Protein Engineering vol.12 no.7 pp.563–571, 1999

The immunoglobulin fold family: sequence analysis and 3D structure comparisons

D.M.Halaby, A.Poupon[1] and J.-P.Mornon

Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS UMR C7590
Universités Pierre et Marie Curie (P6) et Denis Diderot (P7), Tour 16,
Case 115, 4 Place Jussieu, 75252 Paris cedex 05, France.
E-mail: poupon@lmcp.jussieu.fr

[1]To whom correspondence should be addressed

Fifty-two 3D structures of Ig-like domains covering the immunoglobulin fold family (IgFF) were compared and classified according to the conservation of their secondary structures. Members of the IgFF are distantly related proteins or evolutionarily unrelated proteins with a similar fold, the Ig fold. In this paper, a multiple structural alignment of the conserved common core is described and the correlation between corresponding sequences is discussed. While the members of the IgFF exhibit wide heterogeneity in terms of tissue and species distribution or functional implications, the 3D structures of these domains are far more conserved than their sequences. We define topologically equivalent residues in the Ig-like domains, describe the hydrophobic common cores and discuss the presence of additional strands. The disulfide bridges, not necessary for the stability of the Ig fold, may have an effect on the compactness of the domains. Based upon sequence and structure analysis, we propose the introduction of two new subtypes (C3 and C4) to the previous classifications, in addition to a new global structural classification. The very low mean sequence identity between subgroups of the IgFF suggests the occurrence of both divergent and convergent evolutionary processes, explaining the wide diversity of the superfamily. Finally, this review suggest that hydrophobic residues constituting the common hydrophobic cores are important clues to explain how highly divergent sequences can adopt a similar fold.

Keywords: comparative study/hydrophobic core/immunoglobulin fold/multiple sequence alignment/protein folding

Introduction

In a previous paper, we highlighted the considerable variety of the immunoglobulin fold family (IgFF) (Halaby and Mornon, 1998), which contains all sequences or structures having an Ig-like fold (and not only sequences having detectable similarities with immunoglobulins). In fact, many of the structures compared in this paper have no detectable sequence similarity with each other. Many other authors have explored the sequences (e.g. Jones, 1993; Harris and Bajorath, 1995; Smith and Xue, 1997) or the structures (Taylor, 1986; Bork *et al*., 1994; Harpaz and Chothia, 1994) of Ig-like domains. Here we focus on the structural features of the immunoglobulin fold which has been identified in proteins without either apparent sequence identity or functional similarity.

Classical Ig-like domains are composed of 7–10 β-strands, distributed between two sheets with typical topology and connectivity. However, recent structural analyses revealed additional secondary structure elements in this classical scaffold, such as additional strands [SOD, DPA (PapD), RSY] or helices (CTM, HCY) (see Table I for nomenclature). In this paper, we report an all-against-all structural comparison of 52 distinct Ig-like domains (1326 pairs), having less than 55% pairwise sequence identity. The structures considered were selected from the PDB (Table I). Structural-based sequence analysis and comparison of structural features were performed in order to characterize sequence–structure compatibility. Our observations led us to propose a new structural classification within the IgFF.

Methods

The proteins considered in this comparative study possess one or several Ig-like domains (Table I). Secondary structure assignments of the known 3D structures were used to superimpose 52 Ig-like domains found in 34 distinct proteins. Visualization of the structures, distance calculations and observation of the hydrogen bonds were performed using the INSIGHTII 2.3.0 and INSIGHTII 95.0 programs (Biosym, San Diego, CA). A structural phylogenetic tree was built using the program MOLPHY (Saitou and Nei, 1987). Solvent accessible surfaces (SAS) were computed using the algorithm of Lee and Richards (Lee and Richards, 1971; Richards, 1985).

Superimpositions within each group were generated by various automatic programs [COMPOSER (Sutcliffe *et al*., 1987), DALI (Holm and Sanders, 1993) and COMPARER (Sali and Blundell, 1990)] and checked manually. Protein pairs belonging to different groups were superimposed manually using a pseudo-iterative method comparable to that of Hubbard and Blundell (1987). For this type of comparison the use of automatic programs was impossible because of the differences in the orientation of the two sheets and the presence in some structures of extra strands.

Results and discussion

Ig-like domains have similar general shapes, but differ significantly in their sizes, owing to high variability of the loops (Figure 1). While a classical domain contains about 100 residues (Igs), smaller ones (74–90 residues) have been observed in bacterial Ig-like proteins and in several Ig-related molecules (CD2, CD4). Large decorations within loops, sometimes including extra domains, are found in hemocyanin (238 amino acids), transcription factor NFkB (201 amino acids) and cytochrome *f* (214 amino acids).

Topohydrophobic positions were first studied on a bank of fold families, in which all families contain only homolog proteins of known 3D structure with pairwise identity lower than 55% (Poupon and Mornon, 1998). Investigating the PDB, 445 families were constituted, 153 of which contain two or more structures. Only one of these families contains more than 16 members: the immunoglobin superfamily. Consequently, the study of this family appears essential to a better understanding

Table I. Ig-like domains with known 3D structures used in the comparative study[a]

1. Abbreviation		2. Protein	3. Species	4. 3D	5. Pdb	6. Chain	7. Sequence	8. Domain	
								Constant	Variable
Actinoxantin and actinoxantin-like									
ACX	I	Actinoxantin	Sg	2.0	1ACX	—	P01551	ACX	
AKP	I	Kedarcidin	Ac	NMR	1AKP	—	P41249	AKP	
MCM	I	Macromycin	Sma	1.5	2MCM	—	P01549	MCM	
NCO	I	Neocarzinostatin	Sc	1.8	1NCO	—	P01550	NCO	
Bacterial domains									
BGL	I	β-Galactosidase	Ec	2.5	1BGL	—	P00722	BGL	
CELD	I	Cellulase D	Clt	1.9	1CLC	—	P04954	CELD	
ChiA	I	Chitinase A	Ec	2.3	1CTN	—	P07254	ChiA	
CYG	I	Cyclodextrin glycosyl transferase	Bc	2.5	1CYG	—	P31797	CYG	
DPA	I	PapD	Ec	2.5	3DPA	—	P15319	DPA1, DPA2	
Cytokine receptors									
GHR	E	Growth hormone receptor	H	2.8	3HHR	C	P10912	GHR1, GHR2	
HFT	E	Human tissue factor	H	1.69	1HFT	—	P13726	HFT1, HFT2	
Extracellular matrix									
FNA	E	Fibronectin type III	H	NMR + 1.8	1FNA	—	P02751	FNA	
TEN	E	Tenascin	H	1.8	1TEN	—	P24821	TEN	
TLK	E	Telokine	Rb	2.8	1TLK	—	P29294	TLK	
TNM	E	Titin	H	NMR	1TNM	—	X69490	TNM	
Immunoglobulins and related									
CH1	E	Ig CH1 domain	H	2FAB	H	P01857	CH1		
CH2	E	Ig CH2 domain	H	2.9	1FC1	D	P01860	CH2	
CH3	E	Ig CH3 domain	H	2.9	1FC1	D	P01860	CH3	
CL	E	Ig CL domain	H	1.9	2FB4	—	P01842	CL	
α3	E	HLA chain A (α3)	H	3.5	1HLA	A	P10313	ALPHA	
B2MG	E	β2 microglobulin	H	3.5	1HLA	M	P01884	B2MG	
CD2H	E	CD2 (human)	H	2.5	1HNF	—	P06729	CD2HC	CD2HV
CD2R	E	CD2 (rat)	R	2.8	1HNG	A	P08921	CD2RC	CD2RV
CD4	E	CD4 (D1D2)	H	2.2	3CD4	—	P01730	2CD4	1CD4
CD4	E	CD4 (D3D4)	R	2.8	1CID	—	P05540	4CD4	3CD4
CD8	E	CD8	H	2.6	1CD8	—	P01732		CD8
TCR	E	T cell receptor	M	1.7	1BEC	—	P01852	TCRC	TCRV
VL	E	Ig variable domain (λ chain)	H	2.0	7FAB	L	P01703		VL
VK	E	Ig variable domain (κ chain)	H	2.0	1REI	A	P01593		VK
VH	E	Ig variable domain (H chain)	H	2.0	7FAB	H	P01825		VH
Others									
CFB	E	Neuroglian	D	2.0	1CFB	—	P20241	CFB1, CFB2	
CTM	I	Cytochrome *f*	P	2.3	1CTM	—	P36438		CTM
GGT	I + E	Coagulation factor XIII	H	2.65	1GGT	A	P00488	GGT2, GGT3	
GOG	E	Galactose oxidase	Dd	1.9	1GOG	—	Q01745	GOG	
HCY	E	Hemocyanin	Pi	3.2	1HCY	—	P04254	HCY	
NCD	E	Cadherin	M	1.9	1NCI	A	P15116	NCD	
NFkB	I	Transcription factor	H	2.6	1SVC	P	P19838	NFK1, NFK2	
RSY	I	Synaptotagmin I	R	1.9	1RSY	—	P21707	RSY	
SOD	I	Cu, Zn superoxide dismutase	B	2.0	2SOD	O	P00441	SOD	
VCA	E	Vascular cell adhesion molecule	H	1.8	1VCA	A	P19320	VCA1, VCA2	

[a] Column 1: abbreviation used for the studied proteins (I, intracellular; E, extracellular). Column 2: protein name. Column 3: species. Ac, Actinomycetes; B, bovine; Bc, *Bacillus circulans*; Clt, *Clostridium thermocellum*; D, *Drosophila*; Dd, *Dactylium dendroides*; Ec, *Escherichia coli*; H, human; M, mouse; P, plant; R, rat; Rb, rabbit; Sc, *Streptomyces globisporus*; Sma, *Streptomyces macromomyceticus*. Column 4: method used to determine the 3D structure: NMR or crystallography (resolution in Å). Column 5: PDB identifier. Column 6: polypetide chain in the PDB file. Column 7: sequence code in the Swissprot bank (except for titin, X69490 in the GenBank and galactose oxidase, Q01745 in the GenPep). Column 8: Ig-like domains found in the protein.

of the relationships between topohydrophobic positions, size and diversity of a considered protein family, but also it brings new information on the IgFF.

3D superimposition and sequence analysis

Structures were compared by finding the optimal superimposition between the pair considered while avoiding a unique reference structure with which all domains would be compared. The definition of a ‘mean structure’ for the IgFF does not make sense because of the great diversity in structure and function. Such problems in structure superimposition have been widely studied (Lesk *et al*., 1986; Godzik *et al*., 1993; Johnson *et al*., 1993; Maiorov and Crippen, 1994) and cannot be solved by the existing automatic methods. Variations in the lengths of regular secondary structures (strands) and variable loop regions make the alignment of the whole domains impossible (Figure 1). A common core was defined among the 35 structurally equivalent residues, localized in the six strands common to the 52 structures (strands A, B, C, E, F and G). The fourth strand, numbered according to its appearance in the sequence, cannot be aligned for all the domains studied owing to its variable localizations in sheet I (strand D) or in sheet II (strand C′). The variable domains contain both strands C′ and D (Figure 2) and sometimes a ninth strand C″.

The superimposition of the structural core leads to root mean square deviations (r.m.s.d.) between equivalent Cα of

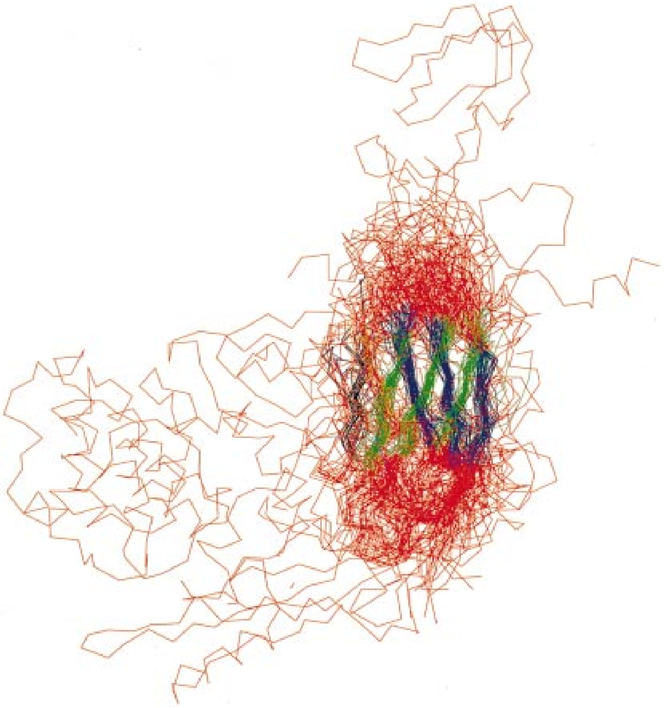


Fig. 1. Exhaustive comparison of 52 domains of the IgFF. The minimal common structural core, formed by about 25% of the domain residues, corresponds to the well superimposed median region. The Ig-like domains have a similar 3D shape, but show wide variability in the length of the loop regions (red). Sheet I is in green, sheet II in blue. The fourth strand of constant domains belongs to sheet I (black) or to sheet II (yellow).

<3.9 Å. The highest values are observed between highly distant domains (such as NCD and SOD or CD4 and DPA) or between domains presenting local divergence of strand conformation, especially in the external strand A′. Figure 3 illustrates the relations between the r.m.s.d. and the sequence identity for each domain pair. Two interesting regions are surrounded: the first one (I) corresponds to high sequence identity and, as expected, low r.m.s.d. (pairs of Ig domains, ACX-like domains), while the second one (II) corresponds to low sequence identity and low r.m.s.d., illustrating once more the fact that structural similarity is not necessarily related to sequence similarity.

The sequences reveal high divergence in the whole set, as shown in Figure 4. As no chemically conserved residues had been detected, it is difficult to propose a consensus sequence from the alignment. In other words, no sequence signature of the Ig fold can be defined. Surprisingly, no sequence identity is observed in the common core for about 2% of the pairs of domains compared. However, sequences can be divided into three subtypes, two with a sequence signature and one that allows residue substitutions but possesses a conserved hydrophobic core.

The first sequence signature is observed for the Ig constant and variable domains (Figure 2). The second concerns the fibronectin type III domains (Fn3). Conserved residues are found in each internal strand, except strand E. This strand shows the best structural fit among the β-strands, but no significant sequence identity can be detected. In several positions, only substitutions conserving hydrophobicity or aliphatic/aromatic amino acid balance are allowed. The more striking conservations of amino acid type are those found in positions A3, B1, C3, E5 and F5 (78, 78, 86, 88 and 67% of VILF residues, respectively). Table II summarizes the different hydrophobic cores found in the IgFF.

Some of the positions of a particular fold are always (in all the proteins adopting this fold) occupied by hydrophobic amino acids. These positions were shown to be key markers of the fold (Poupon and Mornon, 1998, 1999). It has also been demonstrated that the properties of these conserved hydrophobic positions can be enlarged to all the positions occupied by strong hydrophobic amino acids (VILFMYW) in more than 75% of the representatives of the fold and occupied by non-strong loop former amino acids in the remaining representatives (ACTQERK); these positions are called topohydrophobic positions. In the case of the IgFF domains, only position C3 is topohydrophobic for the complete superfamily. A3, B1 and E5 are occupied by strong hydrophobic amino acids in more than 75% of the sequences but are sometimes occupied by amino acids having strong propensities for loops (Callebaut *et al*., 1997) that cannot be integrated in topohydrophobic positions as they were defined. F5 is not topohydrophobic because this position is occupied by strong hydrophobic amino acids in only 67% of the sequences. This result illustrates the great diversity of this super-family.

In order to investigate the similarity between the different groups defined in the IgFF (Figure 2), the topohydrophobic positions in each of them were determined (Table II). A score, which depends on the number of topohydrophobic positions in each group (n_i for the group i, n_j for the group j) and on the number of topohydrophobic positions common to two groups ($N_{i,j}$), was defined:

$$S_{ij} = \frac{N_{i,j}}{n_i} + \frac{N_{i,j}}{N_j}$$

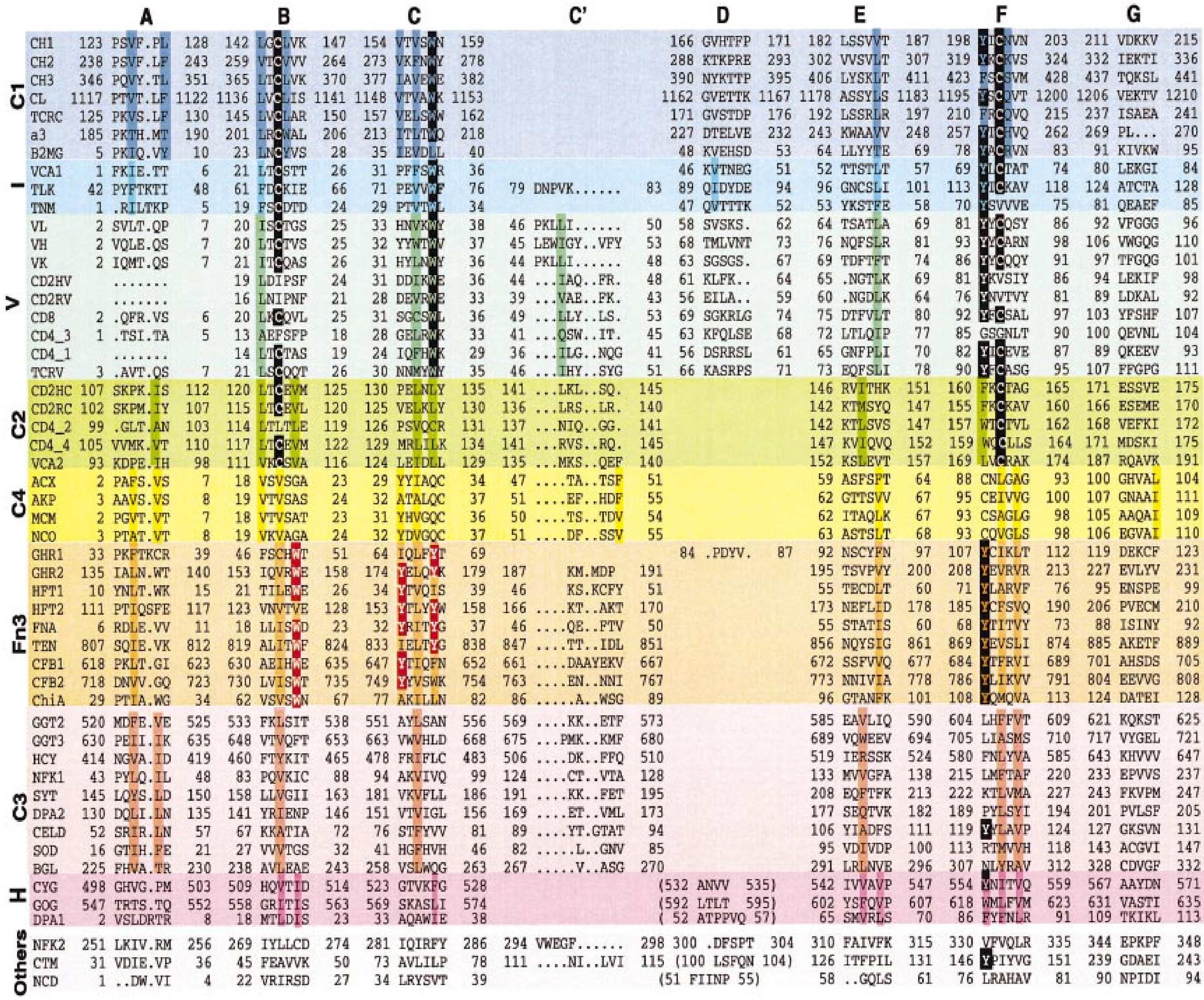
and computed for all possible pairs of groups (Table III). In each group the number of topohydrophobic positions is close to what is expected for an homogeneous family (7–10% of the total amino acids), except for the groups ‘Others’ (which was already known to be heterogeneous) and V.

Hydrogen bonds

Hydrogen bonds in the the Ig-like domains are mainly conserved in the structural common core (Figure 5) (Kabsch and Sander, 1983). However, some domains deviate from this general scheme, in that not all possible H-bonds are formed or alternatively are established between non-equivalent residues. In many domains, the external strands A and G present geometrical distorsions known as β-bulges (Richardson, 1977; Chan *et al*., 1993), which lead to an imperfect general H-bond network. Hydrogen bonds have been extensively shown to be important for the stability and dynamics of protein domains (Pfuhl *et al*., 1997; Vogt and Argos, 1997) and probably play an important role in Ig domains.

Burial of conserved hydrophobic residues

Based on the structural alignment a typical hydrophobic core of the Ig fold can be described. Solvent accessibility surfaces (SAS) were calculated for each residue (Figure 6A). These calculations show that the structural core can be virtually divided into three parts: buried positions (SAS < 20 Å^2), internal positions (20 Å^2 $<$ SAS < 50 Å^2) and exposed positions (SAS > 50 Å^2). The most buried positions are B3, C3 and F3 (SAS < 10 Å^2). For strands B, C, E and F, amino acids with side chains pointing towards the interior of the protein have SAS < 20 Å^2 and amino acids with side chains



			A			B			C			C′			D			E			F			G	
C1	CH1	123	PSVF.PL	128	142	LGCLVK	147	154	VTVSWN	159				166	GVHTFP	171	182	LSSVVT	187	198	YICNVN	203	211	VDKKV	215
	CH2	238	PSVF.LF	243	259	VTCVVV	264	273	VKFNWY	278				288	KTKPRE	293	302	VVSVLT	307	319	YKCKVS	324	332	IEKTI	336
	CH3	346	PQVY.TL	351	365	LTCLVK	370	377	IAVEWE	382				390	NYKTTP	395	406	LYSKLT	411	423	FSCSVM	428	437	TQKSL	441
	CL	1117	PTVT.LF	1122	1136	LVCLIS	1141	1148	VTVAWK	1153				1162	GVETTK	1167	1178	ASSYLS	1183	1195	YSCQVT	1200	1206	VEKTV	1210
	TCRC	125	PKVS.LF	130	145	LVCLAR	150	157	VELSWW	162				171	GVSTDP	176	192	LSSRLR	197	210	FRCQVQ	215	237	ISAEA	241
	a3	185	PKTH.MT	190	201	LRCWAL	206	213	ITLTWQ	218				227	DTELVE	232	243	KWAAVV	248	257	YTCHVQ	262	269	PL...	270
	B2MG	5	PKIQ.VY	10	23	LNCYVS	28	35	IEVDLL	40				48	KVEHSD	53	64	LLYYTE	69	78	YACRVN	83	91	KIVKW	95
I	VCA1	1	FKIE.TT	6	21	LTCSTT	26	31	PFFSWR	36				46	KVTNEG	51	52	TTSTLT	57	69	YLCTAT	74	80	LEKGI	84
	TLK	42	PYFTKTI	48	61	FDCKIE	66	71	PEVVWF	76	79	DNPVK......	83	89	QIDYDE	94	96	GNCSLI	101	113	YICKAV	118	124	ATCTA	128
	TNM	1	.RILTKP	5	19	FSCDTD	24	29	PTVTWL	34				47	QVTTTK	52	53	YKSTFE	58	70	YSVVVE	75	81	QEAEF	85
V	VL	2	SVLT.QP	7	20	ISCTGS	25	33	HNVKWY	38	46	PKLLI......	50	58	SVSKS.	62	64	TSATLA	69	81	YYCQSY	86	92	VFGGG	96
	VH	2	VQLE.QS	7	20	LTCTVS	25	32	YYWTWV	37	45	LEWIGY..VFY	53	68	TMLVNT	73	76	NQFSLR	81	93	YYCARN	98	106	VWGQG	110
	VK	2	IQMT.QS	7	21	ITCQAS	26	31	HYLNWY	36	44	PKLLI......	48	63	SGSGS.	67	69	TDFTFT	74	86	YYCQQY	91	97	TFGQG	101
	CD2HV		....		19	LDIPSF	24	31	DDIKWE	36	44	...IAQ..FR.	48	61	KLFK..	64	65	.NGTLK	69	81	YKVSIY	86	94	LEKIF	98
	CD2RV		....		16	LNIPNF	21	28	DEVRWE	33	39	...VAE..FK.	43	56	EILA..	59	60	.NGDLK	64	76	YNVTVY	81	89	LDKAL	92
	CD8	2	.QFR.VS	6	20	LKCQVL	25	31	SGCSWL	36	49	...LLY..LS.	53	69	SGKRLG	74	75	DTFVLT	80	92	YFCSAL	97	103	YFSHF	107
	CD4_3	1	.TSI.TA	5	13	AEFSFP	18	28	GELRWK	33	41	...QSW..IT.	45	63	KFQLSE	68	72	LTLQIP	77	85	GSGNLT	90	100	QEVNL	104
	CD4_1		....		14	LTCTAS	19	24	IQFHWK	29	36	...ILG..NQG	41	56	DSRRSL	61	65	GNFPLI	70	82	YICEVE	87	89	QKEEV	93
	TCRV	3	.AVT.QS	7	21	LSCQQT	26	30	NNMYWY	35	46	...IHY..SYG	51	66	KASRPS	71	73	EQFSLI	78	90	YFCASG	95	107	FFGPG	111
C2	CD2HC	107	SKPK.IS	112	120	LTCEVM	125	130	PELNLY	135	141	...LKL..SQ.	145				146	RVITHK	151	160	FKCTAG	165	171	ESSVE	175
	CD2RC	102	SKPM.IY	107	115	LTCEVL	120	125	VELKLY	130	136	...LRS..LR.	140				142	KTMSYQ	147	155	FKCKAV	160	166	ESEME	170
	CD4_2	99	.GLT.AN	103	114	LTLTLE	119	126	PSVQCR	131	137	...NIQ..GG.	141				142	KTLSVS	147	157	WTCTVL	162	168	VEFKI	172
	CD4_4	105	VVMK.VT	110	117	LTCEVM	122	129	MRLILK	134	141	...RVS..RQ.	145				147	KVIQVQ	152	159	WQCLLS	164	171	MDSKI	175
	VCA2	93	KDPE.IH	98	111	VKCSVA	116	124	LEIDLL	129	135	...MKS..QEF	140				152	KSLEVT	157	169	LVCRAK	174	187	RQAVK	191
C4	ACX	2	PAFS.VS	7	18	VSVSGA	23	29	YYIAQC	34	47	....TA..TSF	51				59	ASFSPT	64	88	CNLGAG	93	100	GHVAL	104
	AKP	3	AAVS.VS	8	19	VTVSAS	24	32	ATALQC	37	51	....EF..HDF	55				62	GTTSVV	67	95	CEIVVG	100	107	GNAAI	111
	MCM	2	PGVT.VT	7	18	VTVSAT	23	31	YHVGQC	36	50	....TS..TDV	54				62	ITAQLK	67	93	CSAGLG	98	105	AAQAI	109
	NCO	3	PTAT.VT	8	19	VKVAGA	24	32	YDVGQC	37	51	....DF..SSV	55				63	ASTSLT	68	93	CQVGLS	98	106	EGVAI	110
Fn3	GHR1	33	PKFTKCR	39	46	FSCHWT	51	64	IQLFYT	69				84	.PDYV.	87	92	NSCYFN	97	107	YCIKLT	112	119	DEKCF	123
	GHR2	135	IALN.WT	140	153	IQVRWE	158	174	YELQYK	179	187	KM.MDP	191				195	TSVPVY	200	208	YEVRVR	213	227	EVLYV	231
	HFT1	10	YNLT.WK	15	21	TILEWE	26	34	YTVQIS	39	46	KS.KCFY	51				55	TECDLT	60	71	YLARVF	76	95	ENSPE	99
	HFT2	111	PTIQSFE	117	123	VNVTVE	128	153	YTLYYW	158	166	....KT..AKT	170				173	NEFLID	178	185	YCFSVQ	190	206	PVECM	210
	FNA	6	RDLE.VV	11	18	LLISWD	23	32	YRITYG	37	46	....QE..FTV	50				55	STATIS	60	68	YTITVY	73	88	ISINY	92
	TEN	807	SQIE.VK	812	819	ALITWF	824	833	IELTYG	838	847	....TT..IDL	851				856	NQYSIG	861	869	YEVSLI	874	885	AKETF	889
	CFB1	618	PKLT.GI	623	630	AEIHWE	635	647	YTIQFN	652	661	....DAAYEKV	667				672	SSFVVQ	677	684	YTFRVI	689	701	AHSDS	705
	CFB2	718	DNVV.GQ	723	730	LVISWT	735	749	YYVSWK	754	763	....EN..NNI	767				773	NNIVIA	778	786	YLIKVV	791	804	EEVVG	808
	ChiA	29	PTIA.WG	34	62	VSVSWN	67	77	AKILLN	82	86	.....A..WSG	89				96	GTANFK	101	108	YQMQVA	113	124	DATEI	128
C3	GGT2	520	MDFE.VE	525	533	FKLSIT	538	551	AYLSAN	556	569	....KK..ETF	573				585	EAVLIQ	590	604	LHFFVT	609	621	KQKST	625
	GGT3	630	PEII.IK	635	648	VTVQFT	653	663	VWVHLD	668	675	...PMK..KMF	680				689	VQWEEV	694	705	LIASMS	710	717	VYGEL	721
	HCY	414	NGVA.ID	419	460	FTYKIT	465	478	FRIFLC	483	506	....DK..FFQ	510				519	IERSSK	524	580	FNLYVA	585	643	KHVVV	647
	NFK1	43	PYLQ.IL	48	83	PQVKIC	88	94	AKVIVQ	99	124	....CT..VTA	128				133	MVVGFA	138	215	LMFTAF	220	233	EPVVS	237
	SYT	145	LQYS.LD	150	158	LLVGII	163	181	VKVFLL	186	191	....KK..FET	195				208	EQFTFK	213	222	KTLVMA	227	243	FKVPM	247
	DPA2	130	DQLI.LN	135	141	YRIENP	146	151	VTVIGL	156	169	....ET..VML	173				177	SEQTVK	182	189	PYLSYI	194	201	PVLSF	205
	CELD	52	SRIR.LN	57	67	KKATIA	72	76	STFYVV	81	89	....YT.GTAT	94				106	YIADFS	111	119	YYLAVP	124	127	GKSVN	131
	SOD	16	GTIH.FE	21	27	VVVTGS	32	41	HGFHVH	46	82	.....L..GNV	85				95	VDIVDP	100	113	RTMVVH	118	143	ACGVI	147
	BGL	225	FHVA.TR	230	238	AVLEAE	243	258	VSLWQG	263	267	.....V..ASG	270				291	LRLNVE	296	307	NLYRAV	312	328	CDVGF	332
H	CYG	498	GHVG.PM	503	509	HQVTID	514	523	GTVKFG	528				(532	ANVV	535)	542	IVVAVP	547	554	YNITVQ	559	567	AAYDN	571
	GOG	547	TRTS.TQ	552	558	GRITIS	563	569	SKASLI	574				(592	LTLT	595)	602	YSFQVP	607	618	WMLFVM	623	631	VASTI	635
	DPA1	2	VSLDRTR	8	18	MTLDIS	23	33	AQAWIE	38				(52	ATPPVQ	57)	65	SMVRLS	70	86	FYFNLR	91	109	TKIKL	113
Others	NFK2	251	LKIV.RM	256	269	IYLLCD	274	281	IQIRFY	286	294	VWEGF......	298	300	.DFSPT	304	310	FAIVFK	315	330	VFVQLR	335	344	EPKPF	348
	CTM	31	VDIE.VP	36	45	FEAVVK	50	73	AVLILP	78	111	....NI..LVI	115	(100	LSFQN	104)	126	ITFPIL	131	146	YPIYVG	151	239	GDAEI	243
	NCD	1	..DW.VI	4	22	VRIRSD	27	34	LRYSVT	39				(51	FIINP	55)	58	..GQLS	61	76	LRAHAV	81	90	NPIDI	94

Fig. 2. Multiple structural alignment of the Ig-like sequences. This alignment is based on structural conservation of β-strands. Most pairwise identities are lower than 25%. Numbers represent the sequences of regular secondary structures. Strand limits are not those found in the 3D structures, but those common to all domains. The domains can be classified into distinct subtypes on the basis of the similarity of their hydrophobic cores. Identical residues are found in positions B3, C5, F3 (black background, in C1, I, V and C2 subtypes), in B5, C1, C5 (Fn3 subtype, red background) and in F1 (black background, in C1, I, V, Fn3). Topohydrophobic positions are indicated for each group in darker color. No sequence signature can be observed for the whole family. However, positions B1 and F1 are occupied by aliphatic and aromatic residues, respectively. Residues A3, B1, B3, C3, E5 and F5 are mostly hydrophobic (see Figure 6A). Fragments of sequences in parentheses are those of H domains, where the fourth strand is located between the two sheets.

pointing towards the exterior of the protein have SAS between 20 and 50 Å^2. For other strands (A, C′, D and G), SAS of residues oriented towards the interior (between 20 and 50 Å^2)is still lower than that of residues oriented towards the exterior (>50 Å^2), but the values are higher.

These observations lead to internal (B, C, E and F) and external (A, C′, D and G) strands being defined. It is interesting that C3 is a topohydrophobic positions, B3 and F3 are topohydrophobic positions in three groups (C3, C4 and Fn3) and form disulfide bridges in two others (C1 and C2).

The hydrophobic core described here is a key signature with an impact on the structural behavior of the Ig fold.

The disulfide bonds

A dominant feature of canonical Ig domains is the disulfide bridge connecting strands B and F. However, in several members of the family, disulfide bonds, when they exist, involve two strands, a strand and a loop or two loops. The disulfide bridge of Ig molecules is invariably located between positions B3 and F3. Only ~0.5% of the sequences of the KABAT bank (Martin, 1996) lack these cysteine residues, with consequent loss of their biological activities (Proba *et al.*, 1997) in the same manner as Ig mutants with substitutions at the cysteine residues. However, a functional antibody lacking the disulfide bridge has been observed (Rudikoff and Pumphrey, 1986).

The usual method for identifying Ig-like domains consists in checking the length of the fragment between the two cysteine residues. Indeed, the number of residues within the disulfide bridge varies significantly as the domain type changes: 63–76 (V), 54–64 (C1), 38–43 (C2) and 44–51 (I) (Smith and Xue, 1997).

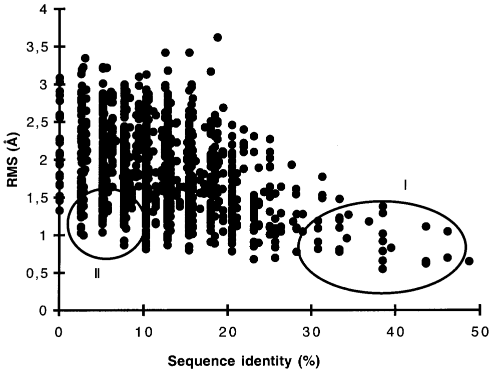


Fig. 3. Relation between divergence of sequence (% identity) and conservation of structure (r.m.s.d. values) in the 52 compared Ig-like domains (1326 pairs). The interesting regions are circled. Zone I corresponds to high identity and low r.m.s.d. (structural and sequence similarities, domains of C1 and C4 subtypes). Zone II corresponds to low identity and low r.m.s.d. (structural but no sequence similarities).

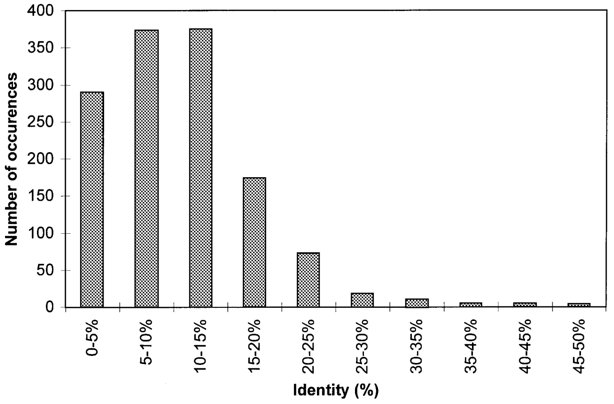


Fig. 4. Sequence identity between the 1326 pairwise compared domains. Sequence identities are frequently low (<25%). Most of the domains present only 5–10% sequence identity.

However, the disulfide bridge is not a common feature of the IgFF. Although expected to be important for the correct and stable folding of an Ig domain, many Ig-like domains lack the B3–F3 disulfide bond, raising the question of its actual involvment in the folding pathway. Furthermore, some proteins, such as TLK, NFK1, NFK2, BGL, CTM and GGT3, exhibit free cysteine residues. The impact of the classical disulfide bridge occurring within Ig-like domains can be evaluated as follows. (a) The absence of the disulfide bridge or the occurrence of a bridge in an atypical position explains the larger separation of the two sheets in comparison with classical domains, as reflected by the distance between the Cα positions B3 and F3 (6.1–7.1 and 7.2–11.7 Å for domains with and without the canonical disulfide bridge, respectively) (Table II). The role of this disulfide bridge in the compactness of the domain has been confirmed by mutagenesis of an Ig-variable domain and demonstrates the ability of the protein to accommodate the absence of the cysteine residues to maintain its fold (Proba *et al*., 1997). (b) In domains lacking the disulfide bridge, the cysteine residue is replaced by a strong hydrophobic residue, the side chains of which maintain the hydrophobic core formation.

In conclusion, the disulfide bond may have more of a functional than a structural role. The absence of this structural constraint in many domains may allow adaptation to specific biological functions or to particular structural features, such as the insertion of additional secondary structure in the domain, and may enhance the assembly of many Ig-like domains such as those of the fibronectin type.

Structural classification of the IgFF

Williams and Barclay (1988) divided classical Ig domains into three topological domain subtypes: C1 (constant 1), C2 (constant 2) and V (variable). The resolution of many structures of Ig-like domains has revealed new topological subtypes including subtype I (intermediate) (Harpaz and Chotia, 1994), S (switched) and H (hybrid) types (Bork *et al*., 1994). The present analysis is in agreement with these studies and extends the comparison to Ig-like domains possessing additional strands, such as the structures of SOD (superoxide dismutase), hemocyanin, DPA (PapD) domain 2 and cytochrome *f*. The only criterion required is the occurrence in the domain of a topology and connectivity similar to those of immunoglobulins (Halaby and Mornon, 1998). Domains that are distant in terms of angles between sheets, twists in some strands or difficult superimposition are also included in our study. The extension of the previous structural classifications to the newly identified structures, combined with a sequence analysis of the Ig-like domains led us to define two new subtypes: C3 (constant 3) and C4 (constant 4) (Table II). The discrimination of these two groups is justified by the differences in sequence (Fn3 and C4 have different hydrophobic cores, the only common feature between them being the presence of a tyrosine in position C1, Fn3 proteins have a tryptophan residue in position B5, an aromatic residue in position C5 and a tyrosine residue in position F1, and none of these is found in proteins of the C4 sub-family), together with structural characteristics (proteins of the C4 sub-family have two conserved disulfide bonds, none of them is found in the proteins of the Fn3 sub-family; proteins of the C4 sub-family have two β-strands forming a small sheet perpendicular to the two canonical ones, implicated in the active site) (Table II). This discrimination can also be justified by the values obtained for the $S_{i,j}$ parameters: the highest value is obtained when comparing the sub-families I and C1, the discrimination between these two being largely accepted, so it seems reasonable to split Fn3–C4 and Fn3–C3, as in both cases the value obtained is much lower.

The information contained in the structural distance matrix (r.m.s.d. values) is illustrated through hierarchical clustering [using the program MOLPHY (Saitou and Nei, 1987)] as shown in Figure 7. The distance between the 52 proteins studied, measured by the r.m.s.d. values, is coherent with the classification in subgroups. However, the tree reported in this study was established on the basis of structural similarity and should not be directly compared with trees constructed on the basis of sequence comparison. Cross-comparison of the 52 Ig-like domains reveals a coherent clustering into subclasses, which together with the sequence analysis results in a new classification of Ig-like domains.

An Ig-like domain invariably contains six strands, A, B, C, E, F and G, which constitute the common structural core. Burried amino acids of strands B, C, E and F constitute the common hydrophobic core. Strands A, C′, C″, D and G are the external strands. The presence or absence of these strands in a domain, except for strand G, determines its appearance

Table II. Structural classification of constant domains of the IgFF: structure and sequence characteristics of each group

Subtype	Proteins	Topology	Disulfide bonds	Conserved residues	Topohydrophobic positions	Structural characteristics	B3–F3 distances (Å)
C1	Igs, HLA, TCR	ABED/CFG	B3–F3	C5 (W), F1 (Y) B1, E5 (aliphatic residues), B3, F3 (C)	A3, A7, B1, B4, C1, C3, C5, E5, F1, F5	Symmetry of the two sheets relative to the molecule axis	6.1–7.1
C2	Igs-related molecules (CD2, CD4, . . .), cell adhesion molecules	ABE/C′CFG	B3–F3 Loop AB–strand G (2CD2) C5–F3 positions (2CD4)	Similar to C1 domains, except positions C5. Residue Y in F1 is replaced by an aromatic residue	A6, B1, B5, C3, C5, E3, F1	Small domains (74–90 aa)	6.5–6.6 except 2CD4 (9.8)
C3[a]	Bacterial domains[b] HCY, NFK1, RSY, SOD	ABE/C′CFG	No disulfide bond, or: —strand–loop bond (HCY, SOD) —loop–loop (HCY), strand–strand bond (DPA2)	No sequence signature. Conserved hydrophobic positions, a variation aliphatic/aromatic remain allowed	A3, A6, B3, C3, E3, F3, F5	Domains are less compact, no symmetry between the sheets. Cylindrical aspect of the domains	7.5–11.2
C4	Actinoxanthine-like domains	ABE/C′CFG	Loop–strand bonds	No sequence signature, but a hydrophobic core formed by C1 (Y), D4 (V or F), F5	A3, A6, B1, B3, C1, C3, C′11, E5, F3, F5, G5	Similar to C3 domains. A small third β-sheet perpendicular to the first two	7.5–9.4
Fn3[c]	Cytokines receptor[d] extracellular matrix, neuroglian, bacterial chitinase	ABE/C′CFG	No disulfide bond, or: —strand–loop bond (CFB1) —loop–loop (ChiA) —strand–strand (HFT2)	Hydrophobic core mainly aromatic: B5 (W) residue Y in C1, C5, F1	A3, B3, B5, C1, C3, C5, E5, F1, F3, F5	Similar to C2 domains.	7.7–9.2
H	GOG, CYG, DPA1	ABE (C′)CFG[e]	No disulfide bond	Similar to C3	A3, B3, B5, C3, C5, E3, E5, F1, F3, F5	Similar to C3, except the C′ strand	10.0–10.9

[a] The C3 and the H domains have similar hydrophobic core characteristics.
[b] Except the chitinase A domain, which is an Fn3 type.
[c] The hydrophobic core of the Fn3 domains is similar to that of the C1, C2 and V domains: it has aromatic residues in positions C5 and F1.
[d] The GHR1 domain has sequential similarities with the Fn3 domain (sequence signature in positions B5, C5, F1), and structural similarities with C1 domains (a disulfide bond and a C1 topology).
[e] The fourth strand C′ is localized between the two sheets.

Table III. Topohydrophobic scores[a]

	I	V	C2	C4	Fn3	C3	Others
C1	1.46	1.21	0.97	1.15	1.40	0.80	1.33
I		1.37	1.14	0.94	1.21	0.62	1.55
V			1.03	0.87	0.90	0.37	1.47
C2				0.70	0.97	0.62	0.93
C4					1.34	1.29	1.03
Fn3						1.33	1.07
	C3						0.33

[a] For each possible pair of groups in the IgFF, the score (defined in the text) has been computed. This score theoretically ranges from 0 to 2 for unrelated or fully related groups, respectively.

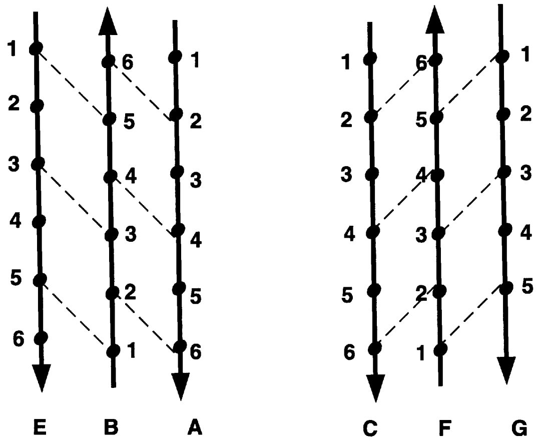


Fig. 5. General hydrogen bond diagram observed in the Ig-like domains. The hydrogen bonds between strands A and G are not shown, because they are less conserved between domains. Numbers represent residue positions, as shown in Figure 2.

in the C, V, I, S or H sets. The greatest variability occurs in the fourth strand, numbered as it appears in the sequence. This strand belongs to the first sheet (strand D, domain C1) or to the second sheet (strand C′, domain S: domains C2, C3, C4). Variable domains contain both strands D and C′ (Figure 6B). The H domains are hybrid forms between the C and S types, the fourth strand lying between the two sheets. Type I corresponds to domains presenting sequence signatures of the C1 domains (in positions B3, C5, F1 and F3) and structural features of variable domains (number and topology of strands). Table IV summarizes the different subtypes described here and their topologies.

As the number of distinct subclasses in the IgFF increases, many questions arise, such as how the subtypes are similar or which subtype could be the first domain from which different subclasses may have evolved. Structural and sequence considerations lead us to cluster the different subclasses into similar groups [(((((C1, V) C2) I) Fn3) C4) C3]. The pairs of compared subclasses are clustered in a manner so as to maximize the sequence identity and to minimize the r.m.s.d. values. From left to right, the sequence identity decreases between pair of subclasses and the r.m.s.d. values increase. The score used for the determination of the above classification is defined as follows: for the subgroups G1 and G2, the score *S*(G1,G2) is

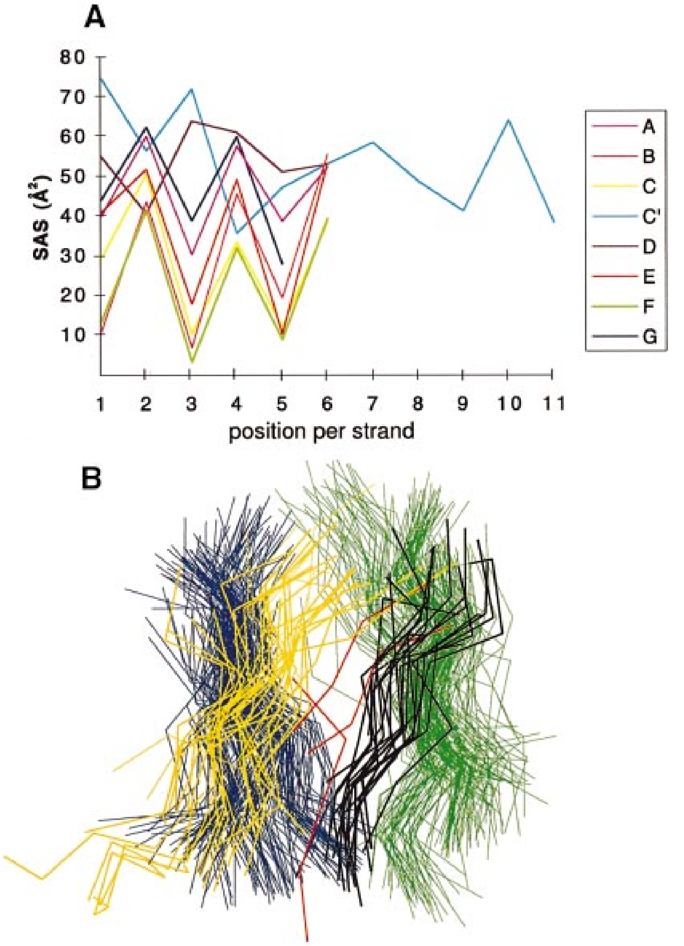


Fig. 6. (**A**) Average of solvent accessibility surfaces (SAS). The mean solvent accessibility surface for each position in each strand was computed using the algorithm of Lee and Richards (Lee and Richards, 1971; Richardson, 1977). (**B**) The fourth strand of constant domains. Sheet I is in green, sheet II in blue. The classification of the Ig-like domains partly depends on the occurrence of the fourth strand in sheet I or II: C1 domains (black, in sheet I), C2, C3, C4, S and Fn3 domains (yellow, in sheet II). The subtype H constitutes hybrid form between these subtypes with the fourth strand located between the two sheets (red).

$$S(\mathrm{G1,G2}) = \frac{1}{n_1 n_1}\left[\sum_{\substack{\mathrm{P1}\in \mathrm{G1}\\ \mathrm{P2}\in \mathrm{G2}}} S_{\mathrm{r.m.s.}}(\mathrm{P1,P2}) + \sum_{\substack{\mathrm{P1}\in \mathrm{G1}\\ \mathrm{P2}\in \mathrm{G2}}} S_{\mathrm{id}}(\mathrm{P1,P2})\right]$$

where n_1 and n_2 are the number of members in subgroups G1 and G2, respectively, with

$$S_{\mathrm{r.m.s.}}(\mathrm{P1,P2}) = \begin{cases} 0 \text{ if } 0 & \leqslant \mathrm{r.m.s.(P1,P2)} \leqslant 1\,\text{Å} \\ 1 \text{ if } \quad \text{Å} & < \mathrm{r.m.s.(P1,P2)} \leqslant 2\,\text{Å} \\ 2 \text{ if } 2\ \text{Å} & < \mathrm{r.m.s.(P1,P2)} \leqslant 3\,\text{Å} \\ 3 \text{ if } 3\ \text{Å} & < \mathrm{r.m.s.(P1,P2)} \end{cases}$$

$$S_{\mathrm{id}}(\mathrm{P1,P2}) = \begin{cases} 0 \text{ if } 0\% & \leqslant \mathrm{id(P1,P2)} \leqslant 10\% \\ 1 \text{ if } 10\% & < \mathrm{id(P1,P2)} \leqslant 25\% \\ 2 \text{ if } 25\% & < \mathrm{id(P1,P2)} \leqslant 35\% \\ 3 \text{ if } 35\% & < \mathrm{id(P1,P2)} \end{cases}$$

where r.m.s.(P1,P2) and id(P1,P2) are the root mean square deviation and the sequence identity, respectively, between the two proteins P1 and P2 belonging to groups G1 and G2.

Different hypotheses have been made mainly to explain how the primordial domain might have gained or lost a strand, leading to intermediate structures. Depending on the authors, the original domain might be the V domain (Williams and Barclay, 1988) or the C2 domain (Hunkapiller and Hood, 1989; Smith and Xue, 1997).

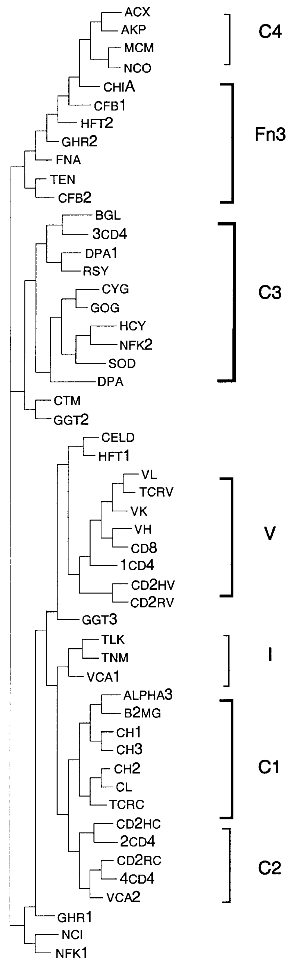


Fig. 7. Structural tree of the IgFF. Multiple cross-comparisons of the Ig-like domains led to a coherent clustering of the domains into subclasses. Comparison of this classification, based on structural criteria (r.m.s.d. values), with those derived from the sequence analysis led to a new classification of the IgFF as indicated on the right. At the bottom of the tree, the domains NCI, NFK1 form a separate cluster, owing to their particular characteristics (see text). Most proteins of a same cluster have similar functions (C1, C2, V) or unknown functions (C3).

Since several Ig-like domains did not cluster with any of the structural sets described above (NCD, CTM, NFK2), additional subclasses of the Ig fold must exist and should be more documented when new 3D structures are solved. The NCD differs from a V domain by the localization of strand A between the two sheets and the absence of hydrogen bonds between strands A and B. The CTM domain presents nine strands as a variable domain, but the connectivity between C and E strands is atypical: the topology of the CTM domain is AA′BCDC′EFG (instead of AA′BCC′DEFG for a variable domain). The second domain of NFK could be described as intermediate between variable domains (same number of strands) and constant domain (a maximum of 14% sequence identity with C4 domains and with bacterial chitinase within the whole superfamily).

Conclusion

In a previous paper, we showed that the immunoglobulin fold family (IgFF) comprises a heterogeneous group of proteins

Table IV. Topology of IgFF subclasses[a]

Type		A	A′	B	C	C′	C′	D	E	F	G	No. of strands
C1	(C)	+		+	+			+	+	+	+	7
C2	(S)	+		+	+	+			+	+	+	7
C3	(S)	+	(+)	+	+	+			+	+	+	7 (8)
C4	(S)	+		+	+	+			+	+	+	7
Fn3	(S)	+		+	+	+			+	+	+	7
V	(V)	+	+	+	+	+	(+)	+	+	+	+	8 (to 10)
I	(I)	+	+	+	+	(+)		+	+	+	+	8 (to 9)
H	(H)	+		+	+	+			+	+	+	7

[a]In column 1 is shown in parentheses the correspondence between the present classification and that of Bork *et al.* (1994). A + indicates the presence of a strand in the corresponding domain. A strand not systematically present in a domain is represented by (+).

sharing structural similarity but exhibiting a wide range of functions, species and tissue distribution. In this paper, 52 Ig-like domains found in the PDB were compared in order to define and characterize sequence and structural constraints of the Ig fold. The structure-based multiple alignment of the sequences revealed low overall sequence identity (often in the 5–15% range) and no functional relationship. Geometrical features, such as secondary structure, hydrogen bonds, disulfide bridges and solvent exposure, were compared through 1326 pairs of Ig-like domains.

Within the compared Ig-like domains, a few residues form the common core. As a general rule, two sequences which share at least 30% sequence identity are considered to fold very similarly (Chothia and Lesk, 1986; Schneider and Sander, 1991). The IgFF is remarkable in that most of the Ig-like domains display <10% sequence identity. Many studies have shown that the folding pattern of a protein is dependent not only on its sequence, but also implicitly on its overall amino acid composition (Nakashima *et al*., 1986; Chou, 1989) and that the size of the protein and the percentage of each amino acid can be used to predict the folding type. In the IgFF domains, most of the residues constituting the common core are, as expected, hydrophobic and are concentrated in a small number of conserved positions, probably responsible for maintenance of the Ig fold. Membership in this continually growing structural family requires specific interactions that stabilize the folded domains: (a) the formation of a typical hydrophobic core coded by the sequence; (b) the occurrence of specific tertiary interactions within the hydrophobic core; (c) in several subtypes, the introduction of disulfide bridges which influence the overall domain shape and also the symmetry between the two sheets.

Although these proteins retain a common fold, structural changes occur as their sequences diverge. Residue substitutions do not change the overall appearance of the β-strands. However, changes in H-bond spacing, twists of strands or in one sheet relative to the other are observed to accommodate the sequence variation. Here we emphasize for a large sample that Ig-like domains have more structural (r.m.s.d. between Cα always <3.9 Å) than sequence similarities (identity mainly <25%). The hydrophobic core probably has a major impact on the uniqueness and stability of the Ig fold. As a general rule, mutations are not disruptive, as we observe a conservation of the properties of amino acids (hydrophobic/hydrophilic) along the alignment. A 29-residue structural core is common to all of the 52 considered domains, defined by the strands B, C, E and F and by six additional residues belonging to strands C′ or D. The external strands A and G are more difficult to align owing to irregularities and distorsions in several domains. The β-bulges occurring in strand A in some domains lead to the appearance of an additional strand A′, such as in Ig-variable domains and many domains distantly related to Ig molecules.

Despite the wide sequence variations in Ig-like domains, the maintenance of the Ig fold seems to be enhanced by a conserved geometry of hydrogen bonds. In addition to sequence analysis of the Ig-like domains, the quantitative evaluation of their structural similarity appears to be important to build models for other members of the IgFF, to elucidate Ig folding principles and to predict new members through sensitive sequence comparisons (e.g. Mornon *et al*., 1997).

The Ig-like domains have been identified in various kingdoms including eukaryotes and prokaroytes, bacteria, viruses, fungi and plants [see Halaby and Mornon (1998) for a review]. Some of these domains lack known biological activities, such as those present in bacterial enzymes. The widespread occurrence of the Ig fold and its appearance in plants (Martinez *et al*., 1994) precludes any species or function exclusivity, i.e. the immune response, and raises the question of the origins of the fold. Is the Ig fold derived from a common ancestor, where in some cases the functional activities have been lost during evolution, or is it a stable structure to which many sequences have converged?

Members of the immunoglobin family are known to be phylogenetically related and Gelfand and Kister (1995) showed that there are 47 similar positions in the Ig sequences of the Kabat bank, eight being strictly conserved. Such identity cannot be extended to the IgFF, illustrated by the fact that no strict topohydrophobic positions can be identified for the whole family. Indeed, the study of tophydrophobic positions in the previously defined groups clearly demonstrated the homogeneity within the groups and the heterogeneity between them. Interestingly, the scores computed for each pair of groups in the IgFF and the phylogenetic tree calculated on the basis of sequence identity and r.m.s.d. values correlate well: the pairs of groups which are close to each other in the phylogenetic tree have high scores and those which are distant in the tree correspond to low scores. This result confirms that topohydrophobic positions are indeed related to structural and sequence features.

The determination of topohydrophobic positions being a very recent technique, it is difficult to quantify it accurately. However, the values for $S_{i,j}$ obtained in the present study fit nearly exactly with structural data: values obtained for two subsets of structures belonging to the same sub-family are always higher than 1.5 (data not shown) and consequently always higher than the values obtained by comparing two different sub-families.

The present study cannot definitely answer the difficult question of whether the IgFF evolved by divergent or convergent processes or both mechanisms. Indeed, structural and sequence conservation are high between subfamilies that are functionally correlated, while they are very low and often completely absent in unrelated proteins within the whole superfamily. At such low levels of sequence identity, it is very difficult to distinguish between convergent or divergent mechanisms of evolution (Burkhard, 1997). However, it appears more likely that both mechanisms may explain the IgFF: convergence of unrelated domains towards a simple and stable fold and divergence within each subtype.

References

Bork,P., Holm,L. and Sander,C. (1994) *J. Mol. Biol.*, **242**, 309–320.

Burkhard,R. (1997) *Fold. Des.*, **2**, S19–S24.

Callebaut,I., Labesse,G., Durand,P., Poupon,A., Canard,L., Chomilier,J., Henrissat,B. and Mornon,J.-P. (1997) *Cell. Mol. Life Sci.*, **53**, 621–645.

Chan,A.W., Hutchinson,E.G., Harris,D. and Thornton,J.M. (1993) *Protein Sci.*, **2**, 1574–1590.

Chothia,C. and Lesk,A.M. (1986) *EMBO J.*, **5**, 823–826.

Chou,P.Y. and Fasman,G.D. (1974) *Biochemistry*, **13**, 222–245.

Gelfand,I.M. and Kister,A.E. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 10884–10888.

Godzik,A., Skolnick,J. and Kolinski,A. (1993) *Protein Engng*, **6**, 801–810.

Halaby,D.M. and Mornon,J.P.E. (1998) *J. Mol. Evol.*, **46**, 389–400.

Harpaz,Y. and Chothia,C. (1994) *J. Mol. Biol.*, **238**, 528–539.

Harris,L. and Bajorath,J. (1995) *Protein Sci.*, **4**, 306–310.

Holm,L. and Sanders,C. (1993) *J. Mol. Biol.*, **233**, 123–138.

Hubbard,T.J.P. and Blundell,T.L. (1987) *Protein Engng*, **1**, 159–171.

Hunkapiller,T. and Hood,L. (1989) *Adv. Immunol.*, **44**, 1–63.

Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) *J. Mol. Biol.*, **231**, 735–752.

Jones,E.Y. (1993) *Curr. Opin. Struct. Biol.*, **3**, 846–852.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Lee,B.K. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.

Lesk,A.M., Levitt,M. and Chothia,C. (1986) *Protein Engng*, **1**, 77–78.

Maiorov,V.N. and Crippen,G.M. (1994) *J. Mol. Biol.*, **235**, 625–634.

Martin,A.C. (1996) *Proteins*, **25**, 130–133.

Martinez,S.E., Huang,D., Szczepaniak,A., Cramer,W.A. and Smith,J.L. (1994) *Structure*, **2**, 95–105.

Mornon,J.-P., Halaby,D., Malfois,M., Durand,P., Callebaut,I. and Tardieu,A. (1997) *Int. J. Biol. Macromol.*, **22**, 219–227.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) *J. Biol. Chem.*, **99**, 153–162.

Pfuhl,M., Improta,S., Politou,A.S. and Pastore,A. (1997) *J. Mol. Biol.*, **265**, 242–256.

Poupon,A. and Mornon,J.-P. (1998) *Proteins*, **33**, 329–342.

Poupon,A. and Mornon,J.-P. (1999) *Theor. Chim. Acta*, **101**, 2–8.

Proba,K., Honegger,A. and Pluckthun,A. (1997) *J. Mol. Biol.*, **265**, 161–172.

Richards,F.M. (1985) The calculation of molecular volumes and areas for structures of known geometry. Acad. Press, Inc.

Richardson,J.S. (1977) *Nature*, **268**, 495–500.

Rudikoff,S. and Pumphrey J.G. (1986), *Proc. Natl Acad. Sci. USA*, **83**, 7875–7878.

Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.

Sali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.

Schneider,R. and Sander,C. (1991) *Proteins*, **9**, 56–68.

Smith,D.K. and Xue,H. (1997) *J. Mol. Biol.*, **274**, 530–545.

Sutcliffe,M.J., Haneef,I., Carney,D. and Blundell,T. (1987) *Protein Engng*, **1**, 377—384.

Taylor,W.R. (1986) *J. Mol. Biol.*, **188**, 233–258.

Vogt,G. and Argos,P. (1997) *Fold. Des.*, **2**, S40–S46.

Williams,A.F. and Barclay,A.N. (1988) *Annu. Rev. Immunol.*, **6**, 381–405.

Received August 7, 1998; revised February 10, 1999; accepted March 16, 1999