The EMBO Journal vol.5 no.4 pp.823–826, 1986

The relation between the divergence of sequence and structure in proteins

Cyrus Chothia[1] and Arthur M.Lesk[2]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, and [1]Christopher Ingold Laboratory, University College London, 20 Gordon Street, London WC1H 0AJ, UK

[2]Permanent address: Fairleigh Dickinson University, Teaneck-Hackensack Campus, Teaneck, NJ 07666, USA

Communicated by M.F.Perutz

Homologous proteins have regions which retain the same general fold and regions where the folds differ. For pairs of distantly related proteins (residue identity ~20%), the regions with the same fold may comprise less than half of each molecule. The regions with the same general fold differ in structure by amounts that increase as the amino acid sequences diverge. The root mean square deviation in the positions of the main chain atoms, Δ, is related to the fraction of mutated residues, H, by the expression: $\Delta(\text{Å}) = 0.40\ e^{1.87H}$.

Key words: evolution/protein homology/model building

Introduction

The comparative analysis of the structures of related proteins can reveal the effects of the amino acid sequence changes that have occurred during evolution (Perutz *et al.*, 1965). Previous work on individual protein families has shown that mutations, insertions and deletions produce changes in three-dimensional structure (Almassy and Dickerson, 1978; Lesk and Chothia, 1980, 1982, 1986; Greer, 1981; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984). Here we report a systematic comparison of structures from eight different protein families. This shows that the extent of the structural changes is directly related to the extent of the sequence changes.

In the work reported here we used the atomic coordinates of 25 proteins (Table I). All these structures have been determined at high resolution (1.4–2.0Å) and refined. The errors in their co-ordinates are 0.15–0.20Å (see references given in Table I). The 25 proteins represent eight different protein families and provide 32 pairs of homologous structures.

Methods and Results

The conserved structural cores and the variable regions of homologous proteins

The structures of homologous proteins can be divided into those regions in which the general fold of the polypeptide chains is very similar and those where it is quite different. In comparing protein structures it is useful to separate the parts that have similar folds from those where the folds differ. We did this using the following quantitative procedure: (i) the main-chain atoms of major elements of secondary structure — helices or two adjacent strands of β-sheet — were individually superposed; and (ii) each superposition was then extended to include additional atoms at both ends. The extension was continued as long as the deviations in the positions of the atoms in the last residue included were no greater than 3 Å. This procedure defined the segments that

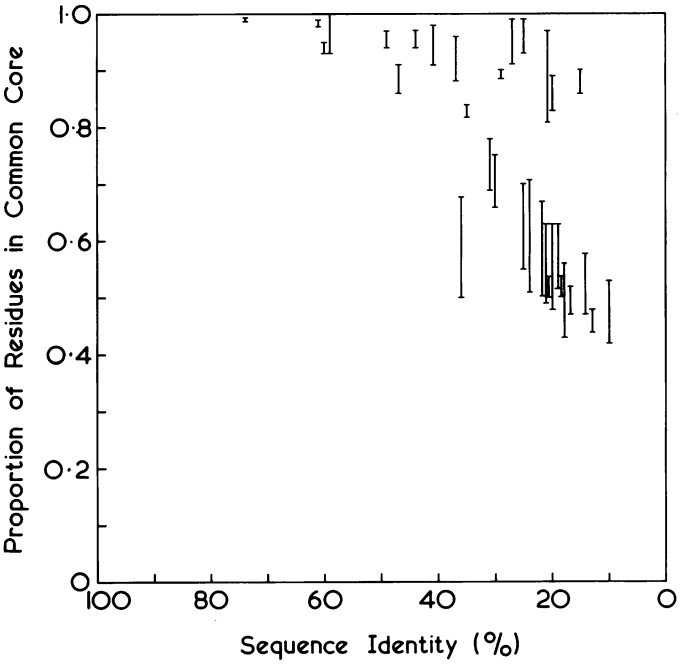

Fig. 1. Size of common cores as a function of protein homology. If two proteins of length n_1 and n_2 have c residues in the common core, the fractions of each sequence in the common core are c/n_1 and c/n_2. We plot these values, connected by a bar, against the residue identity of the core (see Table II).

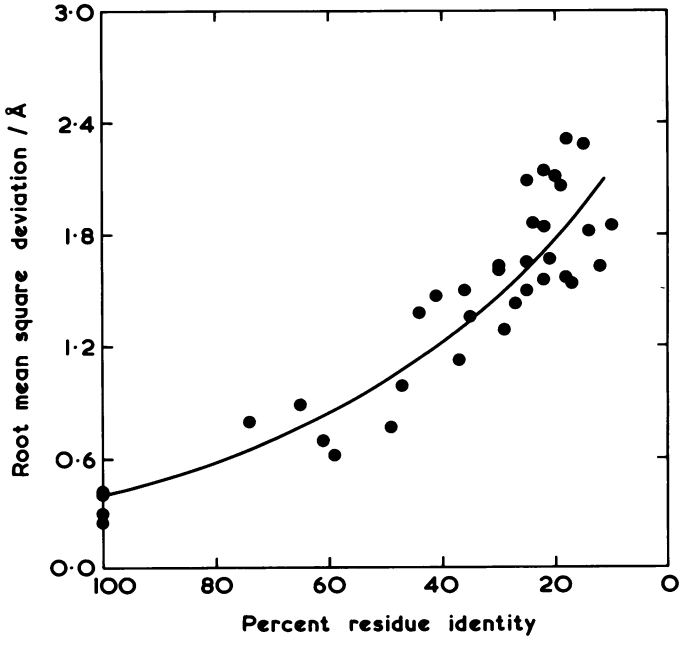

Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

Table I. Homologous proteins determined at high resolution

Family	Protein	Abbreviation	Structure analysis		Reference
			Resolution (Å)	R factor, %	
Globins (deoxy)	Human α subunit	HHBα	1.74	16	Fermi *et al.*, 1984
	Human β subunit	HHBβ			
	Sperm whale myoglobin	1MBD	1.40	14	Phillips, 1980
	Erythrocruorin	1ECD	1.40	18	Steigemann and Weber, 1979
Cytochromes	Tuna c	3CYT	1.50	17	Takano and Dickerson, 1982
	Rice embryo c	1CCR	1.50	19	Ochi *et al.*, 1983
	Bacterial c_2	3C2C	1.68	17	Bhatia, 1981
	Bacterial c_{551}	351C	1.60	19	Matsuura *et al.*, 1982
Serine protease	Bovine γ-chymotrypsin	2GCH	1.90	18	Cohen *et al.*, 1981
	Bovine trypsin	3PTP	1.50	16	Chambers and Stroud, 1979
	S. griseus protease A	2SGA	1.80	14	Sielecki *et al.*, 1979
	S. griseus protease B	3SGB	1.80	14	Read *et al.*, 1983
Dihydrofolate reductase	*L. casei*	3DFR	1.70	15	Bolin *et al.*, 1982
	E. coli	4DFR	1.70	17	Bolin *et al.*, 1982
Cu-electron transport proteins	Bacterial azurin	1AZA	2.00	19	Norris *et al.*, 1983
	Poplar leaf plastocyanin	1PCY	1.60	17	Guss and Freeman, 1983
Sulphydryl protease	Papaya papain	PAP	1.65	16	Kamphuis *et al.*, 1985
	Kiwifruit actinidin	2ACT	1.70	17	Baker, 1980
Lysozyme	Human	1LZ1	1.50	18	Artymiuk and Blake, 1981
	Hen egg white	LZHE	1.60		Grace, 1979
Immunoglobulin domains	Vλ (RHE)	2RHE	1.60	15	Furey *et al.*, 1983
	Vλ (KOL)	KLVL	1.90	19	Marquart *et al.*, 1980
	Vγ (KOL)	KLVH			
	Cλ (KOL)	KLCL			
	$C\gamma_1$ (KOL)	KLCH			

Except for hen egg lysozyme and papain, atomic coordinates were obtained from the protein data bank (Bernstein *et al.*, 1977).

have the same fold in both proteins. They include major elements of secondary structure and peptides that form the active site. We call the collection of such regions the 'common core'. The residues outside the common core are in peripheral elements of secondary structure, in the loops between major elements of secondary structure, at the ends of helices, or in strands at the edges of β-sheets (Lesk and Chothia, 1980, 1982; Greer, 1981; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984).

The results of comparing the 32 pairs of homologous proteins are given in Table II. Pairs whose sequence identity is >50% have 90% or more of the residues of the individual structures within the common cores. Pairs whose residue identity drops to about 20% have common cores that contain betwen 42% and 98% of the residues of individual structures (Table II, Figure 1). Proteins built of β-sheets are at the bottom of this range and proteins built of α-helices are at the top. Compared with helical proteins, β-sheet proteins contain proportionally fewer residues within secondary structures and more in loops, the regions particularly susceptible to local refolding when sequences change.

Structural divergence in the common cores of homologous proteins

Although the core regions retain a common fold, they do undergo structural change as their sequences diverge. Mutations at the interfaces between secondary structures produce changes in the geometry of packing and, in the case of β-sheets, limited local changes in backbone conformation (Lesk and Chothia, 1980, 1982, 1986; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984). The overall extent of the structural divergence of two homologous proteins can be measured by optimally superposing the common cores and calculating the root mean square difference in the positions of their main-chain atoms, Δ. For the 32 homologous pairs of proteins in Table II the values of Δ vary between 0.62 and 2.31 Å (Table II).

The exact value of Δ is, of course, dependent upon the procedure used to define the common cores of homologous proteins. Inspection of the regions not in the common cores shows that they usually have very different conformations. This is especially true of the larger loops. Thus modification of the procedure used here to define the common cores would only produce marginal differences.

Essentially similar results are obtained if, in place of a core derived for each individual homologous pair, we use a core common to all members of a family. For example, in the cytochromes c(rice), c(tuna), c_2 and c_{551}, a 48-residue core is common to all four structures (Chothia and Lesk, 1984). Superpositions of this core in the four structures give the Δ values listed in Table III. Compared with the Δ values for individual core comparisons, these Δ values are somewhat smaller for closely related pairs (in these cases the family core is smaller and more homologous than the pair core), but nearly equal for distantly related pairs (Table III).

The contribution to Δ from experimental error and from differences in molecular environment can be estimated from the comparison of proteins whose structures have been accurately determined in different crystal forms, or in crystals that have more than one molecule in the asymmetric unit. The values of Δ for five such proteins are between 0.25 and 0.40 Å (Table II). The mean is 0.33 Å: one half to one seventh of the Δ values reported here for homologous proteins.

Table II. Common cores of homologous proteins: size, fit and residue identity

Family	Protein pair[a]	Residues in protein pair	Residues in core	r.m.s. difference in core (Å)	Percentage of core residues that are the same in both structures
Globin	HHBα:HHBβ	141:146	137	1.38	44
	HHBα:1MBD	141:153	139	1.43	27
	HHBα:1ECD	151:136	122	2.28	15
	HHBβ:1MBD	146:153	143	1.50	25
	HHBβ:1ECD	146:136	121	2.11	20
	1MBD:1ECD	153:136	132	1.67	21
Cytochrome c	3CYT:1CCR	103:111	103	0.62	59
	3CYT:3C2C	103:112	99	1.13	37
	3CYT:351C	103:82	57	1;65	25
	1CCR:3C2C	111:112	101	1.47	41
	1CCR:351C	111:82	58	1.86	24
	3C2C:351C	112:82	56	1.50	36
Serine protease	2GCH:3PTP	236:222	203	0.99	47
	2GCH:2SGA	236:181	114	2.09	25
	2GCH:3SGB	236:185	116	2.14	22
	3PTP:2SGA	221:181	112	1.84	22
	3PTP:3SGB	222:185	116	2.06	19
	2SGA:3SGB	181:185	172	0.89	65
Immunoglobulin domain	2RHE:KLVL	110:110	108	0.80	74
	2RHE:KLVH	110:125	83	1.63	30
	2RHE:KLCL	110:101	55	1.57	18
	2RHE:KLCH	110:99	48	1.47	13
	KLVL:KLVH	110:125	86	1.61	30
	KLVL:KLCL	110:101	55	1.56	22
	KLVL:KLCH	110:99	52	1.54	17
	KLVH:KLCL	110:101	59	1.82	14
	KLVH:KLCH	110:99	52	1.85	10
	KLCL:KLCH	101:99	83	1.36	35
Dihydrofolate reductase	3DFR:4DFR	159:161	143	1.29	29
Lysozyme	1LZ1:LZHE	130:129	128	0.70	61
Plastocyanin/azurin	1PCY:1AZA	99:129	55	2.31	18
Papain/actinidin	PAP:2ACT	212:218	206	0.77	49

Proteins whose structure has been determined in different environments

				Reference
Trypsin inhibitor	58:58	56	0.40	Wlodawer *et al.*, 1984
Tuna cytochrome c	103:103	103	0.30	Takano and Dickerson, 1981
Azurin	129:129	127	0.37	Norris *et al.*, 1983
Rat protease	224:224	224	0.25	Anderson *et al.*, 1978
Deoxy human haemoglobin	287:287	287	0.30	Fermi *et al.*, 1984

[a]See Table I for abbreviations.

The relationship between the divergence of sequence and structure in the common cores of homologous proteins

The divergence of structure as measured by Δ is a simple function of the fractional sequence identity of the cores (Figure 2). A least squares fit to the data in Table II gives the relationship:

$$\Delta = 0.40\ e^{1.87H}$$

where Δ is measured in Å and H is the fraction of mutated residues. For the 32 pairs of homologous structures in Table II, the values of Δ predicted by this equation are within 20% of the observed values for 23 pairs and within 28% for the other nine.

The exponential form of the relationship arises because proteins accept mutations of surface residues more readily than mutations of buried residues. Closely related proteins differ primarily in surface residues, whereas distantly related proteins differ in both surface and buried residues (Table IV). The mutation of residues buried in the interior usually produces larger structural changes than the mutation of surface residues. Thus the tendency for changes in buried residues to lag behind surface changes results in an exponential relationship between sequential and structural change.

Conclusions

In a previous series of papers we have described the structural differences found in members of individual protein families (Lesk and Chothia, 1980, 1982. Chothia and Lesk, 1982, 1984). The differences in the common cores consist mainly of changes in the relative position and orientation of packed secondary structures and, in the case of β-sheets, some local changes in structure. We have shown here that the overall extent of these changes is directly related to the extent of the sequence differences.

These results imply that the degree of success to be expected in predicting the structure of a protein from its sequence using the known structure of an homologous protein, depends upon the extent of the sequence identity (Lesk and Chothia, 1986). A protein structure will provide a close general model for other proteins with which its sequence homology is >50%. If the homology drops to 20% there will be large structural differences that are at present impossible to predict.

However, the active sites of distantly related proteins can have very similar geometries (Lesk and Chothia, 1980; Chothia and Lesk, 1982; Read *et al.*, 1984). This is because of the coupling of the structural changes that has occurred during evolution (Lesk and Chothia, 1980). Thus the structure of the active site in a protein may provide a good model for those in related proteins even if the overall sequence homologies are low.

Table III. Cytochrome c family. Root mean square difference in the position of main chain atoms of residues in the conserved structural core, Δ

Protein pair[a]	Core determined for individual homologous pairs			Core common to four cytochrome c structures		
	Core size	Δ (Å)	Residue identity in core (%)	Core size	Δ (Å)	Residue identity in core (%)
3CYT:1CCR	103	0.62	59	48	0.38	65
3CYT:3C2C	99	1.13	37	48	0.91	48
1CCR:3C2C	101	1.47	41	48	1.01	56
3C2C:351C	56	1.50	36	48	1.39	35
3CYT:351C	57	1.65	25	48	1.56	31
1CCR:351C	58	1.86	24	48	1.66	27

[a]See Table I for abbreviations.

Table IV. The homology of buried and surface residues

Protein pair	Residue identity (%)		
	Buried residues[a]	Surface residues[a]	Overall
S. griseus proteases A and B	83	52	65
Human and hen egg white lysozyme	77	52	61
Tuna and rice embryo cytochrome c	77	50	59
Human haemoglobin α and *Chironomus* erythrocuorin	21	16	18
IgG Kol domains Vλ and $C\gamma_1$	31	11	17

[a]Buried residues are those with accessible surface areas ≤20 Å².

Acknowledgements

We thank Professor Sir David Phillips for the atomic co-ordinates of hen egg lysozyme, Professor J.Drenth for the atomic co-ordinates of papain, John Cresswell for the figure drawings and The Royal Society, National Science Foundation (PCM83-20171) and the National Institute of General Medical Science (GM25435) for support.

References

Almassy,R.J. and Dickerson,R.E. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 2674–2678.
Anderson,W.F., Matthews,B.W. and Woodbury,R.G. (1978) *Biochemistry*, **17**, 819.
Artymiuk,P.J. and Blake,C.C.F. (1981) *J. Mol. Biol.*, **152**, 737–762.
Baker,E.N. (1980) *J. Mol. Biol.*, **141**, 441–484.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bhatia,G.E. (1981) Ph.D. Thesis, University of California at San Deigo.
Bolin,J.T., Filman,D.J., Matthews,D.A., Hamlin,R.C. and Kraut,J. (1982) *J. Biol. Chem.*, **257**, 13650–13662.
Chambers,J.L. and Stroud,R.M. (1979) *Acta Crystallogr.*, **35B**, 1861–1874.
Chothia,C. and Lesk,A.M. (1982) *J. Mol. Biol.*, **160**, 309–323.
Chothia,C. and Lesk,A.M. (1984) *J. Mol. Biol.*, **182**, 151–158.
Cohen,G.H., Silverton,E.W. and Davies,D.R. (1981) *J. Mol. Biol.*, **148**, 449–479.
Fermi,G., Perutz,M.F., Shaanan,B. and Fourme,R. (1984) *J. Mol. Biol.*, **175**, 159–174.
Furey,W., Wang,B.C., Yoo,C.S. and Sax,M. (1983) *J. Mol. Biol.*, **167**, 661–692.
Grace,D.E.P. (1979) D.Phil. Thesis, Oxford University.
Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027–1042.
Guss,J.M. and Freeman,H.C. (1983) *J. Mol. Biol.*, **169**, 521–562.
Kamphuis,I.G., Drenth,J. and Baker,E.N. (1985) *J. Mol. Biol.*, **182**, 317–329.
Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–342.
Lesk,A.M. and Chothia,C. (1986) *Philos. Trans. R. Soc. Lond.*, **317**, 345–356.
Marquart,M., Deisenhofer,J., Huber,R. and Palm,W. (1980) *J. Mol. Biol.*, **141**, 369–391.
Matsuura,Y., Takano,T. and Dickerson,R.E. (1982) *J. Mol. Biol.*, **156**, 389–409.
Norris,G.E., Anderson,B.F. and Baker,E.N. (1983) *J. Mol. Biol.*, **165**, 501–521.
Ochi,H., Hata,Y., Tanaka,N., Kakudo,M., Sakurai,T., Aihara,S. and Morita,Y. (1983) *J. Mol. Biol.*, **166**, 407–418.
Perutz,M.F., Kendrew,J.C. and Watson,H.C. (1965) *J. Mol. Biol.*, **13**, 669–678.
Phillips,S.E.V. (1980) *J. Mol. Biol.*, **142**, 531–554.
Read,R.J., Fujinaga,M., Sielecki,A.R. and James,M.N.G. (1983) *Biochemistry*, **22**, 4420–4433.
Read,R.J., Brayer,G.D., Jurāsĕk,L. and James,M.N.G. (1984) *Biochemistry*, **23**, 6570–6575.
Sielecki,A.R., Hendrickson,W.A., Broughton,C.G., Delbaere,L.T.J., Brayer, G.D. and James,M.N.G. (1979) *J. Mol. Biol.*, **134**, 781–804.
Steigemann,W. and Weber,E. (1979) *J. Mol. Biol.*, **127**, 309–388.
Takano,T. and Dickerson,R.E. (1981) *J. Mol. Biol.*, **153**, 95–115.
Wlodawer,A., Walter,J., Huber,R. and Sjölin,L. (1984) *J. Mol. Biol.*, **180**, 301–329.

Received on 27 January 1986