CORA—Topological fingerprints for protein structural families

CHRISTINE ANNE ORENGO

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology,
University College, Gower St., London WC1E 6BT, United Kingdom

(RECEIVED July 15, 1998; ACCEPTED December 2, 1998)

Abstract

CORA is a suite of programs for multiply aligning and analyzing protein structural families to identify the consensus positions and capture their most conserved structural characteristics (e.g., residue accessibility, torsional angles, and global geometry as described by inter-residue vectors/contacts). Knowledge of these structurally conserved positions, which are mostly in the core of the fold and of their properties, significantly improves the identification and classification of newly-determined relatives. Information is encoded in a consensus three-dimensional (3D) template and relatives found by a sensitive alignment method, which employs a new scoring scheme based on conserved residue contacts. By encapsulating these critical “core” features, templates perform more reliably in recognizing distant structural relatives than searches with representative structures.

Parameters for 3D-template generation and alignment were optimized for each structural class (mainly- α , mainly- β , α - β), using representative superfold families. For all families selected, the templates gave significant improvements in sensitivity and selectivity in recognizing distant structural relatives. Furthermore, since templates contain less than 70% of fold positions and compare fewer positions when aligning structures, scans are at least an order of magnitude faster than scans using selected structures. CORA was subsequently tested on eight other broad structural families from the CATH database.

Diagnostics plots are generated automatically and provide qualitative assistance for classifying newly determined relatives. They are demonstrated here by application to the large globin-like fold family. CORA templates for both homologous superfamilies and fold families will be stored in CATH and used to improve the classification and analysis of newly determined structures.

Keywords: consensus templates; protein structure classification; residue contacts

Classification of proteins into structural families provides important insights into the nature and mechanisms of evolutionary changes. Apart from scientific curiosity, these data improve our understanding of the relationship between protein sequence and structure and this can be exploited to aid structure prediction algorithms. However, despite the plethora of structure comparison algorithms available (Taylor & Orengo, 1989a; Sali & Blundell, 1990; Holm & Sander, 1993a; Russel & Barton, 1993; Orengo, 1994; Brown et al., 1995), the identification and description of structural families remain somewhat subjective with many different methods and criteria being applied.

In the CATH hierarchic database of protein domains (Orengo et al., 1997; <http://www.biochem.ucl.ac.uk/bsm/cath>), structures are grouped according to similarities in their (C)lass, (A)rchitecture, fold or (T)opology, and whether they are (H)omologous (i.e.,

appear to be evolutionary relatives). Criteria for identifying relatives are based on a pairwise structure similarity score (SSAP) (Orengo et al., 1992), residue overlap between structures and functional information. Several other databases, describing protein structural relationships, are also accessible over the World Wide Web: SCOP (Murzin et al., 1995), FSSP (Holm et al., 1993), 3Dee (Siddiqui & Barton, 1995), DIAL (Sowdhamini et al., 1996), ENTREZ (Hogue et al., 1996), and LPFC (Schmidt et al., 1997). Information within these databases is generated in a variety of ways ranging from manual through to completely automatic, and each approach uses slightly different criteria when clustering proteins into families.

However, as the Protein Data Bank (PDB) (Abola et al., 1987) grows rapidly in size, with more than 13,000 domain structures currently known and nearly 200 domains added each month, all classification methods will encounter similar difficulties in assembling distinct and coherent structural families. Problems are caused in each protein class, by the recurrence of favored motifs (e.g., β -hairpins, β -meanders, α -hairpins, split, and classic $\beta\alpha\beta$ motifs). This recurrence can cause overlap between families when proteins

Reprint requests to: Christine Orengo, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower St., London WC1E 6BT, United Kingdom; e-mail: orengo@biochem.ucl.ac.uk.

have motifs in common but not extensive global similarity. A recent analysis of the CATH database (Orengo et al., 1997) described this phenomenon as the “Russian Doll” effect and showed that it was very pronounced in some of the more common protein architectures (e.g., β -sandwiches, α - β two and three-layer sandwiches).

Although it is important to recognize relationships based on local similarities, data likely to be useful in structure prediction (e.g., substitution matrices, fold libraries) are better extracted from coherent families, containing proteins with significant global similarities. It is particularly important to group proteins whose structural and functional characteristics suggest an evolutionary relationship. Difficulties in consistent classification will become more acute with time and suggest that we will need a richer description of families, based on identifying consensus positions and on describing structural properties for these positions. Reliable, automatic methods are needed for capturing and describing these structural “signatures” and for using consensus data in recognizing newly determined relatives. Improvements in sensitivity will be achieved in an analogous manner to improvements in sequence comparison where conserved residue patterns and consensus sequence templates, identified from a multiple alignment of protein sequences, are more powerful for recognizing distant relatives than using single sequences (Taylor, 1988; Barton & Sternberg, 1990; Bowie et al., 1990; Johnson et al., 1993; Fischer & Eisenberg, 1996).

In a similar way, structural features conserved over a significant evolutionary time scale can be found by comparing distantly related proteins using flexible structure alignment methods. This is possible because analysis of structural families has shown that even at low sequence identities (<10%), a significant proportion of the structure is conserved, particularly in the core (Sander & Schneider, 1991; Flores et al., 1993; Hilbert et al., 1993; Orengo et al., 1993). Several groups have applied structural comparison methods to identify these common features, which can then be used in homology modeling of relatives (Sutcliffe et al., 1987; Sali & Blundell, 1993; Altman et al., 1995; Gerstein & Altman, 1995). Gerstein and Altman (1995) recently described a powerful new method for identifying the structural “core” for a protein family, from multiple superposition of relatives. Using this technique, they have constructed a library of core structures (LPFC) (Schmidt et al., 1997). These provide common frameworks for modeling homologous proteins.

A method is described here for automatically analyzing protein structural families. The CORA suite first selects a set of close structural neighbors, from a family, and performs a multiple structural alignment to identify structurally conserved positions and measure the properties of these positions (residue accessibility, torsional angles, and global geometry as described by inter-residue vectors/contacts). These consensus positions, which are usually associated with the buried core and functional sites, are then encoded in a tertiary template together with their average structural properties. Another program in the suite can be used to scan these templates against structures in the PDB and newly determined structures to identify distant relatives.

The multiple alignment strategy and generation of consensus three-dimensional (3D) templates are described in Materials and methods, together with descriptions of all the parameters used and the optimization of these parameters for each structural class (mainly- α , mainly- β , α - β). In Results and discussion, performance of the method is demonstrated for the three superfold families, used to optimize the parameters; the globin-like fold (mainly-

α), the TIM barrel fold (α - β), and the OB fold (mainly- β). Each family contains many diverse relatives (i.e., with low sequence identities (<25%) and in some cases very different functions). The method is subsequently tested on eight homologous structural families from the CATH database.

Diagnostic plots generated by the suite display average structural properties for conserved positions in a family, and provide a qualitative aid in assessing whether a new structure possesses similar properties and should be assigned to that family. The use of diagnostic plots are illustrated in Results and discussion, by reference to the globin-like fold family.

CORA templates showed increased sensitivity and selectivity in recognizing relatives, compared to scanning the PDB with representative structures. This is achieved partly because “core” templates reduce the noise associated with variable loop regions but also because the alignment method implements a new and highly sensitive scoring scheme that measures the number of conserved residue contacts for the fold, which are also present in the target structure. These consensus contacts are found to be very diagnostic for a particular fold. The CORA suite should therefore improve classification of protein families in CATH, both by enhancing recognition of distant relatives but also by guaranteeing more coherent protein families.

Results and discussion

Assessing the performance of CORA templates for three superfold families

Results are presented below for three superfold families, one from each structural class (mainly- α , mainly- β , α - β). A description of the methods used to generate the 3D template for these families and scan the templates against nonidentical structures from the PDB is given in Materials and methods. Details on the optimization of parameters used is also given in Materials and methods.

Template performance was evaluated by comparing score distributions from data bank scans, with distributions given by the structures selected to make the template. Template scans used the program CORALIGN, while scans with individual structures used the pairwise SSAP program (Taylor & Orengo, 1989a).

CORALIGN calculates two scores (see Materials and methods): a vector based score, which is similar to the scoring scheme developed for SSAP, and a new scoring scheme, which calculates the percentage of completely conserved residue contacts for the family that are present in a new structure.

Figure 1 shows that there are considerable improvements in performance, using the templates, for all three superfold families tested.

Mainly- α class: Globin-like fold family

The globin-like fold family contains 14 nonidentical representatives, six of which were selected for the template (Table 1). Only diphtheria toxin returns low scores against the template. This lacks the large N-terminal helix and a large number of other pairwise interactions in the fold are considerably distorted. CORA analysis for this fold family is described below.

Mainly- β class: OB fold family

The OB fold family consists of a small β -wedge (which can also be described as a distorted β -barrel) formed by two small sheets

Fig. 1. Distribution of scores obtained by scanning selected protein structures or a 3D consensus template across a data set of >1,000 nonidentical structures from the PDB (1,100 for globin templates, 1,500 for TIM and OB templates). From top to bottom, the three rows show plots for the globin-like, OB, and TIM barrel families, respectively. Across the rows, the initial plots show the distribution of SSAP scores for scanning the data sets with the selected structures from each family, performing pairwise structure comparisons with the program SSAP. The next two plots show score distributions (vector score and contact score, respectively) for scanning the CORA template across the data set, using the program CORALIGN. Comparisons with unrelated structures returning zero scores are not included in the plots. For each family tested, all relatives scored above zero.

Table 1. The upper part of the table lists PDB codes for nonidentical representatives for the globin-like, OB, and TIM fold families (CATH release 1.0, N-Level, $\geq 95\%$ identity)^a

Structural family	CATH	PDB codes and CORA contact scores for nonidentical representatives
<i>Superfold families</i>		
Globin-like	1.10.490 <i>mainly-α</i>	1hlm (80), 1hbg (100), [1fslA] (100), [2lhb] (100), 1flp (100), [1emy] (100), [1eca] (90), [1sctA] (100), 1mba (90), 1ash (80), 1lthA (100), [1cpcA] (100), 1colA (70), 1ddt02 (10)
OB-like	2.40.50 <i>mainly-β</i>	1bovA (86), 1mjc (74), 1nmf (49), 1krs (90), [1lylA1] (100), [1asyA1] (100), 1bgh (56), [1prtD] (100), [1prtF] (100), 1tssA2 (82), 1prtB2 (98), [1prtC2] (100), 1ltsD (98), [1chbD] (100), 2sns (67), 1esfA1 (68), 1sebD1 (81), 1se201 (77)
TIM barrel	3.20.20 α - β	1fvpA (18), 1br1B (84), 1tml (50), [1pii02] (98), 1pii01 (97), [1igs] (97), [1wsyA] (100), 1pkm02 (90), [1pkyA2] (100), 1na11 (96), 1fbaA (90), 1nar (87), 2hvm (87), 1xyzA (75), 2ebn (83), 1edt (80), 1ctn02 (84), 1cec (88), 1ghsA (90), 1cbg (83), 1amb01 (81), 1amy01 (88), 2aaa01 (88), 1ppi01 (83), 1cgt01 (87), 1btc (85), 1bglA3 (76), 1tpfA (85), 1ads (85), [1rlcL2] (98), 1ebhA2 (62), 1chrA (83), 2mnr (88), 1gox (97), 2tmdA1 (84), 1oya (90), 1add (79), 1pta (53), 1xis (62), 1dik04 (92), 1pbgA (84)
		PDB codes of representatives selected for templates
<i>Homologous superfamilies</i>		
Cytochrome C5	1.30.370.10 <i>mainly-α</i>	1dvh, 1fcdC1, 1fcdC2, 1cc5, 5cytR, 351c, 1cyi
Interleukin 4	1.20.160.10 <i>mainly-α</i>	1gmfA, 1rcb, 1rhgA, 1lki, 1huw
Lipocalins	2.40.130.10 <i>mainly-β</i>	1sriA, 1hbq, 1epaA, 1mup, 1bbpA, 1hmr
Serine proteases	2.40.80.20 <i>mainly-β</i>	1sgpE1, 1lmwB2, 1try02, 1elg02, 1arb02
Aspartyl proteases	2.170.10.10 <i>mainly-β</i>	2apr, 1lyaB, 1mpp
Periplasmic binding proteins	3.40.190.10 α - β	1dmb02, 1pda02, 1lst02, 1ovt02, 1sbp02, 1poy12
Thioredoxin reductases	3.40.370.20 α - β	1trb01, 3grs01, 1lv101, 1pbe01, 1kifA1
Lipases	3.40.680.10 α - β	1thtA, 1tca

^aMOLSCRIPT drawings for representative structures from each family are shown in Figure 2. Structures in brackets are those selected by CORASET to generate a 3D template for the family. The lower part of the table lists PDB codes for representatives selected to generate templates for eight homologous superfamilies. CATH identification numbers corresponds to (C)lass, (A)rchitecture, (T)opology, (H)omologous superfamily. Details of protein titles, source, and function can be obtained from the CATH web site (<http://www.biochem.ucl.ac.uk/bsm/cath>).

comprising between three and six β -strands (Murzin, 1993). In some structures the barrel is open and in others there is an α -helix connecting strands three and four (Fig. 2). There are 18 nonidentical proteins in the family (Table 1) possessing disparate functions and sequences. CORASET selected six structures for the template, all of which contain the α -helix. A large proportion of OB relatives (>83%) gave high contact scores and good overlap with the template. Three relatives not previously assigned in CATH were identified (the enterotoxins—1esf chain A domain 1, 1seb chain D domain 1, 1se2 domain 1).

Three OB structures gave low contact scores (1bgh, 1nmf, 1rip). They all lack the α -helix, so that strands 4 and 6 hydrogen-bond to form a closed β -barrel. In this sense the secondary structure interactions are different from the wedge-like interactions observed in many other OB structures (see Fig. 2). However, a template generated from another CORASET grouping in the family (1mjc,

1bovA, 1tssA2) gave high contact scores with the outlying structures (1bgh—score 75 and 86% overlap; 1nmf—score 97 and 90% overlap; 1rip—score 83 and 92% overlap) by encapsulating the pair contacts present in these more barrel-like structures.

Alpha-beta class: TIM barrel family

The majority of TIM barrel structures consist of a central eight-stranded β -barrel with α -helices in the loops connecting the β -strands (Fig. 2), although there can be considerable variation in the number and orientation of the α -helices. There are 35 nonidentical structures and more than 12 different functions observed. Five structures were selected by CORASET (Table 1).

More than 90% of nonidentical TIM structures gave high contact scores and good percentage overlap with the template. Furthermore, the template scan is more than 80 times faster than the time required for scanning the data bank with the selected struc-

Fig. 2. MOLSCRIPT representations of structures from the three superfold families used to optimize parameters for CORA and CORALIGN. From top to bottom, the three rows show structures from the globin-like, OB, and TIM barrel fold families, respectively. The first structure on the left is a representative structure. Other structures in the row are outliers, which gave low CORA contact scores against the 3D template for the family.

tures (see footnote to Table 2). Two new members of the family are detected (1pbg chain A and 1dik domain 4).

Two members of the family give poor contact scores—flavoprotein 390 (1fvpA, score 18) and endo-1,4 beta-D-glucanase (1tml, score 50)—though the vector score is reasonably good for the latter. Both structures are different from other members of the family in having only seven β -strands and an open barrel, with no α -helices on one face (Fig. 2). Flavoprotein 390 (1fvpA) is a distant evolutionary relative to bacterial luciferase, which contains eight parallel strands. A new template based on structures selected from the bacterial luciferase homologous superfamily (1nfp, 1brlB) gave good contact score and overlap against 1fvpA (score 94 with 100% overlap).

Although there are currently no close relatives (homologues) of 1tml in the TIM family, in CATH the high vector score for 1tml (75) and 81% overlap with template positions suggest that this structure should tentatively be assigned to the TIM barrel fold family.

In summary, all families tested show significant improvements when scanning with fold templates. It can also be seen that the

contact score is much more discriminating. This is because the vector score expresses average structural similarity and similar motifs tend to increase this score, regardless of their orientation in 3D space. However, conserved contacts occur mostly between buried secondary structures in the core. This property is more dependent on the topology and 3D arrangements of the secondary structures and therefore much more characteristic of a particular fold.

Resolution between related and unrelated structures is best for the globin-like template (Z-score: 13.0, 4.2 for the OB fold and 7.2 for the TIM barrel fold, Table 2). This reflects the degree to which conserved contacts for the globin-like fold are unusual in the mainly- α class. These contacts occupy less than 1% of the consensus contact map (44 contacts). While in the OB (100 contacts) and TIM barrel folds (200 contacts), a much higher proportion of contacts are conserved, many of which reflect common patterns for the class, i.e., contacts between adjacent antiparallel β -strands in the mainly- β class or between α -helices and adjacent parallel β -strands in the α - β class.

Table 2. Parameter optimization for CORA and CORALIGN in the globin-like, OB, and TIM barrel families^a

Family	Parameter	Parameter value	Mean score random	Mean score related	Mean Z score related	F+	R−
Globin fold	select 70, C_beta 8, res_sep 15, coralign_sel 0.15						
	Residue selection (cora_sel)	0.15	2	92	13.4	0	2
		0.20	2	92	13.4	0	2
		0.25	2	89	17.9	0	5
	cora_sel 0.20, C_beta 8, res_sep 15, coralign_sel 0.15						
	Template selection (select)	65	2	92	14.8	0	2
		75	2	93	13.0	0	1
		80	2	91	13.0	0	2
	cora_sel 0.20, select 75, res_sep 15, coralign_sel 0.15						
	Cbeta cutoff (C_beta)	10	4	91	12.0	0	2
		7	1	87	11.5	0	4
OB fold	cora_sel 0.20, C_beta 8, res_sep 15, coralign_sel 0.15						
	Template selection (select)	65	19	79	5.1	1	4
		80	19	79	5.1	0	6
	cora_sel 0.20, select 65, res_sep 15, coralign_sel 0.15						
	Cbeta cutoff (C_beta)	7	17	79	4.4	3	3
		10	19	79	5.4	0	4
	cora_sel 0.20, select 65, C_beta 10, coralign_sel 0.15						
	Residue separation (res_sep)	5	28	84	3.9	4	3
		10	26	83	4.2	1	3
		20	14	75	5.6	0	6
	cora_sel 0.20, select 65, C_beta 10, res_sep 10						
	Residue selection (coralign_sel)	0.1	27	83	4.4	1	3
		0.2	26	84	4.0	2	3
TIM barrel fold	select 70, C_beta 8, res_sep 15, coralign_sel 0.15						
	Residue selection (cora_sel)	0.15	8	81	7.2	0	3
		0.20	8	83	7.2	0	2
		0.25	8	83	7.2	0	2
	cora_sel 0.20, C_beta 8, res_sep 15, coralign_sel 0.15						
	Template selection (select)	65	8	82	7.1	0	3
		75	9	82	7.1	0	4
	cora_sel 0.20, select 70, res_sep 15, coralign_sel 0.15						
	Cbeta cutoff (C_beta)	7	7	77	7.2	0	3
		10	10	84	7.0	0	3
	cora_sel 0.20, select 70, C_beta 8, coralign_sel 0.15						
	Residue separation (res_sep)	5	8	80	7.4	0	2
		10	8	80	7.4	0	2
		20	8	83	7.1	0	2
	cora_sel 0.20, select 70, C_beta 8, res_sep 15						
	Residue selection (coralign_sel)	0.1	9	83	7.2	0	5
		0.2	7	83	7.2	0	2

^aFor each parameter tested, the average CORA contact scores for both related and unrelated structures are shown. Also, the average Z-score for relatives and the number of relatives missed (*R*−) and false positives (*F*+). *R*− and *F*+ were determined using the following cutoffs on the CORA scores: CORA contact score ≥70 globin-like, OB; ≥60 TIM barrel. Lower contact score cutoffs were applied (50—globin-like, TIM; 60—OB) when the CORA vector score was ≥75. CPU hours for scanning the data bank are as follows: (i) globin-like—1.78 (CORA), 46.74 (SSAP); (ii) OB-fold—1.85 (CORA), 32.1 (SSAP); (iii) TIM-barrel—1.69 (CORA), 140 (SSAP). Times are given for scans on an Indigo2.

Applying CORA templates for eight homologous superfamilies

Results for both the OB and TIM barrel families suggest that in some of the broader fold families, it will be necessary to generate more than one CORA template. For the CATH database, we would automatically generate templates for all the homologous superfamilies (H-level) within a given fold family (T-level). Proteins in homologous superfamilies exhibit a higher degree of structural similarity (pairwise SSAP score usually ≥ 80) than analogues, combined with either sequence similarity or functional similarity.

Template performance was tested on eight homologous superfamilies from CATH (see Table 1), using the optimal parameters for CORA and CORALIGN, for each class (see Table 2). The families cover a broad spectrum of different protein architectures and are among the largest in CATH, containing seven or more nonidentical structures. As with the superfolds, it can be seen that for all families tested, the templates considerably improve both sensitivity and selectivity (Fig. 3). The results are particularly impressive for the more unusual architectures of the aspartyl protease and the cytochrome structures. For the serine protease, interleukin-4, and lipocalin families, there is also clear separation of scores for random and related structures. Furthermore, template scans were between 20 and 60 times faster than scanning with the selected structures, depending on the family.

Interestingly, scans with the lipocalin template gave higher contact scores for unrelated proteins (~ 50) than scans for the other mainly- β , barrel-like, OB fold (~ 30). This reflects the fact that consensus contacts for the lipocalin family contain a higher proportion of common contacts for the mainly- β class, i.e., between adjacent β -strands, which will be present in all β -sheet proteins. Similarly, in the α - β families tested, the peak in the range 50–70, for unrelated proteins, corresponds to structures containing a significant proportion of $\beta\alpha$ motifs, often more than 40% of the residues. Proteins returning contact scores above 70 were found to have similar folds.

Using CORA diagnostic plots to assist classification: Application to the globin-like fold family

CORA generates diagnostic plots, which can be used qualitatively to assess whether a new structure should be assigned to a particular structural family. Although any number of different structural properties could be compared once equivalent positions have been identified across a family, plots are currently based on properties shown to be powerful discriminators in protein structure comparison and recognition algorithms (residue accessibility, torsional angles, vector environments, secondary structure interactions, and inter-residue contacts).

There are currently 14 nonidentical globin-like structures in CATH (Table 1). A large proportion function as oxygen binding/transport proteins and possess eight α -helices arranged in two orthogonal layers and labeled A to H by Lesk and Chothia (1980). More recently, other globin-like structures have been determined, having different functions and no significant sequence identity with the common globins. Phycocyanins (1pcpA, 1pcpB) are found in light harvesting bacteria; colicin-A (1colA) is an antibiotic protein and diphtheria toxin is a membrane associated protein with a biological role as a toxin.

A CORA template was generated from five oxygen-binding globins and CORALIGN used to align other O₂-binding globins and

functionally unrelated globins to the template (Fig. 4; Table 3A). It can be seen that structural environments are particularly well conserved for the B,E,G,H helices, even for phycocyanin and colicin-A. Helix pairs A/H, B/E, B/G, and G/H have the highest numbers of residue contacts (Table 3B) suggesting they are the major determinants of fold architecture.

Helix-pair orientations varied by less than 40°, for all pairs. Phycocyanin and colicin-A possess all the major helix pairs and their midpoint separations and interhelix angles are within the ranges for the family, apart for G/H where for phycocyanin and colicin-A the helices pack in another preferred mode, though similar contact patterns are observed (Table 3B; Fig. 5).

In the common globins, more than one-third of contacts in three of the major helix pairs, are completely conserved; 68% for the B/G pair (see Fig. 5; Table 3B). Apart from the A/H pair, a large proportion of these conserved contacts are present in phycocyanin and colicin. This is reflected in the high CORA contact scores: 1pcpA 90, 1pcpB 80, 1colA 70 against the O₂-binding globin template. Vector environment scores are also high: 1pcpA 79, 1pcpB 80, 1colA 82. Taken overall the CORA diagnostics for phycocyanin and colicin-A suggest that it is reasonable to assign these proteins to the globin-like fold family.

By contrast, the diagnostics for diphtheria toxin are less favorable. Comparison with the globin-like template gave only low vector (69) and contact scores (10) (see Table 3A). Diphtheria toxin lacks the large N-terminal A-helix, conserved in all other members of the family and which is responsible for one of the four major helix-pair interactions (A/H). Although the B,E,G helices align weakly in a common three-helix motif with globin-like structures, the residue vector environment scores (coravect(i)) are low. Furthermore, pairwise interactions for these helices are mostly outside the range of angles displayed in other members of the family and few of the conserved contacts for the family are present (Fig. 5; Table 3B).

Therefore, CORA diagnostics showing differences in the major helix pair interactions and low scores against the family template suggest that despite motif similarities with some globin-like structures, diphtheria toxin should be assigned to a separate fold family.

Interestingly, results returned by the pairwise structure comparison method DALI (Holm & Sander, 1993a) show a similar trend. This method also relies on identifying conserved contacts between related structures and calculates a Z-score reflecting the probability that two structures are related. Scores for scans of diphtheria toxin against the globins are in the range of Z-score 2–3, equivalent to the Z-scores returned by small repressor structures and some completely unrelated structures (e.g., methionine synthase), while Z-scores returned by scanning with other globins are generally greater than 15.

Diagnostic plots will be generated for each fold (T-level) and homologous superfamily (H-level) in CATH, both to aid in assigning new structures to a given family and also to illustrate the degree of structural similarity among existing relatives.

Discussion

CORA is a completely automatic method for multiply analyzing proteins in a structural family to identify conserved positions across the family and their structural characteristics. These are described in a 3D template, which consists of average residue attributes (accessibility, torsional angles) and also encodes average residue environments (measured by vectors/contacts) and secondary struc-

Fig. 3. Results of the data bank scans for the eight homologous superfamilies tested. Histograms on the left show the distributions of SSAP scores obtained by data bank scans using the selected structures from the family and the program SSAP. While those on the right show distributions of CORA contact scores from databank scans, using the 3D template for the family and the program CORALIGN. The data sets of nonidentical structures were the same as those used in the OB, TIM barrel, and globin-like template scans. (*Figure continues on the following pages.*)

Fig. 3. Continues.

Fig. 3. Continued.

ture interactions. Consensus contacts plots identify highly conserved contacts, allowing the major fold determinants to be easily recognized. Since the templates encapsulate the critical “core” of the fold, which in the case of homologous proteins has been conserved throughout evolution, they are very good at identifying distant relatives.

Parameters for CORA were optimized using three diverse structural families, one from each protein class (mainly- α , mainly- β , α - β). The families chosen are among the largest known and contain many extremely distant structural relatives. Parameters selected for optimization were those known to affect the performance of methods aligning 3D environments by double dynamic programming (Taylor & Orengo, 1989b). Furthermore, the robustness of the CORA suite, using optimal parameters, was demonstrated by testing the templates for eight other large structural families from the CATH database.

For all three superfold families and the eight homologous superfamilies tested, the CORA templates identify a higher proportion of relatives during data bank scans than could be guaranteed

using individual representatives from the family. Such a consensus approach will become more important as the PDB grows in size, and it becomes harder to choose representative structures for each family. The increased sensitivity and selectivity was achieved by implementing a more reliable scoring scheme, based on comparing residue contacts in a target structure with the most highly conserved contacts for the family.

Improved sensitivity is accompanied by significant increases in speed, as only conserved fold positions are compared against the target structure. Increases in speed varied from ~20 to 80-fold depending on the structural family. However, given the rapid increase in the PDB, classification by 3D templates will represent increasingly larger savings in CPU time for scanning against newly determined structures.

It is unlikely that any single score statistic (e.g., cutoff on SSAP, root-mean-square deviation (RMSD), or Z-score) can be uniformly applied to classify structural relatives. Evolutionary structural changes vary for different protein architectures and topologies, reflecting the different nature of structural constraints stabilizing

Fig. 4. CORA multiple alignment plot for structures from the globin-like fold family. Residues are color-boxed according to the secondary structure assignment (white, coil; red, α -helix; blue, β -strand; orange, 3_{10} helix; green, turn). Hydrophobic residues are printed in white and color boxes for buried residues have a black outline. The shaded bar displays the local conservation score (coravect(i)), which ranges from white (coravect(i) = 0) to black (for a highly conserved structural environment, coravect(i) = 100). Dots reflect the relative number of contacts in which the residue participates. The structures have been divided into three groups: the common globins, phycocyanin, and colicin-A.

the fold. In some families functional constraints also limit evolutionary change. In the common globins, constraints on binding the haem cofactor result in a high degree of structural conservation, even at very low sequence identities (<10%).

Furthermore, Z-scores exhibited by relatives of a given topology, a value commonly used to reflect significance, will also depend on the degree of recurrence of common motifs for the protein class. The TIM barrel fold exhibits high recurrence, with eight repeating $\beta\alpha$ motifs, whilst the globin-like fold can be thought of as three α -hairpin repeats. The $\beta\alpha$ motif is very common in the α - β class and this is reflected in a lower average Z-score from coralign for TIM relatives (~ 7), which is about half that observed for relatives in the globin-like family (~ 13).

The problem of motif recurrence in protein structure classification is further compounded by using simple strategies for grouping structures. For example, methods used in constructing CATH and related databases based on automatic classification (e.g., DALI (Holm & Sander, 1993a); 3Dee (Siddiqui & Barton, 1995)) use single measures of structural similarity (e.g., SSAP score, RMSD) and simple clustering algorithms. CORA templates provide a solution by allowing a broader evaluation of structural similarity to a common core and consensus secondary structures. It follows that since all relatives will share the same common, well-defined features by which they have been identified, families will be more coherent.

The majority of families in CATH (~ 490) are small, coherent groupings, containing only homologous proteins (Orengo et al., 1997). In these families, there is a high level of structural similarity between relatives and it is expected that the templates would not significantly affect the definition of these families but guarantee their continued coherence as increasing numbers of relatives are determined. However, for the superfolds and broader families (~ 20) that constitute more than 30% of nonhomologous structures in CATH (Orengo et al., 1994, 1997), derivation of CORA templates and family analysis is likely to be an iterative procedure. Core features and templates would first be identified for homologues and used to recognize distant relatives. Subsequently, more general templates would be generated for the fold, which describe properties common to both homologues and analogues. Any structures not sufficiently similar to these consensus templates would be removed and compared with templates for other families having similar architectures. In this way, the current CATH classification will be refined to ensure more consistent families.

Data

All coordinates used (see Table 1) were obtained from the Brookhaven Protein Structure Data Bank (PDB) (Abola et al., 1987) and had been determined to 3.0 Å resolution or better. Details of the

Table 3A. Summary of CORA scores and %overlaps for the globin-like fold family^a

	ABCDE	F	GH	Length	CORA SecStr.	Insert SecStr.	%CORA	%Target	Vector score	CORA contacts
1fslA :	HHh	H	HH	143	6	1	95	98	89	100
2lhb :	HH	hH	H	149	6	1	93	92	91	100
1emy :	HH	hH	H	153	6	1	97	94	93	100
1eca :	HHhhH	H	HH	136	6	2	92	100	90	90
1sctA :	hHH	H	hHH	149	6	2	95	93	89	100
1hlm :	HH	H	H	158	6	0	95	89	82	80
1hbg :	HHh	H	HH	147	6	1	95	95	90	100
1flp :	HHh	H	HH	142	6	1	92	95	88	100
1mba :	HH	hH	H	146	6	1	95	95	88	90
1ash :	HH	hH	H	147	6	1	96	96	88	80
1ithA :	HH	H	HHh	141	6	1	94	98	89	100
1cpcA :	h	HH	hHH	162	6	2	89	81	84	100
1cpcB :	hhHH	H	HHh	172	6	3	78	67	79	80
1colA :	hhhHH	H	HHh	197	6	4	80	60	82	70
1ddt02 :	h	h	HH	180	5	8	97	79	67	10

^aConsensus secondary structures for the family are shown in bold. The number of additional secondary structures is listed for each relative, together with CORA scores (vector and contact) and percentage overlap between the structure and the 3D-template.

proteins can be viewed by accessing the CATH web page for the family (<http://www.biochem.ucl.ac.uk/cath>). In addition to PDB files, the programs CORA and CORALIGN require a modified DSSP (WOLF) file (Kabsch & Sander, 1983) containing secondary structure information and residue accessibility, and torsional angles (see Taylor & Orengo, 1989a). These also contain frames of reference for each residue, centered on the C_β atom, which are needed to obtain the vector sets, which define the residue's structural environment or view (see Materials and methods).

Materials and methods

Generation of 3D fold templates describing CORA

CORA consists of a suite of programs for multiply comparing selected proteins from a given structural family using this multiple alignment to analyze different structural characteristics for the family and generating a 3D template from conserved positions. It includes an alignment program for scanning this 3D template against

Table 3B. Consensus pairwise secondary structure torsion angles and contacts—Data for the major helix pairs in the globin-like fold^a

	A/H			B/E			B/G			G/H		
Consensus	13	−84	8	21	104	8	19	127	8	16	157	12
1fslA0	12	−86	9	16	89	9	22	144	9	14	141	9
2lhb00	12	−76	7	22	106	6	13	124	10	22	162	9
1emy00	18	−91	7	28	106	7	20	120	8	21	158	10
1eca00	17	−94	8	16	96	7	21	124	7	11	162	11
1sctA0	18	−93	8	26	106	7	21	118	7	22	150	9
1cpcA0	6	−64	8	19	122	8	20	130	8	11	167	26
1hlm00	15	−84	9	18	109	7	13	125	9	12	146	11
1hbg00	16	−87	8	37	124	6	24	125	8	16	143	9
1flp00	21	−102	8	17	99	8	13	130	9	26	157	8
1mba00	24	−99	7	21	100	8	11	125	9	23	162	9
1ash00	16	−106	8	25	97	6	9	125	10	14	157	10
1ithA0	15	−103	9	24	107	8	17	116	8	30	166	8
1cpcB0	2	−88	9	9	128	9	19	131	9	25	163	5
1colA0	9	−73	9	14	131	7	8	113	10	16	−177	11
1ddt02				9	113	9	8	150	10	19	138	8

^aTotal C_β contacts and pairwise torsional angles are shown for each relative in the family. Average values derived for the 3D-template are shown in the first row.

Fig. 5. Consensus contact maps for two major helix-pair interactions in the globin-like fold (B/E, B/G). The plots compare contacts observed in phycocyanin, colicin-A, and diphtheria toxin with those conserved in the common oxygen binding globins. The axes are labeled with positions from the multiple alignment shown in the CORA alignment plot. Black rectangular boxes along the axes indicate consensus secondary structure regions. Contacts, which are conserved in more than 70% of the common globins, are shaded grey. Colored cells highlight the following: red, the contact is present in 70% or more of the globins and also in phycocyanin, colicin-A, or diphtheria toxin; green, the contact is only present in phycocyanin, colicin-A, or diphtheria toxin.

the data bank of known structures to identify distant relatives for the family.

The three programs comprising the CORA suite are: *CORASET*, which selects a set of sequence-diverse but structurally close representatives for a structural family

CORA, which generates a multiple structural alignment of the selected representatives using information on residue structural environments and residue contacts. This program also produces diagnostic plots illustrating conserved structural properties and generates a 3D template from the most conserved positions.

CORALIGN, which aligns the consensus 3D template against other protein structures from the Brookhaven PDB.

Stage I: Select representatives from a structural family for multiple alignment

Program: *CORASET*. The CATH database of protein domains provides information on all the well-resolved protein structures belonging to a particular fold family. Matrices giving pairwise sequence identities and structural similarity between relatives are generated by a pairwise structural comparison method (SSAP) (Taylor & Orengo, 1989a; Orengo et al., 1992). SSAP returns a score in the range of 0 to 100 (for identical proteins). Homologous proteins generally give scores above 80, while proteins having no clear evolutionary relationship (analogues), but similar folds, usually score between 70 and 80 (Orengo et al., 1992, 1993).

Using the SSAP and sequence identity matrices as input data, the program *CORASET* clusters proteins in a given family into groups having a high SSAP score (>75) but less than 25% sequence identity. Multiple-linkage clustering is used, whereby each protein must have a SSAP score >75 to at least three-quarters of the proteins in a group before being joined to that group. The largest group found is selected to generate a consensus template for the family.

The cutoff of 25% on sequence identity ensures that the most diverse members of the family are included to identify structural features that have been highly conserved over evolution. While the cutoff on SSAP score of 75 is necessary to ensure that the structures are sufficiently structurally similar to guarantee a reliable multiple alignment.

This strategy was chosen to ensure that templates are initially based on clear structural relatives. It is possible that some CATH families contain one or more misclassified structures. Multiple alignments based on all members of the family would therefore be unreliable. By identifying an initial set of “true” relatives, and using the template derived from these to align other probable relatives, it will be possible to validate all the structures within the family. Subsequently, templates could be derived from multiple alignments of all “true” relatives using more sophisticated strategies (e.g., sequence weighting).

Stage II: Multiple structure alignment, fold family analysis, and generation of consensus 3D template

Program: *CORA*.

(1) Multiple structure alignment

The program *CORA* generates a multiple structural alignment using a modified version of SSAPm (Taylor et al., 1994). This is based on a well-established method for structure alignment, which compares residue structural environments between proteins and

employs a double dynamic programming algorithm (SSAP) (Taylor & Orengo, 1989a; Orengo & Taylor, 1990). Residue structural environments or “views” are described as the sets of vectors from C_β positions in a structure to other C_β positions in the same structure. Dynamic programming is applied at two levels: to align the vector sets of residues from each protein and in a final pass to obtain the alignment of residues in the two proteins. Residues compared between the two proteins have similar torsional angles and accessible areas. Detailed descriptions of double dynamic programming and a review of other modified versions of SSAP, developed for data bank scanning, local and multiple alignment, are given in Orengo and Taylor (1996).

CORA uses a simple multiple alignment strategy. Since only highly similar structures are selected (SSAP > 75), a reliable multiple alignment can be obtained by aligning successive members in the order given by scores in the pairwise SSAP similarity matrix.

Consensus structural information is calculated after the addition of each structure to the growing multiple alignment. This is stored in a 3D template (see below) and consists of a description of the average vector environment at each residue position, encapsulated as average vectors to all other positions in the consensus. Information on local structure is also included (average residue torsional angles and average accessibility). During the multiple alignment, all residue positions are included in the consensus. However, the final template generated for data bank scanning only contains structurally conserved positions (see below).

Conservation of the residue environment is taken into account when aligning a new structure to the consensus in that only conserved vectors (standard deviation < 20 Å) are compared to vectors in the new protein. This optimizes the equivalencing of residues in the conserved core regions of the fold.

When aligning structures against the consensus, residues are only compared to consensus positions if they have similar local structure (residue accessibility and phi, psi torsional angles). These attributes were found to be most powerful in discriminating locally similar residues in tests on pairwise structure comparison (see Taylor & Orengo, 1989b). The parameter δ_{res} measures the difference in these local structural properties for residue i in protein A compared to position j in the consensus:

$$\delta_{res} = \text{abs}(\text{access}_i - \text{avgaccess}_j) + \text{abs}(\text{psi}_i - \text{avgpsi}_j) \\ + \text{abs}(\text{phi}_i - \text{avgphi}_j) + (\text{access}_i + \text{avgaccess}_j)$$

where access_i is the accessibility of residue i in protein A and phi_i , psi_i are the torsional angles, while avgaccess_j and avgphi_j , avgpsi_j are the average values for accessibility and torsional angles for position j in the consensus. The final component of δ_{res} biases the selection of residues being compared to positions in the core. During the execution of *CORA*, the parameter δ_{res} is automatically incremented in units of 50, until the percentage of residue pairs selected (cora_sel) reaches the cutoff.

$$\text{cora_sel} = \frac{\text{number of residue pairs selected for comparison}}{\text{total number of residue pairs between proteins A,B}}$$

In addition to having similar accessibility and torsional angles, residues are only compared if they are involved in one or more long-range contacts (C_β - C_β separation ≥ 10 residues). This in-

creates the speed of the method more than fivefold, depending on the fold family and the number of structures being multiply aligned.

(2) Analysis of residue features and generation of 3D template

At each position in the multiple alignment, CORA calculates average values and standard deviations for the following residue properties: (1) accessibility (Lee & Richards, 1971); (2) torsional angles (ϕ , ψ) and secondary structure assignment (DSSP) (Kabsch & Sander, 1983); (3) conservation of the vector environment.

At each position (1), the vector environments are compared across the set of selected structures and a vector conservation score ($\text{coravect}(i)$) is calculated. By averaging over the total number of protein pair comparisons, this automatically penalizes for gaps. $\text{Coravect}(i)$ is normalized to be in the range of 0 to 100, independent of the sizes of the proteins, according to the method described for SSAP (Orengo et al., 1992).

$$\text{coravect}(i) = \frac{1}{N(N-1)/2} \cdot \frac{1}{K} \cdot \sum_{m=1}^N \sum_{n=m+1}^N \sum_{j=1}^K \left(\frac{500}{10 + \text{abs}(\vec{m}_{i \rightarrow j} - \vec{n}_{i \rightarrow j})} \right) \quad (1)$$

$$\text{coravect}(i) = \frac{\ln(\text{coravect}(i)) * 100}{\ln(50)} \quad (2)$$

where N is the number of proteins and K the length of the multiple alignment (e.g., 192 residues for the structures used to generate the 3D template for the globin-like fold family); $\vec{m}_{i \rightarrow j}$ is the vector from position i to position j in protein m .

Positions are selected for the 3D template having a conservation score ($\text{coravect}(i)$) above a defined cutoff, optimized for each structural class (see below). Average structural environments for each selected position are encoded in the template consisting of average vectors to other template positions and standard deviation of these vectors. Average torsional angles and accessibility are also included for each position. Positions involved in conserved $C\beta$ contacts are encoded and secondary structure interactions (average midpoint separations and torsional angles).

(3) Generate consensus contact plots and other diagnostic plots

In distance or contact plots, a 2D matrix, labeled horizontally and vertically with the protein's sequence, is shaded according to whether pairs of residues are in contact in the structure (Phillips, 1970). To compare contact plots between proteins within the same fold family, which may have significant insertions and/or deletions, CORA generates conformant contact plots using the common positions identified from the multiple structural alignment. Contacts are identified if the $C\beta$ atoms from each residue are within a preset distance (default value 8 Å, see below for optimization of this value for CORALIGN). Consensus contact plots can be generated by automatically "overlaying" the conformant plots for each structure in the multiple alignment and highly conserved residue contacts easily detected for a given fold family.

Besides contact plots, CORA generates a number of other plots illustrating structural properties across a family (see Tables 3A,B; Figs. 4, 5). These can be used diagnostically for qualitatively assessing whether a new structure may belong to the family and are discussed above for the globin family.

Stage III: Using the 3D fold template to identify distant relatives

Program: CORALIGN. CORALIGN also uses the same double dynamic programming algorithm (Taylor & Orengo, 1989a) to compare consensus structural environments (i.e., average vectors) for the template with residue vector environments for target structures being aligned against the template. As in CORA, CORALIGN only compares residues in similar local structural environments as determined by accessibility and torsional angles (see above).

Once the target has been aligned against the template, an overall vector score is calculated from the sum of local vector scores ($\text{coravect}(i)$) for each position in the alignment.

$$\text{cora vector score} = \left(\sum_{i=1}^{i=N_{al}} \text{coravect}(i) \right) / N_{al}$$

where N_{al} is the number of positions aligned between target structure and 3D template. Local vector scores ($\text{coravect}(i)$) are calculated in a similar manner as for CORA (see above) by comparing target vectors with equivalent average vectors in the 3D template.

CORALIGN also calculates a contact-based score. This is the percentage of completely conserved contacts in the template structure that is also present in the target structure. Completely conserved contacts in the template are those contacts that are present in all structures selected to make the template. They are easily identified from the consensus contact map.

$$\text{CORA contact score} = \frac{N_{\text{target}} \times 100}{N_{\text{template}}}$$

where N_{template} are the number of completely conserved contacts in the 3D template and N_{target} are the number that are also present in the target structure.

The number of equivalent residues identified between template and target (%overlaps) are also recorded as the following percentages:

$$\text{Equiv}_{\text{template}} = \frac{\# \text{ of equivalent positions}}{\# \text{ of positions in template}} \times 100 (\%)$$

and

$$\text{Equiv}_{\text{target}} = \frac{\# \text{ of equivalent positions}}{\# \text{ of positions in target}} \times 100 (\%).$$

Optimizing parameters for CORA and CORALIGN—Measuring the performance of data bank scans

Parameters for CORA and CORALIGN were optimized for each protein class, using three superfold families from CATH: the mainly- α globin-like fold, the mainly- β OB fold, and the α - β TIM barrel fold (Table 1). For each parameter tested, the CORA template was scanned against a data set of >1,000 nonidentical structures from the PDB, using CORALIGN. A simple optimization strategy was chosen in which parameters likely to affect the performance of CORA and CORALIGN were progressively optimized, to increase both the number of relatives correctly identified and their average alignment score.

Default values for each parameter were used initially to establish suitable cutoffs on CORA scores and %overlaps, which minimized both the number of relatives unrecognized (false negatives R^-) and the number of unrelated proteins with scores above the cutoffs (false positives F^+). Improvements in template performance for different parameter values were then assessed from changes in R^- , F^+ . The average contact score for related structures was also calculated, together with the average Z-score, calculated as follows:

$$Z\text{-score}_{\text{avg}} = \left(\sum_{i=1}^{i=N_{\text{rel}}} Z\text{-score}_i \right) / N_{\text{rel}}$$

where $Z\text{-score}_i$ is the Z-score for i 'th related structure and N_{rel} the number of related structures.

$$Z\text{score}_i = \frac{S_i - S_{\text{unrel_avg}}}{\text{stddev}_{\text{unrel_avg}}}$$

S_i is the CORA contact score for relative i . $S_{\text{unrel_avg}}$ and $\text{stddev}_{\text{unrel_avg}}$ are the average contact score and standard deviation, respectively, for all unrelated structures.

1. Selection of residues for comparison (parameter *cora_sel* and *coralign_sel*)

The double dynamic programming alignment algorithm employed in CORA and CORALIGN is known to be sensitive to the number of residue positions compared (Orengo & Taylor, 1990; Toh, 1997). These values are controlled by the parameters *cora_sel* and *coralign_sel*, which ensure that structurally similar positions are compared (having similar accessibility and torsional angles). For too few positions the alignment can become unreliable in variable regions, while too many positions selected results in excessive noise in the score matrix.

Table 2 shows that for all families, template performance was relatively insensitive to these parameters apart for the TIM template, where performance is slightly affected by decreasing *cora_sel*, as orientations of the α -helices are quite variable across this family.

2. Selection of fold positions from the multiple alignment for the consensus 3d template (parameter *select*)

The number of positions selected for the template is governed by the parameter *select*. This is the cutoff on the required conservation of the vector environment, as measured by *coravect*(i). The optimal values were 65 for the OB fold family (this selects 108 positions, 43% of positions in the multiple alignment), 70 for the TIM barrels (231 positions, 37% of multiple alignment), and 75 for the globin-like fold (147 positions 54% of the multiple alignment).

3. Identification of conserved contacts in a fold family (parameter *c_beta* and *res_sep*)

Conserved contacts are critical in the new scoring scheme used to identify relatives by CORALIGN. Two parameters control the number of conserved contacts identified. *C_beta* sets the allowed $C\beta$ distance for residue contact and the parameter *res_sep* controls the necessary separation along the sequence for two contacting residues. In the mainly- β OB fold family, template performance was very sensitive to both parameters, while for globin-like and TIM barrel families, performance was less affected (Table 2). For the OB family, increasing the *C_beta* cutoff identifies more con-

tacts between the β -sheets. This gives rise to a contact pattern that is very characteristic of the wedge-like architecture of the OB fold rather than the more general β -sheet contact pattern identified at lower values of *C_beta*.

Using CORA templates and consensus data to classify proteins in the CATH structural database

CORA templates have been automatically generated for each fold family and homologous superfamily in CATH. The CORA suite and structural templates will be available via the web address (<http://www.biochem.ucl.ac.uk/bsm/cath>).

Acknowledgments

This work has been possible due to the support of the Medical Research Council. I am indebted to Janet Thornton, David Jones, and Mark Swindells for comments on the manuscript. I would also like to thank Bissan Al Lazaki and Frances Richardson for some preliminary analysis of the globin family using the CORA suite.

References

- Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein data-bank. In: Allen FH, Bergerhoff G, Sievers R. eds. *Crystallographic databases—Information content, software systems, scientific applications*. pp 107–132.
- Altman RB, Hughes C, Gerstein MB. 1995. Methods for displaying macromolecular structural uncertainty: Application to the globins. *J Mol Graphics* 13:142–152.
- Barton GJ, Sternberg MJE. 1990. Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. *J Mol Biol* 212:389–402.
- Bowie JU, Clarke ND, Pabo CO, Sauer RT. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins Struct Funct Genet* 7:257–264.
- Brown N, Taylor WR, Orengo CA. 1995. A protein structure comparison methodology. *Comput Chem* 20:359–380.
- Fischer D, Eisenberg D. 1996. Protein fold recognition using sequence derived predictions. *Protein Sci* 5:947–955.
- Flores TP, Orengo CA, Moss DS, Thornton JM. 1993. Conformational characteristics in structurally similar protein pairs. *Protein Sci* 2:1811–1826.
- Gerstein M, Altman RB. 1995. Average core structures and variability measures: Application to the immunoglobulins. *J Mol Biol* 251:161–175.
- Hilbert M, Bohm G, Jaenicke R. 1993. Structural relationships of homologous proteins as a fundamental principle in homology modelling. *Proteins Struct Funct Genet* 17:138–151.
- Hogue CWV, Ohkawa H, Bryant SH. 1996. WWW-Entrez and the molecular modelling database. *Trends Biochem Sci* 21:226–229.
- Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. 1993. A database of protein structure families with common folding motifs. *Protein Sci* 1:1691–1698.
- Holm L, Sander C. 1993a. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123–128.
- 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.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr* 24:946–950.
- Lee BK, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol* 55:379–400.
- Lesk AM, Chothia CH. 1980. How different amino acids determine similar protein structures: The structure and evolutionary dynamics of the globins. *J Mol Biol* 136:225–270.
- Murzin A. 1993. OB(Oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J* 12:861–867.
- Murzin A, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of the protein database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.

- Orengo CA. 1994. Classification of protein folds. *Curr Opin Struct Biol* 4: 429–440.
- Orengo CA, Brown NP, Taylor WR. 1992. Fast structure alignment for databank searching. *Proteins* 14:139–167.
- Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature* 37:631–634.
- Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH: A hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
- Orengo CA, Taylor WR. 1990. A rapid method for protein structure alignment. *J Theor Biol* 147:517–551.
- Orengo CA, Taylor WR. 1996. SSAP: Sequential structure alignment program for protein structure comparisons. *Methods Enzymol* 266:617–634.
- Phillips DC. 1970. Development of crystallographic enzymology. *Biochem Soc Symp* 31:11–28.
- Russel RB, Barton GJ. 1993. Multiple protein sequence alignment from tertiary structure comparisons: Assignments of global and residue level confidences. *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.
- Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815.
- Sander C, Schneider R. 1991. Database of homology-derived protein structures and structural meaning of sequence alignment. *Proteins Struct Funct Genet* 9:56–68.
- Schmidt R, Gerstein M, Altman RB. 1997. LPFC: An internet library of protein family core structures. *Protein Sci* 6:246–248.
- Siddiqui AS, Barton GJ. 1995. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci* 4:872–884.
- Sowdhamini R, Rufino SD, Blundell T. 1996. A database of globular protein structural domains: Clustering of representative family members into similar folds. *Fold Design* 1:209–220.
- Sutcliffe MJ, Haneef I, Carney D, Blundell TL. 1987. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng* 1:377–384.
- Taylor WR. 1988. Identification of protein sequence homology by consensus template alignment. *J Mol Biol* 188:233–258.
- Taylor WR, Flores TP, Orengo CA. 1994. Multiple protein structure alignment. *Protein Sci* 3:1858–1870.
- Taylor WR, Orengo CA. 1989a. Protein structure alignment. *J Mol Biol* 208: 1–22.
- Taylor WR, Orengo CA. 1989b. A holistic approach to protein structure alignment. *Protein Eng* 2:505–519.
- Toh H. 1997. Introduction of a distance cut-off into structural alignment by the double dynamic programming algorithm. *CABIOS* 13:387–396.