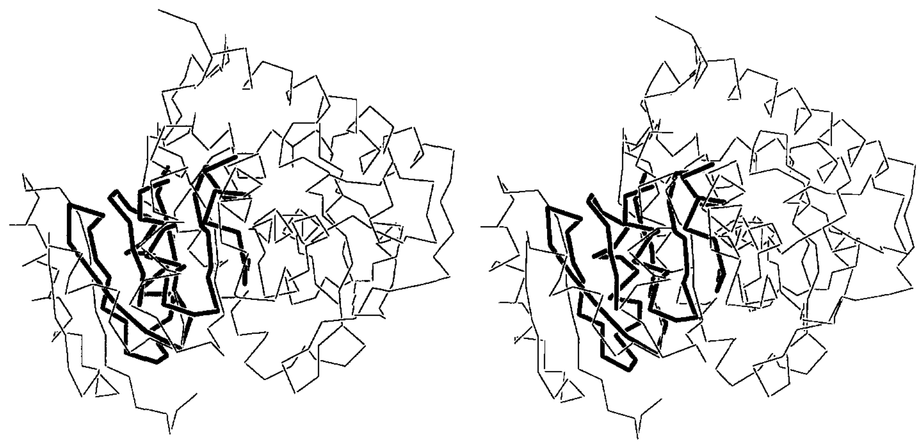

Fig. 8. Superposition of the consensus Rossmann structure (shown in bold) on the Rossmann fold domain of the multidomain protein, malate dehydrogenase (4mdh, chain A). Structures were superposed by the method of Rippmann and Taylor (1991) using residue equivalences generated by a pairwise SSAP alignment.

above). For comparison, Table 4 also shows the pairwise SSAP scores obtained using a representative TIM structure (5timA). It can be seen clearly that the template is a better discriminator of TIM barrel folds

Discussion

The use of the dynamic programming algorithm for all stages in the comparison of protein structures has allowed the methods of multiple sequence alignment to be transposed directly to the problem of multiple structure comparison. The structural equivalent of a position in a multiple sequence alignment (a set of residue types) became a set (bundle) of interatomic vectors. For practical reasons, the bundle was reduced to an average vector and an error term reflecting the coherence of the set. This term was used as a weight, giving reduced emphasis to the comparison of unconserved (incoherent) positions.

Applying the method to typical families of very remotely related proteins, we found that the consensus structures maintained a good core that became increasingly diffuse toward the surface. The consensus structures produced by our method have no distinct boundary between core and loops (ordered and disordered). This should be a great advantage for molecular modeling because the refinement potentials can be specified as target constraints and weighted at a local level by the coherence (error) measure of the vector bundles. This local weight application should overcome the problem often encountered when modeling a new sequence from a family of structures that contain domains in relatively different orientations. In our approach, the interatomic vectors within each domain will be preserved (and so retain their high weighting) irrespective of the relative domain orientations.

When using the consensus structure as a probe against the protein structure databank, we found that the consensus structure was always better than any individual protein that had contributed to it. By analogy with consensus sequence (profile) matching, this might have been expected. However, it was not clear that the effect would have been significant because the information contained in any one structure is much greater than its equivalent sequence (compare the background noise in the dotplots of a sequence comparison and a structural comparison).

Table 3. *Pairwise SSAP scores for TIM-barrel folds*[a]

cons	85.7	84.3	85.6	86.5	84.6
2timA		77.8	75.6	76.9	68.6
1ald			67.4	64.0	74.0
1fcbA_1				61.5	76.4
1pii_1					78.7
1wsyA					

[a] See section on data in the Methods for the correspondence of the PDB codes. The row cons gives the scores of the consensus structure.

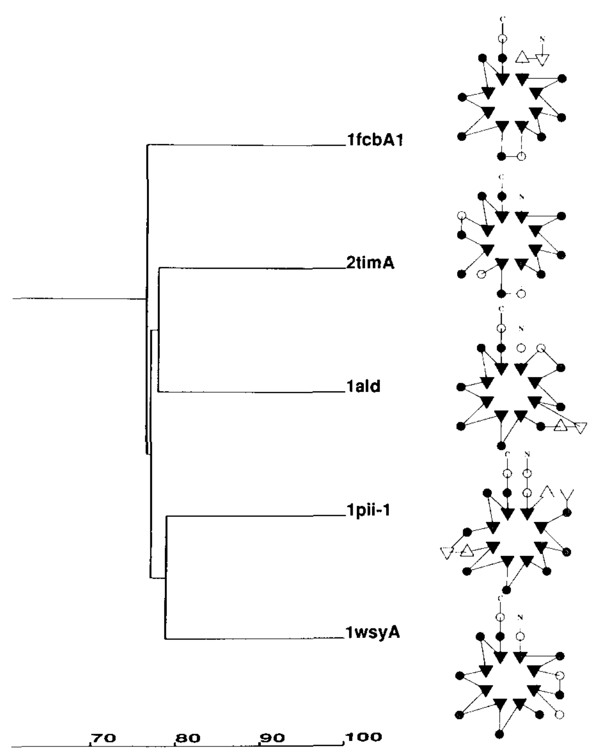

Fig. 9. Dendrogram showing structural relatedness among a set of TIM barrel folds generated from the SSAP pairwise score matrix, as for Figure 1. Schematic TOPS representations of the folds are shown on the far right, adjacent to the corresponding Brookhaven PDB codes.

```
1fcbA1  :                               RKVDISTDMLGS      HVDVPFYVSATA
2timA0  :                                               SKPQPIAAANWKC
1ald00  :                          PYQYPALTPEQKKELSDIAHRIVAPGKGILAADES
1pii01  : MQTVLAKIVADKAIWVEARKQQQPLASFQNEVQPSTRHFYDALQ    GARTAFILECKK
1wsyA0  :                                   MER YENLFAQLNDRREGAFVPFVTL
score   :                                 ..---:++++::   .**#########+
        :    HHHHHHHHHHHHHHHHHHH                HHHHHHHH      EEEE

1fcbA1  : LC             KLGNPLEGEKDVARGCGQGVTKV    PQMISTLAS
2timA0  :   NG               SQQSLSELIDLFNSTSINHDVQCVVASTF
1ald00  : TGSIAKRLQSIGTENTEENRRFYRQLLLTADDRVNP   CIGGVILFH
1pii01  : ASPSKGVI          RDDFDPARIAAIYKHY      ASAISVLTDEKYF     QG
1wsyA0  : GDPGIEQS          L  KIIDTLIDA  GA      DALELGVPFSADGPTIQNAN
score   : +---. .:          --+*########-+#*-+   :++######+--::     ::
        :     HHHHHH      HH     HHHHHH H         EEEEEE        HHHHHH

1fcbA1  :            CSPEEIIEAAPSDKQI QWYQL            YVNSDRKITDD
2timA0  :     VHLA       MTKERLSHPKFVIAAQNA                   IAKSGAFT
1ald00  :     ETLYQKADDGRPFPQVIKSKGGVVGIKVDKGVVPLAGTNGETTTQGLDGLSE
1pii01  :             SFNFLPIVS QIAPQPILCKDFII                 DPY
1wsyA0  : LRAFAAGVTPAQCFEMLALIREKHPTIPIGLLMYANL          VFN NGIDA
score   :     ::::   :--+######-***##+#####:::           ::. :+###
        : HHHHH    HH    HHHHH       EEEE    EHE       EE       HH

1fcbA1  :    LVKNVEKLGVKALFVTVDAPSLGQREKDMKLKFGASRALSKFIDPSLTWKDIEELKK
2timA0  : GEVSLPILKDFGVNWIVLGHSERRAYYG                  ETNEIVADKVAAAV
1ald00  :    RCAQYKKDGADFAKWRCVLKIGEHTPSALAIM           ENANVLARYASICQ
1pii01  :    QIYLARYYQADACLLMLSV                        LDDDQYRQLAAVAH
1wsyA0  :    FYARCEQVGVDSVLVA  D                        VPVEESAPFRQAAL
score   :    ###########@#@##++*:::  .:                 ****##########
        :    HHHHHHH    EEEE            HHHHH                 HHHHHHHH

1fcbA1  : KTKLPIVIKGVQ                  RTEDVIKAAEIG      VSGVVLSNHGGR
2timA0  : ASGFMVIACIGETLQERES     GRTAVVVLTQIAAIAKKLKKADWAKVVIAYEPVWAI
1ald00  : QNGIVPIVEPEILPDGDHDLKRCQYVTEKVLAAVYKALSDHHI  YLEGTLLKPNMVTPG
1pii01  : SLEMGVLTEVS                   NEEEQERAIALG      AKVVGINNR  D
1wsyA0  : RHNIAPIFICPPN                ADDDLLRQVASYG      RGYTYLL
score   : ######@####+::::.:.     ::::::############*: :::####@##+-+--
        : H   EEEEEE   HHHHH HHHHHHHHHHH   HHHHHHH           EEE

1fcbA1  : QLDFSRA PIEVLAETMPIL       EQRNLKDKLEVFVDGGV    RRGTDVLKALCL
2timA0  : GTGKVA  TPQQAQEAHALIRSWVSSKIGADVRGELRILYGGSV     NGKNARTLYQQ
1ald00  : HACTQKFSHEEIAMATVTALRRTV       PPAVTGITFLSGGQSEEEASINLNAINKC
1pii01  : LRDLSID  LNRTRELAPKL    GH      N  VTVISESGI    NTYAQVRELSHF
1wsyA0  :            PLHHLIEKL    KE      YHAAPALQGFGI    SSPEQVSAAVRA
score   : ---:--: +++*########::::::     +#++###@#####    +###########
        :           HHHHHHHHHHHHHHHHH  HH      EEE      HH  HHHHHHHH

1fcbA1  :      GAKGVGLGRPFLYA       NSCYGRNGVEKAIEILRDEIEMSMRLLGVTSIAE
2timA0  :     RDVNGFLVGGASLK                      PEFVDIIKATQ
1ald00  : PLLKPWALTFSYGRALQASALKAWGGKKENLKAAQEEY VKRALANSLACQGKYTPSGQA
1pii01  :       ANGFLIGSALMAH    D            DL HAAVRRVLLGEN
1wsyA0  :      GAAGAISGSAIVKI    IEKNLASPKQMLAEL RSFVSAMK
score   :     :+############+    :  ::::::::::++ +#######-+--
        :         EEEE   HH HHHHHHHH    HHHHHHHH HHHHHHHHH
```

Fig. 10. Structure-derived multiple sequence alignment of 5 TIM barrel folds. Residues in equivalent secondary structure regions, determined from the hydrogen bonding patterns, are shown bold. (See legend to Fig. 2 for details.)

The improvement, as with consensus sequence matching, was probably largely derived from the damped contribution from variable loop regions.

The method described here was based on a global comparison method; however, because of our sole use of the dynamic programming algorithm, any variant of this sequence comparison algorithm can be substituted, including the local alignment method described previously (Orengo & Taylor, 1993). We foresee a useful application of these 2 methods in allowing motifs to be defined and matched against the structure databank – so accumulating a library of consensus fragments suitable for structure prediction and modeling.

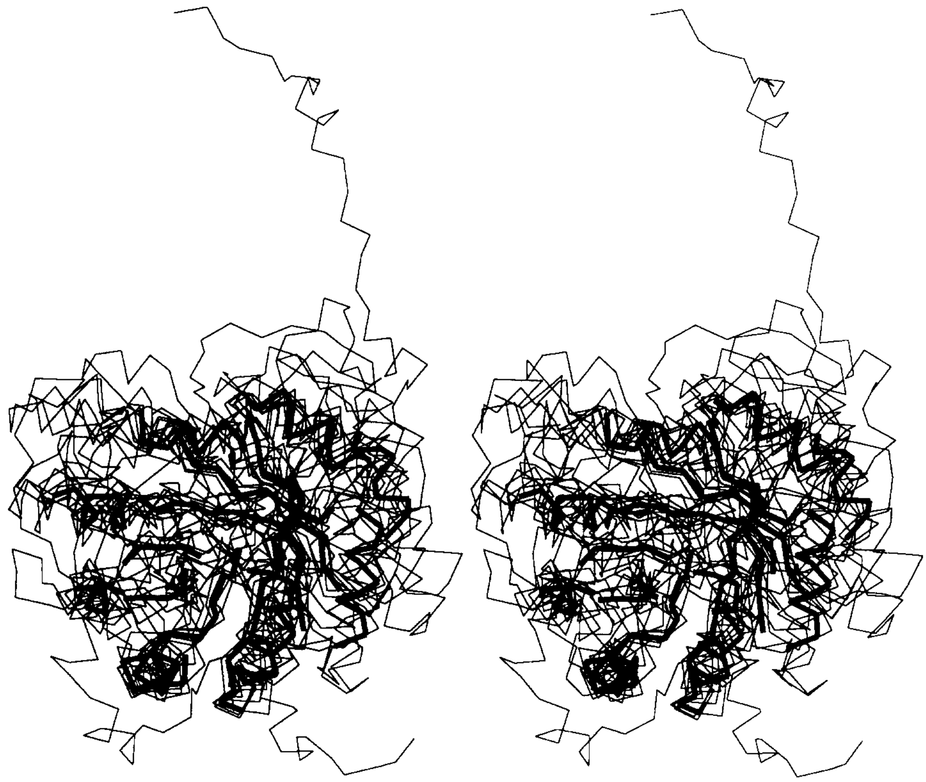

Fig. 11. Multiple structure superposition of TIM barrel folds. The domains were superposed as rigid bodies using the residue equivalences in Figure 10. (See the legend to Fig. 3 and the section on multiple superpositions in the Methods.)

Methods

Outline of structure comparison

The method of Taylor and Orengo (1989b) for the comparison of 2 structures (implemented as the computer program SSAP) was based on the definition of a local structural environment for each position. Each position was characterized by a set of interatomic vectors to all other residues in the structure and by comparing these vector sets, the similarity of 2 positions was quantified. Given a measure for the similarity of pairs of positions, a sequence alignment algorithm can be applied to enforce the expected co-linearity of the resulting equivalence of residues.

Table 4. *Databank hits using TIM-barrel consensus*[a]

PDB code	Title	SSAP	Equivalent
1pl1 452	Anthranilate isomerase	86.48 (76.9)	157
5timA 249	Triosephosphate isomerase	85.74 (100)	157
1wsyA 246	Tryptophan synthase	84.58 (68.8)	157
1ald 363	Aldolase A	84.25 (77.8)	157
5rubA 434	Rubisco	77.36 (68.5)	155
4enl 436	Enolase	75.75 (65.9)	141
2taaA 478	Taka-amylase	74.35 (62.9)	128
1ximA 392	Xylose isomerase	73.78 (70.8)	122
1dri 271	D-Ribose binding protein	69.76 (62.1)	139
1cseE 274	Subtilisin Carlsberg	69.23 (61.5)	122
2cmd 312	Malate dehydrogenase	68.78 (58.7)	133
2liv 344	Leucine binding protein	68.12 (60.6)	148
3grs 461	Glutathione reductase	66.53 (59.6)	133
1ldb 291	Lactate dehydrogenase	66.51 (59.9)	120
5p21 166	Ras p21 protein	65.85 (68.3)	122

[a] Hits with an SSAP score over 70 are all correct (no false positives or negatives). They are followed (after a gap in scores) by alternating α/β-type proteins. This clear discrimination is not achieved using a representative structure (SSAP scores in parentheses).

The method would be simple but for the complication that the equivalence of vectors between each set must be known before 2 sets can be compared and this equivalence is, itself, the final sequence alignment. The problem is apparently circular because the sequence alignment must be known in order to be calculated. This difficulty has led others to stochastic solutions (Sali & Blundell, 1990; Holm & Sander, 1993), however, a more direct solution can be achieved by calculating the difference between all pairs of vectors in each set and performing a sequence alignment at this low level for each pair of residues independently.

Each independent comparison of pairs of vector sets generates a sequence alignment and, although these will undoubtedly differ, a trend will emerge based on the correct equivalences that leads to the best alignment. This trend can be extracted simply by summing all the individual alignments in a "master" (or high-level) matrix and calculating a final consensus alignment by the reapplication of the dynamic programming algorithm. Because the method involves the application of dynamic programming at 2 levels, it has come to be known as the double dynamic programming algorithm.

The pairwise SSAP algorithm generates an overall normalized score in the range of 1–100 that is independent of the sizes of proteins being compared. Empirical trials have shown that values above 80 reflect highly similar folds, whereas values between 70 and 80 suggest related folds having more variation in the loops and orientations of the secondary structures (Orengo et al., 1992, 1993a, 1993b; Orengo & Taylor, 1993).

Outline of multiple sequence alignment

The generalization of the dynamic programming algorithm to multiple sequences, although straightforward, is generally considered impractical. This has led to the development of a number of methods based on the combination of pairwise alignments. These differ only in the order in which they combine the sequences and whether some form of abstraction or consensus is derived at each stage. A robust approach (used in the method of Taylor [1988]) is to combine the most similar sequence pairs independently into a consensus and then to recalculate the similarity between all single and consensus sequences to progressively bring together the most similar of these at each stage. This strategy allows the conserved aspects of each subfamily to become apparent before they are aligned at a later stage.

In the method of Taylor (1988) no abstraction was made and the consensus (or profile) was taken as the multiple alignment itself. A similarity measure between either a single sequence or a pair of alignments was defined for a pair of positions, by the sum of the similarity measure between all pairs of residues in each sequence in the 2 alignments. (A similarity measure for 2 residues can be obtained from an amino acid relatedness matrix such as that of Dayhoff et al. [1978].) This approach retains virtually all the information in each alignment, avoiding the loss that would occur if a more abstract representation were used.

Outline of multiple structure alignment

In sequence alignment, a similarity (or distance) measure is commonly derived from a precalculated matrix (Dayhoff et al., 1978). The equivalent measure in structure comparison is the difference in 2 3-dimensional interatomic vectors. Unlike the generic measure of sequence relatedness, each vector is defined by the structure of the protein in which it occurs. If a consensus sequence retains the identity of all amino acids at 1 position, then an equivalent consensus structure would be a bundle of vectors.

The calculation of the similarity between 2 positions in a pair of alignments depends on the product of the number of sequences in each alignment. Simply, finding the similarity of 2 amino acids through a look-up table is a fast process to calculate and this quadratic dependence is tolerable. By contrast, finding the difference in 2 vectors takes longer to calculate and a quadratic dependence is to be avoided, especially as this occurs at a low level in the comparison of the structural environments. For this reason the representation of a consensus interatomic vector was defined as the average of the component vectors and the comparison of 2 positions as the magnitude of the difference in these average vectors.

Taking an average vector as a consensus solves one problem but raises another. Because some information has been lost, it is impossible to distinguish a coherent bundle of vectors from a divergent bundle of (longer) vectors which, by chance, might have an identical average. An error measure is needed: although this could be defined in a number of ways in terms of both the direction and length of the component vectors, a simple measure that incorporates both these aspects is the magnitude of the vector difference. Again, to avoid a second-order dependence, rather than take the difference between all pairs, the difference of each from their average was calculated.

Implementation details

The multiple alignment program MULTAL (Taylor, 1988, 1990) was combined with the program SSAP, which implements the fast structure comparison algorithm of Orengo and Taylor (1990). The log-normalized score calculated by the later versions of SSAP was used in determining the condensation order in MULTAL because this formulation is almost length-independent and gives a good correlation with other measures of structural similarity (Orengo et al., 1992, 1993a). It also corresponds with more intuitive estimations of similarity among very remotely similar proteins (Orengo & Taylor, 1993; Orengo et al., 1993b).

Consensus vector definition

At each stage in the hierarchic condensation of sequences into a multiple alignment, MULTAL can join more than 2 sequences using the (earlier) algorithm of Taylor (1987b). This feature allows similar proteins to be combined quickly, so saving time through fewer iteration cycles. Although the consensus (average) vectors outlined above are derived at each stage, the method must allow more than 2 structures to be combined in 1 step.

Generally, where a consensus is to be constructed from N aligned structures, an average vector $\vec{r}_{ij}$ was defined between residues i and j using each interatomic vector $\vec{v}_{nij}$ (in protein n), as:

$$\vec{r}_{ij} = \frac{1}{N} \sum_{n=1}^{N} \vec{v}_{nij}. \tag{1}$$

The error associated with this average was then defined as:

$$e_{ij} = \frac{1}{N} \sum_{n=1}^{N} (\vec{r}_{ij} - \vec{v}_{nij})^2. \tag{2}$$

Treatment of gaps

Where a gap occurs in a sequence, there is no vector to be averaged, so the average was taken only over non-gap positions. (Thus, if there is only 1 non-gap residue at an aligned position, its vector will be the average.) However, gap positions should clearly diminish the weight of the vector because positions that can be deleted must be given lesser importance. This was achieved by adding a constant weight g (default 5.0) to the error term e_{ij} in Equation 2 for each gap position.

Basic score definition

The basic scoring method in SSAP is the difference of 2 interatomic vectors between β-carbons, with each defined in a local coordinate frame based on the α-carbon of their residue of origin. If the average vector $\vec{r}_{ij}$ from residue i to j in protein A is being compared to the vector from m to n in protein B, then their difference, δ_{ijmn}, is initially defined as follows:

$$\delta_{ijmn} = (\vec{r}_{ij} - \vec{r}_{mn})^2. \tag{3}$$

This was then converted to a similarity score using the constants a and b — the values of which have been investigated in previous works (Taylor & Orengo, 1989a, 1989b). At this point, the errors associated with each average vector can be introduced to down weight the contribution of variable positions. This was implemented as a product of errors to severely dampen the contribution from the comparison of 2 variable positions:

$$s_{ijmn} = \frac{a}{b + \delta_{ijmn} + w(e_{ij}e_{mn})} \tag{4}$$

where w is an overall weight (default 1.0), which can be specified by the user to adjust the effect of the error contribution.

Residue selection

In the basic SSAP pairwise algorithm, residue pairs are only compared if they have similar buried areas and torsional angles (parameter dtot with default value 150 used by Orengo and Taylor [1990]). Because these selection criteria increase the speed of calculation as well as improve accuracy by reducing noise in the score matrix, they were also applied in the current multiple mode. Additionally, when aligning a consensus or real structure against another consensus structure, positions were selected that have SSAP consensus scores above a cutoff (sel_cut default value 20).

Construction of a consensus template

A multiple sequence alignment derived from structure can easily be portrayed. However, the underlying structural consensus that gave rise to the alignment is less easy to visualize. At the end of the alignment process, there are no atomic coordinates – only a set of average interatomic vectors that are not necessarily mutually consistent (due to, say, relative domain movements).

The most direct method for solving the problem of inconsistent vector sets is to let each vector define a target position and adjust the atomic coordinates to simultaneously minimize differences from the targets, with each vector appropriately weighted by its degree of conservation. A good starting model for this minimization can be obtained by projecting the distance matrix derived from the vector lengths using the technique of distance geometry (Kuntz et al., 1989). This method cannot, however, incorporate weights on individual distances and is therefore not able to give prominence to conserved features. Consequently, a real-space refinement procedure was also used.

Unlike the simple distance constraints minimized in many methods, the the use of vectors provides additional directional information. When located at their origin (on residue i) in the starting model (of coordinates $\vec{A}_1 \ldots \vec{A}_N$), the set of interatomic vectors from residue i specifies target locations for all other atoms. A shift vector, $\vec{s}_{ij}$, can thus be defined for any atom j (with atomic coordinate $\vec{A}_j$) by the jth vector in the set associated with atom i ($\vec{r}_{ij}$):

$$\vec{s}_{ij} = \vec{A}_i - \vec{A}_j + \vec{r}_{ij}. \tag{5}$$

Atoms were then shifted along the s vectors toward their new positions using the simple algorithm described previously (Taylor, 1993). This regularization method is similar to the SHAKE algorithm (van Gunsteren & Berendsen, 1977) used in the program EXPLOR (deVlieg et al., 1988), but follows a Braun and Gō (1985) strategy by applying smaller shifts further from the local center of superposition (by a factor proportional to the square-root of the sequential separation). In the current application, however, the additional weight reflecting the conservation of the vector (e_{ij}) was incorporated, giving the new location ($\vec{A}'_j$) for atom j, as follows:

$$\vec{A}'_j = \vec{A}_j + \frac{\vec{s}_{ij}}{e_{ij}|j-i|^{1/2}}, \; (i \neq j). \tag{6}$$

Data

Test data were extracted from the Protein Data Bank (PDB) (Bernstein et al., 1977). Protein families were selected that exhibited a wide degree of structural similarity and, for comparative purposes, had also been examined extensively in previous pairwise comparisons.

Immunoglobulins

Eleven immunoglobulin domains were selected. All have a common fold consisting of 2 β-sheets held together by a disulfide bridge. However, there are differences in the number of strands in the 2 domains (Fig. 1). The β-sheets in the variable domains contain 5 and 4 strands, respectively, whereas the first sheet of the constant domain contains only 3 strands. The strand labeling (A, B, C, D, E, F, G, H) conventionally adopted is shown in Figure 2. The GH hairpin has a variable length loop and the H strand in the variable type domains contains a β-bulge. The E-F hairpin is similar in both variable and constant domains although a large sequence displacement is needed to align the E-strands (see Fig. 2). When the strands are correctly aligned, the conserved disulfide cysteines (in strands B and G, see Fig. 2) are equivalenced and also a conserved tryptophan (in strand C). Members of the immunoglobulin superfamily were split into their domains, giving 11 data sets as follows (with PDB code in brackets): antigen binding fragment NEW (Saul et al., 1978) [3fab], heavy (H) and light (L) chain variable (1) and constant (2) domains (designated – fabH1, fabH2, fabL1, fabL2, respectively); similarly, for the binding fragment KOL (Marquart et al., 1980) [2fb4] (fb4H1, fb4H2); constant fragment FC (Deisenhofer, 1981) [1fc1] (fc1A1, fc1A2); variable light chain domain [2rhe] (Furey et al., 1983) [2rhe]; histocompatibility factor CD8 (Leahy et al., 1992) [1cd8] (1cd800); histocompatibility factor HLA (Bjorkman et al., 1987) [3hla]; the β-2 microglobulin (hlaB0).

Alternating α/β proteins

Doubly wound domains. Ten doubly wound alternating α/β folds were selected: the *ras* oncogene protein p21 (Pai et al., 1990) [5p21]; ribosomal elongation factor Tu (La-Cour et al., 1985) [1etu]; flavodoxin (*Clostridium mp.*) (Smith et al., 1977) [4fxn]; flavodoxin (*Desulfovibrio vulgaris*) (Watt et al., 1991) [2fx2]; flavodoxin (*Chondrus crispus*) (Fukuyama et al., 1990) [2fcr]; glycinamide ribonucleotide transformylase (A chain) (Chen et al., 1992) [1grc]; lactate dehydrogenase (LDH) (Abad-Zapatero et al., 1987) [6ldh] (N-terminal domain); alcohol dehydrogenase (ADH) (Eklund et al., 1984) [8adh] (N-terminal domain); glyceraldehyde-3-phosphate dehydrogenase (GPD) (Skarzynski et al., 1987) [1gd1] (N-terminal domain); bacterial chemotaxis Y protein (che-Y) (Volz & Matsumura, 1991) [3chy].

No pairs had sequence identity greater than 35%. Seven were single domain structures, whereas the remaining 3 (6ldh01, 8adh01, 1gd1O1) were single domains extracted from multidomain proteins. The doubly wound or Rossmann fold has been described as having 6 parallel β-strands forming a single sheet of strand order bA, bB, bC, bD, bE, bF; with 4 helices in connecting loops between the strands (aB, aC, aE, aF). The chain first forms the A, B, C strands before doubling back on itself to form the D, E, F strands. The fold was initially identified in dinucleotide binding proteins (Rao & Rossmann, 1973) and subsequently, proteins having different functions but similar folds were also identified (Walker et al., 1982). Only the bacterial chemotaxis protein (3chy) of the proteins in this group does not bind a nucleotide.

TIM-barrel domains. This alternating α/β-fold consists of 8 parallel β-strands wound in a single direction to form a barrel with α-helices packed on the outside. Five TIM barrel structures were selected: flavocytochrome *b*2 (A chain) (Xia & Mathews, 1990) [1fcb]; triosephosphate isomerase (A chain) (Wierenga et al., 1987) [2tim]; aldolase (Gamblin et al., 1991) [1ald]; phosphoribosyl anthranilate isomerase (N-terminal domain) (Wilmanns et al., 1992) [1pii]; tryptophan synthase (A-chain) (Hyde et al., 1988) [1wsy].

Presentation of results

Dendrograms were constructed by applying single linkage cluster analysis to the SSAP score matrices, generated from pairwise structural alignments between the folds (Tables 1, 2, 3). Cartoons of the secondary structural elements are presented as schematic representations produced by the program TOPS (Flores et al., 1994). α-Helices are represented by circles and β-strands by triangles. TOPS diagrams of the folds are shown on the far right of the dendrograms, adjacent to the corresponding Brookhaven PDB codes.

Multiple superpositions

Structures are superposed in the same order as depicted in the dendrograms. Each structure was superposed to the structure or average structure found at the point where they join the tree. Rigid body superposition was carried out using the method of McLachlan (1979) with each superposition weighted using the average SSAP scores for the alignment (Rippmann & Taylor, 1991). The overall structure was calculated for the alignment positions that contain residues from each sequence. Kinemages are colored from blue to red according to the average SSAP score at each position.

References

Abad-Zapatero C, Griffith JP, Sussman JL, Rossmann MG. 1987. Refined crystal structure of dogfish M4 apo-lactate dehydrogenase. *J Mol Biol 198*:445–445.

Barton GJ, Sternberg MJE. 1987. A strategy for the rapid multiple alignment of protein sequences. *J Mol Biol 198*:327–337.

Barton GJ, Sternberg MJE. 1988. LOPAL and SCAMP: Techniques for the comparison and display of protein structure. *J Mol Graph 6*:190–196.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, 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.

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature 329*:506–506.

Braun W, Gō N. 1985. Calculation of protein conformations by proton-proton distance constraints—A new efficient algorithm. *J Mol Biol 186*:611–626.

Chen P, Schulze-Gahmen U, Stura E, Inglese J, Johnson D, Marolewski A, Benkovic SJ, Wilson IA. 1992. Crystal structure of glycinamide ribonucleotide transformylase from *Escherichia coli* at 3.0 angstrom resolution: A target enzyme for chemotherapy. *J Mol Biol 227*:283–292.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, ed. *Atlas of protein sequence and structure, vol 5, (Suppl 3).* Washington D.C.: National Biomedical Research Foundation. pp 345–352.

Deisenhofer J. 1981. Human FC fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9- and 2.8-angstroms resolution. *Biochemistry 20*:2361–2361.

deVlieg J, Scheek RM, van Gunsteren WF, Berendsen HJC, Kaptein R, Thomason J. 1988. Combined procedure of distance geometry and restrained molecular dynamics techniques for protein structure determination from nuclear magnetic resonance data: Application to the DNA binding domain of *lac* repressor from *Escherichia coli. Proteins Struct Funct Genet 3*:209–218.

Eklund H, Samama JP, Jones TA. 1984. Crystallographic investigations of nicotinamide adenine dinucleotide binding to horse liver alcohol dehydrogenase. *Biochemistry 23*:5982–5982.

Farber GK, Petsko G. 1990. The evolution of barrel enzymes. *Trends Biochem Sci 15*:228–234.

Flores TP, Moss DS, Thornton JM. 1994. An algorithm for automatically generating protein topology cartoons. *Protein Eng 7*:31–37.

Fukuyama K, Wakabayashi S, Matsubara H, Rogers LJ. 1990. Tertiary structure of oxidized flavodoxin from a eukaryotic red alga *Chondrus crispus* at 2.35-angstroms resolution: Localization of charged residues and implication for interaction with electron transfer partners. *J Biol Chem 265*:15804–15804.

Furey W Jr, Wang BC, Yoo CS, Sax M. 1983. Structure of a novel Bence-Jones protein (RHE) fragment at 1.6 angstroms resolution. *J Mol Biol 167*:661–661.

Gamblin SJ, Davies GJ, Grimes JM, Jackson RM, Littlechild JA, Watson HC. 1991. Activity and specificity of human aldolases. *J Mol Biol 219*: 573–573.

Holm L, Sander C. 1993. Protein-structure comparison by alignment of distance matrices. *J Mol Biol 233*:123–138.

Hyde CC, Ahmed SA, Padlan EA, Miles EW, Davies DR. 1988. Three-dimensional structure of the tryptophan synthase multienzyme complex from *Salmonella typhimurium*. *J Biol Chem 263*:17857–17857.

Johnson MS, Sali A, Blundell TL. 1990a. Phylogenetic-relationships from 3-dimensional protein structures. *Methods Enzymol 183*:670–690.

Johnson MS, Overington JP, Blundell TL. 1993. Alignment and searching for common protein folds using a data-bank of structural templates. *J Mol Biol 231*:735–752.

Johnson MS, Sutcliffe MJ, Blundell TL. 1990b. Molecular anatomy: Phyletic relationships derived from 3-dimensional structures of proteins. *J Mol Evol 30*:43–59.

Jones TA, Thirup S. 1986. Using known substructures in protein model building and crystallography. *EMBO J 5*:819–822.

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

Kuntz ID, Thomason JF, Oshiro CM. 1989. Distance geometry. *Methods Enzymol 177*:159–204.

La-Cour TFM, Nyborg J, Thirup S, Clark BFC. 1985. Structural details of the binding of guanosine diphosphate to elongation factor tu from *E. coli* as studied by X-ray crystallography. *EMBO J 4*:2385–2385.

Leahy DJ, Axel R, Hendrickson WA. 1992. Crystal structure of a soluble form of the human T-cell coreceptor CD8 at 2.6 Å resolution. *Cell 68*: 1145–1162.

Lesk AM, Branden CI, 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.

Marquart M, Deisenhofer J, Huber R, Palm W. 1980. The intact immunoglobulin molecule KOL and its antigen-binding fragment at 3.0 angstroms and 1.9 angstroms resolution. *J Mol Biol 141*:369–369.

McLachlan AD. 1979. Gene duplication in the structural evolution of chymotrypsin. *J Mol Biol 128*:49–79.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

Orengo CA, Brown NP, Taylor WR. 1992. Fast protein structure comparison for databank searching. *Proteins Struct Funct Genet 14*:139–167.
Orengo CA, Flores TP, Jones DT, Taylor WR, Thornton JM. 1993a. Recurring structural motifs in proteins with different functions. *Curr Biol 3*:131–139.
Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993b. Identification and classification of protein fold families. *Protein Eng 6*:485–500.
Orengo CA, Taylor WR. 1990. A rapid method for protein structure alignment. *J Theor Biol 147*:517–551.
Orengo CA, Taylor WR. 1993. A local alignment method for protein structure motifs. *J Mol Biol 233*:488–497.
Pai EF, Krengel U, Petsko GA, Goody RS, Kabsch W, Wittinghofer A. 1990. Refined crystal structure of the triphosphate conformation of h-ras p21 at 1.35 angstroms resolution: Implications for the mechanism of GTP hydrolysis. *EMBO J 9*:2351–2351.
Pickett SD, Saqi MAS, Sternberg MJE. 1992. Evaluation of the sequence template method for protein-structure prediction: Discrimination of the $(\beta/\alpha)_8$-barrel fold. *J Mol Biol 228*:170–187.
Rao ST, Rossmann MG. 1973. Supersecondary structure. *J Mol Biol 76*: 241–256.
Rippmann F, Taylor WR. 1991. Visualization of structural similarity in proteins. *J Mol Graph 9*:3–16.
Russell RB, Barton GJ. 1992. Multiple protein-sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct Funct Genet 14*:309–323.
Sali A, Blundell TL. 1990. The 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.
Saul FA, Amzel LM, Poljak RJ. 1978. Preliminary refinement and structural analysis of the FAB fragment from human immunoglobulin NEW at 2.0 angstroms resolution. *J Biol Chem 253*:585–585.
Skarzynski T, Moody PCE, Wonacott AJ. 1987. Structure of hologlyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 angstroms resolution. *J Mol Biol 193*:171–171.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195–197.
Smith WW, Burnett RM, Darling GD, Ludwig ML. 1977. Structure of the semiquinone form of flavodoxin from *Clostridium mp.*: Extension of 1.8 angstroms resolution and some comparisons with the oxidized state. *J Mol Biol 117*:195–195.
Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins: I. 3-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng 1*:377–384.
Taylor WR. 1987a. Multiple sequence alignment by a pairwise algorithm. *CABIOS 3*:81–87.
Taylor WR. 1987b. Protein structure prediction. In: Bishop MJ, Rawlings CJ, eds. *Nucleic acid and protein sequence analysis: A practical approach.* Oxford: IRL Press. pp 359–385.
Taylor WR. 1988. A flexible method to align large numbers of biological sequences. *J Mol Evol 28*:161–169.
Taylor WR. 1990. Hierarchical method to align large numbers of biological sequences. *Methods Enzymol 183*:456–474.
Taylor WR. 1993. Protein fold refinement: Building models from idealised folds using motif constraints and multiple sequence data. *Protein Eng 6*:593–604.
Taylor WR, Orengo CA. 1989a. A holistic approach to protein structure comparison. *Protein Eng 2*:505–519.
Taylor WR, Orengo CA. 1989b. Protein structure alignment. *J Mol Biol 208*:1–22.
van Gunsteren WF, Berendsen MJC. 1977. Algorithms for macromolecular dynamics and constraint dynamics. *Mol Phys 34*:1311–1327.
Volz K, Matsumura P. 1991. Crystal structure of *Escherichia coli* che-Y refined at 1.7-angstrom resolution. *J Biol Chem 266*:15511–15519.
Walker JE, Saraste M, Runswick WJ, Gay NJ. 1982. Distantly related sequences in the α-subunits and β-subunits of ATP synthase, myosin, kinases and other ATP requiring enzymes and a common nucleotide binding fold. *EMBO J 1*:945–951.
Watt W, Tulinsky A, Swenson RP, Watenpaugh KD. 1991. Comparison of the crystal-structures of a flavodoxin in its 3 oxidation-states at cryogenic temperatures. *J Mol Biol 218*:195–208.
Wierenga RK, Kalk KH, Hol WGJ. 1987. Structure determination of the glycosomal triosephosphate isomerase from *Trypanosoma brucei brucei* at 2.4 angstroms resolution. *J Mol Biol 198*:109–109.
Wilmanns M, Priestle JP, Niermann T, Jan JN. 1992. Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 angstroms resolution. *J Mol Biol 223*:477–477.
Xia Z, Mathews FS. 1990. Molecular structure of flavocytochrome at 2.4 angstroms resolution. *J Mol Biol 212*:837–837.