doi:10.1016/S0022-2836(02)00653-8 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **322**, 205–233

JMB

Sequence Conservation in Families Whose Members Have Little or No Sequence Similarity: The Four-helical Cytokines and Cytochromes

Emma E. Hill*, Veronica Morea and Cyrus Chothia

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

**Corresponding author*

Proteins for which there are good structural, functional and genetic similarities that imply a common evolutionary origin, can have sequences whose similarities are low or undetectable by conventional sequence comparison procedures. Do these proteins have sequence conservation beyond the simple conservation of hydrophobic and hydrophilic character at specific sites and if they do what is its nature? To answer these questions we have analysed the structures and sequences of two superfamilies: the four-helical cytokines and cytochromes c'-b_{562}. Members of these superfamilies have sequence similarities that are either very low or not detectable.

The cytokine superfamily has within it a long chain family and a short chain family. The sequences of known representative structures of the two families were aligned using structural information. From these alignments we identified the regions that conserve the same main-chain conformation: the common core (CC). For members of the same family, the CC comprises some 50% of the individual structures; for the combination of both families it is 30%. We added homologous sequences to the structural alignment. Analysis of the residues occurring at sites within the CCs showed that 30% have little or no conservation, whereas about 40% conserve the polar/neutral or hydrophobic/neutral character of their residues. The remaining 30% conserve hydrophobic residues with strong or medium limitations on their volume variations. Almost all of these residues are found at sites that form the "buried spine" of each helix (at sites i, $i+3$, $i+7$, $i+10$, etc., or i, $i+4$, $i+7$, $i+11$, etc.) and they pack together at the centre of each structure to give a pattern of residue–residue contacts that is almost absolutely conserved. These CC conserved hydrophobic residues form only 10–15% of all the residues in the individual structures.

A similar analysis of the cytochromes c'-b_{562}, which bind haem and have a very different function to that of the cytokines, gave very similar results. Again some 30% of the CC residues have hydrophobic residues with strong or medium conservation. Most of these form the buried spine of each helix and play the same role as those in the cytokines. The others, and some spine residues bind the haem co-factor.

Keywords: four-helix bundle; helix packing; CAFASP; remote homology

Present addresses: E. E. Hill, U.C. Berkeley, Brenner Lab, 561 Koshland Hall, CA 94702-3102, USA; V. Morea, CNR Center of Studies of Molecular Biology c/o Dept. of Biochemical Sciences "A. Rossi-Fanelli", University of Rome "La Sapienza", P.le A. Moro 5, 00185-Rome, Italy.

Abbreviations used: CC, common core; ASA, accessible surface area; PDB, Protein Data Bank; RMSD, root mean square deviation; HMM, hidden Markov model; G-CSF, granulocyte colony-stimulating factor; IL, interleukin; LIF, leukemia inhibitory factor; GH, growth hormone; PL, placental lactogen; CNTF, ciliary neurotrophic factor; OSM, oncostatin M; EPO, erythropoietin; GM-CSF, granulocyte-macrophage colony-stimulating factor; M-CSF, macrophage colony-stimulating factor; SCF, stem cell factor; ECCB, *E. coli* cytochrome b562; RMCP, *R. molischianum* cytochrome c'; CVCP, *C. vinosum* cytochrome c'; RCCP, *R. capsulatus* I cytochrome c'; RPCP, *R. palustris* cytochrome c'; AXCP, *Alcaligenes* sp. cytochrome c'.

E-mail address of the corresponding author: emma@compbio.berkeley.edu

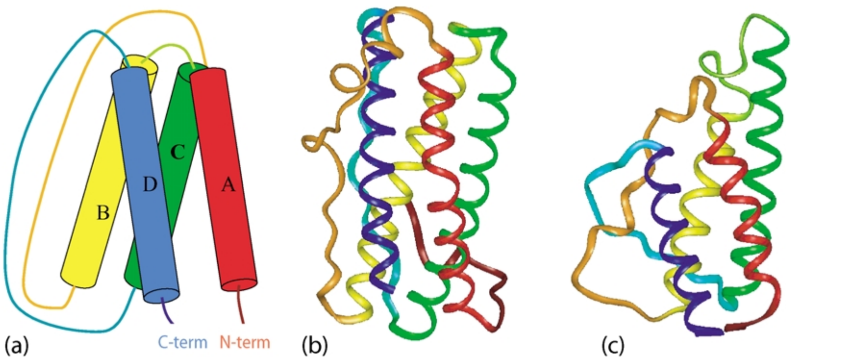


Figure 1. Cartoon and ribbon representations of four-helical cytokines. In each representation, the structure starts at the N-terminal end with helix A that is coloured red, the long crossover between helices A and B is orange, helix B is yellow, helix C is green, the long crossover between helices C and D is coloured turquoise and helix D is blue. (a) A cartoon representation of a four-helical cytokine. (b) Leukaemia inhibitory factor (PDB-ID: 1lki), a long chain cytokine. Long chain cytokines are on average 180 amino acid residues long and have helices of between 20 and 30 amino acid residues in length. They sometimes have small additional helices in the long A–B helix crossover which passes in front of helix D. (c) Interleukin-4 (PDB-ID: 1rcb) a short chain cytokine. Short chain cytokines are on average 140 amino acid residues long with helices of between 10 and 20 amino acid residues in length. They have two small beta strands, one leading directly into helix D and the other in the long crossover between helices A and B which passes behind helix D.

Introduction

Protein structures are more conserved than their sequences.[1–4] Indeed there are cases of proteins for which we have convincing structural, functional and genetic evidence of a common evolutionary origin but for which conventional sequence comparison methods cannot detect significant sequence similarities. Is this because, in spite of their common origin, they have evolved to a point where there is no significant sequence similarity, except for those residues necessary to maintain their hydrophobic interiors and hydrophilic exteriors; or is it because there is sequence conservation in regions of functional and structural importance but, which is hidden by the noise from the regions where the sequence changes have been extensive? If the latter case occurs it raises a second question: which sites are conserved and what is their structural and functional role? Here we provide at least partial answers to these questions by an investigation of the members of two superfamilies: the four-helical cytokines and the cytochromes c'-b_{562}.

For the four-helical cytokines there is good evidence for their evolutionary relationship because, in addition to unique structural similarities, they perform similar functions, bind a subset of homologous receptors and conserve same frame exon–intron boundaries.[5–7] However, with marginal exceptions, their sequences cannot be matched by sequence comparison methods (see Appendix A). The cytochromes c' are more closely related to each other in that their sequences make significant matches to one other but their sequence identities are low, mostly 20–25%, and they do not match cytochrome b_{562}.

Given the difficulty of detecting relationships between these proteins from sequence alone we use here a procedure that makes use of the structures known for members of the two superfamilies. In outline if not in detailed execution, the procedure is straightforward. First an optimal structural alignment of the sequences of known structure is produced. To this alignment we add sequences that are clearly homologous to the sequences of known structure and the alignment of which is unambiguous. We then examine the residue frequencies at each site in the alignment to determine if there is any residue conservation and, if there is, its nature. Last we investigate the role of residues at conserved positions to discover their role in the structure and/or function.

We compare the results found for the four-helical cytokines with those found for the cytochrome superfamily of proteins. This comparison allows us to see if the results observed are generic to four-helix bundles or if they are superfamily specific.

Note that throughout the paper, the terms “family” and “superfamily” are used according to the SCOP[8] definition and not interchangeably.

The four-helical cytokines

The fold of these proteins is a four-helix bundle. The orientation of the helices in these cytokines differs from the conventional four-helical up-and-down bundle in that they have an up-up-down-down topology with long overhand connections between helices A–B and C–D (Figure 1). This topology is so far unique to these cytokines. The four-helical cytokine superfamily is sub-classified in SCOP[8] into three families, the long and short chain cytokines which we are analysing here, and the interferons/interleukin 10 family. The overall topology of the four-helical cytokines is conserved between the families but there are some differences in secondary structure, chain length and topological details (Figure 1).

Cytokines are soluble secreted proteins that act as chemical messengers important in intercellular communication. They regulate the differentiation, proliferation, activation and death of many cell types, with particular involvement in the

Table 1. The structures used in this analysis from the long and short chain families of the four-helical cytokine and the four-helical up-and-down cytochrome superfamilies

Family	Protein	PDB code	Species	Resolution (Å)	*R*-value (*R*-free)	Reference
A. *Four helical cytokine superfamily*						
Long chain	G-CSF	1bgc	Bovine	2.2	0.21	17
	IL-6	1alu (C)	Human	1.9	0.21 (0.27)	18
	LIF	1lki	Mouse	2.0	0.18	19
	GH	1hgu	Human	2.5	0.21	20
	PL	1f6f [a]	Sheep	2.3	0.22 (0.28)	21
	CNTF	1cnt [1234] (C)	Human	2.4	0.19 (0.24)	22
	Leptin	1ax8 (mut)	Human	2.4	0.18 (0.28)	23
	OSM	1evs [a]	Human	2.2	0.20 (0.26)	24
Short chain	EPO	1eer [a] (mut)	Human	1.9	0.24 (0.31)	25
	GM-CSF	2gmf [ab]	Human	2.4	0.23	26
	IL-4	1rcb	Human	2.25	0.21	27
	IL-5[a]	1hul [ab]	Human	2.4	0.21 (0.36)	28
	Flt3[b]	1ete [abcd] (C)	Human	2.2	0.23 (0.28)	29
	SCF	1scf [abcd] (C)	Human	2.2	0.19 (0.24)	30
	IL-2	3ink[cd](mut)	Human	2.5	0.20	31
B. *Cytochrome superfamily*						
b_{562}	ECCB	256b [ab] (C)	*E. coli*	1.4	0.16	43
c'	RMCP	2ccy [ab] (C)	*R. molischianum*	1.67	0.19	42,44
	CVCP	1bbh [ab] (C)	*C. vinosum*	1.8	0.19	45
	RCCP	1cpq (C)	*R. capsulatus*	1.72	0.15	46
	RPCP	1a7v [ab]	*R. palustris*	2.3	0.19	47
	AXCP	1e85 [a] (C)	*Alcaligenes* sp.	1.35	0.19 (0.22)	48

Letters or numbers shown in square brackets after the PDB code refer to chains solved for that structure. Where more than one chain is available the one used is underlined, with the exception of IL-5 (PDB-ID: 1hul), the intertwined dimer for which helices A–C are contributed from chain a and helix D from chain b, in which case both chains are used. (mut) Structures solved in a mutant form. (C) Structures which have been solved in complex with their receptors.

[a] Forms intertwined dimer.

[b] Forms dimer.

regulation of the circulatory system and production of immunity and inflammatory responses. Each four-helical cytokine is recognised by the extracellular domain of a specific membrane spanning receptor.[9] It is the bound receptor–cytokine complex that is then able to initiate the resulting cellular signalling cascade of events involving janus kinases (JAKs), signal transducers and activators of transcription (STAT) and tyrosine kinases (TYKs).[10,11]

Several analyses have been previously carried out on the four-helical cytokine structures. In 1993, Sprang & Bazan[5] carried out an analysis of the nine structures available involving both long and short chain cytokines. They identified a correspondence between the four helices and also those stretches of the loops containing the β-strands in the short chain family. They observed two completely conserved amino acid residues between the four short chain cytokines and none between the long chain cytokines. Inter-helical contact patterns[12–14] can be a valuable tool for the production of optimal superpositions. Indeed, Denesyuk and co-workers carried out an analysis of inter-helical contacts in interferons-beta and -gamma and in dimeric IL-5 in order to produce optimal superpositions.[15] Another previous analysis by Rozwarski and co-workers of the short chain cytokine family[16] identified a CC of 61 residues within the five structures available at that time.

Here we extend greatly the work of these previous groups. We now have more structures for which we produced structural alignments. We present results for both the long and short chain cytokines separately as well as a merger of these two sets of results. Finally we add sequence homologues to the structural alignments in order to obtain more data on residue conservation within core positions. We also analyse inter-helical contacts so as to understand which positions are structurally important. We identify core regions within these proteins and go on to further define conserved sets of residues within these cores that play the major role in determining the three-dimensional structures of the long and short chain cytokines, both separately and together.

Structures used in this work

Using the classification of these two families in the SCOP database,[8] we selected one representative of each type of cytokine in the long and short chain four-helical cytokine families. The criteria taken into account are mode of solution, resolution, *R*-factor, *R*-free, whether solved in complex or as a mutant and the number and positioning of any disordered residues. Details of the eight long chain and seven short chain cytokine proteins[17–31] selected for analysis are given in Table 1. All structures were obtained from the Brookhaven Protein Data Bank (PDB).[32]

Structural alignments

Pairwise structural alignments

It is important to note that the periodicity of the alpha-helix structure makes it possible to align the main-chain atoms of any given pair of alpha-helices in many alternative ways, all with low root mean square deviation (RMSD) values. In the case of a four-helix bundle, it is possible to produce several alternative alignments by shifting all or some of the helices of one structure by the same number of residues, with respect to the other structure.[33] For this reason, to produce correct structural alignments of the four-helical proteins, we take into account not only RMSD values, but also other structural features that are conserved in homologous protein structures.

Initial structural alignments produced using PINQ[34] (an interactive protein structure analysis program) were refined to obtain an exact positioning of secondary structure elements based on their hydrogen bonds, residue accessible surface areas (ASAs), inter-residue contacts and RMSDs of the superimposed atoms. For the pairwise comparisons, we defined as structurally conserved those regions whose main-chain atoms superimpose with an RMSD of less than 3 Å.

In cases when it was impossible to obtain an initial superposition of whole structures automatically, the structures were fitted helix by helix and then pieced back together to obtain the optimal overall superposition. For every pair of proteins the superposition obtained was visually inspected. We were able to align members of the long chain family to the short chain family and *vice versa* by virtue of a member of each family that has chain and helix length most similar to those of the other family: EPO (PDB-ID: 1eer) a short chain family member and leptin (PDB-ID: 1ax8) in the long chain family.

Details of all the pairwise alignments are shown in Table 2. The length of the core (the region that has a conserved main-chain conformation) and the RMSD values vary between different pairs of proteins, with core lengths between 81 and 119, 61 and 86 and 58 and 97 residues and RMSDs ranging from 1.1–2.8 Å, 1.9–2.9 Å and 1.4–2.8 Å for the long chain, short chain and long and short chains together, respectively. The sequence identities of the core regions range between 7 and 31%, 5 and 24%, and 5 and 17% for the long chain, short chain and both families together, respectively.

Multiple structural alignments

We produced the multiple structural alignments of the long chain, short chain and both families together by merging the pairwise alignments. This required a consistency check of all pairwise alignments, including verification of conservation of ASAs and inter-helical contact data. Production of these multiple structural alignments allowed the definition of the common core (CC) regions shared by the proteins within each family and between the two families. Figure 2 shows the resulting multiple structural alignments of the long chain family, the short chain family and both families together.

Long chain cytokines. Figure 3 shows the multiple sequence alignment obtained for the eight members of the long chain family of known structure. There is a consensus CC of 101 residues, which consists of 69 residues of central core that is structurally conserved in all eight structures and there is a peripheral core consisting of 32 residues from either side of any one region of the central core, that are conserved by seven of the eight structures.

Short chain cytokines. The resulting multiple sequence alignment for the short chain family is shown in Figure 4. We identified a 57 residue central core common to all seven short chain proteins and a peripheral core of nine residues common to six of the seven structures. Thus we identified a consensus CC with a total length of 66 residues. A comparison of our resulting alignment with that produced by Rozwarski and co-workers[16] shows that the only disagreement is a shift of one turn of helix C for one structure (IL-5 PDB-ID: 1hul). They align two cysteines, but we find a better overall agreement of the different structural features (such as hydrogen bond patterns, inter-residue contacts, and ASA and RMSD values) is given by our alignment.

Long and short chain cytokines. The multiple structural alignment for all long and short chain cytokines in this study shows a central core of 38 residues and a peripheral core of ten residues and therefore a consensus CC of a total of 48 residues. The matching regions from the long and short chain alignments together are summarised in Table 3 and shown in Figures 3 and 4. Sprang & Bazan[5] presented alignments for the A and D helices of the long and short chain cytokines. A comparison of their results with our final alignment shows that we agree for the alignments involving only the long or short chain proteins but that we have a different final alignment for the long and short chain proteins together. We believe that the availability of more structures and the fact that we carried out calculations for all helices and exploited several structural features (see above), which allowed us to produce an improved alignment.

Solvent inaccessible positions

The alignments (Figures 3 and 4) indicate the sites in the different proteins that are structurally equivalent. We determined the extent to which these sites are similarly buried or in contact with the solvent by calculating the solvent ASA[35] of each residue in each structure. We defined buried residues as those with an ASA of 20 Å^2 or less. In certain borderline cases, the residues were visually

Table 2. Root mean square deviations (RMSDs) in the common cores of superposed cytokines

A. *Long chain family*

		PDB-ID (chain length in amino acids)						
PDB-ID		1evs (187)	1f6f (197)	1ax8 (143)	1lki (171)	1hgu (188)	1alu (165)	1bgc (184)
1cnt	RMSD	1.5	1.9	1.1	1.6	1.8	1.1	1.5
	#/core	21/116	8/119	10/94	17/105	13/88	12/89	25/102
1bgc		2.0	2.4	1.4	2.4	2.8	1.5	–
		23/97	16/112	13/90	13/101	9/95	18/105	
1alu		1.8	1.5	1.1	2.1	1.8	–	
		16/99	12/110	16/83	14/96	10/89		
1hgu		1.9	1.3	2.0	1.9	–		
		10/86	32/104	7/81	9/84			
1lki		1.7	2.1	2.0	–			
		22/93	11/110	9/87				
1ax8		1.6	1.5	–				
		8/86	11/90					
1f6f		1.8	–					
		14/104						

B. *Short chain family*

		PDB-ID (chain length in amino acids)					
PDB-ID		1scf (147)	2gmf (120)	3ink (128)	1hul (107)	1eer (166)	1ete (137)
1rcb (129)	RMSD	2.4	2.0	2.0	2.3	2.5	2.5
	#/core	10/77	17/71	12/80	11/70	7/86	8/84
1ete		1.8	2.1	2.8	2.4	2.9	–
		9/85	7/61	7/72	8/68	12/78	
1eer		2.8	2.3	2.7	2.6	–	
		11/75	9/63	14/81	9/75		
1hul		2.8	2.2	2.8	–		
		4/78	14/71	11/81			
3ink		2.6	1.9	–			
		13/76	7/71				
2gmf		2.4	–				
		8/66					

C. *Matches between the short chain cytokine EPO (PDB-ID: 1eer), which has helices similar in length to a long chain cytokine, and the long chain cytokines*

PDB-ID		1evs	1f6f	1ax8	1lki	1hgu	1alu	1bgc	1cnt
	RMSD	2.0	1.6	2.2	2.2	1.7	2.3	1.9	1.8
1eer	#/core	14/97	11/73	11/77	6/76	13/81	8/85	8/63	5/91

D. *Matches between the long chain cytokine leptin (PDB-ID: 1ax8), which has helices similar in length to a short chain cytokine, and the short chain cytokines*

PDB-ID		1scf	2gmf	3ink	1hul	1eer	1ete	1rcb
1ax8	RMSD	1.6	2.3	1.4	1.6	1.7	2.5	2.8
	#/core	10/65	7/60	11/63	11/73	11/63	6/58	11/72

RMSD: root mean square deviation in Å between the pair of structures in question for the core main chain atoms. #: number of identical residues within the common core. Core: the number of residues the main-chain atoms of which can be superimposed with an RMSD of less than 3 Å for the pair of structures in question.

inspected within the structures to determine to what extent it was the backbone or side-chain that is solvent accessible. Data on residue ASA values is given in Figures 3 and 4.

In the long chain alignment (Figure 3) there are 26, nine and one equivalent positions at which all eight, seven and six of the structures, respectively, have a buried residue. Similarly, in the short chain alignment (Figure 4) we see 17, seven and three positions at which all seven, six and five of the structures, respectively, have a buried residue. From the alignment of the long and short chain families together we see 12, six and one positions at which all 15, 14 and 13 of the structures, respectively, have a solvent inaccessible residue.

Collection and alignment of sequence homologues

In order to obtain more general data on the conservation of residues present at positions within the CC regions, homologues were collected by matching the sequences of the structures against the non-redundant database NRDB90[36] by running FASTA.[37] Hits were considered to be significant if the E-value was ≤ 0.01. They were filtered to exclude sequences of $>90\%$ or $<40\%$ sequence identity with the query: this ensured removal of any redundancy between matches and the exclusion of distant homologues. We exclude distant homologues, firstly because we cannot be certain

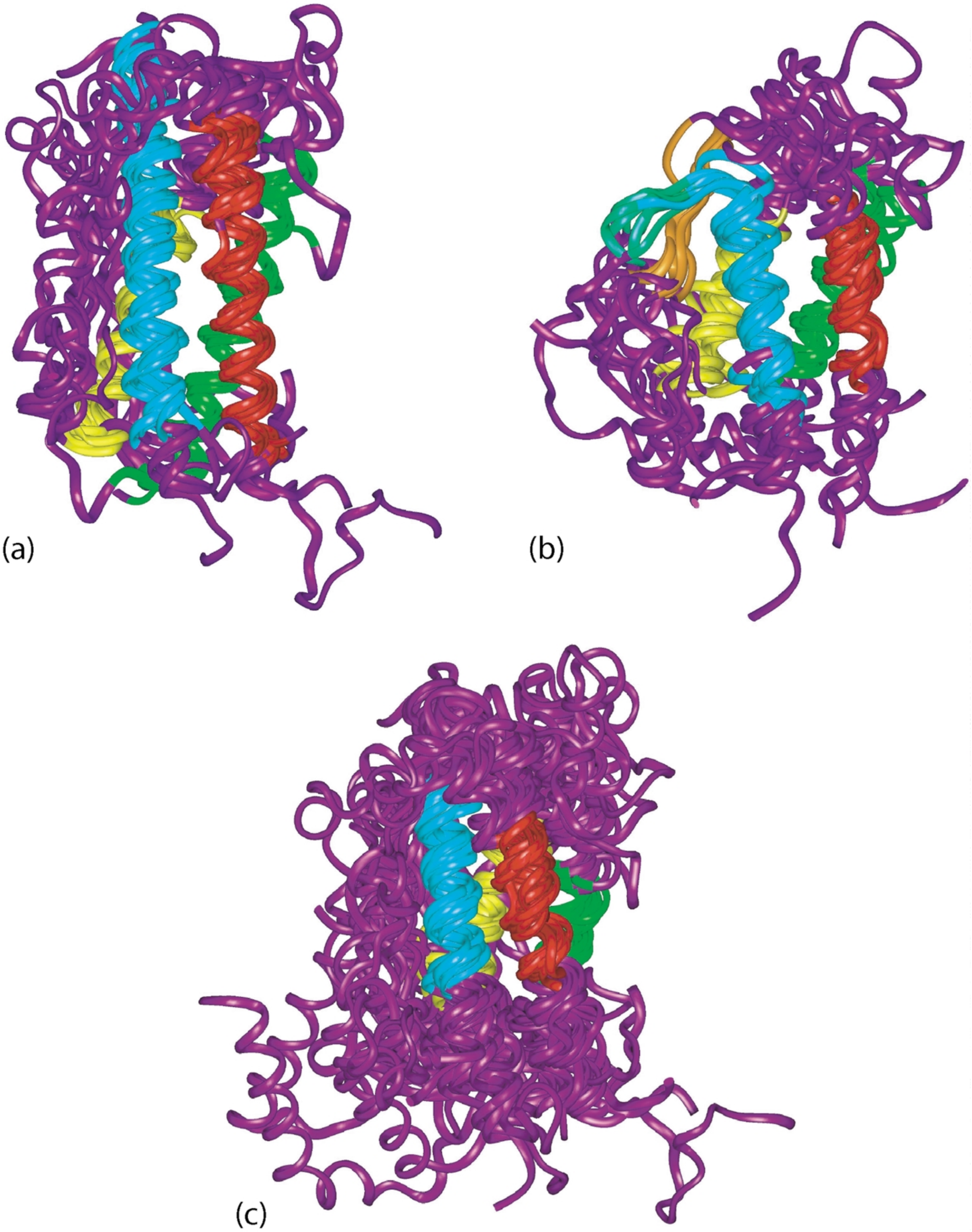


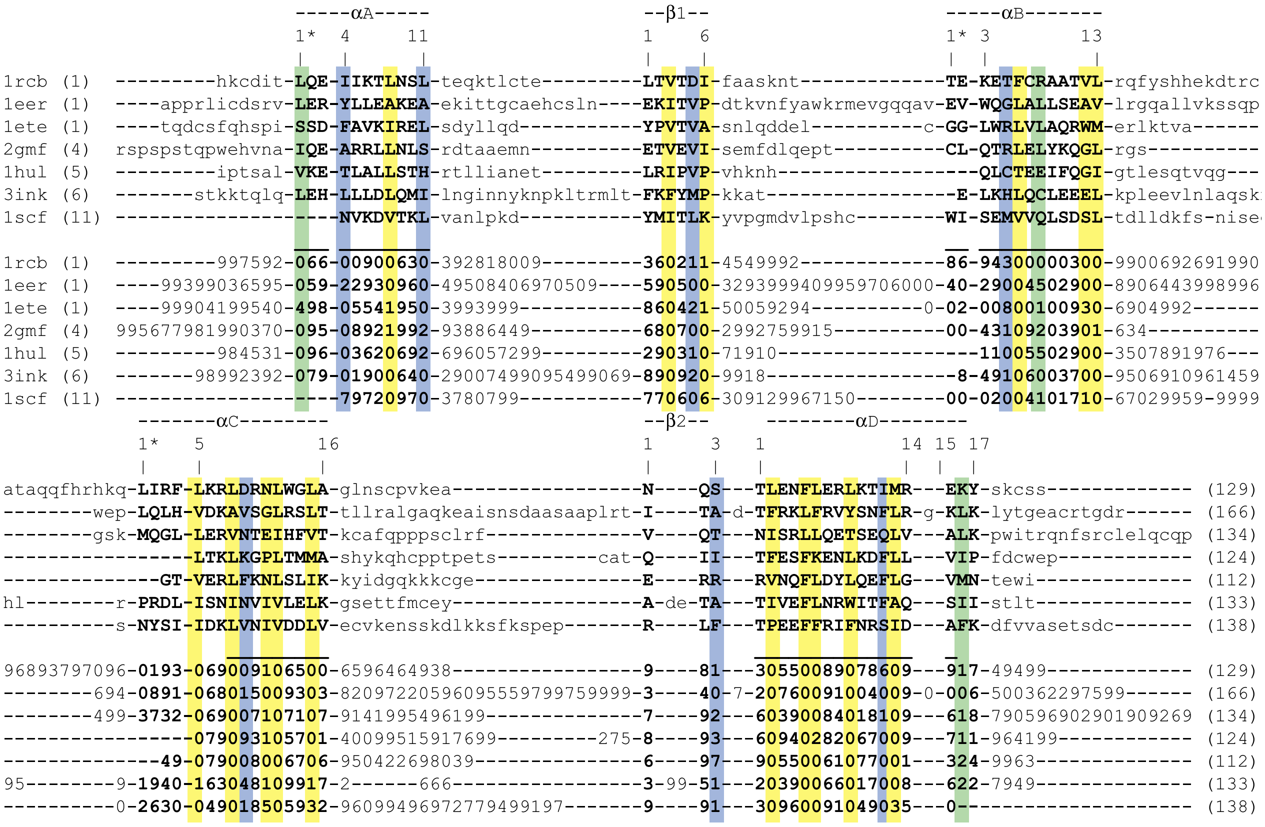
Figure 2. Multiple structure superpositions of the four-helical cytokines. Structural superpositions of (a) the eight long chain cytokines, (b) the seven short chain cytokines and (c) the eight long chain and seven short chain cytokines. Non-core regions are coloured purple. Helices are coloured as in Figure 1. The beta strands of the short chain cytokines in Figure (b) are coloured turquoise and orange.

of the accuracy of their alignment to the sequences of known structure and, secondly, because they may have regions that are not structurally equivalent to the query. Twenty-eight sequences were removed from a total of approximately 325 sequence homologues (mostly sequences of $<40\%$ sequence identity, fragmented sequences, or sequences containing gaps within regions matching

```
                              -------------A---------------                                         ------------B------------
                              1            15  16*       24*                                        1* 4*5            19 20*23*
                              |             |   |         |                                         |  | |             | |  |
1cnt (11) ---------------phrr-DLCSRSIWLARKIRS---DLTA-L-TESY-vkhqglws-----------------------------el-TEAE-RLQENLQAYRTFHVL-LARL
1bgc (9)  ---------------slpq-SFLLKCLEQVRKIQA---DGAE-L-QERL-caahklchpeelmllrhslgipqaplsscssq-----sl-QLRG-CLNQLHGGLFLYQGL-LQAL
1alu (19) ------------------l-TSSERIDKQIRYILD---GISA-L-RKET-cnksnmcen--lnlpkm-aekdg-----cfqs-----gf-NEET-CLVKIITGLLEFEVY-LEYL
1hgu (2)  -------------ptipls-RLFQNAMLRAHRLHQ---LAFD-T-YEEF-eeayipkeqkysflqapqaslcfsesiptpsnreqaqqk-SNLQ-LLRISLLLIQSWLEP-VGFL
1lki (9)  -------natcairhpchg-NLMNQIKNQLAQLNG---SANA-L-FISY-ytaqgepfpnnveklcapnmtdfpsfhgng----------TEKT-KLVELYRMVAYLSAS-LTNI
1ax8 (3)  -----------------iq-KVQDDTKTLIKTIVT---RIND-I------shtqsvsskqkvtgldfipglhpil--------------------TLSKMDQTLAVYQQI-LTSM
1evs (1)  --------aaigscskeyr-VLLGQLQKQTDLMQDtsr-----L-LDPY-iriqgldvpklrehcrerpgafp--------seetlrgl-GRRG-FLQTLNATLGCVLHR-LADL
1f6f (1)  aqhppycrnqpgkcqiplq-SLFDRATTVANYNSK---LAGE-M-VNRF-deqygqginseskvinchtssittpnskaeaint------EDKI-LFKLVISLLHSWDEP-LHHA
                                       ______   ____ _                                                      ____________ _
1cnt (11) ---------------9959-600690291097097---4086-1-2940-39606895-----------------------------95-7227-106600410251644-0490
1bgc (9)  ---------------9059-920990592099036---1056-0-5990-36529249489269719923138183610279-----55-9691-003401400930320-0920
1alu (19) ------------------9-754880392099029---1063-0-7930-199193199--383191-49910-----0698-----56-6846-005400200170301-0700
1hgu (2)  -------------999573-952750294059039---3092-1-0633-397923--9596929797887022943950499890897-3614-006501200800280-2494
1lki (9)  -------905179599496-924770771059056---4075-0-1710-482129505953991037699811749293----------9585-302201800100141-0282
1ax8 (3)  -----------------99-923580582095016---3199-9--------------------79929119159--------------------302700400110250-0962
1evs (1)  -----------89495969-600380593048066456-----0-0300-09015076993898093997513--------45970966-5972-009304710160199-1552
1f6f (1)  9995972999989695546-220590183059029---1017-0-1991-6999---------939020882911969790992------9278-006100101501580-0950

                    --------------C-------------                            ---------------D--------------
                    1*  5*6               23 24*26*                         1* 4*5                  25 26*28*
                    |   | |                | | |                            |  | |                   | | |
ledqqvhftpteg----df-HQAIH-TLLLQVAAFAYQIEELMI-LLE-ykiprneadgm--------------l-FEKK-LWGLKVLQELSQWTVRSIHDL-RFI-sshqtgipó- (187)
agispe------------l-APTLD-TLQLDVTDFATNIWLQME-DLG-aapamptft---------------sa-FQRR-AGGVLVASQLHRFLELAYRGL-RYL-a--------- (173)
qnrfes------------s-EEQAR-AVQMSTKVLIQFLQKKAK-----nldaittp-dpttnaslltklqaqnq-WLQD-MTTHLILRSFKEFLQSSLRAL-RQM----------- (184)
rsvfanslvygasds-----------DVYDLLKDLEEGIQTLMG-RLE-dgsprtgqafkqtyakfdanshn----DDAL-LKNYGLLYCFRKDMDKVETFL-RIV-qcrsvegscg (190)
trdqkvlnpt-----avsl-QVKLN-ATIDVMRGLLSNVLCRLC-NKY-rvghvdvppvpdhsdk--------ea-FQRK-KLGCQLLGTYKQVISVVVQAF--------------- (180)
p-------------------SRNVI-QISNDLENLRDLLHVLAF-SKS-chlpeasgletldslggvleasgy--------STEVVALSRLQGSLQDMLWQL-DLS-pgc------- (146)
eqrlpkaqdlersglnied-LEKLQ-MARPNILGLRNNIYCMAQ-LLD-nsdtaeptkagrgasqpptptpasda-FQRK-LEGCRFLHGYHRFMHSVGRVF-SKW----------- (187)
vtelanskgtsp-al-----LTKAQ-EIKEKAKVLVDGVEVIQK-RIH-pgeknepypvwseqssltsqde-----NVRR-VAFYRLFHCLHRDSSKIYTYL-RIL-kcrltcset- (199)
                           _________                                              _______________
1800698317994----91-29308-501910130092047009-529-69338790979--------------9-9963-870390087029504701900-970-389274996- (187)
820199------------0-28108-603920570064098204-973-729991818---------------44-9667-000000022079008602900-992-4--------- (173)
375199------------1-79709-402910850092089508-----95983994-79983574199068468-9813-600060071057005200401-955----------- (184)
949328498949685-----------945600640192086046-989-98679483899998319609599----9813-591221060167002000110-710-1364594969 (190)
0930994299-----1960-29707-500830920262010400-895-9339095394596898--------93-7744-550022000069204904925--------------- (180)
9-------------------59309-702910860473076009-569-492692865965981266194989--------537500113032008304921-997-144------- (146)
7991395960997647880-18718-909730620663090029-629-9---------------------9395-9596-160032060006007000810--------------- (187)
281359-9-----21-----46408-603861980182047009-853-9739599939199272063989-----9679-800840050079015515400-920-3595------ (197)
```

Figure 3. Structure-based sequence alignment and solvent ASA of the long chain four-helical cytokines. The CC region is shown in upper case bold type and the alignment is only meaningful for these regions. Core positions are indicated by the numbers above the alignment, those marked with an asterisk belong to the peripheral core. Corresponding ASA data is given in the equivalent alignment below that of the sequences. The numbers are approximations calculated by considering an ASA of between 0 and 9.99 Å^2 as 0; of between 10 and 19.99 Å^2 as 1; of between 20 and 29.99 Å^2 as 2 and finally any ASAs of greater than 90 Å^2 as 9. We consider residues with number 0 or 1 (i.e. those positions at which the ASA is less than 20 Å^2) as being inaccessible to solvent. In borderline cases, visual inspection was carried out in order to determine whether specific residues should be considered solvent accessible or buried. Positions highlighted in yellow are those at which all structures have an amino acid that is solvent inaccessible. Those positions highlighted in blue are those at which all but one of the structures have an amino acid that is solvent inaccessible. Positions highlighted in green are those at which all but two of the structures have an amino acid that is solvent inaccessible. At certain positions gaps have been incorporated so as to separate main and peripheral core regions from one another and from the rest of the sequences. Horizontal lines represent the regions within the core that correspond to the core regions for the alignment of long and short chain cytokine proteins together.



```
                          -----αA-----                   --β1--                     -----αB-------
                          1*  4     11                   1    6                     1* 3        13
                          |   |      |                   |    |                     |  |         |
1rcb (1)  ---------hkcdit-LQE-IIKTLNSL-teqktlcte---------LTVTDI-faasknt-------------TE-KETFCRAATVL-rqfyshhekdtrc
1eer (1)  ----apprlicdsrv-LER-YLLEAKEA-ekittgcaehcsln----EKITVP-dtkvnfyawkrmevgqqav-EV-WQGLALLSEAV-lrgqallvkssqp
1ete (1)  ----tqdcsfqhspi-SSD-FAVKIREL-sdyllqd-----------YPVTVA-snlqddel----------c-GG-LWRLVLAQRWM-erlktva------
2gmf (4)  rspspstqpwehvna-IQE-ARRLLNLS-rdtaaemn----------ETVEVI-semfdlqept----------CL-QTRLELYKQGL-rgs----------
1hul (5)  ---------iptsal-VKE-TLALLSTH-rtllianet---------LRIPVP-vhknh------------------QLCTEEIFQGI-gtlesqtvqg---
3ink (6)  -------stkktqlq-LEH-LLLDLQMI-lnginnyknpkltrmlt-FKFYMP-kkat-----------------E-LKHLQCLEEEL-kpleevlnlaqsk
1scf (11) --------------------NVKDVTKL-vanlpkd-----------YMITLK-yvpgmdvlpshc--------WI-SEMVVQLSDSL-tdlldkfs-nise

1rcb (1)  ---------997592-066-00900630-392818009---------360211-4549992-------------86-94300000300-9900692691990
1eer (1)  ----99399036595-059-22930960-49508406970509----590500-3293999409959706000-40-29004502900-8906443998996
1ete (1)  ----99904199540-498-05541950-3993999-----------860421-50059294----------0-02-00800100930-6904992------
2gmf (4)  995677981990370-095-08921992-93886449----------680700-2992759915----------00-43109203901-634----------
1hul (5)  ---------984531-096-03620692-696057299---------290310-71910------------------11005502900-3507891976---
3ink (6)  -------98992392-079-01900640-29007499095499069-890920-9918-----------------8-49106003700-9506910961459
1scf (11) --------------------79720970-3780799-----------770606-309129967150--------00-02004101710-67029959-9999
            --------αC--------                           --β2--     --------αD--------
            1*   5         16                            1     3   1           14 15 17
            |    |          |                            |     |   |            |  |  |
ataqqfhrhkq-LIRF-LKRLDRNLWGLA-glnscpvkea-----------------N----QS---TLENFLERLKTIMR---EKY-skcss------------- (129)
--------wep-LQLH-VDKAVSGLRSLT-tllralgaqkeaisnsdaasaaplrt-I----TA-d-TFRKLFRVYSNFLR-g-KLK-lytgeacrtgdr------ (166)
--------gsk-MQGL-LERVNTEIHFVT-kcafqpppsclrf--------------V----QT---NISRLLQETSEQLV---ALK-pwitrqnfsrclelqcqp (134)
-----------------LTKLKGPLTMMA-shykqhcpptpets---------cat-Q----II---TFESFKENLKDFLL---VIP-fdcwep------------ (124)
--------------GT-VERLFKNLSLIK-kyidgqkkkcge---------------E----RR---RVNQFLDYLQEFLG---VMN-tewi-------------- (112)
hl--------r-PRDL-ISNINVIVLELK-gsettfmcey-----------------A-de-TA---TIVEFLNRWITFAQ---SII-stlt-------------- (133)
----------s-NYSI-IDKLVNIVDDLV-ecvkensskdlkksfkspep-------R----LF---TPEEFFRIFNRSID---AFK-dfvvasetsdc------- (138)

96893797096-0193-069009106500-6596464938-----------------9----81---30550089078609---917-49499------------- (129)
--------694-0891-068015009303-82097220596095559799759999-3----40-7-20760091004009-0-006-500362297599------ (166)
--------499-3732-069007107107-9141995496199--------------7----92---60390084018109---618-790596902901909269 (134)
-----------------079093105701-40099515917699---------275-8----93---60940282067009---711-964199------------ (124)
--------------49-079008006706-950422698039---------------6----97---90550061077001---324-9963-------------- (112)
95--------9-1940-163048109917-2------666-----------------3-99-51---20390066017008---622-7949-------------- (133)
----------0-2630-049018505932-96099496972779499197-------9----91---30960091049035---0--------------------- (138)
```

Figure 4. Structure-based sequence alignment and solvent ASA of the short chain four-helical cytokines. See legend to Figure 3.

to secondary structure elements). Between three and 111 sequence homologues were collected for each cytokine structure. The sequence homologues were added to the structure-based sequence alignments and the residue frequencies at each core position determined. Because the number of homologues for each structure is so variable, residue frequencies were normalised in order to give an unbiased representation of the residues at any one site.

Classification scheme for residues at equivalent sites

We put individual amino acid residues into three broad classes on the basis of their intrinsic hydrophobicity and the extent to which they are found at the interior and exterior of proteins.[38,39] These classes are:

(i) hydrophilic residues (*s*, surface): R, K, E, D, Q and N;

(ii) neutral residues (*n*): P, H, Y, G, A, S and T;

(iii) hydrophobic residues (*b*, buried): C, V, L, I, M, F and W.

An examination of the amino acid residues occurring at all positions within the sequences of the cytokine structures shows that their residues are exactly 33% hydrophilic, 33% neutral and 33% hydrophobic (data not shown). Here we consider that a set of residues with related properties occurring at a site in 80% or more of the sequences is a good indication of significant conservation. This criterion for conservation at a site is somewhat broader than that used in a study of more closely related proteins[38] and is dictated by the high sequence divergence of this family.

Sites that conserve these residue classes, or subsets of them, are labelled *s*, *n*, or *b*; sites that have combinations of hydrophilic and neutral residues or of hydrophobic and neutral residues are labelled *sn* and *bn*, respectively.

Sites that conserve closely related sets of residues are classified on the basis of the extent of the variation that they present:

(i) Strongly conserved sites are those where the volume variations are small and their *s*, *n*, or *b* character is conserved (e.g. V, L, I and M or L, I, M and F) in more than 80% of sequences.

Table 3. Corresponding regions of structure in the long and short chain cytokines

Family	Helix A	Helix B	Helix C	Helix D
Long chain	A10–A20*	B8–B20*	C7–C15	D6–D20
Short chain	A1*–A11	B1*–B13	C8–C16	D1–D15
Consensus	A′1–A′11	B′1–B′13	C′1–C′9	D′1–D′15

See also those regions marked in Figures 3 and 4.

Table 4. Conserved positions in the long chain cytokine family

A. *Positions of strong and medium conservation, solvent inaccessible sites conserving residue class and the number of helices each structure maintains contacts with at each position*

			No. of helices contacted by each structure									
Helix position	Classification from structures	Classification from sequence homologues	1lki	1cnt	1alu	1ax8	1bgc	1evs	1f6f	1hgu	Total	#
Sites that have strong conservation												
B6 ⑧	L 88/LF 100	L 96	2	2	1	1	2	2	2	2	14	6
A13 ⑧	LIM 88	VLIM 90	2	2	2	2	2	2	1	2	15	7
A20* ⑧	LIM 88	VLIM 92	2	1	2	2	2	2	2	1	14	6
B20* ⑧	L 88/LV 100	VLI 96	2	2	2	2	2	1	2	1	14	6
B23*	*bn* 100	VLIM 99	2	2	2	2	2	1	1	0	12	5
C11 ⑧	AVLIM 88	VLIM 80	3	2	3	3	3	3	1	3	21	7
C18 ⑧	VLI 100	VLIM 99	1	2	3	1	2	2	2	2	15	6
A24* ⑧	LFY 88	LIMFY 94	2(1)	1(1)	1	–	1(1)	3	1	1	9(3)	2(2)
C14 ⑧	LF 100	LIF 99	2	3	3	3	2	3	3	2	21	8
D11 ⑧	ALF 100	LIMF 88	2	2	2	3	1	2	1	1	14	5
D14 ⑧	LFY 100	LFY 95	3	2	3	3	3	3	1	3	21	7
D25 ⑧	LF 100	LIF 96	2	1	3	2	2	2	2	2	16	7
C21 ⑧	LIM 63 RKQ 37	LIM 63 + RKEQ 35	2	2	3	2	1	3	2	2	17	7
B9 ⑧	*bn* 88	VLIM 72 + N13	3	2	2	2	2	2	2	0	15	7
Sites that have medium conservation												
A6 ⑧	*bn* 100	ASTCLI 89	3	2	2	2	2	2	1	1	15	6
A10 ⑧	AVLI 88	AVLI 83	3	2	2	2	2	2	2	2	17	8
A17* ⑦	*bn* 100	ASTVLI 89	1	2	1	2	0	–	2	1	9	3
C4* ⑦	AVLI 100	AVLIM 94	1	2	2	2	1	1	1	–	10	3
C7 ⑧	AVLI 88	AVLIM 87	2	3	3	3	3	2	2	0	18	7
C22 ⑧	AIM 88	AVLIM 92	1	1	1	1	1	1	1	1	8	0
D18 ⑧	*bn* 100	ASTLIM 100	1	2	3	2	2	2	1	3	16	6
D21 ⑧	*bn* 100	ASTVLIM 100	2	1	1	2	1	2	2	1	12	4
A2	*bn* 100	VLMF 84	1	1	0	1	1	1	1	1	7	0
A3 ⑥	*bn* 88	CLMF 71 + Q 13	3	2	1	0	1	2	2	1	12	4
B13 ⑧	VLIY 100	VLIMY 99	1	2	2	2	2	2	2	2	15	7
B16 ⑧	VLFWY 100	VLIFYW 100	2	2	1	2	1	2	1	1	12	4
D10 ⑧	AVLIF 100	AVLIMF 95	1	1	1	1	1	1	1	1	8	0
Sites that conserve residue class at positions inaccessible to solvent												
B12 ⑧	*bn* 100	*bn* 100	1	1	1	0	1	1	1	1	7	0
B19 ⑧	*bn* 88	*bn* 91	1	1	1	1	0	1	1	0	6	0
C10 ⑧	*sn* 75	*sn* 87	1	1	1	1	1	0	1	1	7	0

B. *Positions that conserve residue class (*s, sn *or* bn)

Helix position	Classification from structures	Classification from sequence homologues	Helix position	Classification from structures	Classification from sequence homologues
B15	*bn* 88	*bn* 83	C2*	*sn* 88	*sn* 92
D5	*bn* 88	*bn* 82	C3*	*sn* 100	*sn* 98
D8	*bn* 100	*bn* 97	C5*	*sn* 88	*sn* 82
D27*	*bn* 75	*bn* 83	C6	*sn* 88	*sn* 87
D28*	*bn* 100	*bn* 95	C8	*sn* 75	*sn* 81
A1	*sn* 88	*sn* 92	C12	*sn* 88	*sn* 87
A4	*sn* 88	*sn* 83	C13	*sn* 75	*sn* 92
A5	*s* 100	*s* 95	C16	*sn* 100	*sn* 97
A11	*sn* 100	*sn* 98	C26*	*sn* 100	*sn* 97
A12	*sn* 88	*sn* 83	D2*	*s* 75	*sn* 84
A15	*sn* 100	*sn* 93	D3*	*s* 88/*sn* 100	*sn* 99
A18*	*sn* 88	*sn* 97	D12	*sn* 100	*sn* 95
A19*	*sn* 100	*sn* 98	D15	*sn* 100	*sn* 99
A22*	*s* 88	*sn* 82	D16	*s* 88/*sn* 100	*sn* 99
A23*	*sn* 100	*sn* 87	D19	*sn* 88	*sn* 87
B1*	*sn* 100	*sn* 100	D23	*sn* 88	*sn* 89
B2*	*s* 88	*sn* 83	D26*	*s* 88/*sn* 100	*s*91
B3*	*sn* 88	*sn* 88			
B4*	*sn* 88	*sn* 87			
B8	*sn* 75	*sn* 80			
B11	*sn* 88	*sn* 91			
B21*	*sn* 100	*sn* 91			
B22*	*sn* 100	*sn* 90			

Numbers shown in parentheses ( ) are the number contacts made to a helix residue that is not in the common core. #: number of structures that makes 2+ contacts. *bn*: conserved hydrophobic/neutral site. *sn*: conserved hydrophilic/neutral site. ⑧: position at which all structures are inaccessible to solvent. ⑦: position at which 7/8 structures are inaccessible to solvent. ⑥: position at which 6/8 structures are inaccessible to solvent.

(ii) Sites of medium conservation are those presenting either (a) small or medium sized hydrophobic residue (e.g. A, V, L, I and M) or (b) medium or large sized hydrophobic residues (e.g. V, L, I, M and F).

Note that Ser and Thr can be found on the buried faces of helices. Their polar –OH groups can be neutralised by the formation of hydrogen bonds to main-chain oxygen atoms so that they can effectively act as hydrophobic residues. We also see certain cases where sites normally occupied by medium or large hydrophobic residues are occupied by Arg, Lys, Glu or Gln. In these cases, the long hydrophobic portion of their side-chains is tucked into the interior of the structure with the charged group on the surface. This gives them a pseudo-hydrophobic role.

Conserved positions

The residue conservation found at structurally equivalent sites in the alignments are described in Tables 4–6.

Long chain cytokines. Examination of residue frequencies (Table 4) within the core of the long chain family shows that:

Of the 101 sites in the common structural core, 70 have significant conservation of some kind and 31 show little or no conservation.

Of the 70 conserved sites, 14 are strongly conserved, 13 have medium conservation and 43 sites conserve the general class of residue (seven *bn*, 34 *sn* and two *s*).

Thirteen of the 14 strongly conserved sites and 12 of the 13 sites with medium conservation are buried in almost all of the structures. The other two sites (αB23* and αA2) are both situated close to the end of a helix and are buried in some of the structures.

Of the 43 sites conserving residue class, 40 are situated on the external surface of the helices and are partly accessible to solvent in all or most of the structures. The other three are covered by one of the two long crossovers which pass over the external surface of helices B and D.

Table 4(a) shows the 14 positions of strong conservation, the 13 positions of medium conservation and the three sites of residue class conservation that are inaccessible to solvent. These residues are more interesting to investigate further in that their volume and properties are more constrained than those of solvent accessible residues. We include within this table data on the number of inter-helical contacts being made in each structure by residues at every structurally equivalent site. Table 4(b) shows the 40 positions of residue class conservation that are solvent accessible, 33 *sn*, two *s* and the remaining five *bn* (these five are all situated within helices B and D and are buried by the long crossover peptides).

Short chain cytokines. Residue frequencies within the core of the short chain family are shown in Table 5.

Of the 66 residues forming the CC, 44 positions show some form of conservation and 22 no conservation.

Of the 44 conserved positions, 14 show strong conservation, four show medium conservation and 26 have residue class conservation.

Of the 21 sites with residue class conservation that are in helices, 16 are solvent accessible and five are mostly solvent inaccessible. These five include one *sn* and four *bn* sites. Three of the *bn* sites are covered by one of the long crossover regions between helices A–B or C–D.

Thirteen of the 14 strongly conserved positions and all four sites with medium conservation are solvent inaccessible.

Table 5(a) lists the 13 strongly conserved positions situated within helices, the four sites with medium conservation and the five helical sites conserving residue class that are inaccessible to solvent. Data on the inter-helical contacts that structurally equivalent sites make within each representative structure are also shown. Table 5(b) includes the data on the strongly conserved site located within the first beta strand and all 21 helical positions conserving residue class that are solvent accessible (of which four are *bn*, 13 are *sn* and four are *s*).

Long and short chain cytokines. The frequency data for the long and short chain family was merged and the results are shown in Table 6. The CC regions for the long and short chain families combined correspond to those core regions marked with triangles on the long and short chain structure-based sequence alignment (Figures 3 and 4).

Of the 48 sites within the CC, 29 show some form of conservation and 19 show no conservation.

Of the 29 conserved sites, five have strong conservation, nine have medium conservation and the remaining 15 have residue class conservation.

Of the 15 sites conserving residue class, five are inaccessible to solvent, these include four *bn* (that are all covered by the long peptide crossovers) and one *sn* site.

The five strongly and nine medium conserved sites are solvent inaccessible.

Table 6(a) lists the five sites with strong conservation, the nine sites with medium conservation and also the five sites that conserve residue class and are solvent inaccessible. Table 6(b) shows the classification of the ten *sn* sites conserving residue class that are solvent accessible.

Table 5. Conserved positions in the short chain cytokine family and their interhelical contacts

A. *Conservation at positions of strong and medium conservation and at solvent inaccessible sites conserving residue class*[a]

			Number of helices contacted by each structure								
Helix position	Classification from structures	Classification from sequence homologues	3ink	1eer	1hul	1rcb	2gmf	1ete	1scf	Total	#
Sites that have strong conservation											
A1* ⑤	VLI 86	VLIM 86	3	2	2	2	2	1	–	12	5
B6 ⑦	VLI 86	VLI 87	3	2	1	2	2	2	2	14	6
B13 ⑦	VLIM 100	VLIM 98	1	1	2	2	2	2	2	12	5
C5 ⑦	VLI 100	VLIM 98	1(1)	1(1)	1	1(1)	3	1(1)	2	10(4)	2(4)
C8 ⑦	VLIA	VLIM86	1	2	1	3	3	3	3	16	5
C12 ⑦	VLI 100	VLI 98	2	2	2	2	1	2	2	13	6
C15 ⑦	VLIM 100	VLIM 98	2	2	2	2	2	1	1	12	5
D5 ⑦	FL 100	LF 98	1	1	1	1	1	1	1	7	0
D13 ⑦	ALIM 100[b]	LIM 87	1	3	2	2	2	3	1	14	5
A5	VLI 71/bn 86	VLI 82 + RKQ 13	1	1	1	1	1	0	0	5	0
D6 ⑦	FL 86	LIF 87 + KQ 13	1	2	0	1	2	2	3	11	4
D12 ⑥	IF 71	IF 71 + Q 14	1	1	1	0	1	1	1	6	0
D16 ⑥	LIMF 86	LIMFY 82 + K14	1	2	1	2	2	1	2	11	4
Sites that have medium conservation											
D9 ⑦	LFWY 86	LFWY 84	2	2	1	2	3	2	2	14	6
A4 ⑥	*bn* 100	ATLIFY 88	2	1	2	2	2	2	1	12	5
A8 ⑦	VLIA100	AVLIF 100	1	1	2	2	2	2	2	12	5
B9 ⑦	*bn* 100	AVLIFY 100	2	2	1	2	2	2	1	12	5
Sites that conserve residue class at positions inaccessible to solvent											
A11 ⑥	*bn* 100	*bn* 100	2	2	1	2	2	2	2	13	6
B5 ⑥	*bn* 71	*bn* 80	1	1	0	0	1	1	1	5	0
B12 ⑦	*bn* 86	*bn* 84	1	1	1	1	1	1	1	7	0
C11 ⑦	*sn* 71	*sn* 81	1	1	1	1	0	1	1	6	0
D2 ⑦	*bn* 100	*bn* 100	0	1	0	1	1	0	0	3	0

B. *Table showing positions of residue class conservation* (bn *or* sn) *plus conserved positons within the beta strands*

Helix or strand position	Classification from structures	Classification from sequence homologues
β1.3 ⑦	VIF 100	VLIF 98
β1.5 ⑥	VLM 86	*bn* 86
β1.6 ⑦	*bn* 86	*bn* 85
β2.3 ⑥	*bn* 86	*bn* 90
C4*	*bn* 100	*bn* 96
A2*	*sn* 100	*sn* 100
A3*	*s* 86/*sn* 100	*s* 92
A9	*sn* 100	*sn* 100
β1.4	*n* 86/*sn* 100	*sn* 97
B11	*s* 86/*sn* 100	*s* 81
C3*	*sn* 86	*sn* 82
C6	*sn* 100	*sn* 98
C7	RKN 100	*sn* 98
C10	*sn* 86	*sn* 90
C16	*sn* 86	*sn* 86
β2.2	*sn* 71	*sn* 84
D1	T 71/*sn* 100	*sn* 93
D3	*s* 71/*sn* 86	*sn* 81
D4	*s* 86/*sn* 100	*s* 82
D7	*s* 100	*s* 94
D11	*sn* 100	*sn* 92
D17	*sn* 86	*sn* 92

[a] Positions within β-sheets are excluded here due to the different nature of their helical contacts. See also the legend to Table 4.
[b] C has been mutated to A in the structure.

Conservation in the four-helical cytokines

We can summarise the conservation of structure and sequence in four-helical cytokines in the following terms:

Within the long and short chain families, the different members have close to half of their structures in the same conformation. When the long and short families are considered together the region that has a common conformation comprises a quarter of the largest structure (PL, PDB-ID: 1f6f) and just under half the smallest (IL-5, PDB-ID: 1hul).

In the three CCs, i.e. those of the long chain family, the short chain family and both families together, close to two-thirds of the residue sites have some type of residue conservation. There are 27, 18 and 14 sites with strong or medium

Table 6. Conserved positions in the long and short chain cytokine family

A. *Positions of strong and medium conservation and any solvent inaccessible sites conserving residue class and the number of helices each structure maintains contacts with at each position*[a]

				Number of helices contacted by each structure																
Helix position	Position in long chain/shortchaincore	Classification from structures	Classification from sequence homologues	1lki	1cnt	1alu	1ax8	1bgc	1evs	1f6f	1hgu	3ink	1eer	1hul	1rcb	2gmf	1ete	1scf	Total	#
Sites that have strong conservation																				
B′6 (15)	B13/B6	VLIFY 93	VLI 87	1	2	2	1	2	2	1	2	3	2	1	2	2	2	1	26	10
B′13 (15)	B20*/B13	VLIM 100	VLIM 98	2	2	2	2	2	1	2	1	1	1	2	2	2	2	2	26	11
C′5 (15)	C11/C12	VLIM 87	VLIM 89	3	2	3	3	3	3	1	3	2	2	2	2	1	2	2	34	13
C′8 (15)	C14/C15	VLIMF 100	LIMF 93	2	2	3	3	2	3	2	2	2	2	2	2	2	1	1	31	13
D′6 (15)	D11/D6	ALF 93 + K 7	LIF 88	2	3	1	3	1	0	1	1	1	1	0	1	1	2	3	21	5
Sites that have medium conservation																				
A′1 (13)	A10/A1*	AVLI 86	AVLIM 84	2	2	2	2	2	2	1	1	2	1	2	1	1	1	–	22	8
C′1 (15)	C7/C8	AVLI 93	AVLIM 93	2	3	3	3	3	2	2	0	1	2	1	3	3	3	3	34	12
D′13 (15)	D18/D13	ALIM 87[b]	CLIM 85	1	2	3	2	2	2	1	3	1	3	2	2	2	2	1	29	11
D′9 (15)	D14/D9	LFYW 93	LFYW 92	3	3	3	3	3	3	1	3	2	2	1	2	2	2	2	35	13
A′4 (14)	A13/A4	*bn* 87	AVLIMFY 84	2	2	2	2	2	2	1	2	2	1	2	2	2	2	0	26	12
A′8 (14)	A17*/A8	AVLI 93	AVLIF 89	1	2	1	2	0	–	2	1	1	1	1	2	1	2	2	19	6
A′11 (14)	A20*/A11	ALIM 80	AVLIM 83	2	1	2	2	1	1	2	1	2	1	0	1	1	2	2	21	7
B′9 (15)	B16/B9	AVLIFYW 100	AVLIFYW 100	2	2	1	2	1	2	1	1	2	2	1	2	2	2	1	24	9
D′5 (15)	D10/D5	AVLIF 100	AVLIF 97	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	14	0
Sites that conserve residue class at positions inaccessible to solvent																				
B′5 (14)	B12/B5	*bn* 87	*bn* 92	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	13	0
B′12 (15)	B19/B12	*bn* 87	*bn* 88	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	13	0
C′4 (15)	C10/C11	–	*Sn* 84	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	13	0
D′2 (14)	D7/D2	*bn* 87	*bn* 85	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
D′12 (14)	D17/D12	*bn* 80	*bn* 80	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	14	0

B. *Positions with residue class conservation*

Helix position	Position in long chain/shortchaincore	Classification from structures	Final classification
A′2	A11/A2*	*sn* 100	*sn* 99
A′3	A12/A3*	*sn* 93	*sn* 91
A′9	A18*/A6	*sn* 93	*sn* 99
A′10	A19*/A7	*sn* 87	*sn* 89
B′11	B18/B11	*sn* 87	*sn* 83
C′6	C12/C13	*sn* 80	*sn* 85
D′1	D6/D1	*sn* 87	*sn* 85
D′7	D12/D7	*sn* 100	*sn* 97
D′10	D15/D10	*sn* 93	*sn* 89
D′11	D16/D11	*sn* 100	*sn* 96

[a] See legend to Table 4.
[b] C has been mutated to A in the structure.

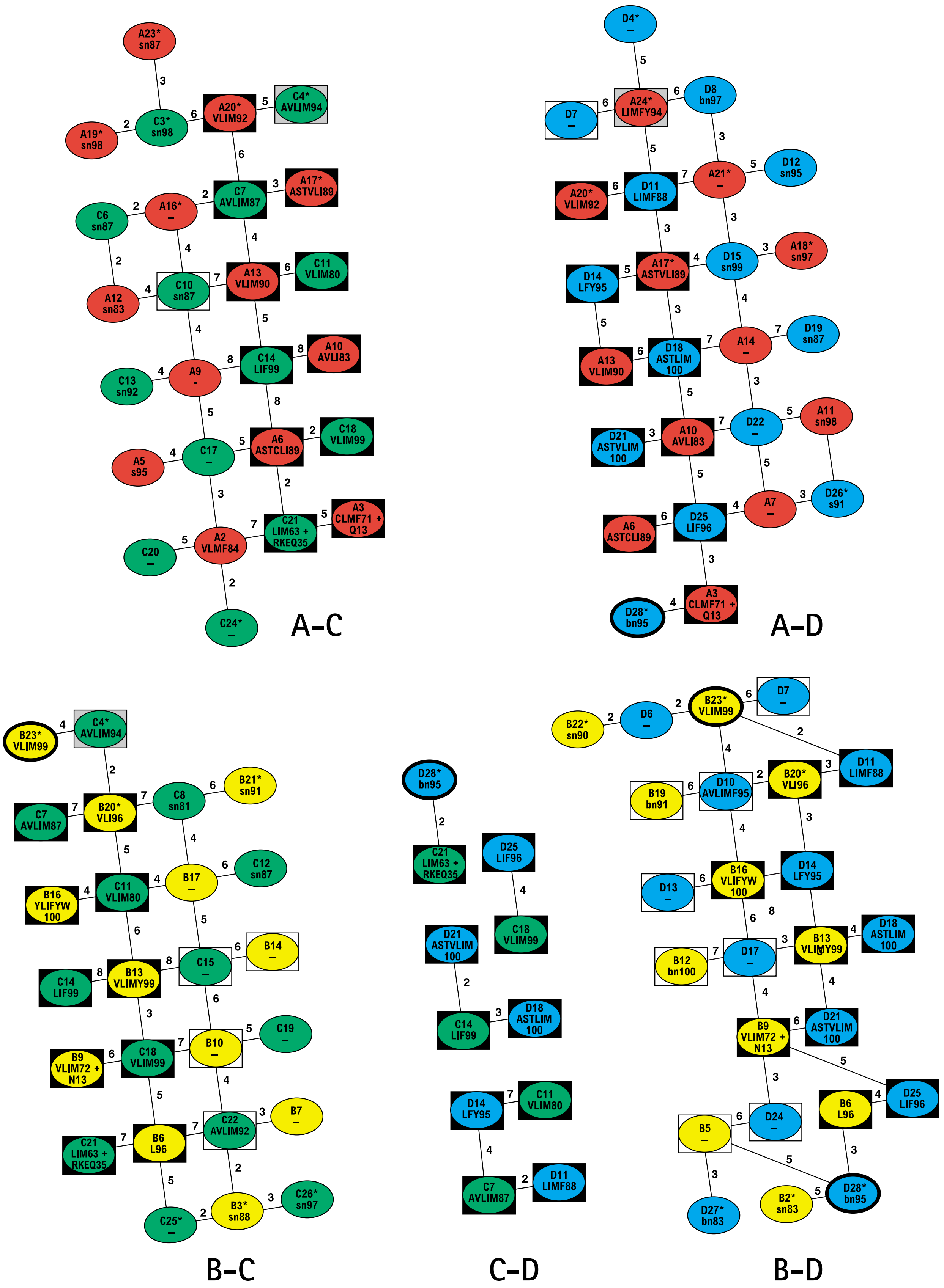


Figure 5. Helical contact nets for the CC of the long chain cytokines. Numbers shown in above lines indicate the number of structures in which that contact is made. 🟥 Helix A; 🟨 helix B; 🟩 helix C; 🟪 helix D; ○ solvent accessible position of hydrophobic conservation; ▢ position inaccessible to solvent; ■ position inaccessible to solvent and involved in contacting more than one helix; ▨ as above with exceptions.
]

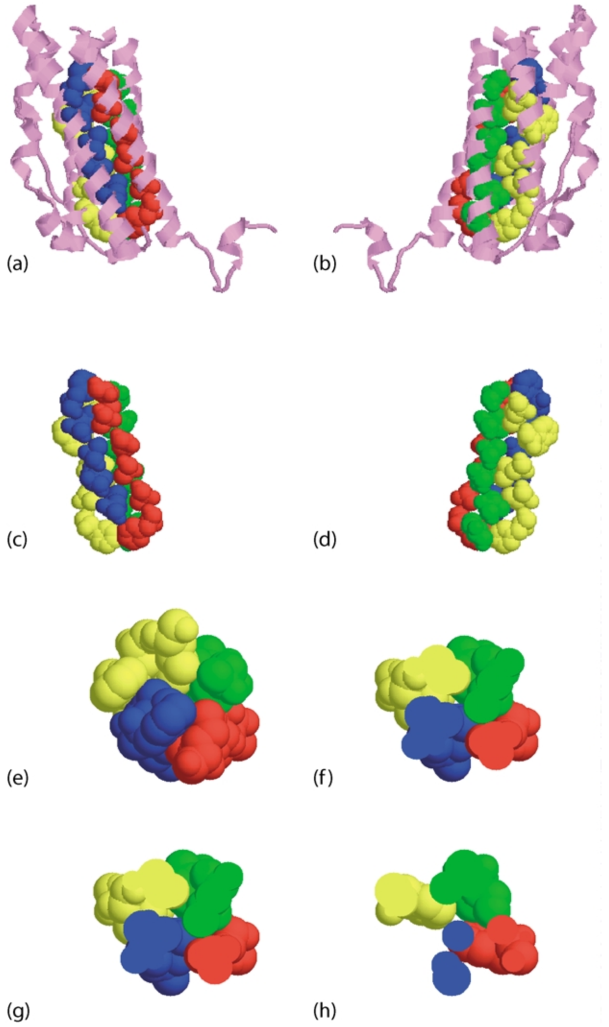


Figure 6. Representation of the 23 conserved sites involved in making multiple inter-helical contacts within placental lactogen (PDB-ID: 1f6f), a long chain cytokine. Figures (a) and (b) are front and back views of a ribbon representation with the 23 conserved positions shown by space filling and colour coded according to Figure 1. Figures (c) and (d) are the same views as in Figures (a) and (b) but only the 23 positions are shown. Figures (e)–(h) show the 23 positions viewed from the top and slabbed progressively down at levels of approximately 0, 30, 60 and 90%.

conservation in the three CCs, respectively, and in each case they form just over one quarter of the total number of residues in the CCs.

The structural role of residues with strong and medium conservation: inter-helical contacts

We carried out a detailed analysis of the four-helical cytokine structures to determine the structural role of the residues at sites of strong and medium conservation. This analysis showed that their principal role is making multiple helix–helix contacts at the centre of the structures.

The inter-helical contacts made by residues in the CC regions of the structures were determined for each member of the long and short chain families. We represented the results of these calculations using "helical nets" (Figures 5, 8 and 11). Helical nets are two-dimensional representations of alpha helices; if we cut down the centre of an alpha-helix and open it out we would see the residues situated in ascending slanted lines of three or four residues per line. Vertical strips down the "nets" therefore represent one side of the helix in question. Here we use the "nets" to show the regions and residues involved in making contacts between pairs of helices in multiple structures.

The helix–helix contacts in the structures of the long chain family are shown in Figure 5. In Table 4(a), we list the number of inter-helical contacts made by each residue at sites of strong or medium conservation. Comparison of the Table and Figure shows that although in different structures homologous residues can have differences in the number of contacts (Table 4(a)) the *pattern* of residue contacts is almost absolutely conserved (Figure 5).

Examination of the contact maps clearly shows that 23 of the 27 most conserved sites lie along one side of each helix at sites i, $i+3$, $i+7$, $i+10$, etc. or at i, $i+4$, $i+7$, $i+11$, etc. These 23 positions (αA3, αA6, αA10, αA13, αA17, αA20, αB6, αB9, αB13, αB16, αB20, αB23, αC7, αC11, αC14, αC18, αC21, αD11, αD14, αD18, αD21, αD25 and αD28) form one vertical strip down one side of the helix to which they belong. These residues are present in two or three helical nets in Figure 5, which means that they are involved in multiple inter-helical contacts in several, if not all of the structures. An analysis shows the 23 residues make-up the deeply buried core at the centre of each protein (Figure 6). The four helices contribute an almost equal number of residues that pack together to make an inter-locking core (see Walshaw & Woolfson,[14] and references therein, for a discussion of the nature of residue packing at helix interfaces).

Of this set of 23 positions, 11 have strong conservation (αA13, αA20, αB6, αB20, αB23, αC11, αC14, αC18, αD11, αD14 and αD25). Figure 7 shows that the residues at these 11 positions form a central line running down the centre of the core of these proteins around which the other 12 positions wrap. This arrangement explains why they are slightly more constricted in terms of amino acid composition than the other positions at which structural conservation occurs.

An analysis of the conserved residues and inter-helical contacts in the short chain family and in the combination of the two families gives an identical picture of the role of residues at sites that have strong and medium conservation.

Inter-helical contacts of the short chain family are shown in Figure 8 and the numbers of inter-helical contacts made by residues at conserved core positions are reported in Table 5(a). A set of 15 positions of strong and medium conservation (αA1*, αA4, αA8, αA11, αB6, αB9, αB13, αC5, αC8,

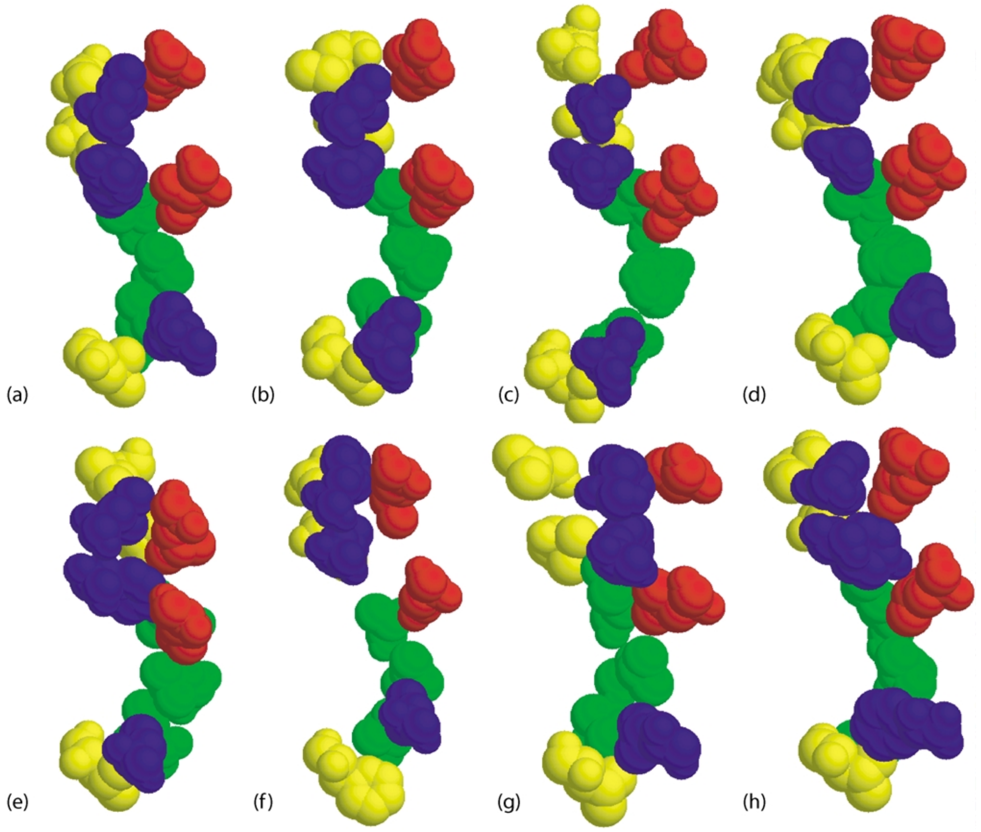


Figure 7. Representation of the 11 strongly conserved sites involved in making multiple inter-helical contacts within the eight structures of the long chain cytokines. Figures (a)–(h) show the 11 equivalent positions in the structures with PDB-IDs: 1alu, 1ax8, 1bgc, 1cnt, 1evs, 1f6f, 1hgu and 1lki, respectively. Sites within different helices are colour coded according to Figure 1.

αC12, αC15, αD6, αD9, αD13 and αD16) form a strip on one side of each helix and make multiple contacts to form the deeply buried central core of the protein (Figure 9). Of these residues, eight have strong conservation (αA1, αB6, αB13, αC5, αC8, αC12, αC15 and αD13), and they form a central structure around which the other seven residues of medium conservation wrap (Figure 10).

Helical nets were produced for the merged data of the long and short chain families together (Figure 11). From these and the data included in Table 6(a), we identified a set of 13 sites with strong or medium conservation (αA′1, αA′4, αA′8, αA′11, αB′6, αB′9, αB′13, αC′1, αC′5, αC′8, αD′6, αD′9 and αD′13). Once again residues at these positions form a strip on one side of each helix and make multiple inter-helical contacts (Figure 12). There are five positions of strong conservation (αB′6, αB′13, αC′5, αC′8 and αD′6). These sites are positioned at the centre of proteins belonging to both the long and short chain families (Figure 13) and are the sites subjected to the strongest constraints in this superfamily.

The four-helical cytochromes: c′ and b_{562}

These cytochromes are found in many bacteria and are probably involved in nitric oxide transfer.[40] They are similar to the cytokines in having divergent sequences[41] and being formed by four alpha-helices A–D, but they differ in that the helices have a conventional up-and-down topology. They also differ from the cytokines in binding a cofactor: a haem group which has helices A and C packed against one face and helix D against the other face (Figure 14).

Structures are known for the cytochromes c′ from five different species and cytochrome b_{562} from one species (Table 1). Representative structures[42–48] for the six species were chosen according to the same criteria used for the long and short chain cytokines (see above).

Using FASTA,[37] significant matches can be found between most of the five cytochromes c′ sequences of known structure but they make no significant match to the sequence of cytochrome b_{562}. Using the HMMs in SUPERFAMILY[49] significant matches are made between all six of these sequences.

Structural alignments

Using the procedures described above, we obtained a structural alignment of the sequences of the six four-helical cytochromes of known structure (Figures 14 and 15). Some comparisons of the cytochrome structures have been made

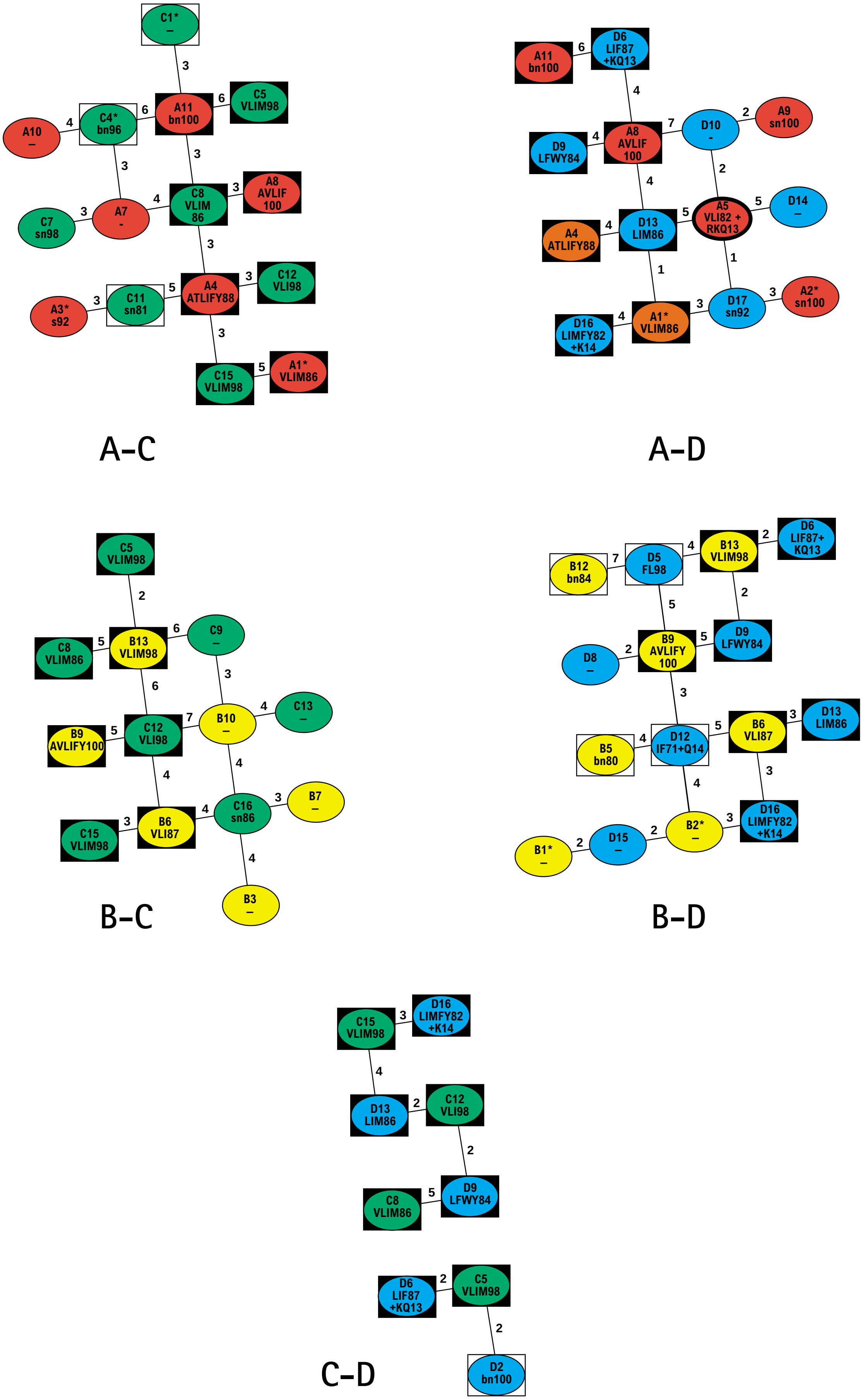


Figure 8. Helical contact nets for the CC of the short chain cytokines. See key to Figure 5.

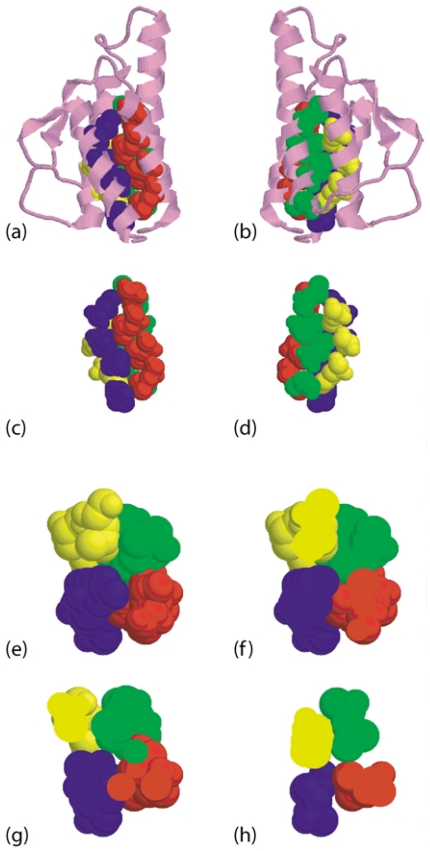


Figure 9. Representation of the 15 conserved sites involved in making multiple inter-helical contacts within IL-4 (PDB-ID: 1rcb) a short chain cytokine. Figures (a) and (b) are front and back views of a ribbon representation with the 15 positions shown by space filling and colour coded according to Figure 1. Figures (c) and (d) are the same views as in Figures (a) and (b) but only the 15 positions are shown. Figures (e)–(h) show the 15 positions viewed from the top and slabbed progressively down at levels of approximately 0, 30, 60 and 90%.

previously; Weber and Salemme first identified the relationship between cytochrome c′ and cytochrome b_{562}.[44] Since then several alignments of the cytochromes c′ have been made at various times as new structures became available. Two of these alignments are essentially the same as ours,[45,50] and two of them agree except for a shift of one turn in helix B.[46,47]

From the structural alignments, we identified a CC of 82 amino acid residues which are split approximately equally across the four helices (see Figure 15). The main-chain atoms of the 82 residues in the CCs of the five cytochromes c′ have RMSDs between 1.1 Å and 2.0 Å. Their residue identities are between 21 and 29%. Matches of the CC residues of cytochrome b_{562} to those of the five cytochromes c′ have RMSDs of between 1.6 Å and 2.0 Å and residue identities between 13 and 22%. The ASAs for all positions in each structure were calculated and the results are given in Figure 15.

Sequence homologues and conserved positions

At present only a few homologues are known for the cytochromes c′ and b_{562}: matching the sequences of the six structures against the sequences in NRDB90[36] only collected between zero and nine homologues per structure. Removing redundant matches gave just 15 unique sequences to add to the six of known structure. These 21 sequences were used to calculate the residue frequencies at each site in the CC. Because all the cytochrome c′ sequences are orthologous and there are very few sequences homologous to cytochrome b_{562} in comparison, we did not normalise the resulting numbers.

Using the classification system and criteria described above for the cytokines, we derived the conservation at each site and the results are shown in Table 7.

Of the 82 positions in the CC, eight have little or no conservation, 48 conserve the residue class (39 are *sn*, one is *n* and eight are *bn*) and 26 have medium or strong conservation. The proportion of sites with some conservation, 90%, is higher than the two-thirds found for the cytokines but the proportion of sites with medium or strong conservation is almost exactly the same: just over a quarter.

The structural role of conserved positions

We determined, for each of the cytochrome structures, the residue contacts formed between the helices and between the proteins and the haem group. The results of these calculations are given in the helical net drawings in Figure 16. Inspection of these nets and of the ASA data shown in Figure 15 shows that 18 of the 26 sites with medium and strong conservation lie along the buried spine of each helix. Of these, eight have strong conservation: αA9, αA16, αB20, αC5, αD10, αD14, αD17 and αD21. The other ten have medium conservation: αA6, αA13, αB6, αB9, αC12, αC16, αC19, αC23, αD3 and αD7. There are another five sites that lie along these spine positions that have *bn* conservation but are nonetheless involved in making multiple inter-helical contacts: αA5, αB2, αB13, αB16 and αC9. Figure 17(a) and (b) shows the first strongly conserved eight positions within one cytochrome c′ structure, Figure 17(c) and (d) shows the positions of the ten medium conserved and five *bn* conserved positions. These residues are involved in two or three interactions with residues in adjacent helices and/or the haem group (Figures 15 and 16).

The A–B, B–C and C–D helix interfaces are similar to those in the four-helical cytokines in their extent and in the buried spine residues of helices B and C being involved in the packing against both adjacent helices. The A–D interface is less extensive: this is because the haem group packs between the A and D helices (Figures

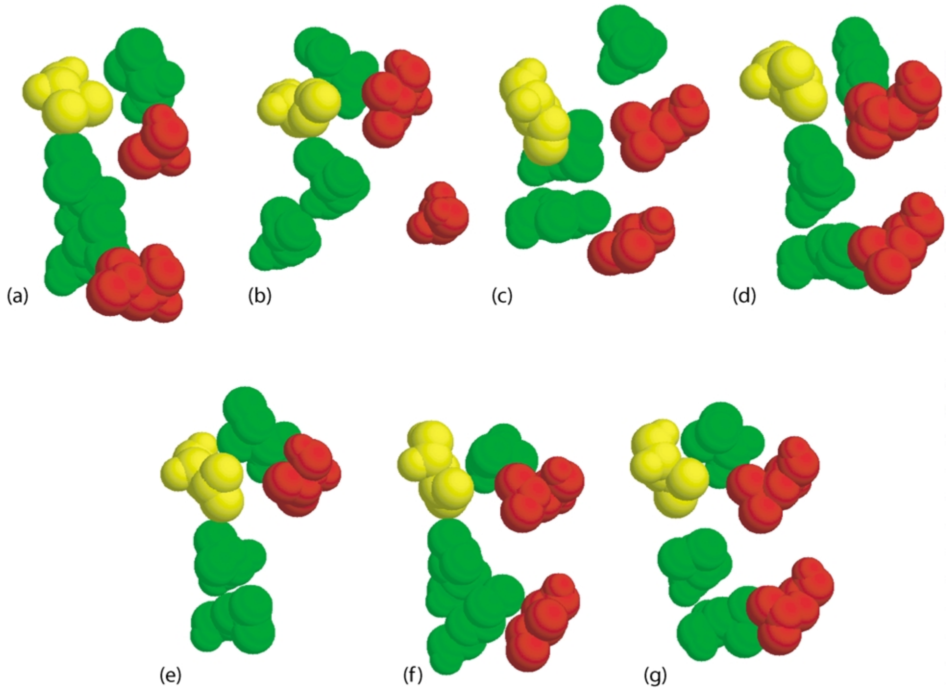


Figure 10. Representation of the eight strongly conserved sites involved in making multiple inter-helical contacts within the seven structures of the short chain cytokines. Figures (a)–(g) show the equivalent positions within the structures with PDB-IDs: 1eer, 1ete, 1hul, 1rcb, 1scf, 2gmf and 3ink, respectively. Sites within different helices are colour coded according to Figure 1. Note that in Figure (e) aA1* is missing as it is the structure lacking that position within the peripheral core.

14–17) and they are only in direct contact at one end.

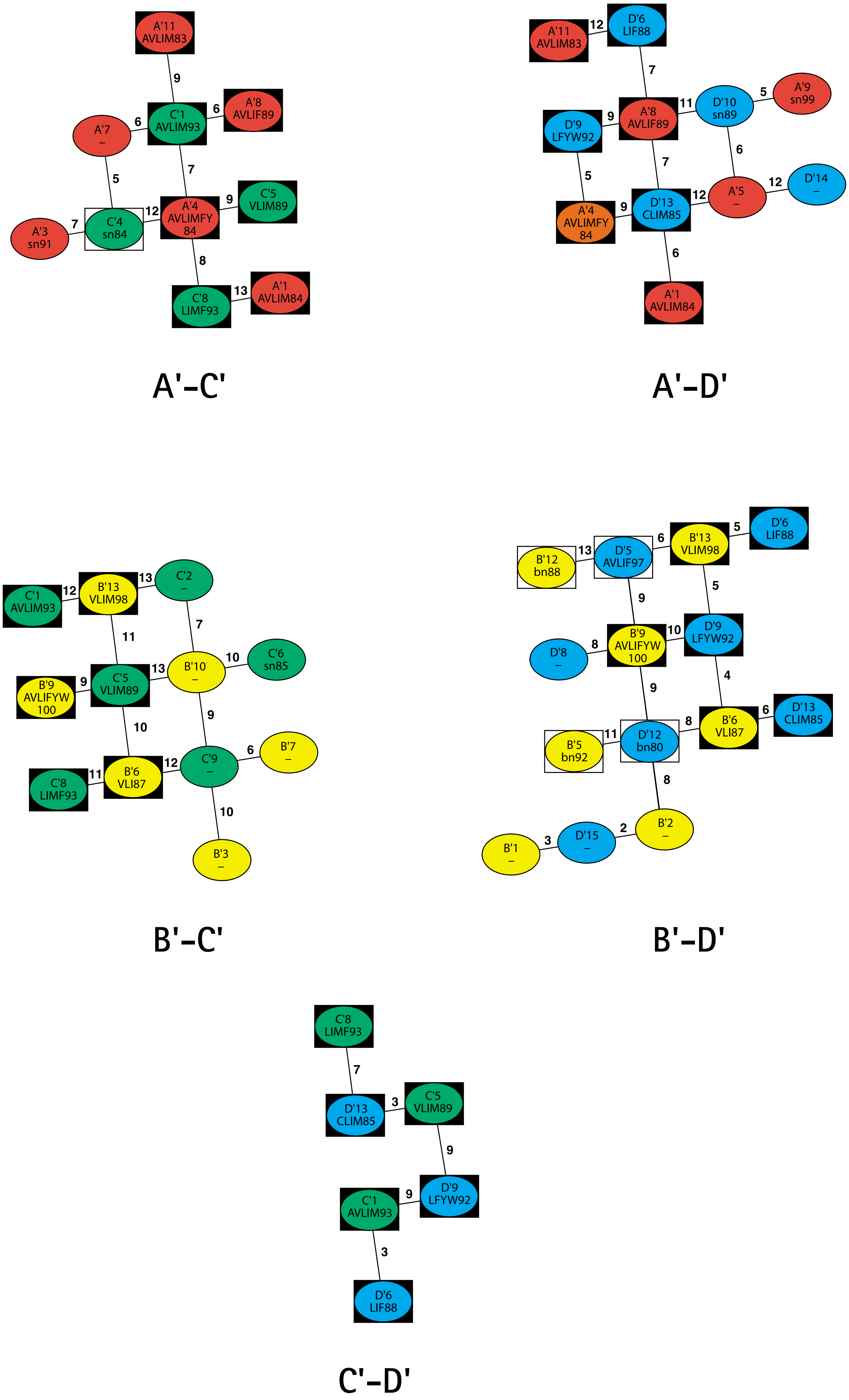
Of the remaining eight sites with medium or strong conservation, five are hydrophilic and three are neutral in nature rather than hydrophobic as one might expect to find at the internal spines of the helices. The character of six of these eight sites can be explained by the fact that they are concerned, directly or indirectly, with the packing around the haem group: αA2, αD11, αD16, αD18, αD19 and αD22. Of this six, three are also involved in making inter-helical contacts with one other helix. The remaining two out of the eight sites, αC1 and αC4, are involved in contacting at least one other helix (Figure 16). These eight sites are shown in Figure 17(e)–(h) along with the other well conserved positions within a cytochrome c′ structure.

Discussion

At the beginning of this paper, we raised questions regarding proteins that have evolutionary relationships but the sequences of which have little or no similarity when measured by conventional methods. What is the actual relationship between their sequences? Have they evolved to a point where there is no significant sequence similarity, except for that necessary to maintain their hydrophobic interiors and hydrophilic exteriors; or is there still some sequence conservation in regions of functional and structural importance? It has been known for some time that in very divergent proteins it is common for the peripheral regions to have different conformations and, therefore, no significant sequence relationships. These peripheral regions can comprise half or more of the structure.[3] This means that the questions raised above only have meaning for the parts of related proteins that retain the same conformation.

To answer these questions for two superfamilies of four-helix bundle proteins, we determined the regions that retain the same conformation and examined the residues found at equivalent sites in the sequences of the structures and of their sequence homologues. Although there are differences in the details, the general results from the two analyses are very similar and present a coherent picture.

In the proteins discussed here just over a quarter of the sites that form the CC have medium or strong conservation. All but a few of these occur at sites that form the buried spine of each helix, i.e. at sites i, $i+3$, $i+7$, $i+10$, etc. or i, $i+4$, $i+7$, $i+11$, etc. These residues pack together at the centre of the structure with homologous residues making homologous contacts. They do not include all the buried residues. In different CCs, 35%, or a little more, of residues are buried but only 25%, or a little less, have medium or strong conservation and occupy sites on the helix spine.



A'-C'

A'-D'

B'-C'

B'-D'

C'-D'

Figure 11. Helical contact nets for the CC of the long and short chain cytokines together. See key to Figure 5.

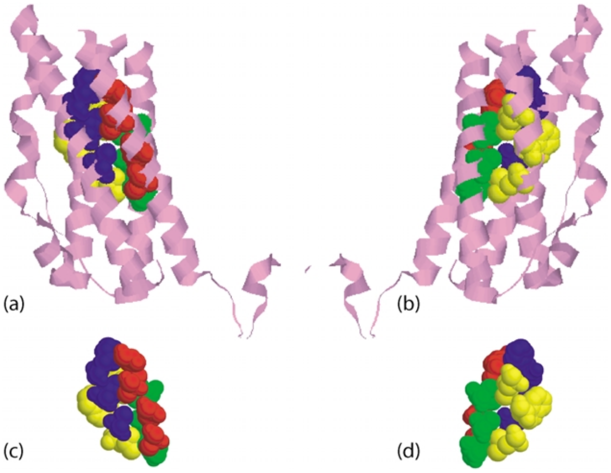


Figure 12. Representation of the 13 conserved sites involved in making multiple inter-helical contacts within placental lactogen (PDB-ID: 1f6f) a long chain cytokine. Figures (a) and (b) are front and back views of a ribbon representation with the 13 positions shown by space filling and colour coded according to Figure 1. Figures (c) and (d) are the same views as in Figures (a) and (b) but only the 13 positions are shown.

In the cytochromes there are eight additional sites of medium, strong or absolute conservation which, together with many of the spine residue, are involved in haem binding.

On the surface of the cytochromes most of the sites are in the *sn* category, i.e. they conserve residues that are hydrophilic or neutral. On the surfaces of the cytokines the proportion of conserved sites is smaller. This may be because their surfaces are involved in binding cell surface receptors and different cytokines have different specific receptors.

The high sequence variability of the four-helical cytokine superfamily and fold could be due to at least two reasons. First, four-helical cytokines have been duplicated and selected to create new and altered functions, i.e. the ability to bind to different receptor subtypes. In higher eukaryotic species multiple cytokines coexist, therefore the divergence between them must be sufficient for them to have high affinity for their own receptor and no significant affinity for the receptors of the other cytokines. Secondly, the fact that many of these cytokines are receptor specific implies that some co-evolution of cytokine and receptor has taken place as they are under fewer constraints than proteins that have a cofactor/substrate that does not evolve (see also Ref. 16). This can be seen in those cases where cytokine receptor recognition is actually species specific.

Given that long chain cytokines are approximately 180 amino acid residues in length and short chain cytokines are approximately 140 amino acid residues, we see that the CCs only comprise 55, 45 and 30% of the entire chain length for the long chain, the short chain and both families together, respectively (Table 8). The size of these CCs is typical of proteins that have diverged to the extent found in this family.[3] This means that the sites that conserve residue class correspond to approximately 25–15% of the entire chain length of the long and short chain sequences, and the sites with strong or medium conservation

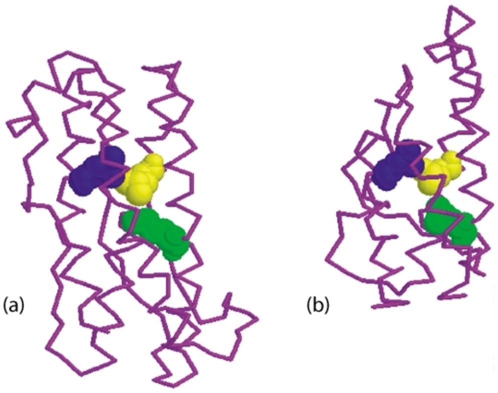


Figure 13. A space fill representation of the five strongly conserved sites involved in making multiple inter-helical contacts within one long chain structure and one short chain structure. Positions are coloured according to Figure 1. (a) Leukemia inhibitory factor (PDB-ID: 1lki), a long chain cytokine. (b) Interleukin-4 (PDB-ID: 1rcb), a short chain cytokine.

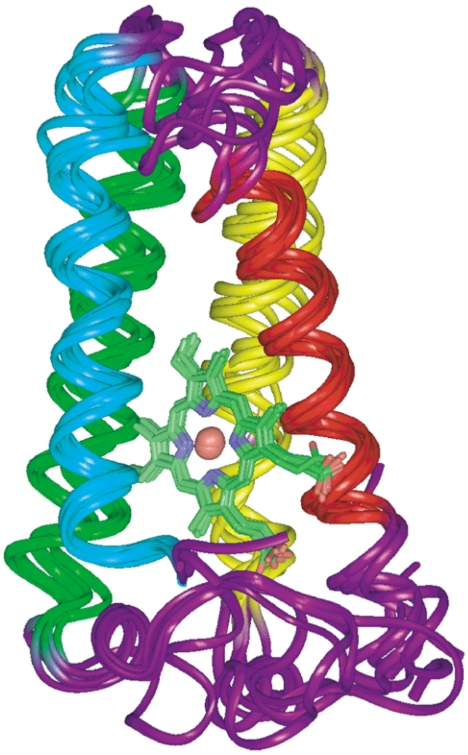
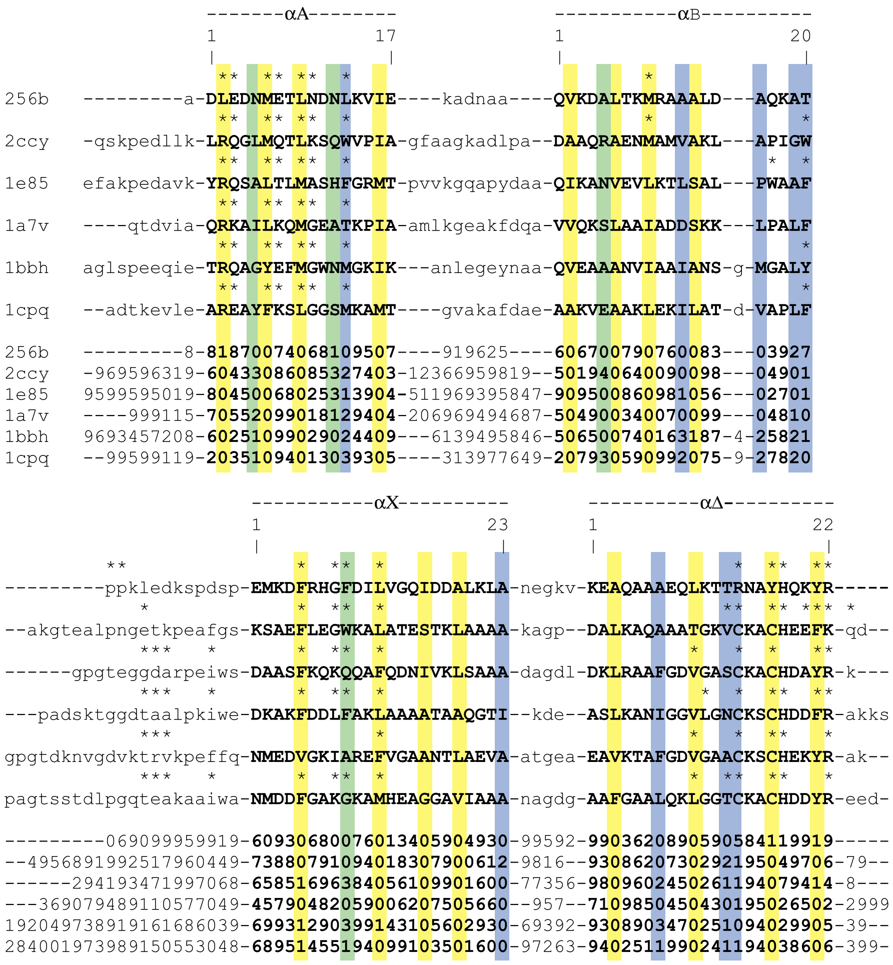


Figure 14. Multiple structure alignment of the six four-helical cytochromes. Non-core regions are coloured purple. Helices are coloured as in Figure 1. The haem co-factor is shown in the centre of the four helices and is coloured light-green (carbon), blue (nitrogen), red (oxygen) and orange (iron).



```
                  -------αA--------              -----------αB----------
                  1              17              1                    20
                  |               |              |                     |
                   **  ** **  *                          *
256b   ---------a-DLEDNMETLNDNLKVIE----kadnaa----QVKDALTKMRAAALD---AQKAT
                   **  ** **  *                          *             *
2ccy   -qskpedllk-LRQGLMQTLKSQWVPIA-gfaagkadlpa--DAAQRAENMAMVAKL---APIGW
                   **  ** **  *                                     *  *
1e85   efakpedavk-YRQSALTLMASHFGRMT-pvvkgqapydaa-QIKANVEVLKTLSAL---PWAAF
                   **  ** **  *
1a7v   ----qtdvia-QRKAILKQMGEATKPIA-amlkgeakfdqa-VVQKSLAAIADDSKK---LPALF
                   **  ** **  *                                        *
1bbh   aglspeeqie-TRQAGYEFMGWNMGKIK---anlegeynaa-QVEAAANVIAAIANS-g-MGALY
                   **  ** **  *                                        *
1cpq   --adtkevle-AREAYFKSLGGSMKAMT----gvakafdae-AAKVEAAKLEKILAT-d-VAPLF

256b   ---------8-81870074068109507----919625----606700790760083---03927
2ccy   -969596319-60433086085327403-12366959819--501940640090098---04901
1e85   9599595019-80450068025313904-511969395847-909500860981056---02701
1a7v   ----999115-70552099018129404-206969494687-504900340070099---04810
1bbh   9693457208-60251099029024409---6139495846-506500740163187-4-25821
1cpq   --99599119-20351094013039305----313977649-207930590992075-9-27820

                      -----------αX----------       ----------αΔ----------
                      1                    23       1                   22
                      |                     |       |                    |
         **               *  **  *                               *  **  **
---------ppkledkspdsp-EMKDFRHGFDILVGQIDDALKLA-negkv-KEAQAAAEQLKTTRNAYHQKYR-----
            *             *  **  *                              **  ** *** *
--akgtealpngetkpeafgs-KSAEFLEGWKALATESTKLAAAA-kagp--DALKAQAAATGKVCKACHEEFK-qd--
            ***   *       *  **  *                           *   *  *   **
------gpgteggdarpeiws-DAASFKQKQQAFQDNIVKLSAAA-dagdl-DKLRAAFGDVGASCKACHDAYR-k---
            ***   *       *  **  *                            *  *  **  **
---padsktggdtaalpkiwe-DKAKFDDLFAKLAAAATAAQGTI--kde--ASLKANIGGVLGNCKSCHDDFR-akks
            ***                  *                           *   *  **   *
gpgtdknvgdvktrvkpeffq-NMEDVGKIAREFVGAANTLAEVA-atgea-EAVKTAFGDVGAACKSCHEKYR-ak--
            ***   *       *  **  *                           *  **  **   *
pagtsstdlpgqteakaaiwa-NMDDFGAKGKAMHEAGGAVIAAA-nagdg-AAFGAALQKLGGTCKACHDDYR-eed-

---------069099959919-60930680076013405904930-99592-9903620890590584119919-----
--4956891992517960449-73880791094018307900612-9816--9308620730292195049706-79--
------294193471997068-65851696384056109901600-77356-9809602450261194079414-8---
---369079489110577049-45790482059006207505660--957--7109850450430195026502-2999
192049738919161686039-69931290399143105602930-69392-9308903470251094029905-39--
284001973989150553048-68951455194099103501600-97263-9402511990241194038606-399-
```

Figure 15. Structure-based sequence alignment and solvent ASA of the four-helical cytochromes. See legend to Figure 3. Haem-binding positions are marked with an asterisk in the line above the sequence.

Table 7. Conservation at core positions within the cytochromes

	Helix			
Site	A	B	C	D
1	*sn* 90	*sn* 81	KEDN 100	*sn* 100
2	RD 81	*bn* 90	*bn* 86	*sn* 100
3	*sn* 95	*sn* 86	*sn* 100	VLIMF 95
4	*sn* 95	*sn* 90	KED 81	*sn* 100
5	*bn* 90	*sn* 100	FY 90	*n* 81
6	LMFY 95	AVLI 90	*sn* 95	*sn* 100
7	*sn* 100	*sn* 90	*sn* 95	VLIMFY 90
8	–	*sn* 90	*sn* 86	*sn* 100
9	VLIM 86	LIMF 81	*bn* 86	*sn* 100
10	*sn* 90	–	*sn* 86	VLI 90
11	–	*sn* 95	*sn* 90	G 81
12	*sn* 95	*bn* 90	VLMF 90	*sn* 100
13	LMFWY 90	*bn* 90	*sn* 81	*sn* 90
14	*sn* 90	*sn* 81	*sn* 100	C 90
15	–	*sn* 81	*sn* 100	*sn* 100
16	VLIM 95	*bn* 100	AVI 81	AST 81
17	–	–	*sn* 86	C 90
18		*sn* 95	*sn* 95	H 100
19		*bn* 86	AVLM 95	EDQ 90
20		FYW 90	–	*sn* 100
21			*sn* 100	YF 100
22			–	RK 100
23			AVLI 100	

correspond to approximately 10–15%. Even at these sites, a certain degree of variation is still permitted and if we further restrict the sites to those that have strong conservation (i.e. those sites showing one of a small subset of residues with similar properties) we are left with only 11, eight and five sites within the long chain, short chain and both families together, respectively. This equates to only 6, 6 and 3% of the average chain lengths. The small proportion of sites that are actually conserved in the two families explains why simple sequence homology detection methods fail, in most cases, to detect any relationship between proteins belonging to these families.

The cytochromes have more extensive conservation than is found in the cytokines. The CC comprises just over 65% (Table 8) of the cytochrome sequences, and within the CC, there are very few sites of no conservation. However, the percentage of strong and medium conserved positions that are involved in making multiple inter-helical contacts and/or are involved in binding haem when calculated is very similar to that of the four-helical cytokines.

This work extends our current understanding of the relationships between the sequences of

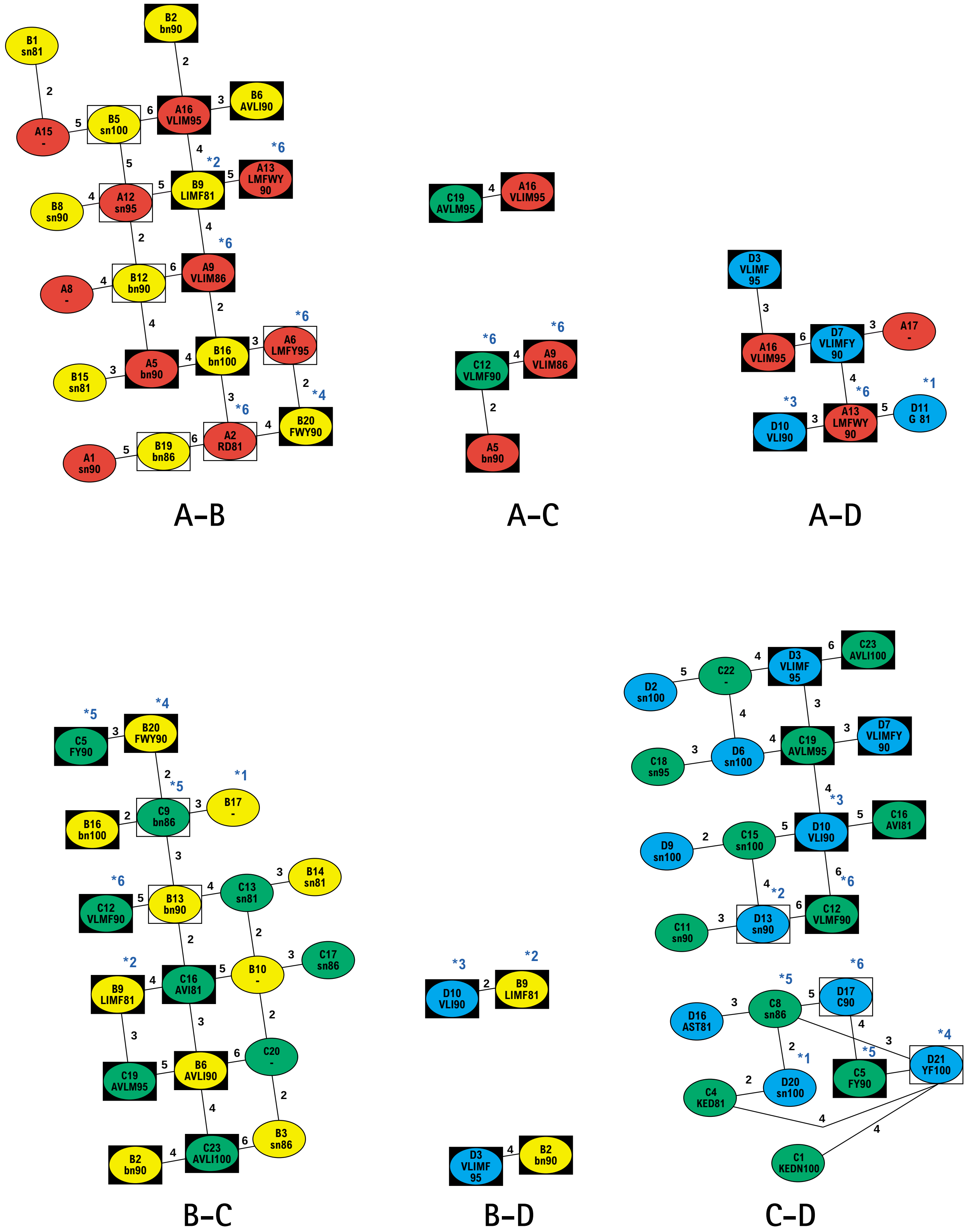


Figure 16. Helical contact nets for the CC of the cytochromes. See key to Figure 5. Haem-binding positions are marked by an asterisk shown in blue to the top right of that position. The number next to the asterisk represents the number of structures in which this position binds haem.

evolutionarily related proteins and demonstrates how homology may not be easily detectable by simple sequence comparisons. In particular, the different degree of conservation in the sequences of two structurally similar superfamilies, the four-helical cytokines and the cytochromes c′-b_{562}, shows how a signal in the sequence is only present in those proteins, the cytochromes, that are

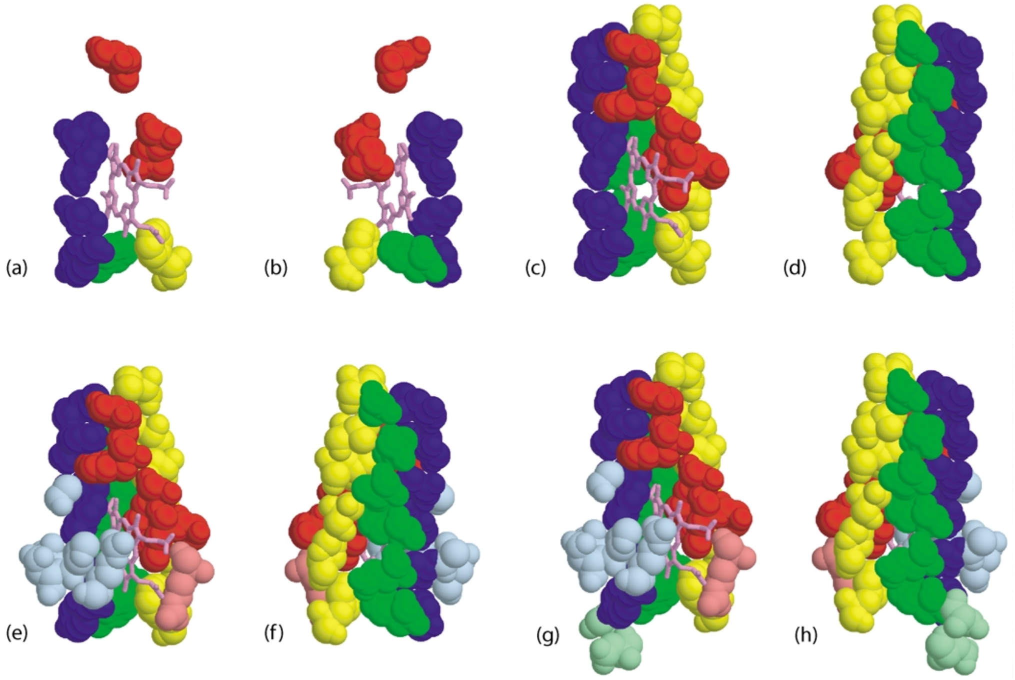


Figure 17. Space fill representations of the strong and medium conserved positions within AXCP (PDB-ID: 1e85) a four-helical cytochrome. Positions are coloured according to the helix they are in using colours as in Figure 1. The haem co-factor is shown in violet. Figures (a) and (b) are front and back views showing the eight positions of strong *b* conservation involved in making multiple inter-helical contacts or contacting the haem. Figures (c) and (d) are as in Figures (a) and (b) with additional ten positions of medium *b* conservation and five *bn* conserved positions. Figures (e) and (f) are as in Figures (c) and (d) with the additional six sites of strong or medium *n* or *s* sites that are involved in packing around the haem. These six sites are within helices A and D and are coloured in pink and light-blue, respectively. Figures (g) and (h) are as in Figures (e) and (f) plus the two strong or medium conserved sites at the end of the C helix that make inter-helical contacts with one other helix and are coloured here in light-green.

Table 8. Proportion of the structures formed by the common core and conserved positions

Superfamily	Family	Average chain length in amino acids	Number of residues in common core (% of average chain length)	Number of sites with some conservation (% of average chain length)	Number of sites with strong/ medium conservation involved in multiple inter-helical contacts (% of average chain length)
Cytokines	Long	180	101 (56)	70 (39)	23 (13)
	Short	140	66 (47)	44 (31)	15 (11)
	Long and short	160	48 (30)	29 (18)	13 (8)
Cytochromes	c′ and b_{562}	124	82 (66)	74 (60)	23 (19)

constrained by the necessity of binding the haem cofactor. Conversely, the cytokines, and their specific receptor proteins, have co-evolved, thus their sequences have diverged beyond the point of recognition by commonly used homology detection methods. These findings stress the necessity of exploiting, in such difficult cases, structural information in the comparison processes.

Acknowledgements

We thank Arthur Lesk for the use of PINQ and Julian Gough for the use of SUPERFAMILY and his alignment statistics program. We also thank colleagues for useful discussions and feedback. VM was supported by the EC (TMR Marie Curie Research Training Scheme) contract B104CT985056. Whilst revisions were carried out EEH was supported by the Department of Energy (DE-AC03-76SF00098).

References

1. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270.
2. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin: II. Some

relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678.

3. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
4. Bajaj, M. & Blundell, T. (1984). Evolution and the tertiary structure of proteins. *Annu. Rev. Biophys. Bioeng.* **13**, 453–492.
5. Sprang, S. R. & Bazan, J. F. (1993). Cytokine structural taxonomy and mechanisms of receptor engagement. *Curr. Opin. Struct. Biol.* **3**, 815–827.
6. Bazan, J. F. (1990). Haemopoietic receptors and helical cytokines. *Immunol. Today*, **11**, 350–354.
7. Betts, M. J., Guigo, R., Agarwal, P. & Russell, R. B. (2001). Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution. *EMBO J.* **20**, 5354–5360.
8. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
9. Nicola, N. A. & Hilton, D. J. (1998). General classes and functions of four-helix bundle cytokines. *Advan. Protein Chem.* **52**, 1–65.
10. Gadina, M., Hilton, D., Johnston, J. A., Morinobu, A., Lighvani, A., Zhou, Y. J. *et al.* (2001). Signaling by type I and II cytokine receptors: ten years after. *Curr. Opin. Immunol.* **13**, 363–373.
11. Ihle, J. N. (1995). Cytokine receptor signalling. *Nature*, **377**, 591–594.
12. Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215–250.
13. Crick, F. H. C. (1953). The packing of alpha-helices: simple coiled coils. *Acta Crystallog.* **6**, 689–697.
14. Walshaw, J. & Woolfson, D. N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427–1450.
15. Denesyuk, A. I., Zav'yalov, V. P. & Korpela, T. (1994). Modelling of interhelical contacts in interferons-beta, -gamma, and dimeric interleukin-5. *Biochem. Biophys. Res. Commun.* **201**, 1401–1405.
16. Rozwarski, D. A., Gronenborn, A. M., Clore, G. M., Bazan, J. F., Bohm, A., Wlodawer, A. *et al.* (1994). Structural comparisons among the short-chain helical cytokines. *Structure*, **2**, 159–173.
17. Lovejoy, B., Cascio, D. & Eisenberg, D. (1993). Crystal structure of canine and bovine granulocyte-colony stimulating factor (G-CSF). *J. Mol. Biol.* **234**, 640–653.
18. Somers, W., Stahl, M. & Seehra, J. S. (1997). 1.9 Å crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling. *EMBO J.* **16**, 989–997.
19. Robinson, R. C., Grey, L. M., Staunton, D., Vankelecom, H., Vernallis, A. B., Moreau, J. F. *et al.* (1994). The crystal structure and biological function of leukemia inhibitory factor: implications for receptor binding. *Cell*, **77**, 1101–1116.
20. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. (1998). Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* **277**, 1111–1128.
21. Elkins, P. A., Christinger, H. W., Sandowski, Y., Sakal, E., Gertler, A., de Vos, A. M. & Kossiakoff, A. A. (2000). Ternary complex between placental lactogen and the extracellular domain of the prolactin receptor. *Nature Struct. Biol.* **7**, 808–815.
22. McDonald, N. Q., Panayotatos, N. & Hendrickson, W. A. (1995). Crystal structure of dimeric human ciliary neurotrophic factor determined by MAD phasing. *EMBO J.* **14**, 2689–2699.
23. Zhang, F., Basinski, M. B., Beals, J. M., Briggs, S. L., Churgay, L. M., Clawson, D. K. *et al.* (1997). Crystal structure of the obese protein leptin-E100. *Nature*, **387**, 206–209.
24. Deller, M. C., Hudson, K. R., Ikemizu, S., Bravo, J., Jones, E. Y. & Heath, J. K. (2000). Crystal structure and functional dissection of the cytostatic cytokine oncostatin M. *Struct. Fold Des.* **8**, 863–874.
25. Syed, R. S., Reid, S. W., Li, C., Cheetham, J. C., Aoki, K. H., Liu, B. *et al.* (1998). Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature*, **395**, 511–516.
26. Diederichs, K., Boone, T. & Karplus, P. A. (1991). Novel fold and putative receptor binding site of granulocyte-macrophage colony-stimulating factor. *Science*, **254**, 1779–1782.
27. Wlodaver, A., Pavlovsky, A. & Gustchina, A. (1992). Crystal structure of human recombinant interleukin-4 at 2.25 Å resolution. *FEBS Letters*, **309**, 59–64.
28. Milburn, M. V., Hassell, A. M., Lambert, M. H., Jordan, S. R., Proudfoot, A. E., Graber, P. & Wells, T. N. (1993). A novel dimer configuration revealed by the crystal structure at 2.4 Å resolution of human interleukin-5. *Nature*, **363**, 172–176.
29. Savvides, S. N., Boone, T. & Andrew Karplus, P. (2000). Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots. *Nature Struct. Biol.* **7**, 486–491.
30. Jiang, X., Gurel, O., Mendiaz, E. A., Stearns, G. W., Clogston, C. L., Lu, H. S. *et al.* (2000). Structure of the active core of human stem cell factor and analysis of binding to its receptor kit. *EMBO J.* **19**, 3192–3203.
31. Brandhuber, B. J., Boone, T., Kenney, W. C. & McKay, D. B. (1987). Three-dimensional structure of interleukin-2. *Science*, **238**, 1707–1709.
32. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
33. Feng, Z. K. & Sippl, M. J. (1996). Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.* **1**, 123–132.
34. Lesk, A. M. (1986). *Integrated Access to Sequence and Structural Data. Biosequences: Perspectives and User Services in Europe* (Saccone, C., ed.), EEC, Bruxelles.
35. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
36. Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.
37. Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.
38. Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* **278**, 457–479.
39. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.
40. Bartsch, R. G. (1978). In *The Photosynthetic Bacteria* (Clayton, R. K. & Sistram, W. R., eds), pp. 249–280, Plenum Press, New York.
41. Ambler, R. P., Bartsch, R. G., Daniel, M., Kamen, M. D., McLellan, L., Meyer, T. E. & Van Beeumen, J.

(1981). Amino acid sequences of bacterial cytochromes c′ and c-556. *Proc. Natl Acad. Sci. USA*, **78**, 6854–6857.

42. Finzel, B. C., Weber, P. C., Hardman, K. D. & Salemme, F. R. (1985). Structure of ferricytochrome c′ from *Rhodospirillum molischianum* at 1.67 Å resolution. *J. Mol. Biol.* **186**, 627–643.
43. Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. D. & Matthew, F. S. (1981). Improvement of the 2.5 Å resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* **148**, 427–448.
44. Weber, P. C., Salemme, F. R., Mathews, F. S. & Bethge, P. H. (1981). On the evolutionary relationship of the 4-alpha-helical heme proteins. The comparison of cytochrome b562 and cytochrome c′. *J. Biol. Chem.* **256**, 7702–7704.
45. Ren, Z., Meyer, T. & McRee, D. E. (1993). Atomic structure of a cytochrome c′ with an unusual ligand-controlled dimer dissociation at 1.8 Å resolution. *J. Mol. Biol.* **234**, 433–445.
46. Tahirov, T. H., Misaki, S., Meyer, T. E., Cusanovich, M. A., Higuchi, Y. & Yasuoka, N. (1996). High-resolution crystal structures of two polymorphs of cytochrome c′ from the purple phototrophic bacterium *Rhodobacter capsulatus*. *J. Mol. Biol.* **259**, 467–479.
47. Shibata, N., Iba, S., Misaki, S., Meyer, T. E., Bartsch, R. G., Cusanovich, M. A. *et al.* (1998). Basis for monomer stabilization in *Rhodopseudomonas palustris* cytochrome c′ derived from the crystal structure. *J. Mol. Biol.* **284**, 751–760.
48. Lawson, D. M., Stevenson, C. E., Andrew, C. R. & Eady, R. R. (2000). Unprecedented proximal binding of nitric oxide to heme: implications for guanylate cyclase. *EMBO J.* **19**, 5661–5671.
49. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.
50. Dobbs, A. J., Anderson, B. F., Faber, H. R. & Baker, E. N. (1996). Three-dimensional structure of cytochrome c′ from two Alcaligenes species and the implications for four-helix bundle structures. *Acta Crystallog. sect. D*, **52**, 356–368.

Appendix: Four-helical cytokines, the detection of their homology and the alignment of their sequences by automatic procedures

With only marginal exceptions, conventional sequence comparison methods neither detect the homology of the four-helical cytokines nor accurately align their sequences. The homology of the four-helical cytokines was inferred from the inspection of their structures (see references in the main text) and is supported by the similarity of their functions and gene structures. Here we have aligned their sequences using structural information. There are a number of web-based servers that use sophisticated procedures both attempt detection of homology and for the production of sequence and structure-based alignments. The homologies and sequence align-ments we have described here for the cytokines can be used to get a rough indication of how successful automated servers are for determining evolutionary relationships and producing accurate sequence alignments of this protein superfamily. A full assessment of these servers would of course require tests on many different protein superfamilies.

The sequences of the 15 four-helical cytokine structures (see the main text) used in this work were submitted to the servers listed in Table A1 to see how many matches were made both between proteins of the same family (long chain or short chain) and between proteins belonging to the two different families (Table A1). The sequence alignments produced by these servers and those available from databases of sequence alignments were also compared to those described here.

Homology detection

The methods chosen were: FASTA[A1-1] as a good example of a widely used conventional sequence homology detection method; SUPERFAMILY[A1-2] an implementation of the SAM-T99[A1-3] HMM procedure which is optimised on SCOP superfamilies; 3dPSSM,[A1-4] Fugue,[A1-5] FFAS,[A1-6] Bioinbgu[A1-7] and mGenThreader[A1-8] (the latter five were chosen as they perform well in blind tests such as CAFASP2[A1-9] and Livebench[A1-10]). These methods take a variety of approaches—some incorporate sequence-based profiles, some include structure profiles based on SCOP and others take a threading approach. We will not give here accounts of these procedures: they can be found in articles that describe the individual methods[A1-2–A1-8] and are reviewed in the CAFASP2 and Livebench papers.[A1-9,A1-10]

Each of the methods requires a significance threshold—when available we used those suggested by the authors of the different methods. For FASTA, SUPERFAMILY and FFAS we take matches with *e*-value ≤ 0.01 as significant. For 3D-PSSM our *e*-value cut-off includes anything up to 1.0. Fugue gives its matches as either CERT, LIKELY or MARG, we include all three of these categories (*Z*-score ≥ 4.7) in our analysis. For the Bioinbgu server we used a threshold cut-off of 25 for significant matches. This is the score at which Livebench found that the server starts to give false positives (Ref. A1-10 and a personal communication). For mGenThreader we include matches made with either CERT, HIGH or MED confidence levels (i.e. *e*-value ≤ 0.1).

Results of the submission of the 15 cytokine sequences to the different programs and servers are shown in Table A1. Four of the procedures (FASTA, SUPERFAMILY, FFAS and Bioinbgu) detect only a small number of the 210 possible relationships: between four and eight and most of these involve reciprocal pairs, i.e. 1alu to 1bgc and 1bgc to 1alu (see Table A1). Fugue finds a total of 21 relationships (including six pairs of reciprocal matches). 3dPSSM and mGenThreader do better

Table A1.

A. *Summary of the results produced by automatic servers in matching four-helical cytokine sequences*

Query sequences	Target sequences														
	Long chain family								Short chain family						
Protein (PDB-ID)	G-CSF (1bgc)	CNTF (1cnt)	LIF (1lki)	Leptin (1ax8)	IL-6 (1alu)	OSM (1evs)	PL (1f6f)	GH (1hgu)	Flt3 (1ete)	EPO (1eer)	GM-CSF (2gmf)	IL-4 (1rcb)	IL-5 (1hul)	IL-2 (3ink)	SCF (1scf)
G-CSF (1bgc)		GT 3d FU	3d FU	GT	FA SF GT 3d FU FF BI	GT 3d FU	GT	GT 3d				GT			
CNTF (1cnt)	3d FU		SF GT 3d FU FF		3d FU	FU				GT FU		GT			
LIF (1lki)		3d FF FU			3d	FU FF BI 3d									
Leptin (1ax8)	3d										GT				
IL-6 (1alu)	FA SF GT 3d FU FF BI	GT 3d	GT 3d			GT	GT	GT		GT				FU	
OSM (1evs)	FU	FU	GT 3d FU FF									GT			
PL (1f6f)	GT 3d	GT 3d	GT 3d FU		GT 3d	GT FU		FA SF GT 3d FU FF BI				GT 3d			
GH (1hgu)	GT 3d	GT 3d	GT 3d	GT 3d	GT 3d	GT	FA SF GT 3d FF BI								
Flt3 (1ete)												GT			GT
EPO (1eer)															GT
GM-CSF (2gmf)													3d		
IL-4 (1rcb)			3d						GT						
IL-5 (1hul)															
IL-2 (3ink)												3d			
SCF (1scf)									GT	GT		GT			

B. *Summary table of the number of matches made by each different method*[a]

	Intra-family matches		Inter-family matches		
Method	Long chain total (reciprocal)	Short chain total (reciprocal)	Long to short	Short to long	Total
GT	27 (6)	7 (3)	7	0	41
3d	27 (8)	2 (0)	1	1	31
FU	19 (6)	0	2	0	21
FF	8 (4)	0	0	0	8
SF	5 (2)	0	0	0	5
BI	5 (2)	0	0	0	5
FA	4 (2)	0	0	0	4
Total possible matches	56 (28)	42 (21)	56	56	210

[a] FA: FASTA; SF: SUPERFAMILY; GT: mGenThreader; 3d: 3dPSSM; FU: Fugue; FF: FFAS; BI: Bioinbgu.

finding 31 and 41 relationships, respectively. 3dPSSM has eight pairs of reciprocal relationships for the long chain family. mGenThreader has six pairs of reciprocal matches for the long chain family and three for the short chain family. We note, however, that 3dPSSM incorporates automatically produced structural alignments of proteins according to their SCOP classification, which are then searched against and therefore this method has a bias towards the identification of relationships between sequences belonging to some SCOP families and superfamilies when the sequences of one of the structures is represented within one of their multiple structural alignments (see below).

As is apparent from Table A1, all procedures performed best in the detection of long chain to long chain matches. Other types of matches, between the sequences of different short chain family members and between long and short chain sequences, were only found by mGenThreader (14 matches), 3dPSSM (four matches) and Fugue (two matches).

Alignment accuracy

We compared the alignments produced by these methods and alignments available from public databases with those produced by our work. Two points need to be made about these comparisons.

First, the alignments produced by the automatic procedures do not distinguish regions that have the same structural conformation from those where the conformation differs. In this assessment we compared only the alignments of the sequence regions that our work has shown to have the same conformation; i.e. the regions in bold, upper case letters in Figures 3 and 4 from the main part of this paper; the regions of lower case letters correspond to regions that do not have a meaningful structural alignment and are ignored. Secondly, many of the alignments obtained from the various servers are pairwise rather than multiple alignments and there are frequently inconsistencies between different pairs.

Three of the servers that detect homology produced alignments for between 21 and 41 pairs of sequences (see Table A1) and each of their resulting pairwise alignments were analysed:

(i) Fugue[A1-5] made 21 alignments: 19 between long chain sequences and two between a long and a short chain sequence. Four of the matches agree with our alignment for all four helices; ten agree for three helices, five agree for two helices (these five include one of the long–short chain matches) and the remaining two only agree for one helix. Of the six reciprocal matches none of the alignments agreed for all four helices either to ours or to one another.

(ii) 3dPSSM[A1-4] makes 27 long chain–long chain matches, two short chain–short chain and two inter-family matches. Three of the long chain sequences (1bgc, 1hgu and 1f6f) make matches to either six or all of the remaining seven long chain sequences. For the 31 alignments, 15 are in complete agreement with ours (six of these represent three of the reciprocal pairs), 13 agree for three helices (including the two short–short chain matches, one long–short match and one of the reciprocal pairs), two agree for two helices and finally the short–long match only agrees for one helix.

(iii) mGenThreader[A1-8] produced matches between 41 pairs of sequences. Of these 27 are long chain–long chain matches. Four sequences: 1bgc, 1alu, 1f6f and 1hgu match six or all of the other seven long chain sequences. The other four sequences make one or no matches (Table A1). In 14 cases the alignment is the same as ours for all four helices; in nine cases it is the same for three helices and in four cases it is the same for two helices. There are six pairs of reciprocal matches, of which four have an alignment the same as ours for one or both matches. In the other reciprocal pairs different mismatches are made. Seven of the 41 matches are short chain to short chain and these agree with our alignments for all, or for three, of the helices. The long chain to short chain matches give a most interesting result. The three matches made to 1eer or 2gmf (see Table A1) agree with ours for all four helices. The four matches made to 1rcb completely disagree with ours; agree with each other, and involve a concerted seven-residue shift in the alignment of each helix. Inspection of this second alternative shows that it is plausible in terms of the fit of main-chain atoms as measured by RMSD but, unlike our alignment, it is not as well supported as the first alternative by other structural features such as solvent ASA or conservation of contacts.

Various sets of stored alignments were analysed and compared and the results are described below:

(i) The CATH dictionary of homologous superfamilies[A1-11] includes alignments for five of the long chain (1bgc, 1lki, 1hgu, 1cnt and 1ax8) and three of the short chain (1hul, 1rcb and 3ink) four-helical cytokines. Of the long chain sequence alignments, three (1bgc, 1lki and 1cnt) agreed for helices A and C, two of these also agreed for helix B but all three misaligned helix D. The alignments of the other two structures also agreed with our alignment of helices A and B, but the other two helices were not in agreement with our alignment, nor with the other three CATH alignments. The short chain family alignments have no agreement with ours and there are even places where they align non-corresponding helices with one another. This means therefore that the long–short alignments have no correspondence to ours.

(ii) Homstrad[A1-12] contains for the four-helical cytokines three different small alignments. One, called cyto, is of three different PDB entries for a single protein G-CSF. The second is called Hormone and aligns the different PDB chains for the growth hormone protein. The third alignment is for 1lki and 1evs (LIF_OSM) and this alignment agrees with ours entirely.

(iii) FSSP[A1-13] has alignments for 13 of the structures discussed here, and for the other two structures it has different PDB representatives (1huw instead of 1hgu and 1exz instead of 1scf). For the long chain family all alignments agree with ours except for helices A and D of one structure (1huw). For the short chain family they agree for five out of the seven structures entirely. The two other structures, however, are completely different (2gmf and 1hul). The FSSP alignment of both the long and short chain families together contains several inconsistencies with respect to its alignments for the separate families. Although several parts of the FSSP global alignment match up with our alignment, some parts of it have jumped out of the positions they have in the separate alignments for the two families. As an example, IL-4 (1rcb) was among those that agreed with our alignment for the short chain family, however, within the global alignment it is wrongly aligned to the other short chain sequences for all four helices.

(iv) DALI[A1-14] contains a multiple alignment of four long chain (1ax8, 1alu, 1bgc and 1huw) and four short chain cytokine sequences (1eer, 1rcb, 3ink and 2gmf) in DC_3_35_56. In this alignment three of the long chain sequences (1ax8, 1alu and 1huw) agree for all four helices. For the alignment of the four short chain sequences, three agree for helix A (1eer, 1rcb and 3ink). Two pairs of helices agree with our alignment of helices B, C and D. However, the alignment of the two pairs with each other is different to ours. For the DALI alignment of the long and short chain sequences there are several regions that match ours, i.e. the three long chain structures and one short chain structure that agree for the A helix, the same three long chain structures and two different short chain sequences that agree for helices B, C and D. However, there are two copies of the protein G-CSF (1cd9 and 1bgc) that are not aligned to one another, nor do they align well to the closely related sequence of IL-6 (1alu).

(v) 3dPSSM[A1-4] incorporates within its method a pre-calculated fold library of 3dMSAs (multiple structural alignments) and the MSA representing the four-helical cytokines was analysed. Within this MSA the long chain alignment includes five of our eight structures (1cnt, 1alu, 1bgc, 1ax8 and 1lki) and one equivalent PDB chain (1huw instead of 1hgu). These six structures agree with our alignment for helices A and B but the alignment does not extend any further than this. For the short chain alignment five of the structures (3ink, 1eer, 2gmf, 1ete and 1hul) are included along with one equivalent PDB chain (1iar instead of 1rcb). For these structures the alignment of helices A and D agreed with ours for three structures (1eer, 1ete and 1hul). The alignment of helix B also agrees within structures 1eer and 1ete. For the comparison between the two families, helices A and B match between the long chain family and the two structures of the short chain family which agreed for these two helices also. The rest of the alignment is in disagreement with ours.

Overall, we see that the consensus of the results of the different automatic servers and databases supports the alignments of the four-helical cytokines that we derived using manual procedures. However, their inconsistent and partial nature mean that they could not have produced simply and directly the sequence alignments that are the basis of the work described here. It is clear, however, that the results produced by mGenThreader, 3dPSSM, Fugue and FSSP would have provided significant help to the manual procedures which were used to produce the structural alignments and to their verification.

References

A1. Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.

A2. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.

A3. Hughey, R. & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.* **12**, 95–107.

A4. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.

A5. Shi, J., Blundell, T. L. & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257.

A6. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232–241.

A7. Fischer, D. (2000). Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130.

A8. Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.

A9. Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. R. *et al.* (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **45**, 171–183. suppl. 5..

A10. Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. (2001). LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **45 (Suppl. 5)**, 184–191.

A11. Bray, J. E., Todd, A. E., Pearl, F. M., Thornton, J. M. & Orengo, C. A. (2000). The CATH dictionary of homologous superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.* **13**, 153–165.

A12. Mizuguchi, K., Deane, C. M., Blundell, T. L. & Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471.

A13. Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure–structure alignment of proteins. *Nucl. Acids Res.* **24**, 206–209.

A14. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucl. Acids Res.* **29**, 55–57.

Edited by J. Thornton

(Received 28 February 2002; received in revised form 10 June 2002; accepted 19 June 2002)