J Mol Evol (1998) 46:389–400

JOURNAL OF MOLECULAR EVOLUTION

The Immunoglobulin Superfamily: An Insight on Its Tissular, Species, and Functional Diversity

D.M. Halaby, J.P.E. Mornon

Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS URA 09, Université Pierre et Marie-Curie, Case 115, 4 Place Jussieu, 75252 Paris cedex 05, France

Received: 4 June 1997 / Accepted: 15 September 1997

Abstract. The immunoglobulin superfamily (IgSF) is a heterogenic group of proteins built on a common fold, called the Ig fold, which is a sandwich of two β sheets. Although members of the IgSF share a similar Ig fold, they differ in their tissue distribution, amino acid composition, and biological role.

In this paper we report an up-to-date compilation of the IgSF where all known members of the IgSF are classified on the basis of their common functional role (immune system, antibiotic proteins, enzymes, cytokine receptors, etc.) and their distribution in tissue (neural system, extracellular matrix, tumor marker, muscular proteins, etc.), or in species (vertebrates, invertebrates, bacteria, viruses, fungi, and plants). The members of the family can contain one or many Ig domains, comprising two basic types: the constant domain (C), with seven strands, and the variable domain (V), with eight, nine, or ten strands. The different overviews of the IgSF led to the definition of new domain subtypes, mainly concerning the C type, based on the distribution of strands within the two sheets.

The wide occurrence of the Ig fold and the much less conserved sequences could have developed from a common ancestral gene and/or from a convergent evolutionary process. Cell adhesion and pattern recognition seem to be the common feature running through the entire family.

Key Words: Proteins families — Immunoglobulin fold — Domains — Fibronectin type III — Evolution — Function — Constant domains — Variable domains

Introduction

The term "immunoglobulin superfamily" (IgSF) initially referred to Igs and other proteins involved in the immune response and sharing the same 3D topology. The subsequent discovery of the Ig fold in proteins not functionally related to Igs led to the definition of new functional families, structurally similar to the Igs, such as that of the cytokine receptors (Thoreau et al. 1991) or of the bacterial proteins (Bork and Doolittle 1992) containing the fibronectin type III module. In recent years, increasing numbers of novel members of the IgSF have been identified.

Membership of the IgSF of a newly identified protein is usually based on the conservation in folding and in sequence of specific features found within the Ig molecules. These criteria include domain size (~100 amino acids), the number of strands, and the general topology of the Ig fold (Williams and Barclay 1988). The evaluation of sequence similarities uses statistical tests, with many distinct sequences if possible. In some cases, however, such as that of superoxide dismutase (code PDB 1SOD) or the hemocyanins (code PDB 1HCY), a significant sequence similarity failed to be detected, although the Ig fold of these proteins was confirmed by their crystallographic structures.

In this paper, we deal with the IgSF via a global

Abbreviations: C, constant domain; Fn3, fibronectin III; Ig, Immunoglobulin; IgSF, Immunoglobulin superfamily; V, variable domain

Correspondence to: D.M. Halaby; e-mail Halaby@lmcp.jussieu.fr

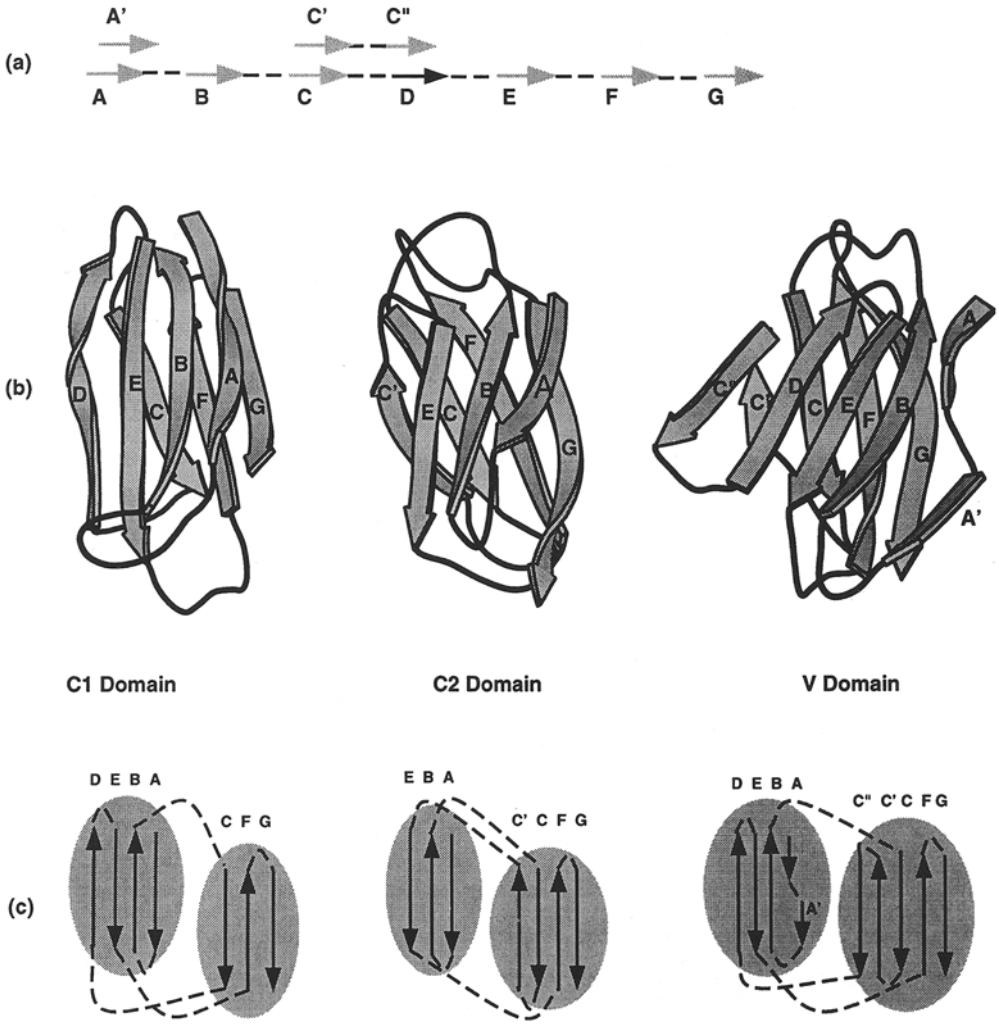

Fig. 1. Linear, spatial, and topological representation of the Ig fold. **A** Linking of β strands in sequences. **B** Ribbon representation (Molscript, Kraulis 1991) of Ig domains showing the spreading of the strands in the two antiparallel sheets. Note that the fourth strand belongs to sheet I in C1-type domains and to sheet II in C2-type domains. V domains contain three additional strands: A′, C′, and C″. **C** Schematic representation of the spatial connectivities of β strands in 3D structures.

approach. We have taken into account all proteins containing one or several domains similar to that of an Ig unit (described in Fig. 1). The many previous studies (Kuma et al. 1991; Bork and Doolittle 1992; Bork et al. 1994; Brummendorf and Rathjen 1994; Gerstein and Altman 1995; Jones et al. 1993; Williams and Barclay 1988) only dealt with specific sections of the IgSF. Kuma et al. (1991) excluded from their study members of type C1 and viral proteins. Jones et al. (1993) reviewed the structure and functions of antibodies and molecules involved in the immune system. Bork et al. (1994) proposed a comparison of certain known structures. Other researchers explored the characteristics of a unique subgroup, such as immunoglobulins (Gerstein and Altman 1995), cell adhesion molecules (Brummendorf and Rathjen 1994) or bacterial Fn3 domains (Bork and Doolittle 1992). Thus, none of these studies provided an overview of the IgSF. We have therefore attempted to build an up-to-date compilation of proteins containing one or more Ig-like domains. They are described and classified according to their major functional characteristics and by tissue and species distribution.

Methods

Research on members of the IgSF was performed in the Swiss-Prot, EMBL, and PIR sequence databases using key-word or sequence-pattern searches. The PDB (Protein data bank) was searched for the known 3D structures of members of the IgSF. At the same time, a literature research was undertaken. We encountered two levels of difficulty while searching for members of the IgSF:

1. Redundancy in the sequence banks. Many of the proteins were identified in different species or by different researchers. Although Table 1 shows more than 150 proteins, not counting species redundancies, we found more than 2,000 sequences in data banks.
2. The variety of terminologies used to design an Ig fold. Many terms are found in the literature that reflect a constant domain, such as C type, C1 type, C2 type, H type, S type, I type, Fn3 (fibronectin III)

Table 1. Proteins containing one or many Ig-like domains[a]

Abbreviations	Species[b]	Name	Domains	3D (PDB code)	Sequence code
Antibiotic proteins					
Acx	Sg	Actinoxantin	1 C	y(1acx)	P01551
Akp	Ac	Kedarcidin	1 C	y(1akp)	P41249
Mcm	Sma	Macromycin	1 C	y(1mcm)	P01549
Ncs	Sc	Neo-carzinostatin	1 C	y(1nco)	P01550
Bacterial proteins					
*Enzymes					
Apu	Ct	α-amylase/Pullulanase	2 Fn3	n	P38939
Bgl	H, M, Ec	β-galactosidase	1 C	y(1bgl)	P00722
CelD	Clt	Cellulase D	1 C	y(1clc)	P04954
Cenb	Cf	Endoglucanase E (cellulase)	3 Fn3	n	P26225
ChiA1	Bc, Sl, Sol	Chitinase A1	2 Fn3	n	P20533
Cgtase	Bc, Kp	Cyclodextrin glycosyl-transferase	1 C	y(1cyg)	P31797
EndC	Cf	Endoglucanase C	2 C	n	P14090
Pehx	Ech	Exo-poly-α-D-galacturonosidase	1 Fn3	n	P15922
Phb	Af	Depolymerase	1 Fn3	n	P12625
Sialidase[c]	Ac	Sialidase	2 C	y(1eur)	Q02834
*Chaperones					
Aggd	Ec	Aggregative adherence fimbria I	2 C	n	P46004
Caf1M	Yp	Chaperone protein	2 C	n	P26926
ClpE	Ec	Chaperone protein	2 C	n	Q05433
Cs3-1	Ec	Chaperone protein	2 C	n	P15483
EcpD	Ec	Chaperone protein	2 C	n	P33128
FaeE	Ec	Chaperone protein	2 C	n	P25401
FanE	Ec	Chaperone protein	2 C	n	P25402
FimB (FhaD)	Bp	Chaperone protein	2 C	n	P33409
FimC	St, Ec	Chaperone protein	2 C	n	P31697
FocC	Ec	Chaperone protein	2 C	n	P46008
HifB	Hi	Chaperone protein	2 C	n	P35757
LpfB	St	Chaperone protein	2 C	n	P43661
MrkB	Kp	Chaperone protein	2 C	n	P21646
MyfB	Ye	Chaperone protein	2 C	n	P33407
NfaE	Ec	Chaperone protein	2 C	n	P46738
PapD	Ec	Chaperone protein	2 C	y(3dpa)	P15319
PrsD	Ec	Chaperone protein	1 C	n	P42183
PsaB	Yp	Chaperone protein	2 C	n	P31523
SefB	Se	Chaperone protein	2 C	n	P33387
YcbF	Ec	Chaperone protein	1 C	n	P40876
YehC	Ec	Chaperone protein	2 C	n	P33342
YhcA	Ec	Chaperone protein	2 C	n	P28722
YraI	Ec	Chaperone protein	2 C	n	P42914
*Others					
FcαR	Sa	IgA Fc receptor	1 C	n	X58470**
CEA Antigens-related					
CEA	H	Carcinoembryonic antigen	6 C + 1 V	n	P06731
NCA	H	Normal cross-reacting antigen	2 C + 1 V	n	P31997
PSβG	H	Pregnancy-specific β1 glycoprotein	3 C	n	P11462
T1	M	Oncoprotein	1 C	n	M24843**
*Bgp et Bgp-like					
Bgp	H	Biliary glycoprotein 1	1 (or 3)C + 1 V	n	P13688
C-CAM/ecto-ATPase	R	Cell-CAM	3 C + 1 V	n	P16573
CGM2 (MHVR)	M	CEA gene family member 2	1 C + 1 V	n	P31809
Cytokines receptors					
CNTFRα	H,R	Ciliary neurotrophic factor receptor α	3 (or 4) C	n	P26992
CRFB4	H	Cytokine receptor class II	2 C	n	Q08334
EPOR	H,M,R	Erythropoietin receptor	2 C	y(1ebp)	P19235
G-CSFR	H,M	G-CSFR receptor	4 C	y(1cto)	Q99062
GHR	H,P,R,Rb,Ch	Growth hormone receptor	2 C	y(3hhr)	P10912
GM-CSFR	H	GM-CSF receptor	3 (or 4) C	n	P15509
IL1R	M	Interleukin 1 receptor	3 C	y(1itb)	P14778
IL2R	H,M,Dg	Interleukin 2 receptor	2 C[d]	n	P31785
IL3R	M,H	Interleukin 3 receptor	3 C	n	P26951

Table 1. Continued.

Abbreviations	Species[b]	Name	Domains	3D (PDB code)	Sequence code
IL4R	M,H	Interleukin 4 receptor	2 C	n	P24394
IL5R	M,H	Interleukin 5 receptor	3 C	n	Q01344
IL6R (gp130)	H,M,R	Interleukin 6 receptor	3 C	n	P08887
IL7R	H	Interleukin 7 receptor	2 C	n	P16871
Il9R	H,M	Interleukin 9 receptor	2 C	n	Q01113
Il10R	H	Interleukin 10 receptor	2 C	n	L12120**
Il12R	H	Interleukin 12 receptor	3 C	n	P42701
INF α[e]	B,H,M	Interferon α receptor	2 (or 4) C	n	P17181
INF β	H,M	Interferon β receptor	2 (or 4) C	n	P48551
INF γ	H,M	Interferon γ receptor	2 C	y (on hold)	P15260
LIFR	H,M	Leukemia inhibitory factor receptor	5 C	n	P42702
Pr1R[f]	H,R,M,Rb,Ch	Prolactin hormone receptor	2 C	y (on hold)	P16471
TF	H	Tissue factor	2 C	y (1hft)	P13726
TPOR	H, M	Thrombopoietin receptor	2 (or 4) C	n	P40238
Enzymes					
*Phosphatase					
GLEPP1	H, Rb	Glomerular epithelial protein 1	8 Fn3	n	U09490**
HPTPeta	H	Protein tyrosine phosphatase	3 Fn3	n	D37781**
LAR (PTPRF)	H	Leukocyte antigen related	3 C + 8 Fn3	n	P10586
LCA (CD45, PTPRC)	H,M,R	Leukocyte common antigen	2 Fn3	n	P08575
PTPK	H,M	Protein-tyrosine phosphatase κ	1 C + 4 Fn3	n	P35822
*Tyrosine kinase					
c-Kit	H	Proto-Oncogene	4 C + 1 V	n	P10721
Csf1R (*c-fms*)[g]	H	Colony stimulating factor-1 receptor	4 C + 1 V	n	P07333
Eck	H	Epithelial cell kinase	2 Fn3	n	P29317
FGFR	H,F,M,Ch,R	Fibroblast growth factor receptor	2 (or 3) C	n	P11362
					P21802,
Jtk, bek, Ksam	H				P22607, P22455
cek	Ch				P18460
mfr	M				P16092
IGF1R	H	Insulin-like growth factor I receptor	2 Fn3	n	P08069
InSR	H,M,R	Insulin receptor	2 Fn3	n	P06213
IRR	H,Gp,F	Insulin-related receptor	2 Fn3	n	P14616
Klg	Ch	Kinase-like gene	7 C	n	M63437
PDGFR α	H,M,R,F	Platelet-derived growth factor receptor	4 C + 1 V	n	P16234
β	H,M				P09619
STK-1	H	Stem cell tyrosine kinase	4 C + 1 V	n	P36888
Flk-2/Flt-3	M	tyrosine kinase receptor			
tie-2	B,H,M	Endothelial cell glycoprotein	2 C + 3 Fn3	n	Q02763
Tif	H	Receptor tyrosine kinase	1 C + 1 Fn3	n	U02566**
TrkB	H	Protein tyrosine kinase	2 C	n	X75958**
Ufo and Ufo-like					
Dtk	M	Developmental tyrosine kinase	2 C + 2 Fn3	n	U18933**
Sky (Nyk/Mer)	H	Receptor tyrosine kinase	2 C + 2 Fn3	n	D17517**
UFO/AXL	H,M	Tyrosine-protein kinase receptor	2 C + 2 Fn3	n	P30530
VGR (FLK1, FLT1, FLT4)	H, M	Vascular endothelial growth factor receptor	7 C	n	P35918
*Others					
F13	B,H	Coagulation factor XIII	2 C	y (1ggt)	P00488
SOD	—[h]	Superoxyde dismutase	1 C	y (1sod)	P00441
Extracellular matrix					
Col4 (MMP, CLG4)	H,M,Rb	Collagen type IV	3 Fn3	n	P08253
Col6A3	H, Ch	Collagen type VI α 3	1 Fn3	n	P12111
Col7A1	H	Collagen type VII α 1	9 Fn3	n	Q02388
Col14A1	Ch	Collagen type XIV α 1	6 Fn3	n	P32018
Fn3	H,B,R,Ch,M	Fibronectin type III	16 Fn3	y (1fna)	P02751
G1	R	Cartilage proteoglycan domain G1	1 V	n	P03994
HSPG	H	Heparan sulfate proteoglycan	22 C	n	P34741
	M	Perlecan, BPG	15 C	n	Q05793
Protein "link"	R	Cartilage link protein	1 V	n	P03994
TN	H,M,Ch	Tenascin	15 Fn3	y (1ten)	P24821
Unc 52	Ce	Perlecan homolog	16 C	n	L13458**
Undulin	H	Undulin	3 or 7 Fn3	n	M64108**
Versican	H	Large fibroblast proteoglycan	1 V	n	P13611

Table 1. Continued.

Abbreviations	Species[b]	Name	Domains	3D (PDB code)	Sequence code
Immune system					
*Immunoglobulins					
H	H[i]	Chain H	4 C1 + 1 V	y (2fab)	P01857
κ	H[i]	Chain L Kappa	1 C1 + 1 V	y (1dfb)	P01834
λ	H[i]	Chain L Lambda	1 C1 + 1 V	y (2fb4)	P01842
*Ig Receptors					
Poly IgR	H, Rb, R	Transepithelial transport receptor of Ig	1 C + 4 V	n	P01833
FcγR	H, M, R	IgG Fc receptor	2 C[j]	n	P12315
FCεR	H, R, M	IgE receptor	2 C	y (1hli)	P06734
FcαR	H	IgA Fc receptor	2 C	n	P24071
CMRF35	H	Fc receptor (?)	1 V	n	X66171**
CD79a	B, H, M	IgM receptor-associated glycoprotein	1 C	n	P11912
*CMH					
β2mg	H[i]	β2-microglobulin	1 C1	y (1hla)	P01884
B system	Ch	B-G antigen	1 V	n	M27663
Cl I	H[i]	MHC class I	1 C1	y (1hla)	P10313
CL II	H[i]	MHC class II	1 C1	y (1seb)	P01903
Zn-α2 gp	H, R	Zn-α2 glycoprotein, HLA-like antigen	1 C	n	P25311
*β2mg-associated antigen					
CD1a chain H	H, M, Rb	CD1 antigen	1 C1	n	P06126
TL chain H	M	Thymus Leukaemia antigen	1 C1	n	M16810**
Qa chain H	M	Qa antigen	1 C1	n	L00606**
*TCR	H, M, Rb	T cell receptor	1 C1 + 1 V	y (1bec)	P01852
*Cell surface antigens					
Alcam	H	Activated leukocyte cell adhesion molecule	3 C + 2 V	n	L38608**
B7 (CD80)	H, M, Rb	B lymphocyte activation antigen 7	1 C + 1 V	n	P33681
B7-2 (CD86)	H, M, Rb	B lymphocyte activation antigen 7-2	1 C + 1 V	n	P42081
B29 (CD79b, OSPA)	H, M, Sp	B cell antigen	1 V	n	P40259
CD2 (LFA2)	H, R, M, Ho	T cell erythrocyte receptor	1 C2 + 1 V	y (1hnf)	P06729
CD3	H,M,R,Sh,Dg	T3 antigen	1 C	n	P04234
CD4	H[i]	T4 antigen	2 C2 + 2 V	y (3cd4)	P017730
CD7 (GP40)	H	T cell Leukaemia antigen	1 V	n	P09564
CD8 (Lyt-3)	H[i]	T8 antigen	1 V	y (1cd8)	P01732
CD19	H, Gp, M	B cell antigen	2 C	n	P15391
CD22	H, M	B cell antigen	6 C + 1 V	n	P20273
CD28	H,M,R,C,Rb	T cell antigen	1 V	n	P10747
CD48 (Bcm-1, Blast 1)	H, M, R	Lymphocyte activation marker	1 C + 1 V	n	P09326
CTLA4	H, M, Rb	Cytolytic T cell antigen 4	2 V	n	P16410
HB15 (CD83)	H	Lymphocyte cell antigen	1 V	n	Q01151
ICAM-1	H, M, R, Dg	InterCellular Adhesion Molecule 1	5 C	n	P05362
ICAM-2	H, M	InterCellular Adhesion Molecule 2	2 C	y (1zxq)	P13598
ICAM-3	H	InterCellular Adhesion Molecule 3	5 C	n	P32942
LAG3 (FDC)	H	Lymphocyte activation gene-3 protein	3 C + 1 V	n	P18627
LFA3 (CD58)	H	Lymphocyte Function-Associated antigen	1 C + 1 V	n	P19256
Ly-9 (Lgp-100)	H, M	Lymphocyte antigen	2 C + 2 V	n	Q01965
p58	H	NK cell receptor	2 C	n	P43627
Tactile	H	T cell activation antigen	1 C + 2 V	n	P40200
VpreB	H, M	Pre-B lymphocyte antigen	1 V	n	P12018
V7	H	Leukocyte surface protein	7 V	n	Z33642**
Invertebrates[k]					
α-agglutinin	Y	α-agglutinin	1 V	n	P20840
Ama	D	Amalgam	2 C + 1V	n	P15364
ApCAM	A	Aplysia cell adhesion molecule	5 C + 2 Fn3	n	AC42632*
DLAR	D	Protein tyrosine phosphatase	3 C + 9 Fn3	n	P16621
DPTP	D	Protein tyrosine phosphate	2 C + 2 Fn3	n	P16620
DTRK (FR1)	D	Tyrosine kinase receptor	4 C + 2V	n	S19247*
Fasciclin II	G, D	Fasciclin II	5 C + 2 Fn3	n	P22648
Fasciclin III	D	Fasciclin III	2 C + 1 V	n	P15278
GCTK	Gc	Sponge receptor tyrosine kinase	1 C	n	P42159
Hcy	Pi	Hemocyanin	1 C	y (1hcy)	P04254
HIG	D	Locomotion-related protein	1 C	n	Q09101
IMP-L2	D	Neural/ectodermal development factor	2 C	n	Q09024
IRRE-C	D	Irregular chiasm C-roughest	5 C	n	S34129
Lachesin	D	Lachesin	2 C + 1 V	n	L13255**

Table 1. Continued.

Abbreviations	Species[b]	Name	Domains	3D (PDB code)	Sequence code
MDM	Ls	Molluscan defence molecule	5 C	n	U58769**
MSP	N	Major Sperm protein	1C	y (1msp)	P27440
Neuromusculin	D	Neuromusculin	8 C + 1 V	n	L23146**
Nrg	D	Neuroglian	6 C + 5 Fn3	y (1cfb)	P20241
OncA	Tt	Oncosphere antigen A	1 Fn3	n	P22080
p4	Sm,Cm,Thm	Hemolin	4 C	n	P25033
REGA-1	G	REGA glycoprotein	3 C + 1 Fn3	n	X93601**
RSP5	Y	Ubiquitin-protein ligase	1 C	n	P39940
7LES	D	Sevenless	7 Fn3	n	P13368
UNC-5	Ce	Uncoordinate movement-5	1 C + 1 ID	n	AB44294
UNC-13	Ce	Phorbol ester/DAG binding protein	2 C	n	P27715
Muscle proteins					
CaVPT	Cl	Ca^{2+} vector protein target	2 C	n	P05548
C-protein	Ch	C-protein	6 C + 3 Fn3	n	P16419
	H		7 C + 4 Fn3	n	Q00872
Kettin	D	Kettin	4 C	n	X72709
MLCK	Ch, Rb	Myosin light-chain kinase	3 C + 1 Fn3	n	P11799
M-Protein	Ch	Protein M	5 C + 6 Fn3	n	Q02173
Projectin	D	Projectin	1 C + 3 Fn3	n	X66018**
Skelemin	M,B	Skelemin	7 C + 5 Fn3	n	Z22866**
Telokin	Rb	Telokin	1 C	y (1tlk)	P29294
Titin	H, R, Ch, P, B	Connectin	19 C + 50 Fn3	y (1tnm)	D16541**
Unc 22	Ce	Twitchin	26 C + 31 Fn3	n	X15423**
Neural system					
ARIA	Ch	Acetylcholine receptor-inducing activity	1 C	n	P27177
BIG-1 (Pang)	R, M	Brain-derived IgSF molecules 1	6 C + 4 Fn3	n	U11031**
DM-GRASP(BEN, Scl)	Ch	Axonal surface protein	3 C + 2 V	n	JH0506*
F11	H	Neuro-1 antigen	6 C + 4 Fn3	n	U07819**
F3	M				P12960
Contactin	Ch				P10450
Gicerin	Rb, Ch	Gicerin	3 C + 2 V	n	D38559**
Gp135	H	Membrane glycoprotein		n	
L1	H,M	Neural cell recognition molecule L1	6 C + 5 Fn3	n	P32004
NILE	R	Nerve growth-factor inducible large external			Q05695
NgCAM (NrCAM)	Ch	Neuron–glia cell adhesion molecule			Q03696 (P35331)
MAG	H,M,R	Myelin-associated glycoprotein	4 C + 1 V	n	P20916
MOG	R, H	Myelin/oligodendrocyte glycoprotein	1 V	n	P23515
MRC-OX2	R, H	Lymphoid/neuronal membrane gp	1 C + 1 V	n	P41217
NCAM	H,R,M,Ch,F,B	Neural cell adhesion molecule	5C + 1(or 2)Fn3	n	P13591
Neurofascin	R,Ch	Ankyrin-binding glycoprotein	6 C + 5 Fn3	n	S26180*
Neurolin	Fi	Neurolin	3 C + 2 V	n	L25056
obCAM	B	Opioid binding cell adhesion molecule	2 C + 1 V	n	P11834
P_o	H,B,M,R,Ch,S	Myelin P_o protein	1V	n	P25189
Rab3A	B, M	Rabphilin-3A	2 C	n	P47708
ScnB2	R	Brain sodium channel B2 subunit	1 V	n	U37026**
SMP	Ch	Schwam cell myelin protein	1 V	n	S83711**
Synaptotagmin (p65)	—[1]	Synaptotagmin	2 C	y (1rsy)	P21579
TAG-1	H	Transiently expressed axonal surface gp-1	6 C + 4 Fn3	n	Q02246
Snap	R				P22063
Axonin-1	Ch				P28685
Thy-1 (CDw90)	R,Ch,H,M	Thymocyte antigen	1 V	n	P04216
Tumor markers					
AAMP	H	Angio-associated migratory cell protein	2 C	n	M95627**
Basigin (gp42)	Ch,H,M,R	Blood-Brain barrier HT7 antigen	1 C + 1 V	n	P35613
B-CAM	H	Cell surface gp of epithelial cancers	3 C + 2 V	n	X80026**
DCC	H	Deleted in colorectal carcinoma	4 C + 6 Fn3	n	A54100*
gp70	M	Embigin embryonic cells	1 V	n	P21995
IAP (CD47, OA3)	H, M	Integrin associated protein	1 V	n	Q08722
MUC18	H	Melanoma-associated antigen	3 C + 2 V	n	A34507
PANG	M	Plasmocytoma-associated neuronal gp	6 C + 4 Fn3	n	L01991**
pE4 (TuAg.1)	R	Rat carcinoma-associated antigen	2 C + 1 V	n	L12025**
Viral proteins[e]					
GIII	Pv, Hsv	Glycoprotein GIII	2 C + 1 V	n	P14378

Table 1. Continued.

Abbreviations	Species[b]	Name	Domains	3D (PDB code)	Sequence code
GPV (p14)	Vzv	Glycoprotein GPV	2 C + 1 V	n	P09256
HA	Vv, Rcn, Vav	Vaccinia virus hemagglutinin	1 V	n	P08714
Polg	TBVE	Genome polyprotein	1C	y (1svb)	P14336
T = 4	NWV	Capsid protein	1 C[m]	y (1isj)	S43937**
VB16	Cpv, Vv, Vav	Interleukin 1 binding protein	3 C	n	P21116
VB19	Vav, Vv	S antigen	2 C + 1 V	n	P21077
UL44	Hsv	Glycoprotein C	2 C + 1 V	n	P10228
Others					
α_1GP	H,Do,Ho,P	α_1 B-glycoprotein	5 V	n	P04217
CD	B, Ch, H, M, F	Cadherin	5 C	y (1nci)	P19022
CD33	H	Myeloid cell surface antigen (GP67)	1 C + 1 V	n	P20138
CTM	Br	Cytochrome F	1V	y (1ctm)	P36438
CTX	F	Cortical thymocyte marker	1 C + 1 V	n	V43330**
FIP-fve	Fu	Fungal immunomodulatory protein	1 V	n	P80412
gaoA	Dd	Galactose oxidase	1 C	y (1gog)	Q01745
gp49	M	Mouse mast cell glycoprotein 49	2 C	n	U05264**
Hrg	H	Heregulin α	1 C	n	M94165**
Kal	H, Ch	Kallmann syndrome protein homolog	4 fn 3	n	P23352
Lu gp	H	Lutheran glycoprotein	3 C + 2 V	n	X83425**
LZ8	Fu	Immunomodulatory protein Ling ZHI-8	1V	n	P14945
MADCAM-1	M	Mucosal addressin cell adhesion molecule	2 C + 1V	n	S33601*
NAR	S	Nurse Shark antigen receptor	5 C + 1 V	n	L16765***
NDF	R	Neu differentiation factor	1 C	n	M92430**
NF-kB (p50)	M	Transcription factor	1 C + 1 V	y (1svc)	P25799
PD-1	M	Programmed death 1	1 V	n	Q02242
PECAM-1	H, M	Platelet endothelial cell adhesion molecule	6 C	n	P16284
PVR	H, M	Poliovirus receptor	2 C + 1 V	n	P15151
RAGE	B	Receptor for advanced glycation end products	2 C + 1 V	n	M91212**
RhoGDIs	H	Guanine nucleotides dissociation inhibitors	1 C	y (1rho)	P52565
Sialoadhesin	H	Macrophage receptor	16 C + 1 V	n	Z36293**
ST2	H, M	Stimulated state 2	3 C	n	Q01638
VCAM-1	H, M, R	Vascular cell adhesion molecule 1	3 (or 7) C	y (1vca)	P19320

[a] Classification is made on the basis of species, tissue distribution, or biological activity. Homonyms indicating identical or similar proteins in different species are in italics. The sequence code is that of the Swiss-Prot sequence bank, except where indicated by * (PIR), ** (EMBL), or *** (GENBANK). Since the type of constant domain is assigned rather arbitrarily, we have chosen not to specify the subtype of constant domains, except for those of the immune system when known. However, when the 3D is not yet known, the constant domain is often designated as C2 type, to differentiate it from the C1 domain, found only in the immune system. Determination of the 3D structure of some of these proteins may confirm or deny the attribution of the Ig fold to their spatial architecture

[b] Abbreviations used: A, *Aplysia;* Ac, Actinomycete; Af, *Alcaligenes faecalis;* B, bovine; Bc, *Bacillus circulans;* Bp, *Bordetella pertussis;* Br, *Brassica rapa;* C2, constant domain; Ch, chicken; C, constant; Ce, *C. Elegans;* Cf, *Cellulomonas fimi;* Cm, *Cecropia* moth; C1, common Lancelet; Clt, *Clostridiuum thermocellum;* Cpv, cowpox virus; Ct, *Clostridium thermohydrosulfuricum;* D, *Drosophila;* Dd, *Dactylium dendroides;* Dg, dog; Do, donkey; Ec, *E. coli;* Ech, *Erwinia chrysanthemi;* F, frog; Fi, fish; Fn3, fibronectin type III domain; Fsv, feline sarcoma virus; G, grasshopper; Gc, *Geodia cydonium;* gp, glycoprotein; Gp, guinea pig; H, human; Hi, *Haemophilus influenzae;* Ho, horse; Hsv, *Herpex simplex* virus; IgSF, *immunoglobulin superfamily;* Kp, *Klebsiella pneumonia;* M, mouse; N, nematode; Nwv, *Nudaurelia* ω *capensis* virus; n, not determined; P, pig; Pi, *Panulinus interruptus;* Pv, *Pseudorabies* virus; R, rat; Rb, rabbit; Rcn, raccoon poxvirus; S, shark; Sa, *streptococcus agalactiae;* Sc, *Streptomyce carzinostaticus;* Se, *Salmonella enteritidis;* Sg, *Streptomyces globisporus;* Sh, sheep; Sl, *Streptomyces lividans;* Sm, silk moth; Sma, *Streptomyces macromomyceticus;* Sol, *Streptomyces olivaceoviridis;* Sp, spirochetes; St, *Salmonella typhimunium;* TBEB, tick-borne encephalitis virus; Thm, tobacco hawkmoth; Tt, *Taenia taeniaformis;* V, variable domain; Vav, *Variola* virus; Vv, *Vaccinia* virus; Vzv, varicella-zoster virus; Y, yeast; y, determined; Ye, *Yersinia enterocolitica;* Yp, *Yersinia pestis*

[c] Also in parasites (P23253)

[d] One Fn3 domain is found in the dog IL2γR and in IL2RβR

[e] In the vaccinia virus, the IL1βr homolog is B15R, that of INF receptor type I is B18R

[f] Four domains in the chicken sequence (Q04594)

[g] Also in viruses (v-fms)

[h] Found in several vertebrates, invertebrates, plants, nematodes, and bacteria

[i] Found in several other vertebrates

[j] FcγR1 contains 3 C2

[k] Invertebrate muscle proteins are included in the "muscle proteins" group

[l] Found in several vertebrates, invertebrates, and nematodes

[m] This domain is a new variant of the C type in that it contains additional strands

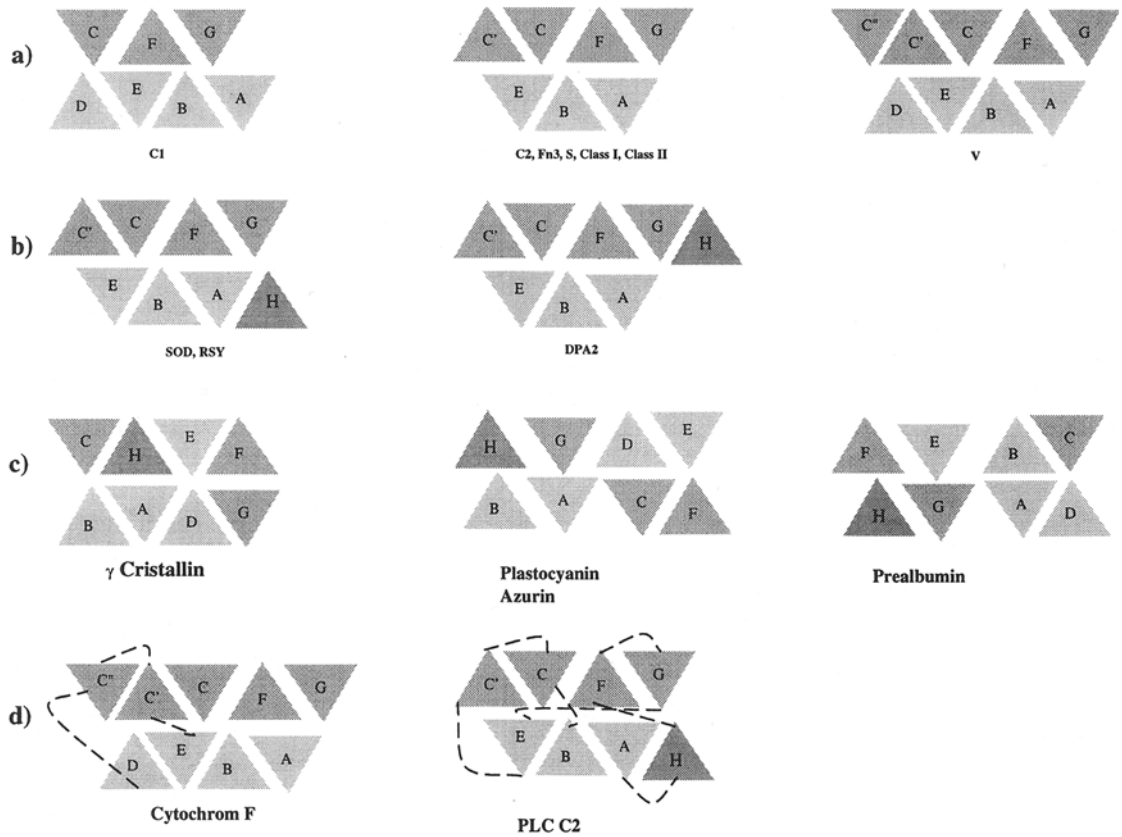

Fig. 2. The Greek Key motif: schematic representation of observed topologies and connectivities. Strands are distributed in two sheets and are spatially connected as they appear in the sequence *A, B, C, C′, C″, D, E, F, and G*, except for **D,** where strand connectivity is shown explicitly. Some domains contain an additional strand *H*. **A** Ig and Ig-like domains. **B** Superoxide dismutase (SOD), synaptotagmin (RSY), and the second domain of PapD (DPA2) contain an additional strand H, next to strand A in SOD and RSY, and to strand G in DPA2. **C** Examples of domains having a sandwich architecture similar to that of Igs. These domains show similar connectivity but different topology. **D** The domains of cytochrom F and of phospholipase C (PLC C2) show similar topology but different connectivity to that of an Ig domain.

type, motif class I, motif class II, and CRH (cytokine receptor homolog) type. In some papers the term C2 was in fact "local" nomenclature and did not correspond to an Ig-like domain. We therefore verified membership of the IgSF for each protein. All the terms cited above refer to seven β strands (named A–G) folded in two antiparallel sheets—namely, the constant domain. These terms were based on structural comparison of known 3D structures of the Ig-like domains, the distribution of the strands in the two sheets, and the presence or absence of a disulphide bond. For the sequences of unknown structures, assignment of secondary structure, and consequently of the subtype of the fold, is rather arbitrary for the C type.

Results

Table 1 shows members of the IgSF. For each protein, we indicate the name, the species in which it was identified, the number and type of Ig-like domains, the PDB code for its structure (when determined), and its code in the sequence data banks. For modular proteins, the number of Ig-like domains is that found in the sequenced proteins, such as Kettin (Lakey et al. 1993), where 30 Ig-like domains may be found in the native proteins.

Structural Characteristics of the Ig Fold

The immunoglobulin fold (Ig fold) constitutes a class of β proteins presenting a sandwich architecture where the strands, distributed in two sheets, are connected in a typical way, ABCDEFG (each letter refers to a strand in order of its appearance in the sequence), as shown in Fig. 1. The connectivity of the strands is used to differentiate between the Ig fold and other similar Greek Key β architectures (Fig. 2). Based on the number and location of the strands, and with reference to the IgSF, two kinds of domains were defined: variable (AA′BED/CC′C″FG) and constant (ABED/CFG). The first main structural difference between the C and V domains is an extra loop in the V domain between strands C and D. Second, there is a conserved disulphide bond which presents different characteristics in the two domains. In a V or V-related domain, there are about 65–75 residues between the two cysteines. In a C domain, this segment contains 55–60 residues (Williams and Barclay 1988), or even less in certain Ig-related molecules, such as in the sequence of human CD2 (Jones et al. 1992). In attributing a type of domain, C or V, to a newly identified sequence, the occurrence of cysteines is thus very helpful when available.

Analysis of the growing number of members identified, however, revealed new subtypes of Ig domains, especially regarding the C domain.

Classification of IgSF Members

All members identified are listed in Table 1. Whenever possible, the domain type or subtype has been indicated as explained above. The *C* domain represents classical seven-stranded β sheets. Domain *C1* is a variant of C in that it is found in the molecules involved in the immune system: Ig molecules, T-cell receptors, and MHC antigens. Previously, the *C2* domain was described on the basis of sequential criteria as being similar to both the V and C domains. It has a C fold, but strands E and F are similar to those found in variable domains (Williams and Barclay 1988). The term was later used to designate a switch in strand D (ABE/DCFG). Other subtypes of the constant-like domains are not indicated in the table unless they are explicitly given in the literature. The *S type* is topologically similar to the C2 type (Bork et al. 1994). The *Fn3* domain is similar to the C2 type, except that it lacks the disulphide bond (Brummendorf and Rathjen 1994), although IgSF membership of Fn3 domains is controversial (Bork and Doolittle 1992; Little et al. 1994). An exception is the disulphide bonds found in Fn3 domains of proteins F11, fasciclin II, and neuroglia (Brummendorf and Rathjen 1994). The *I type* is an intermediate between the V and C1 domains (Gerstein and Altman 1995). The *H type* is a variant of the S type: A bulge occurs in strand D which thus belongs to sheet I and to sheet II (ABED/D'CFG) (Bork et al. 1994). The *Class I* domain refers to the motif Fn3, while that of *class II* refers to the C2 motifs. The terms class I and class II are used exclusively for muscle proteins (Price and Gomer 1993). Finally, the *CRH type* indicates a C2-like domain containing the major characteristic of cytokine receptors, i.e., the WS × WS motif (Fukunaga et al. 1991).

Functional Diversity of the Ig Superfamily

In Table 1, IgSF members are listed according to their major functional characteristics, which we have classed in eight groups as follows:

1. Molecular transport: Antibiotic proteins actively transport chromophores through different cellular compartments. In vertebrates, hemocyanins transport oxygen molecules. The Ig receptor (Poly IgR) transports Igs through the epithelial wall.
2. Morphoregulation: The proteins of the extracellular matrix are involved in the architectural organization and elasticity of a large number of tissues. Fibronectin acts as an "adhesive" that conserves tissue integrity. Cell adhesion molecules exposed on the embryonic cell surfaces (such as cadherin molecules) favor the organization of these cells into differentiated tissues.
3. Cell phenotype markers: e.g., tumor-cell markers and surface molecules of hematopoietic cells. The latter allow us to evaluate the cell type (B or T cells of the immune system) or the state of cell differentiation (B or pre-B cells, activated T cells, transformed cells, etc.).
4. Cell adhesion molecules: The cellular immune response requires many cell adhesion molecules of the IgSF, such as CD2, CD4, CD8, and MHC (major histocompatibility complex molecules). Some cell phenotype markers are also involved in cell adhesion, e.g., B7, B29, CD19, CD3, CD7, etc.
5. Virus receptors: e.g., PVR, CD4, ICAM-1, and Bgp molecules, which are the respective receptors of poliovirus, HIV, rhinovirus, and MHV virus, in addition to their constitutive functions.
6. Shape recognition and toxin neutralization: The property of sequence polymorphism and structural variability of the immunoglobulin molecules allows the immune system to adapt for different antigens. The immune response is based on capacity for shape recognition and cell-cell adhesion.
7. Viral and bacterial molecules: In bacteria, IgSF members are domains within enzymes. They also can be involved in pili assembly and/or synthesis. In viruses, they act as receptors of cellular mediators (interferons, interleukins, etc.), or, as surface proteins, they may enhance virus virulence and dissemination.
8. Others: Some other original functions have been observed in the IgSF, such as the regulation of gene transcription (NF-kB), cell migration (VCAM, PECAM-1, etc.), or cell death marking (PD-1). Finally, a number of Ig-like domains do not possess any as-yet-identified biological function.

Characteristics of IgSF Members

We have stated that members of the IgSF present a common fold and diverse functions based on a common feature which is mainly relevant to molecular recognition processes. However, the proteins also possess their own individual characteristics, leading to the functional diversity described above.

1. The ligand: The ligands of the Ig-like domains range from small molecules (antigens, chromophores), to hormones (growth hormone, interferons, prolactin, etc.), up to giant molecules (muscle proteins). Whatever the size, the ligand/Ig-like domain interaction can be homophilic (CEA, NCA, NCAM in vertebrates, fascilin II and III in invertebrates) or heterophilic. In the latter case, the two partners can either be members of the IgSF (CD2/LFA3) or not (extracellular matrix). Some domains show a double activity, both homo- and heterophilic (NCAM, CEA, etc.).
2. Binding sites are localized either in the loop regions (the most variable part of Igs) or in strands. For instance, distinct areas of the sheets are used to bind the ligands of the MHC, CD8, CD4, and PapD molecules

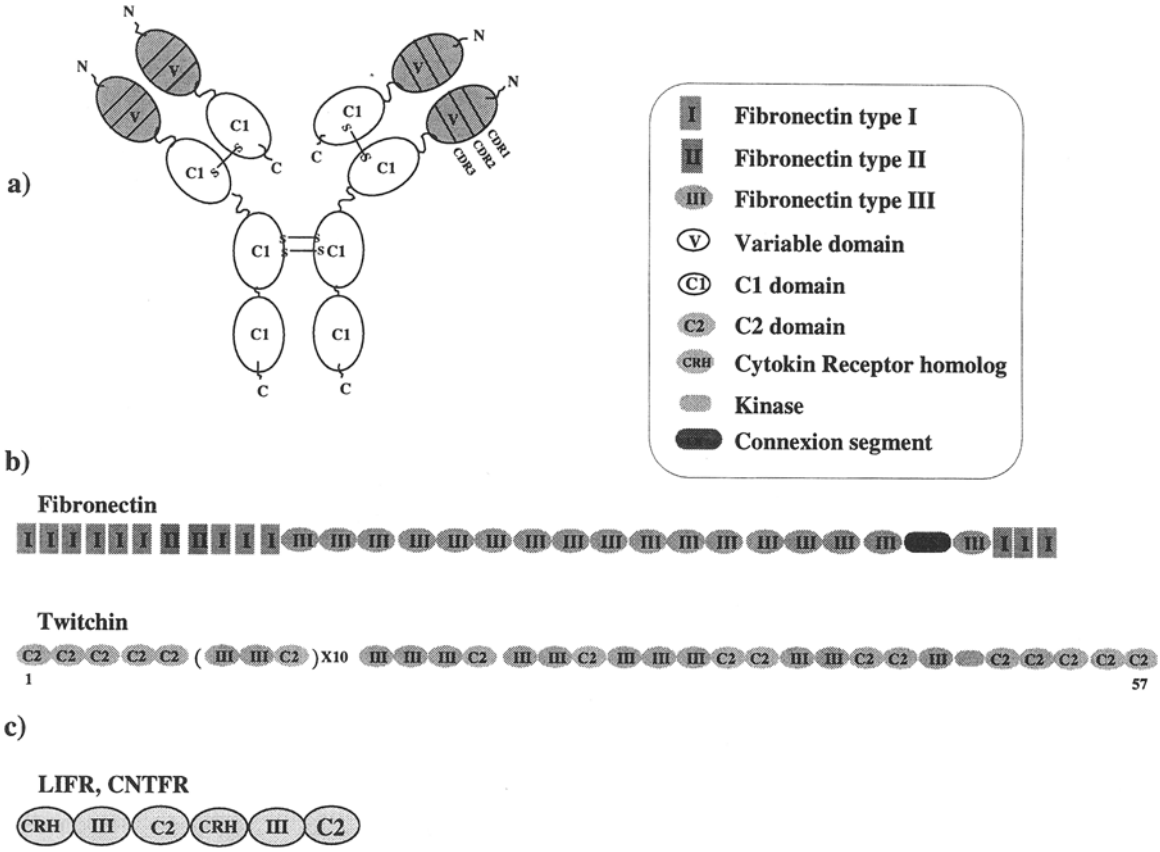

Fig. 3. Example showing different types of assembly of Ig-like domains in some proteins. **A** The C1 and V domains in the Ig molecule. The disulphide bridge (S-S) and the hypervariable regions (CDR1, CDR2, CDR3) are shown. **B** Modular association of Fn3 domains with non Ig-like domains (Fn I and Fn II) in fibronectin, and with C2 domains in the twitchin. **C** Assembly of three different kinds of domain (CRH, Fn III, and C2) in the receptor of cytokines LIF and CNTFR.

or of the growth hormone receptor (GHR). These binding sites may be formed by a single chain (CD2, CD4), by homodimers (GHR, CD8), or by heterodimers.

3. The number of domains involved in biological function may be one (CD8, P_0, antibiotic proteins), two (some cytokine receptors, bacterial proteins, etc.), or more (cell adhesion molecules, giant proteins, etc.). The polypeptide chain may form only one domain, or else several domains of possibly different types (V-C, C-Fn3, C-CRH-Fn3). Interestingly, the association of Fn3 and C2 domains, which confers complexity in binding and recognition processes (Rathjen 1991), is found in enzymes and in neural cell adhesion molecules (Table 1). In these latter molecules, Fn3 is invariably located close to the membrane (Brummendorf and Rathjen 1994). The Ig domain may be part of a more complex architecture, such as the bacterial β-galactosidase (Bgl) molecule for example. In the Bgl domain, however, which contains five domains, including one Ig-like fold, the exact role of the Ig domain remains unclear. Some of these assemblies are shown in Fig. 3.

Species Distribution of the Ig-Like Domains

Our search for the IgSF members led us to the impressive finding that the Ig-like domains are widely distributed in nature. Ig-like folds are present in eukaryotes and prokaryotes, in vertebrates and invertebrates, and in fungi, parasites, bacteria, viruses, and plants. Within species, variations are, however, seen. While the Ig-fold shows wide tissue distribution in vertebrates, its expression is more limited to the neural system of invertebrates. The wide distribution of the Ig fold therefore raises the question of whether these proteins are homologous (arising from a common ancestor) or analogous (developing through convergent evolution of distinct proteins towards the same stable fold). If a common ancestor exists, sufficient sequence divergence was maintained to produce a variety of functional products, while retaining a common structural core. In contrast, analogous proteins would mean that the Ig fold, which supports diverse biological activities, is a rather stable structural framework.

Conclusion

We give below a list and classification of all available members of the immunoglobulin superfamily (IgSF).

Their remarkable diversity in terms of biological activity and tissue and species distribution has been reviewed. The Ig fold is found in vertebrates and invertebrates, as well as in bacteria, viruses, fungi, and plants. Members of the IgSF are involved in distinct and independent functions such as the immune response, growth factor receptors, tumoral markers, enzymatic activity, neural cell development, antitumoral activity, maintenance of tissue integrity, as virus receptors, in viral and bacterial pathogenesis, or simply in cell adhesion. Although very different, most of these biological functions share a common functional pathway, based on homophilic or heterophilic recognition and adhesion process.

The finding of a similar folding pattern in various species and for various functions is relevant to its enigmatic origin and to the question of whether it is encoded by a common ancestor or constitutes an energetically stable fold. Among vertebrates, the most distant member of this class to share with mammals a gene coding for immunoglobulins is the shark, the genome of which is markedly different in the organization of gene segments (Bartl et al. 1994). In contrast to the three types of MHC found in shark, no evidence of a TCR is yet available, nor for a T-cell-mediated immune response (Kasahara et al. 1992). As regards invertebrates, insects possess their own immune system, which, although based on cellular mechanisms of adhesion and recognition, such as phagocytosis, does not involve immunoglobulins but secreted bactericidal, bacteriolytic, or bacteriostatic protein (Lindstrom-Dinnetz et al. 1995).

The identification of prokaryotic domains raises the question as to when the immunoglobulin superfamily arose. Sequence comparison of virus domains and other members of the IgSF failed to reveal significant similarities. This finding suggests that the domain had been captured by the virus from the immune system of the host eucaryotic cell and converted to the virus through evolution (Jin et al. 1989). This hypothesis finds further support in the divergence observed in molecular specificity of the IFN (interferon) type I receptor expressed by vaccinia virus and mammals. While the virus receptor shows relative affinities for different IFN type I species, the human receptor presents strict human specificity, with the exception of the bovine cytokine (Symons et al. 1995). These different species provide natural hosts, with variable affinity, for the vaccinia virus. A similar scheme of acquisition of the Ig-fold domains by bacteria has been proposed (Bork and Doolittle 1992), according to which the occurrence of these domains could be the result of horizontal gene transfers.

It is very difficult today to prove whether the immunoglobulin fold has arisen from a common ancestor gene, was acquired by different species during evolution by vertical or horizontal gene transfer, or if it is a stable structural framework that serves as a vehicle for a variety of biological functions, allowing many subtle variations in the key area of protein-protein interaction. The number of known 3D structures with Ig-like domains, which is in itself small in relation to the number of sequences identified, is rapidly growing. To address the question of the common ancestor, we are now pursuing sequence analysis and comparison of known 3D structures of Ig-like domains. The analysis of the common structural core should shed light on the degree of structural conservation of the IgSF members and on the stable fold hypothesis.

Acknowledgments. We gratefully thank Annette Tardieu for reading this manuscript and for many helpful discussions and suggestions during the preparation of this account.

References

Bartl S, Baltimore D, Weissman IL (1994) Molecular evolution of the vertebrate immune system. Proc Natl Acad Sci USA 87:6934–6938

Bork P, Doolittle RF (1992) Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci USA 89:8990–8994

Bork P, Holm L, Sander C (1994) The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 242:309–320

Brunmendorf T, Rathjen F (1994) Cell adhesion molecules 1: Immunoglobulin Superfamily. Protein profile, 1, issue 9. Academic Press

Fukunaga R, Ishizaka-Ikeda E, Pan C-X, Seto Y, Nagata S (1991) Functional domains of the granulocyte colony-stimulating factor receptor. EMBO J 10:2855–2865

Gerstein M, Altman RB (1995) Average core structures and variability measures for protein families: application to the immunoglobulins. J Mol Biol 251:161–175

Jin DY, Li ZL, Jin Q, Hao YW, Hou YD (1989) Vaccinia virus hemagglutinin. A novel member of the immunoglobulin superfamily. J Exp Med 170:571–576

Jones CH, Pinkner JS, Nicholes AV, Slonium LN, Abraham SN, Hultgren SJ (1993) FimC is a periplasmic PapD-like chaperone that directs assembly of type 1 pili in bacteria. Proc Natl Acad Sci USA 90:8397–8401

Jones EY, Davis SJ, Williams AF, Harlos K, Stuart DI (1992) Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. Nature 360:232–239

Kasahara M, Vasquez M, Sato K, McKinney C, Flajnik MF (1992) Evolution of the major histocompatibility complex: isolation of class II A cDNA clones from the cartilaginous fish. Proc Natl Acad Sci USA 89:6688–6692

Kraulis P (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 24:946–950

Kuma KI, Iwabe N, Miyata T (1991) The immunoglobulin superfamily. Curr Opin Struct Biol 1:384–393

Lakey A, Labeit S, Gautel M, Ferguson C, Barlow DP, Leonard K, Bullard B (1993) Kettin, a large modular protein in the Z-disc of insect muscles. EMBO J 12:2863–2871

Lindstrom-Dinnetz I, Sun S-C, Faye I (1995) Structure and expression of Hemolin, an insect member of the immunoglobulin superfamily. Eur J Biochem 230:920–925

Little E, Bork P, Doolittle RF (1994) Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. J Mol Evol 39:631–643

Price MG, Gomer RH (1993) Skelemin, a cytoskeletal M-disc periphery protein, contains motifs of adhesion/recognition and intermediate filament proteins. J Biol Chem 268:21800–21810

Rathjen FG (1991) Neural cell contact and axonal growth. Curr Opin Cell Biol 3:992–1000

Symons JA, Alcami A, Smith GL (1995) Vaccinia virus encodes a soluble type I interferon receptor of novel structure and broad species specificity. Cell 81:551–560

Thoreau E, Petridou B, Kelly PA, Djiane J, Mornon J-P (1991) Structural symmetry of the extracellular domain of the cytokine/growth hormone/prolactin receptor family and interferon receptors revealed by Hydrophobic Cluster Analysis. FEBS Lett 282:26–31

Williams AF, Barclay AN (1988) The immunoglobulin superfamily. Domains for cell surface recognition. Annu Rev Immunol 6:381–405