Protein Science (1993), *2*, 1811-1826. Cambridge University Press. Printed in the USA.

Comparison of conformational characteristics in structurally similar protein pairs

T.P. FLORES,[1,2,3] C.A. ORENGO,[1] D.S. MOSS,[2] AND J.M. THORNTON[1]

[1] Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology,
University College, Gower Street, London WC1E 6BT, United Kingdom

[2] Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College,
University of London, Malet Street, London WC1E 7HX, United Kingdom

(Received May 18, 1993; Accepted August 10, 1993)

Abstract

Although it is known that three-dimensional structure is well conserved during the evolutionary development of proteins, there have been few studies that consider other parameters apart from divergence of the main-chain coordinates. In this study, we align the structures of 90 pairs of homologous proteins having sequence identities ranging from 5 to 100%. Their structures are compared as a function of sequence identity, including not only consideration of Cα coordinates but also accessibility, Ooi numbers, secondary structure, and side-chain angles. We discuss how these properties change as the sequences become less similar. This will be of practical use in homology modeling, especially for modeling very distantly related or analogous proteins. We also consider how the average size and number of insertions and deletions vary as sequences diverge. This study presents further quantitative evidence that structure is remarkably well conserved in detail, as well as at the topological level, even when the sequences do not show similarity that is significant statistically.

Keywords: accessibility; homology modeling; Ooi number; protein structure; secondary structure; side-chain angles; structural alignment

It is well known that the three-dimensional structure of a protein is much better conserved during evolution than is sequence, to the extent that homologous proteins with insignificant sequence similarities retain very similar topologies. Previous studies have concentrated mainly on how the positions of the α carbons change as sequences diverge, measured by root mean square (RMS) overlaps between structures (Chothia & Lesk, 1986, 1987; Hubbard & Blundell, 1987; Orengo et al., 1992). However, many different methods are used to characterize structures (e.g., secondary-structure content, accessibility, and torsion angles), and given the use of highly diverged families of sequences to predict protein structure, it is pertinent to ask how these other parameters change as a protein evolves. For example, multiple sequences are now used regularly to predict secondary structure and accessibility (Benner & Gerloff, 1991; Rost et al., 1993). If these properties change substantially in distantly related proteins, then this fact should be taken into account during prediction. In addition, it is of interest to ask how multiple sequence changes—such as those that have occurred in distantly related structures—are accommodated in a structure. For example, are the side chains altered radically in their packing by changing χ^1 values, or are residues buried by side-chain groups more accessible when the side chains mutate?

A successful ab initio method for the determination of protein structure from amino acid sequence has yet to be discovered. Currently, the most accurate method for the prediction of protein structure is model building from a protein or proteins of known structure that have been identified as homologous from sequence analysis. The first attempts at model building were conducted with an α-lactalbumin, on the basis of the hen egg white lysozyme coordinates, and with mammalian serine proteases (Browne et al., 1969; Greer, 1981). The early studies were carried out predominantly by hand, but this process has now been automated (Sutcliffe et al., 1987a,b; Blundell et al., 1988; Šali et al., 1990). The method involves four fundamental steps: (1) determine an accurate alignment between the protein sequences; (2) from this alignment, replace the core residues from the known structure with those of the unknown; (3) replace the loops of the known

Reprint requests to: T.P. Flores, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom.

[3] Present address: National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom.

structure with plausible conformations that the unknown sequence could occupy; and (4) build the side-chain conformations.

With the success of this method and the realization that proteins with very little detectable sequence identity may fold into very similar structures, this technique is being extended to the modeling of ever more distantly related proteins. Several recent publications have described algorithms that allow the most likely fold for a given sequence to be identified from a data base of known folds (Bowie et al., 1990, 1991; Overington et al., 1990; Jones et al., 1992) or that select the most appropriate topology from plausible models (Finkelstein & Reva, 1991; Taylor, 1991). Once a possible candidate has been identified as the most likely fold, a model of the sequence of unknown structure must be constructed, based on the known fold and in a manner identical to that described previously. With this increase in the prediction of structure by homology modeling, it is useful to be aware of the extent of divergence expected in these parameters, such as χ^1 values, to improve the modeling protocols. The extension of these methods to the more difficult problem of modeling analogous proteins makes this information even more relevant.

We can also ask the question whether deviations in coordinate positions, torsion angles, or accessibility can be used to distinguish analogous and homologous structures. Proteins with the same fold and function have presumably evolved from the same ancestor and may show better conservation of detail (e.g., side-chain packing) than analogous structures (i.e., those that have the same topology but no apparent functional or evolutionary relationship).

With the increased size of the protein structure data base (Bernstein et al., 1977) and the abundance of automatic methods for the comparison of protein structures (Taylor & Orengo, 1989a,b; Orengo & Taylor, 1990; Šali & Blundell, 1990; Rose & Eisenmenger, 1991; Orengo et al., 1992; see Orengo, 1992, for review), it is now possible to consider a larger data base of structures that have been aligned more reliably and a wider set of characteristics. In this paper we have aligned a data base of 90 protein pairs with pairwise identities ranging from 5 to 100%. We consider a wide variety of characteristics that are crucial to prediction and modeling of protein structures, including the divergence of Cα main-chain atoms, solvation parameters, secondary-structure content, and side-chain conformations.

Results

Comparison of independent solutions of the same structure

Table 1 summarizes the comparisons between identical proteins, i.e., proteins with the same name and from the same species, that have been solved in different labora-

Table 1. *Structure comparisons of identical proteins determined independently*[a]

		Comparative percent agreement								
Protein		RMS deviation of Cα atoms (Å)	Accessibility (%)	Ooi radius 8 Å	Ooi radius 14 Å	Secondary structure	χ^1		χ^2	
1	2						All	<15%	All	<15%
1cd4	2cd4	0.77	12.0	0.8	2.0	84.4	55.3	23.3	46.5	23.6
1p2p	4p2p	0.57	5.6	0.5	1.2	89.5	30.1	17.0	46.4	21.6
1i1b	**2i1b**	**0.36**	**5.2**	**0.2**	**0.9**	**96.0**	**16.3**	**12.0**	**18.4**	**7.4**
1i1b	**4i1b**	**0.32**	**4.2**	**0.2**	**0.5**	**95.4**	**7.5**	**15.7**	**20.2**	**12.5**
1i1b	5i1b	0.34	4.6	0.3	0.7	98.0	17.1	14.7	20.0	5.7
2i1b	**4i1b**	**0.39**	**5.2**	**0.3**	**1.0**	**95.4**	**35.1**	**17.3**	**25.7**	**8.7**
2i1b	5i1b	0.33	4.5	0.3	0.9	95.4	26.3	15.2	24.9	5.7
4i1b	5i1b	0.39	4.6	0.4	0.9	94.7	34.0	18.6	33.2	8.8
1utg	**2utg(A)**	**0.51**	**7.1**	**0.3**	**0.8**	**94.3**	**33.0**	**10.4**	**47.8**	**170.5**
2cna	3cna	1.00	7.5	0.9	2.3	94.1	50.0	45.1	43.1	37.4
2fgf	**3fgf**	**0.68**	**6.1**	**0.5**	**1.1**	**94.4**	**30.1**	**17.8**	**22.7**	**2.1**
3est	**6est**	**0.30**	**2.7**	**0.3**	**0.8**	**99.2**	**21.1**	**6.8**	**22.8**	**4.2**
3pep	4pep	0.79	4.9	0.5	1.2	92.6	30.0	17.8	27.7	15.0
3pep	5pep	0.88	5.5	0.5	1.4	93.6	38.8	24.6	34.7	16.9
4pep	5pep	0.55	4.6	0.4	1.2	96.0	36.4	20.0	23.8	12.4
4cpv	**5cpv**	**0.43**	**5.4**	**0.3**	**0.7**	**96.3**	**24.9**	**5.3**	**32.1**	**3.2**
Residue mean		**0.40**	**4.7**	**0.3**	**0.8**	**96.3**	**22.8**	**12.1**	**25.2**	**12.6**
		(995)	**(995)**	**(995)**	**(995)**	**(995)**	**(663)**	**(198)**	**(185)**	**(31)**

[a] All values except those of RMS deviation of Cα backbone atoms and secondary structure are the average difference between the aligned proteins. Boldface type is used for those pairs of structures that are both refined at a resolution of 2.0 Å or better; the residue mean and (in parentheses) the number of residues compared are given for these pairs.

tories by independent groups. These pairs allow an estimate of the amount of difference between structures that would be expected by experimental error. Three of the protein pairs in this table are not identical (see Table 2), but are included because they are from the same species and would be expected to have 100% sequence identity. In these cases there are up to four residues that are not the same. The RMS deviation between the Cα atoms for structures defined to resolutions better than 2 Å ranges from 0.30 to 0.68 Å. When all protein pairs are considered, regardless of their resolution, this increases to 1.0 Å. This is in agreement with the work of Hubbard and Blundell (1987), who found that the RMS deviation decreases as resolution is improved.

The solvation and contact parameters compare well within protein pairs, generally with a small overall difference in values. The secondary-structure assignments match less well, with structures at lower resolution giving values less than 85% consistent. As expected, the side-chain conformation angles agree much more in the buried regions of the protein, where the electron density is usually adequate for confident placement of side chains. The minima in which the side chains fall are in good agreement, with 81.7% of all χ^1 angles and 95.8% of buried χ^1 angles occupying the same minima. For χ^2 angles whose χ^1 occupy the same minima, these values are 86.7% and 97.1%, respectively. This has implications for testing routines devised for generating side-chain conformations given the backbone atoms: many of the accessible side chains of the X-ray structures are defined poorly, and therefore when comparing predicted and observed side-chain angles of all residues, only about 80% agreement can be expected.

For comparison, the protein pair 1paz and 2paz involved molecular replacement using 1paz as the starting model for the refinement of 2paz. The RMS deviation for this pair is much lower, 0.14 Å. Similarly, the other deviations are also found to be much lower.

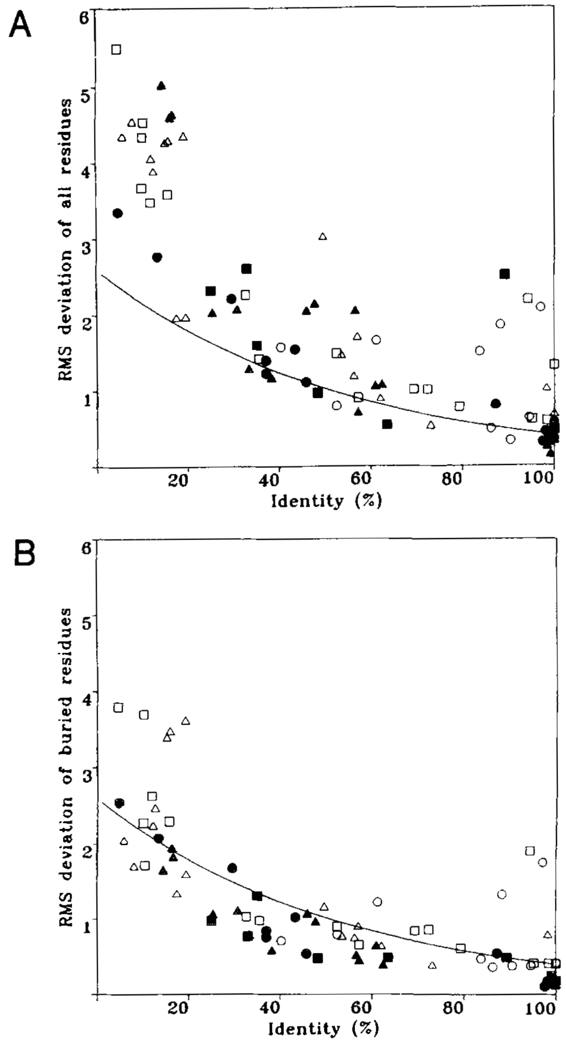

Fig. 1. The RMS deviation of Cα backbone coordinates as a function of sequence identity. **A:** All equivalent residues. **B:** Equivalent residues whose side-chain solvent accessibility is <15%. Symbol shape represents protein class: ○, predominantly α; △, predominantly β; □, α/β. Filled symbols represent pairs of structures refined at a resolution of 2 Å or better. The line in each panel is that due to analysis of Chothia and Lesk (1986).

Conservation of backbone conformation in structurally similar proteins

Probably the simplest measure of structural relatedness involves consideration of backbone conformation and of how the Cα coordinates compare between structures as sequence identity diverges. For all equivalent residues, the RMS deviation of the Cα atoms increases steeply in a nonlinear fashion with decreasing sequence identity (Fig. 1A). Equivalent residues whose side-chain accessibility is less than 15% exhibit a similar trend, but the upturn for protein pairs of low sequence identity is not as pronounced (Fig. 1B).

Previous studies have defined the criteria for equivalence differently. Chothia and Lesk (1986, 1987) calculated the RMS deviation for residues in the common core. This common core is obtained by a series of superpositions of the main-chain atoms of the major secondary-structure elements, gradually including additional residues until the Cα of the last residue to be included deviates by no more than 3 Å. After all of the equivalent residues have been determined, the overall superposition of these residues is calculated. For proteins with less than 20% sequence identity, this accounted for approximately half the number of residues in each sequence. Using this definition, Chothia and Lesk (1986, 1987) showed that the variation of RMS deviation in Cα coordinates with sequence identity may be fitted with an exponential curve. In contrast, Hubbard and Blundell (1987) obtained equivalent residues by a succession of superpositions using the whole

Table 2. *Pairs of aligned protein structures used in this study*

Name	Code	Chain	Length	Resolution	Name	Code	Chain	Length	Resolution	% Identity
Leucine aminopeptidase *Bos taurus*	1lap		480	2.7	Carboxypeptidase A *Bos taurus*	5cpa		307	1.54	4.6
Cytochrome B562 *Escherichia coli*	256b	A	106	1.4	Myohemerythrin *Themiste zostericola*	2mhr		118	1.7	4.7
Erythrina trypsin inhibitor *Erythrina cafera*	1tie		165	2.5	Interleukin-1 beta *Homo sapiens*	1i1b		151	2.0	7.9
Actin *Oryctolagus cuniculus*	—	A	372	2.8	ATPase fragment of heat shock cognate protein *Bos taurus*	—		382	2.2	9.9
Cytoplasmic malate dehydrogenase *Sus scrofa*	4mdh	A	333	2.5	Flavodoxin *Clostridium* MP	4fxn		138	1.8	10.1
Rhodanese (domain 1; res. 1-148) *Bos taurus*	1rhd		148	2.5	Rhodanese (domain 2; res. 149-293) *Bos taurus*	1rhd		145	2.5	10.3
Glycolate oxidase *Spinacia oleracea*	1gox		349	2.0	Tryptophan synthase *Salmonella typhimurium*, strain TB2211/PSTH8	1wsy	A	246	2.5	11.8
CD4 (res. 1-98) *Homo sapiens*	1cd4		98	2.3	CD4 (res. 99-176) *Homo sapiens*	1cd4		75	2.3	12.0
Azurin *Pseudomonas aeruginosa*	1azu		124	2.7	Pseudoazurin *Alcaligenes faecalis*, strain S-6	1paz		120	1.55	12.5
Hemoglobin III (erythrocruorin) *Chironomus thummi thummi*	1ecd		136	1.4	Hemoglobin *Homo sapiens*	4hhb	A	141	1.74	13.2
Elastase *Sus scrofa*	3est		240	1.65	Proteinase A *Streptomyces griseus*, strain K1	2sga		181	1.5	14.4
Immunoglobulin IgAκ Fab fragment J539 (res. 1-118) *Mus musculus*	2fbj	H	118	1.95	Immunoglobulin IgG1 Fc fragment (res. 238-341) *Homo sapiens*	1fc1	A	103	2.9	14.6
Azurin *Alcaligenes denitrificans*	2aza	A	129	1.8	Plastocyanin *Populus nigra* var. *italica*	1pcy		99	1.6	15.2
Lactate dehydrogenase *Sus scrofa*	5ldh		333	2.7	Cytoplasmic malate dehydrogenase *Sus scrofa*	4mdh	A	333	2.5	15.6
Azurin *Alcaligenes denitrificans*	2aza	A	129	1.8	Pseudoazurin *Alcaligenes faecalis*, strain S-6	1paz		120	1.55	15.8
Retinol binding protein *Homo sapiens*	1rbp		174	2.0	Bilin binding protein *Pieris brassicae*	1bbp	A	173	2.0	16.2
Proteinase A *Streptomyces griseus*, strain K1	2sga		181	1.5	Trypsin beta *Bos taurus*	4ptp		223	1.34	16.6
CD4 (res. 1-98) *Homo sapiens*	1cd4		98	2.3	Immunoglobulin IgG1 Fab fragment (res. 1-109) *Homo sapiens*	2fb4	L	216	1.9	17.3
Troponin C *Meleagris gallopavo*	5tnc		161	2.0	Calmodulin *Bos taurus*	3cln		143	2.2	18.9
Azurin *Pseudomonas aeruginosa*	1azu		124	2.7	Plastocyanin *Populus nigra* var. *italica*	1pcy		99	1.6	19.2
CD4 (res. 1-98) *Homo sapiens*	1cd4		98	2.3	Immunoglobulin Fab Bence/Jones protein *Homo sapiens*	2rhe		114	1.6	19.4
Porcine pancreatic secretory trypsin inhibitor *Sus scrofa*	1tgs	I	56	1.8	Ovomucoid third domain *Lophura nycthemera*	2ovo		56	1.5	25.0

Rous sarcoma virus protease Rous sarcoma virus, strain PR-C	2rsp	A	114	2.0	HIV-1 protease HIV I	5hvp	A	99	2.0	25.3
Hemoglobin V *Petromyzon marinus*	2lhb		149	2.0	Myoglobin *Aplysia limacina*	4mba		146	2.0	29.5
Acid proteinase penicillopepsin *Penicillium janthinellum*	3app		323	1.8	Pepsin *Sus scrofa*	4pep		326	1.8	30.7
L-Lactate dehydrogenase *Lactobacillus casei*	1llc		320	3.0	Lactate dehydrogenase *Mus musculus*	2ldx		331	2.98	32.5
Proteinase K *Tritirachium album limber*	2prk		279	1.5	Subtilisin Carlsberg *Bacillus subtilis*	1cse	E	274	1.2	32.8
Rat mast cell protease II *Rattus rattus*	3rp2	A	224	1.9	Trypsin beta *Bos taurus*	4ptp		223	1.34	33.2
Chymotrypsin inhibitor 2 *Hordeum vulgare*, hiproly strain	2ci2	I	65	2.0	Eglin C *Hirudo medicinalis*	1cse	I	63	1.2	34.9
L-Lactate dehydrogenase *Bacillus stearothermophilus*	1ldb		291	2.8	Lactate dehydrogenase *Squallus acanthias*	6ldh		329	2.0	35.4
Alpha lactalbumin *Papio cynocephalus*	alc		122	1.7	Lysozyme *Homo sapiens*	1lz1		130	1.5	36.9
Cytochrome C2 *Rhodospirillum rubrum*	3c2c		112	1.68	Cytochrome C5 *Katsuwonis pelamis* Linnaeus or *Thunnus alalunga*	cyt	R	103	1.5	36.9
Trypsin beta *Bos taurus*	4ptp		223	1.34	Elastase *Sus scrofa*	3est		240	1.65	38.1
Phospholipase A2 *Crotalus atrox*	1pp2	R	122	2.5	Phospholipase As *Bos taurus*	1bp2		123	1.7	40.2
Hemoglobin *Equus caballus*	2mhb	A	141	2.0	Hemoglobin *Equus caballus*	2mhb	B	146	2.0	43.3
Myohemerythrin *Themiste zostericola*	2mhr		118	1.7	Hemerythrin *Themiste dyscrita*	2hmq	A	114	1.66	45.6
Immunoglobulin IgG1 Fab fragment (res. 1-109) *Homo sapiens*	2fb4	L	109	1.9	Immunoglobulin IgA Fab fragment (res. 1-113) *Mus musculus*	1mcp	L	113	1.7	45.9
Immunoglobulin Fab Bence/Jones protein *Homo sapiens*	2rhe		114	1.6	Immunoglobulin IgGκ Fab Bence/Jones REI *Homo sapiens*	1rei	A	107	2.0	47.7
Actinidin *Actinidia chinensis*	2act		218	1.7	Papain *Carica papaya*	1ppd		212	2.0	48.1
Immunoglobulin IgG1 pFc′ fragment *Cavia porcellus*	1pfc		111	3.13	Immunoglobulin IgG1 Fc fragment *Homo sapiens*	1fc1	A	206	2.9	49.5
D-Glyceraldehyde-3-phosphate dehydrogenase *Bacillus stearothermophilus*	1gd1	O	334	1.8	D-Glyceraldehyde-3-phosphate dehydrogenase *Homo sapiens*	3gpd	R	334	3.5	52.4
N-terminal domain of 434 repressor protein Phage 434	1r69		63	2.0	434 Cro protein Phage 434	2cro		65	2.35	52.4
Immunoglobulin IgG1 Fab fragment (res. 1-118) *Homo sapiens*	2fb4	H	118	1.9	Immunoglobulin IgAκ Fab fragment (res. 1-122) *Mus musculus*	1mcp	H	122	1.7	56.6
Phosphofructokinase *Bacillus stearothermophilus*	4pfk		319	2.4	Phosphofructokinase *Escherichia coli*	2pfk	A	301	2.4	57.1
Plastocyanin *Populus nigra* var. *italica*	1pcy		99	1.6	Plastocyanin *Enteromorpha prolifera*	7pcy		98	1.8	57.1
Immunoglobulin IgG Fab new (res. 1-117) *Homo sapiens*	3fab	H	117	2.0	Immunoglobulin IgG1κ Fab fragment (res. 1-116) *Mus musculus* and *Gallus gallus*	3hfm	H	116	3.0	59.5

(continued)

Table 2. *Continued*

Name	Code	Chain	Length	Resolution	Name	Code	Chain	Length	Resolution	% Identity
Carbonic anhydrase (form B) *Homo sapiens*	2cab		256	2.0	Carbonic anhydrase II *Homo sapiens*	3ca2		256	2.0	60.9
Cytochrome C *Oryza sativa*	1ccr		111	1.5	Cytochrome C *Katsuwonis pelamis* Linnaeus or *Thunnus alalunga*	1cyc		103	2.3	61.2
Azurin *Pseudomonas aeruginosa*	1azu		124	2.7	Azurin *Alcaligenes denitrificans*	2aza	A	129	1.8	62.1
Proteinase B *Streptomyces griseus*, strain K1	3sgb	E	185	1.8	Proteinase A *Streptomyces griseus*, strain K1	2sga		181	1.5	62.4
Rubredoxin *Clostridium pasteurianum*	5rxn		54	1.2	Rubredoxin *Desulfovibrio gigas*	1rdg		52	1.4	63.5
Subtilisin Carlsberg *Bacillus subtilis*	1sbc		274	2.5	Subtilisin BPN (Novo) *Bacillus amyloliquefaciens*	2sbt		274	2.8	69.3
D-Glyceraldehyde-3-phosphate dehydrogenase *Homo sapiens*	3gpd	R	334	3.5	D-Glyceraldehyde-3-phosphate dehydrogenase *Homarus americanus*	1gpd	R	333	2.9	72.4
Immunoglobulin IgG1 Fab fragment (res. 1-105) *Mus musculus*	2hfl	L	105	2.54	Immunoglobulin IgG2bκ Fab R19.9 (res. 1-108) *Mus musculus*	1f19	L	108	2.8	72.6
Trypsin *Rattus rattus*	2trm		223	2.8	Trypsin beta *Bos taurus*	1ntp		223	1.8	73.1
Leucine binding potein *Escherichia coli*, strain K12	2lbp		346	2.4	Leucine binding protein *Escherichia coli*, strain K12	2liv		344	2.4	79.4
Phospholipase A2 *Sus scrofa*	1p2p		124	2.6	Phospholipase A2 *Bos taurus*	1bp2		123	1.7	83.7
Myoglobin *Sus scrofa*	1pmb	A	153	2.5	Myoglobin *Physeter catodon*	1mbd		153	1.4	86.3
Hemoglobin *Equus caballus*	2mhb	A	141	2.0	Hemoglobin *Homo sapiens*	4hhb	A	141	1.74	87.2
Myoglobin *Phoca vitulina*	1mbs		153	2.5	Myoglobin *Sus scrofa*	1pmb	A	153	2.5	88.2
Ovomucoid third domain *Lophura nycthemera*	2ovo		56	1.5	Ovomucoid third domain *Coturnix coturnix japonica*	1ovo	A	56	1.9	89.3
Insulin *Bos taurus*	2ins	A	21	2.5	Insulin *Sus scrofa*	4ins	A	21	1.5	90.5
Hexokinase B *Saccharomyces cerevisiae*	2yhx		457	2.1	Hexokinase A *Saccharomyces cerevisiae*	1hkg		457	3.5	94.3
Lysozyme *Meleagris gallopavo*	2lz2		129	2.2	Lysozyme *Gallus gallus*	8lyz		129	2.5	94.6
Human class I histocompatibility antigen AW 68.1 *Homo sapiens*	2hla	A	270	2.6	Human class I histocompatibility antigen A2.1 *Homo sapiens*	3hla	A	270	2.6	95.2
TRP repressor *Escherichia coli*	1wrp	R	102	2.2	TRP repressor *Escherichia coli*	3wrp		101	1.8	97.0
Mutant lysozyme Phage T4 in *Escherichia coli*	1l35		164	1.8	Lysozyme Phage T4 in *Escherichia coli*	3lzm		164	1.7	97.6

Troponin C *Gallus gallus*	4tnc		160	2.0	Troponin C *Meleagris gallopavo*	5tnc		161	2.0	98.1
Trypsin inhibitor (Arg 15 analogue) *Bos taurus*	4tpi	I	58	2.2	Trypsin inhibitor *Bos taurus*	5pti		58	1.8	98.3
Concanavalin A *Canavalia ensiformis*	2cna		237	2.0	Concanavalin A *Canavalia ensiformis*	3cna		237	2.4	98.3
Eglin C *Hirudo medicinalis*	1tec	I	63	2.2	Eglin C *Hirudo medicinalis*	1cse	I	63	1.2	98.4
Erabutoxin *Laticauda semifasciata*	3ebx		62	1.4	Erabutoxin *Laticauda semifasciata*	5ebx		62	2.0	98.4
Ferredoxin *Azotobacter vinelandii*	4fd1		106	1.9	Ferredoxin *Azotobacter vinelandii*	1fd2		106	1.9	99.1
Pseudoazurin *Alcaligenes faecalis*, strain S-6	1paz		120	1.55	Pseudoazurin *Alcaligenes faecalis*, strain S-6	2paz		123	2.0	99.2
Calcium-binding parvalbumin B *Cyprinus carpio*	4cpv		108	1.5	Calcium-binding parvalbumin B *Cyprinus carpio*	5cpv		108	1.6	100.0
CD4 *Homo sapiens*	1cd4		173	2.3	CD4 *Homo sapiens*	2cd4		176	2.4	100.0
Elastase *Sus scrofa*	3est		240	1.65	Elastase *Sus scrofa*	6est		240	1.8	100.0
Fibroblast growth factor *Homo sapiens*	2fgf		126	1.77	Fibroblast growth factor *Homo sapiens*	3fgf		124	1.6	100.0
Interleukin-1 beta *Homo sapiens*	4i1b		151	2.0	Interleukin-1 beta *Homo sapiens*	5i1b		151	2.1	100.0
Interleukin-1 beta *Homo sapiens*	1i1b		151	2.0	Interleukin-1 beta *Homo sapiens*	5i1b		151	2.1	100.0
Interleukin-1 beta *Homo sapiens*	1i1b		151	2.0	Interleukin-1 beta *Homo sapiens*	2i1b		153	2.0	100.0
Interleukin-1 beta *Homo sapiens*	1i1b		151	2.0	Interleukin-1 beta *Homo sapiens*	4i1b		151	2.0	100.0
Interleukin-1 beta *Homo sapiens*	2i1b		153	2.0	Interleukin-1 beta *Homo sapiens*	4i1b		151	2.0	100.0
Interleukin-1 beta *Homo sapiens*	2i1b		153	2.0	Interleukin-1 beta *Homo sapiens*	5i1b		151	2.1	100.0
Pepsin *Sus scrofa*	3pep		326	2.3	Pepsin *Sus scrofa*	4pep		326	1.8	100.0
Pepsin *Sus scrofa*	3pep		326	2.3	Pepsin *Sus scrofa*	5pep		326	2.34	100.0
Pepsin *Sus scrofa*	4pep		326	1.8	Pepsin *Sus scrofa*	5pep		326	2.34	100.0
Phospholipase A2 *Sus scrofa*	1p2p		124	2.6	Phospholipase A2 *Sus scrofa*	4p2p		123	2.4	100.0
Uteroglobin *Oryctolagus cuniculus*	1utg		70	1.34	Uteroglobin *Oryctolagus cuniculus*	2utg	A	70	1.64	100.0
Wheat germ agglutinin *Triticum vulgaris*	9wga	A	170	1.8	Wheat germ agglutinin *Triticum vulgaris*	9wga	B	170	1.8	100.0

protein. The first superposition is calculated using a few residues that are known to be equivalent. This set of equivalent residues is then expanded by including more residues that are close to each other in the previous superposition. This process is repeated until the numbers of equivalent residues converge. The common core is taken as equivalent residues that have less than 7% side-chain accessibility. This gives a core of 20–35% of residues per protein. Hubbard and Blundell (1987) suggested a linearly increasing RMS deviation with decreasing sequence identity, with sequence identity given as the number of identical residues in the core divided by the total number of residues in the core. When separating α and β proteins, they suggested that two different equations are needed to fit these data adequately. Orengo et al. (1992) considered the RMS deviation of the Cα atoms that have a non-zero score from the structural alignments, as described later. Their results showed a nonlinear upturn of RMS deviation with diverging sequence identity similar to that seen by Chothia and Lesk (1986, 1987), but their values are generally higher than the exponential curve described earlier.

For the purpose of comparison, the best-fit line calculated by Chothia and Lesk (1986) has been included in Figure 1A and B. However, this line does not fit the data in either of these figures. This is not surprising because the results presented here are not constrained by the 3-Å cutoff imposed by Chothia and Lesk (1986). What is clear is the nonlinear increase in structural divergence with decreasing sequence identity.

The three azurin with plastocyanin structural alignments are seen to be outliers in Figure 1 because one of the β-sheets in the β-sandwich is shifted (Taylor & Orengo, 1989a). The other outliers have different intervening secondary structures, which produce rigid-body shifts of the secondary-structure elements, particularly with α-helices. The outliers at the high homology end of the scale are caused predominantly by large deviations at either the N or C terminal. Exceptions to this are the alignments of yeast hexokinase A and B, which have two different substrates bound. Here the structure is kidney-shaped, with the active site found in the cleft. The two lobes of these proteins are shifted apart and hinged around the middle, causing this large deviation.

As expected, for protein pairs in which one or both structures have a nominal resolution better than 2 Å, the RMS deviations are larger. This is a reflection of the difficulty in placing atoms in the electron density at low resolutions, particularly in loop regions. Also, comparison of Figure 1A and B shows that it is the accessible residues that differ most as the sequence identity becomes very low. These accessible areas, where the differences occur on the whole, tend to be limited to the loop regions. This is particularly noticeable for the immunoglobulin fragments, which have large variation in their loops required for diversity of antigen recognition.

The method of Chothia and Lesk (1986, 1987) using a 3-Å distance for determining residue equivalences cannot be applied during modeling because only sequence equivalences are known. Hubbard and Blundell (1987) tried to overcome this by considering residue accessibility, so that only buried residues are used as the common core. We see no difference in behavior between protein classes (Fig. 1A,B) because these data seem to be dispersed evenly, independent of the secondary-structure content of the proteins.

Conservation of solvation and contact values

Solvation is an important factor in determining protein structure, and many of the methods that match sequences to folds contain a solvation term. This term assumes that the exposure of a given residue and its structurally equivalent residue will be the same. Therefore, it is important to determine to what extent this is affected by sequence identity.

There is an approximately linear relationship between conservation of accessibility and percent identity, i.e., as the sequences diverge, the average difference in side-chain accessibility increases (Fig. 2B,C). For identical residues, with sequence identity above 20%, the average difference in accessibility values is below 10%. Including all residues, accessibility is slightly less well conserved. This reflects the basic conservation of the core and the fact that although loops change their conformation, they are generally accessible. Accessibility is equally well conserved in all types of secondary structure.

The conservation of Ooi numbers, expressed as the number of α carbon atoms found within a defined radius around the α carbon of the residue of interest, is very similar to the trend found for accessibilities, i.e., the increase in difference is linear with decreasing sequence identity (Fig. 3). The average difference of the Ooi value for the 8-Å radius is up to about 2, which is one-fifth of the average Ooi value for this radius. However, the average difference for the 14-Å radius, although larger, is only one-eighth of the average Ooi number for this radius. As expected, the Ooi value for the 14-Å radius, which measures the gross environment of each residue, is conserved better than that of the smaller radius, which is more sensitive to local variations.

There are only a few noticeable outliers in these comparisons, found at the low sequence identity end and involving comparisons of proteins of differing size, e.g., 4mdh with 4fxn.

Conservation of secondary structure

The trend for secondary structure is again approximately linear, similar to that of accessibility (Fig. 4A,B), i.e., as the percent identity decreases, the secondary structure is less well conserved, dropping to 90% at 60% sequence identity. None of the secondary structures was conserved

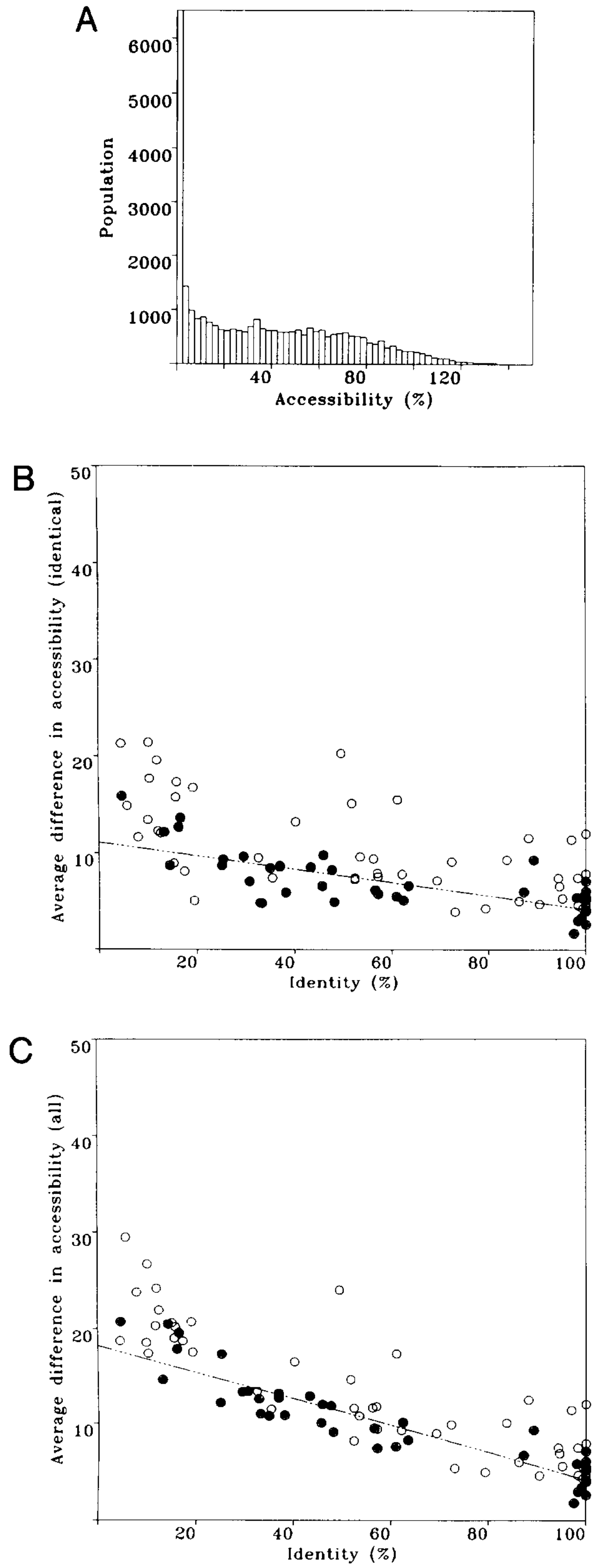

Fig. 2. **A:** Distribution of side-chain solvent accessibilities expressed as a percentage calculated over the whole data set. **B,C:** Average difference in percent accessibility as a function of sequence identity for identical residues (B) and for all equivalent residues (C). Lines fitted by linear regression to the filled symbols, which represent pairs of structures refined at a resolution of 2 Å or better.

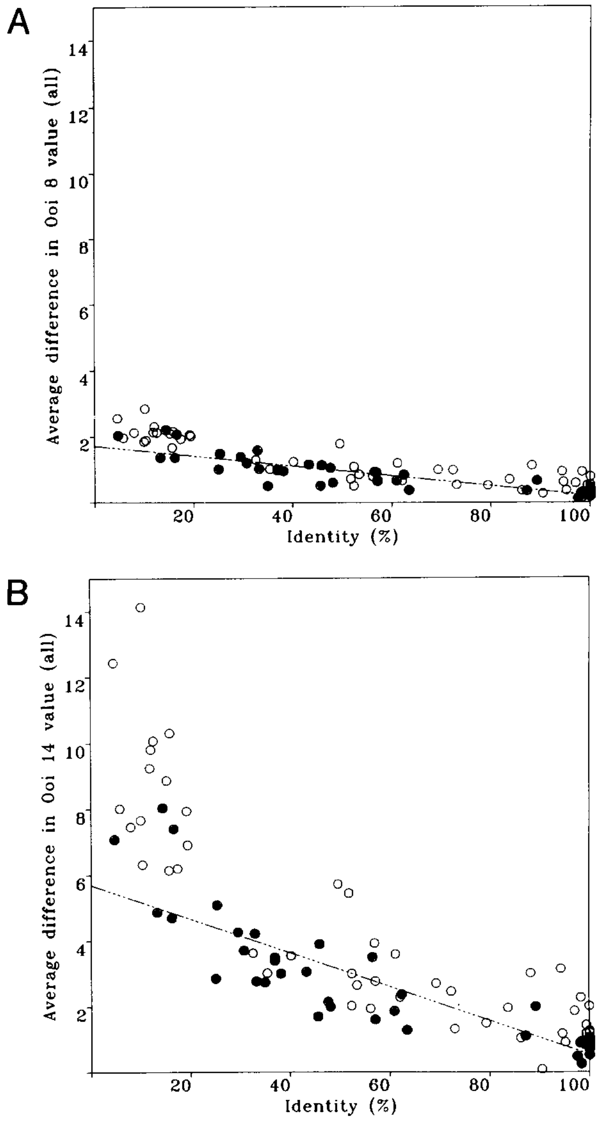

Fig. 3. Average difference in Ooi number for all equivalent residues as a function of sequence identity. **A:** Ooi 8-Å radius. **B:** Ooi 14-Å radius. Lines fitted by linear regression to the filled symbols, which represent pairs of structures refined to 2 Å or better.

preferentially in comparison to others. The spread of points is greater than that of the solvation values. This is due in part to low-resolution structures, for which automatic assignment of secondary structure is less accurate (Morris et al., 1992). In our experience, we find that the automatic assignment of β-strand residues in low-resolution structures is often difficult, and that the number of these residues increases with increasing resolution.

Conservation of χ^1 and χ^2 side-chain conformations

One of the most important parts of modeling by homology is building in the side chains. Methods involve build-

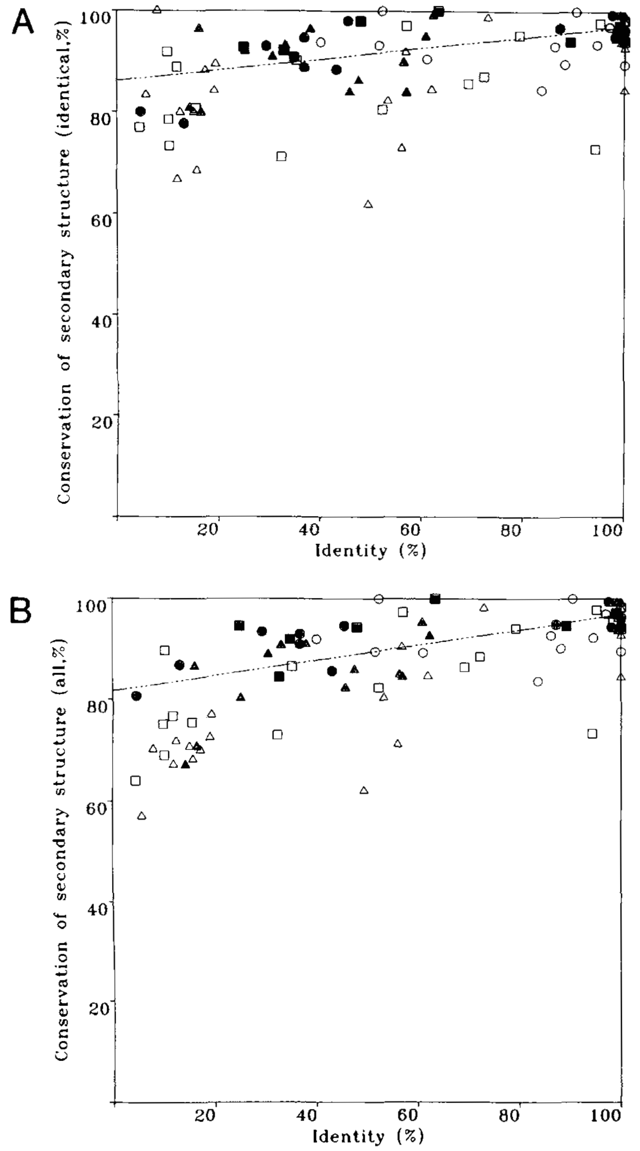

Fig. 4. Conservation of residue secondary structure as a function of sequence identity for (**A**) identical and (**B**) all equivalent residues. Symbol shape represents protein class: ○, predominantly α; △, predominantly β; □, α/β. Lines fitted by linear regression to the filled symbols, which represent pairs of structures refined at a resolution of 2 Å or better.

ing the new side chain in a conformation similar to that adopted by the equivalent residue. Summers et al. (1987) determined rules for modeling side chains based on proteins aligned structurally by hand. Their study was limited to seven proteins from three protein families, with sequence identities in the range of 16–60% and, unlike the present study, their data set was not large enough to investigate how conservation of side-chain conformation varies with sequence identity.

The average difference for both χ^1 and χ^2 angles increases approximately linearly with decreasing sequence identity. In the case of χ^1 angles, those for identical residues increase by the order of 10–15% for all solvation values, whereas those for residues that are buried increase by less than 10%. When all residues that are equivalent structurally are considered, this increase is more pronounced (Fig. 5). For χ^2 angles, the values are more scattered due to reduced numbers of examples per protein compared to χ^1 angles. Except for a few cases, residues that are buried are likely to have very similar χ^2 values, within 10% of each other. When all equivalent residues are considered, there is a gradual increase of the order of 20–25% (Fig. 5D).

As percentage identity decreases, there is an increasing number of side chains that change conformation from one well, or "conformer," to another (Fig. 6). In a manner similar to the average changes in χ angles, these changes are influenced predominantly by accessibility of the side chain. The more buried the side chain, the more likely it is that the side-chain conformations will agree. If the accessibility is less than 15%, the conservation of χ^1 side-chain conformation of all identical residues is independent of sequence identity: 95% of these χ^1 side-chain angles are conserved, suggesting that simply transferring the χ^1 from one structure to another is a good first approximation. The slope in Figure 6C is due to the side chains of different residue types that are slightly less well conserved, depending on the overall sequence identity. This difference is never less than 60% and on average is near 90%. The numbers of occurrences of χ^2 described previously were inevitably much decreased, and the points exhibited greater scatter. It is not possible to see clearly whether there is a trend in χ^2 angles for all of the categories described for χ^1 angles. These data suggest, however, that if the χ^1 angle agrees for buried residues, then χ^2 is 95% likely to agree also, regardless of sequence identity.

The more accessible side chains are much less likely to adopt the same conformation. This is due largely to the actual disorder of the surface side chains. However, from Figures 5 and 6 we can see that when both structures are at high resolution, the side-chain conformations are more likely to agree. Structure pairs related by molecular replacement generally have higher agreement than those solved independently.

Insertions and deletions

Pascarella and Argos (1992) have considered the occurrence of insertions and deletions (indels) in protein pairs aligned using a variety of methods. They made three main observations about the size and occurrence of indels with decreasing percent residue identity: (1) the average length increases exponentially from slightly more than 2 to slightly less than 5 residues; (2) the average length of intervening residues between indels decreases exponentially from just under 60 to 7 or 8; and (3) the average number

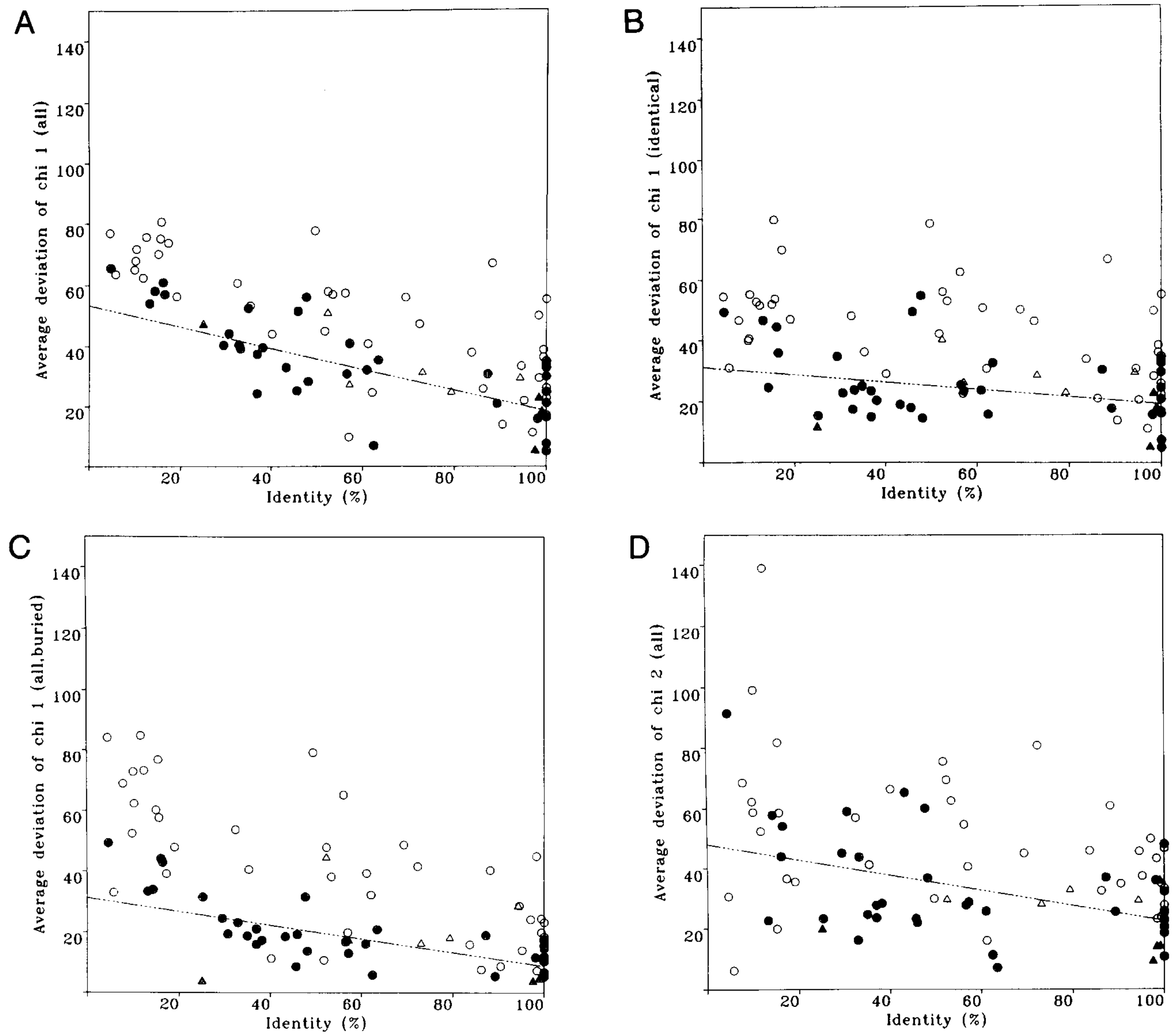

Fig. 5. Average difference of χ angles as a function of sequence identity. **A:** χ^1 for all equivalent residues. **B:** χ^1 for all identical residues. **C:** χ^1 for residues whose side-chain accessibility is <15%. **D:** Average difference of χ^2 angles whose χ^1 angles are conserved. Symbol shape represents the structure determination method: △, molecular replacement; ○, all other methods. Lines fitted by linear regression to the filled symbols, which represent pairs of structures refined to 2 Å or better.

of indels per aligned residue increases from 1 to just over 5 per 100 residues.

In our data set, the average length of indels is seen to increase with diverging sequence identity from one to around six residues (Fig. 7A). These values are similar to those reported by Pascarella and Argos (1992), who also suggested that the number of indels per residue versus sequence identity does not exhibit exponential behavior, but becomes saturated at low sequence identities. However, by considering the number of indels per residue based on the shortest sequence, we see an exponential relationship (Fig. 7B). This difference is probably due to the fact that the number of aligned residues is inherently linked to the number of indels, which is reflected in the plot of two correlated parameters. These findings suggest that the initiation and propagation of a gap should be less penalized for less related sequences.

Discussion

The results presented here are dependent on the quality of the structural alignments calculated by SSAP (structure and sequence alignment program). In our experience, the alignments obtained are very reliable. For example,

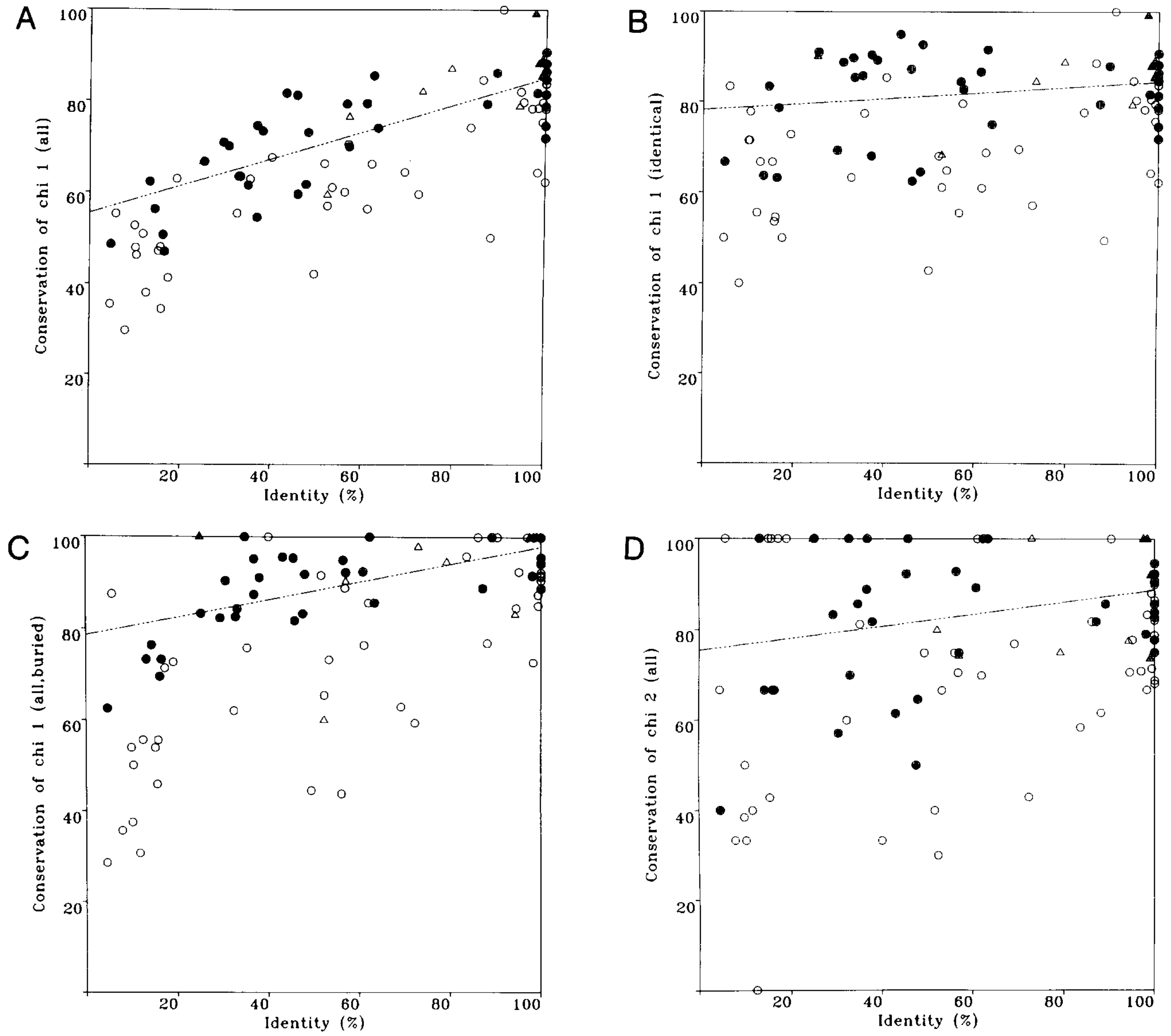

Fig. 6. Conservation of χ angle conformer as a function of sequence identity. **A:** χ^1 for all equivalent residues. **B:** χ^1 for all identical residues. **C:** χ^1 for residues whose side-chain accessibility is <15%. **D:** Conservation of χ^2 angles whose χ^1 angles are conserved. Symbol shape represents refinement method: △, molecular replacement; ○, all other methods. Lines fitted by linear regression to the filled symbols, which represent pairs——ctures refined at a resolution of 2 Å or better.

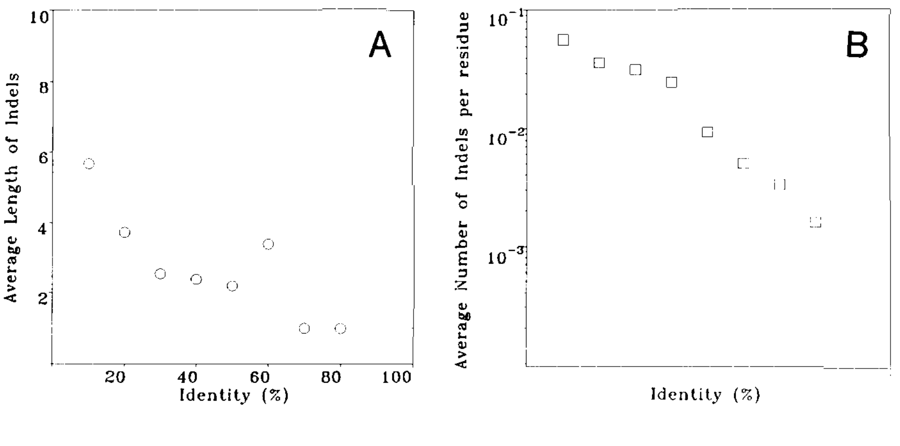

Fig. 7. A: Change in the average length of insertions and deletions per residue with diverging sequence identity. **B:** Change in number of insertions and deletions per residue with diverging sequence identity. The large indel seen in the comparison between 1cms and 1psg has been removed from the calculations because this is formed by proteolytic cleavage of pro-chymosin.

the SSAP alignment with the lowest sequence identity (leucine amino peptidase and carboxypeptidase A) successfully identifies the same equivalences as those determined recently by graph theoretic techniques (Artymuik et al., 1992). The SSAP alignment method is different from many other procedures in that it assigns residues as those that have equivalent environments, whereas other methods determine equivalences from a three-dimensional superposition. This leads to improved alignments for pairs of structures where secondary structural elements have rigid body shifts, giving large Cα deviations yet maintaining the same topology.

In most cases, the conformational characteristics described in this paper remain well conserved even at very low sequence identities (<20%). This provides quantitative evidence that structure is much more conserved than sequence. These findings reinforce the validity of current methods for modeling proteins even for very distant sequences. Despite shifts in the Cα backbone of homologous proteins, the solvation state, secondary structure, and side-chain conformations are all well conserved. The largest changes in homologous protein structures are found to be restricted to residues in and around indels. This also explains why confident placement of loops and the changes associated with different loop conformations is one of the major difficulties for molecular modelers. The divergence between all these characteristics shows a clear correlation with percent identity. This allows the derivation of confidence limits that may be applied to validate the results of homology modeling. For instance, it is unlikely that structures with <20% sequence identity will have an RMS deviation of the α carbon coordinates of less than 1 Å for buried residues.

We also find a clear difference in conservation values between pairs of proteins whose structures are refined using high-resolution data compared with the other pairs. This difference suggests that, as the resolution of the structures of the proteins in the pairs increases, the observed characteristics are much better conserved, i.e., variation is partially due to uncertainty in atomic coordinates.

Are the weakly related proteins formed by divergence from a common ancestor or by convergence from two unrelated ancestors? It seems likely that pairs of proteins with a common activity exhibit a similar structure due to divergent evolution. A recent publication from Ollis et al. (1992) provides some compelling arguments for divergent evolution of the α/β hydrolase fold adopted by five hydrolytic proteins with weak sequence homology. These have a wide variety of substrates yet maintain common catalytic residues even though the binding site has diverged. However, the question is more open for structures that have a similar topology yet have different functions. For instance, the evolutionary relationship between the eight-stranded α/β-barrels (TIM barrels) is still unclear. On one hand it is argued that it is divergent evolution by circular rearrangement of the gene (Farber & Petsko, 1990), whereas others argue for convergent evolution due to the existence of two distinct modes of side-chain packing in these structures (Lesk et al., 1989). This has been questioned further with the crystal structure of a seed storage protein that also has a TIM barrel structure but has no apparent enzymic activity (Hennig et al., 1992). Is it then a protein that has lost its enzymic activity from divergent evolution, or is the TIM barrel structure a common building block to which many sequences converge?

In our data base of protein pairs, only the first six pairs in Table 2 can be considered to have the "same" structure but little or no functional relatedness. It could be argued that if these proteins have converged rather than diverged, then the details of their structures (e.g., side-chain packing, accessibilities, etc.) might show a discontinuity compared to proteins that have diverged over time (i.e., on the plots presented in this paper, these proteins might be expected to be outliers). The plots all suggest that changes in structure are more common below 20% sequence identity. However, some of the pairs having clearly related functions show less similarity than those of unrelated functions. Thus, these data provide no evidence that the "analogous" structures are any less similar than homologous structures with equivalently low sequence identities. This cannot prove that these proteins have evolved by divergent evolution, however, because the constraints of topology may dictate this level of similarity in detailed structure and packing.

Methods

Below a threshold of around 25% sequence identity it is generally very difficult to align proteins optimally from a knowledge of their amino acid sequences alone (Sander & Schneider, 1991). It is necessary to incorporate structural information to obtain the best alignment. For the purpose of this investigation, we aligned 90 pairs of structures using the structure and sequence alignment program SSAP (Taylor & Orengo, 1989a,b; Orengo & Taylor, 1990; Orengo et al., 1992). Throughout this paper, sequence identity is taken as the number of identical residues that can be aligned divided by the length of the shorter sequence; this value is expressed as a percentage.

Protein pairs used for analysis

Protein pairs were chosen in one of two ways: (1) Pairs of proteins that were identified from the literature as being structurally homologous with little sequence homology were aligned with SSAP. (2) A selection of pairs was chosen from an all-pairwise sequence alignment of the Brookhaven Protein Data Bank (D.T. Jones, pers. comm.). No more than two pairs were chosen from any cluster of sequences, thereby ensuring that the data, as far as possible, were not biased to any family of structures. Also included are pairs of protein structures with identical sequences that have been determined independently, in order to determine the variability one would expect from experimentation. All these pairs of sequences are summarized in Table 2 and were aligned using SSAP.

Protein structure alignment and superposition

SSAP uses comparison of internal residue–residue vector distances to determine the optimal alignment between two or more structures. A detailed description of this algorithm was given by Taylor and Orengo (1989a). This residue separation is taken as the vector distance between Cβ atoms in preference to Cα atoms because the former contain more geometric information. For example, the Cβ position defines the side of a β-strand on which the side chain is found. Pairs of residues from each structure having similar main-chain angles and accessibility are selected, and the internal vectors of these residues to all the other residues within their structures are calculated. Using a function that is inversely proportional to the difference between these calculated vectors, it is possible to score a two-dimensional matrix whose axes correspond to the sequences of the proteins being compared. The optimal path representing the alignment of the structures may be traced through this matrix using a dynamic programming algorithm. Scores along the alignment paths generated for each residue pair are accumulated in another matrix of the same dimensions. The optimal path is then traced through this matrix and is equivalent to a consensus alignment. This alignment is then improved further by considering 20 of the highest scoring pairs in this matrix and repeating the process. This last stage may be repeated. The main advantage of this method, in addition to its speed, is that (unlike some other methods) no previous equivalences between the structures are required.

All pairs of proteins in our data set were superimposed based on the alignment obtained from SSAP using the method of Rippmann and Taylor (1991). In summary, this method performs a weighted superposition of the two proteins using the least-squares method of McLachlan (1979). The weights are obtained from the comparison scores generated by SSAP, which give a measure of residue similarity based on the local structural environment between the two proteins.

Residue-by-residue comparisons

The following comparisons, except those for insertions and deletions, are made for all residues that are structurally equivalent, as defined by a non-zero SSAP score. This criterion gives >90% of residues for comparison even at the lowest sequence identity, which is considerably larger than in the previous comparisons. However, this is more appropriate when the problem of assessing structural divergence and modeling is being addressed. Throughout the following discussions, the results at the residue level are considered in two groups where applicable. One group is made up of identical residue pairs, and the other consists of all pairs of residues. However, it is important to remember that these data are not equally numerous for identical and nonidentical residues, because those that have a higher percent identity will obviously have more identical residues, and vice versa at the other end of the scale. Consideration is also given to the solvation state of the side-chain atoms. Residues whose side chains are <15% accessible are taken as buried, resulting in an average of 30% of all side chains considered to be buried for structures that are <20% identical sequentially. The 15% cutoff was chosen as the value that separates the large peak associated with buried residues from those that are accessible (see Fig. 2A). We find very little difference between the results using the 15% cutoff or any other value below 20% residue accessibility.

Higher resolution data allow for more confident placement of main chains and side chains into the observed electron density. For the purpose of this investigation, we consider primarily those pairs of structures whose resolutions are 2.0 Å or better; these points are identified in the figures by filled symbols. All lines are fitted to these pairs only, although the remaining pairs are included for the purpose of comparison. If a related protein structure is available, molecular replacement techniques may have been used to identify the location and orientation of the protein molecule in the crystal. This protein structure is then used as the model in the early stages of refinement, and the final structure may be biased toward the original model. As far as possible we have tried to identify pairs of proteins that have been solved by molecular replacement.

The pairs of proteins are considered to belong to one of three classes of protein structures: predominantly α, predominantly β, and the remaining proteins, which we group as $\alpha\beta$. The term "predominantly" is used to mean that no more that 15% of the other secondary structure occurs (Taylor & Thornton, 1984). Where applicable, these classes have been represented in the figures by different symbols to denote properties that may influence the results.

Calculation of characteristics

The RMS deviations of equivalent Cα coordinates were calculated from the superpositions. Values were calculated for all residues and separately for residues that were <15% accessible.

Monomeric side-chain accessibilities were calculated by the method of Lee and Richards (1971) using a 1.4-Å probe radius and expressed as the percentage of the accessible surface area compared with that of the same residue in the extended tripeptide Ala-residue-Ala. The absolute difference in accessibility was averaged over all equivalent residues. A similar descriptor of solvation used commonly is the Ooi number (Nishikawa & Ooi, 1986). Unlike accessibility, this is very simple and quick to compute. The Ooi number of a residue is simply the number of Cα atoms within a given radius. This radius is usually 14 Å, although 8 Å is sometimes used. In an manner identical to that of accessibility, the average difference in Ooi values is computed. For both accessibilities and Ooi numbers, the variations within α, β, and coil regions are calculated.

Secondary structure was assigned by the method of Kabsch and Sander (1983). The conservation of α-helix, β-strand, or coil state was considered for each residue. Residues in the 3_{10}-helical state were grouped with other helical residues. Residue secondary structure was counted as conserved if the category (α, β, or coil) did not change.

Several studies on proteins have identified the most favored conformations of side chains (Janin et al., 1978; Bhat et al., 1979; McGregor et al., 1987). Due to the wealth of structurally equivalent residues, discussion is limited to residues whose torsion angles are strictly comparable (for χ^1 these are Leu, Phe, Met, Trp, Cys, Ser, Asn, Gln, Tyr, His, Asp, Glu, Lys, and Arg, and for χ^2 these are Met, Gln, Glu, Lys, and Arg). Any residues that are branched at the Cβ or Cγ side-chain atoms are excluded. The side-chain conformations of these residues in well-ordered structures cluster closely around three minima: $g+ = -60°$, $g- = 60°$, and $t = 180°$, corresponding to staggering of substituents (Morris et al., 1992). The $\Delta\chi$ between equivalent residues is calculated and averaged over all structurally equivalent residues in each protein pair. In addition, conservation of the conformation of the side chains is considered in terms of the minima in which they fall. If the angles of the structurally equivalent residues fell within 60° of the same well, they were considered to agree. Such a generous criterion allows for some flexibility with the less refined structures. The χ^2 side-chain angles were considered only for those equivalent residues whose χ^1 angles agree. The effect of accessibility on the conservation of side-chain angle was also considered. To determine exactly to what extent the molecular replacement model influences the results, pairs solved using this method were, as far as possible, identified from the literature.

Insertions and deletions

The numbers of indels were calculated from the alignments, and these were used to calculate the length and frequency of occurrence with respect to sequence identity. Because the length and conformation of loops are often very different for proteins with weak sequence identity, SSAP often fragments the alignments at this point. For the purpose of investigation of indels, this ambiguity was overcome by considering fragmented alignments in loops as one indel if the length of sequence between two gaps was less than four residues. This value was based on visual inspection of the sequence alignments obtain by SSAP.

Acknowledgments

We thank D. Jones, C. Mason, M. Swindells, and S. Hubbard for helpful discussions. This work was funded by the Science and Engineering Research Council.

References

Artymuik, P.J., Grindley, H.M., Park, J.E., Rice, D.W., & Willet, P. (1992). Three-dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph-theoretical techniques. *FEBS Lett. 303*, 48–52.

Benner, S.A. & Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary structure: A prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul. 31*, 121–181.

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

Bhat, T.N., Sasisekharan, V., & Vijayan, M. (1979). An analysis of side-chain conformation in proteins. *Int. J. Pept. Protein Res. 13*, 170–184.

Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L., & Sutcliffe, M. (1988). Knowledge-based protein modelling and design. *Eur. J. Biochem. 172*, 513–520.

Bowie, J.U., Clarke, N.D., Pabo, C.O., & Sauer, R.T. (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.

Bowie, J.U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three dimensional structure. *Science 253*, 164–170.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., & Hill, R.L. (1969). A possible three dimensional structure of bovine α lactalbumin based on that of hen egg white lysozyme. *J. Mol. Biol. 42*, 65–86.

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823–826.

Chothia, C. & Lesk, A.M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol. 52*, 399–405.

Faber, G.K. & Petsko, G.A. (1990). The evolution of α/β barrel enzymes. *Trends Biochem. Sci. 15*, 228–234.

Finkelstein, A.V. & Reva, B. (1991). A search for the most stable folds of protein chains. *Nature 351*, 497–499.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J. Mol. Biol. 153*, 1027–1042.

Hennig, M., Schlesier, B., Dauter, Z., Pfeffer, S., Betzel, C., Höhne, W.E., & Wilson, K.E. (1992). A TIM barrel protein without enzymatic activity? *FEBS Lett. 306*, 80–84.

Hubbard, T.P.J. & Blundell, T.L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng. 1*, 159–171.

Janin, J., Wodak, S., Levitt, M., & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol. 125*, 357–386.

Jones, D.T., Taylor, W.R., & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature 358*, 86–89.

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

Lee, B.K. & Richards, F.M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol. 55*, 379–400.

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

McGregor, M.J., Islam, S.A., & Sternberg, M.J.E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol. 198*, 295–310.

McLachlan, A.D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol. 128*, 49–79.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., & Thornton, J.M. (1992). Stereochemical quality of protein structure coordinates. *Proteins Struct. Funct. Genet. 12*, 345–364.

Nishikawa, K. & Ooi, T. (1986). Radial locations of amino acid residues in a globular protein: Correlation with the sequence. *J. Biochem. 100,* 1043–1047.

Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J., Sussman, J.L., Verschueren, K.H.G., & Goldman, A. (1992). The α/β hydrolase fold. *Protein Eng. 5,* 197–211.

Orengo, C.A. (1992). A review of methods for protein structure comparison. *Springer Ser. Biophys. 7,* 159–188.

Orengo, C.A., Brown, N.P., & Taylor, W.R. (1992). Fast structure alignment for protein databank searching. *Proteins Struct. Funct. Genet. 14,* 139–167.

Orengo, C.A. & Taylor, W.R. (1990). A rapid method of protein structure alignment. *J. Theor. Biol. 147,* 517–551.

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

Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol. 224,* 461–471.

Rippmann, F. & Taylor, W.R. (1991). Visualization of structural similarity in proteins. *J. Mol. Graph. 9,* 169–174.

Rose, J. & Eisenmenger, F. (1991). A fast unbiased comparison of protein structure by means of the Needleman-Wunsch algorithm. *J. Mol. Evol. 32,* 340–354.

Rost, B., Schneider, R., & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem. Sci. 18,* 120–123.

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

Šali, A., Overington, J.P., & Blundell, T.L. (1990). Knowledge-based protein modelling. *Trends Biochem. Sci. 15,* 235–240.

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

Summers, N.L., Carlson, W.D., & Karplus, M. (1987). Analysis of side-chain orientations in homologous proteins. *J. Mol. Biol. 196,* 175–198.

Sutcliffe, M.J., Haneef, I., Carney, D., & Blundell, T.L. (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.

Sutcliffe, M.J., Hayes, F.R.F., & Blundell, T.L. (1987). Knowledge based modelling of homologous proteins, part II: Rules for the conformations of substituted side-chains. *Protein Eng. 1,* 385–392.

Taylor, W.R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Protein Eng. 4,* 853–870.

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

Taylor, W.R. & Orengo, C.A. (1989b). A holistic approach to protein structure alignment. *Protein Eng. 2,* 505–519.

Taylor, W.R. & Thornton, J.M. (1984). Recognition of super-secondary structure in proteins. *J. Mol. Biol. 173,* 487–514.