Immunogenetics (2002) 53:857–883
DOI 10.1007/s00251-001-0408-6

ORIGINAL PAPER

Manuel Ruiz · Marie-Paule Lefranc

IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures

Received: 18 September 2001 / Revised: 26 November 2001 / Published online: 30 January 2002

Abstract A new database, IMGT/3Dstructure-DB, was developed and implemented in the IMGT (international ImMunoGeneTics database) information system (http://imgt.cines.fr) to provide a unique expertised resource on immunoglobulin and T-cell receptor structural data. Corresponding protein sequences were annotated with IMGT tools, which allow the precise identification of the genes expressed in these proteins, and the description of framework and complementarity determining regions according to the IMGT standardized nomenclature and IMGT unique numbering. Two-dimensional graphical representations of the V-DOMAINs, designated as Colliers de Perles, are automatically produced. A query Web interface allows interactive search of the IMGT/3D structure-DB data. In this article, IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures in the Protein Data Bank are presented.

Keywords IMGT · Immunoglobulin · 3D structure · Variable region · Database

Introduction

Many studies use results from sequence–structure relationship analysis and large protein structure comparison to optimize immunoglobulin (IG) and T cell receptor (TR) engineering methods, including humanization, mutagenesis, and phage display technologies. IMGT, the international ImMunoGeneTics database (http://imgt.cines.fr) (Ruiz et al. 2000; Lefranc 2001a), contains more than 54,000 IG and TR sequences from human and 104 other vertebrate species and provides exhaustive and high-quality annotations of IG and TR sequences, based on the IMGT Scientific chart rules and IMGT-ONTOLOGY concepts (Giudicelli and Lefranc 1999), and on the standardized IMGT unique numbering (Lefranc 1997, 1998, 1999). In recent years all the human germline IG and TR genes have been characterized (Lefranc and Lefranc 2001a, b). The IMGT nomenclature was approved by the HUGO (HUman Genome Organization) Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/) in 1999, and links have been established between IMGT, GDB (Genome DataBase, Toronto, http://www.gdb.org/), and LocusLink (NCBI, http://www.ncbi.nlm.nih.gov/LocusLink/). The lists of the IG genes (Barbié and Lefranc 1998; Lefranc 2001b; Pallarès et al. 1998, 1999; Ruiz et al. 1999; Scaviner et al. 1999) and TR genes (Folch and Lefranc 2000a, b; Scaviner and Lefranc 2000a, b) have recently been published. Whereas the interoperability between sequence and genome databases has been realized, it is still difficult to compare sequence data with structural data. Indeed, the Protein Data Bank (PDB, http://www.rcsb.org/pdb/) (Berman et al. 2000) is the major provider of protein structural data. However, querying across the complete PDB is limited by missing, erroneous, and inconsistently reported experimental data, nomenclature, and functional annotation (Bhat et al. 2001). Whereas PDB data standardization is in progress, a detailed and complete annotation of IG and TR data remains difficult for the generalist PDB. Sequential numbering of the amino acids for each IG and TR protein chain is not uniform in the different PDB files, which does not allow automatic large sequence and structure comparison.

To fill in the gap, a new database, IMGT/3Dstructure-DB, was developed and implemented in the IMGT information system (http://imgt.cines.fr). This database provides a unique expertised resource on IG and TR structural data extracted from PDB. Currently 403 IG and 22 TR structures are available. Corresponding PDB sequences are annotated with IMGT tools, which allow the precise identification of the genes expressed in these proteins, and the delimitation of important functional re-

M. Ruiz · M.-P. Lefranc (✉)
IMGT, the international ImMunoGeneTics database,
http://imgt.cines.fr, Laboratoire d'ImmunoGénétique Moléculaire,
LIGM, Université Montpellier II, UPR CNRS 1142, IGH,
141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
e-mail: lefranc@ligm.igh.cnrs.fr
Tel.: +33-4-99619965, Fax: +33-4-99619901

Table 1 Overview of the human immunoglobulin 3D structures

IMGT protein name[a]	PDB code	Fragment[b]	Number of fragments	Ligands	Experimental technique	Resolution	PDB release date	References[c]
17B	1g9m	Fab	1	HIV-1 gp120 (envelope protein)/CD4	X-ray diffraction	2.20	27 Dec 2000	46
	1g9n	Fab	1	HIV-1 gp120 (envelope protein)/CD4	X-ray diffraction	2.90	27 Dec 2000	46
	1gc1	Fab	1	HIV-1 gp120 (envelope protein)/CD4	X-ray diffraction	2.50	19 Aug 1998	30
3D6	1dfb	Fab	1		X-ray diffraction	2.70	31 Oct 1993	12
	1obe	Fab	1	HIV-1 gp41 synthetic peptide	theoretical model		15 May 1997	19
9E	1dx3	Fv	1		theoretical model		15 Dec 2000	40
B12	1hzh	IgG1	1		X-ray diffraction	2.70	15 Aug 2001	55
B7-15A2	1aqk	Fab	1		X-ray diffraction	1.84	04 Feb 1998	28
BO2C11	1iqd	Fab	1	Factor VIII C2 domain [Human]	X-ray diffraction	2.00	15 Aug 2001	56
Bre	1b0w	V-KAPPA	3		X-ray diffraction	1.80	22 Dec 1999	32
	1bre	V-KAPPA	6		X-ray diffraction	2.00	15 Oct 1995	18
	1qp1	V-KAPPA	3		X-ray diffraction	2.06	28 Jun 1999	39
CH2E-1	1g84	CH2	1		NMR, 15 structures		16 May 2001	53
Cle	1lil	L-LAMBDA	2		X-ray diffraction	2.65	15 May 1997	22
Del	1b6d	L-KAPPA	2		X-ray diffraction	2.74	30 Mar 1999	38
Fab-12	1cz8	Fab	2	Vascular endothelial growth factor	X-ray diffraction	2.40	20 Mar 2000	34
FabM	1hez	Fab	2	Protein L [*Peptostreptococcus magnus*]	X-ray diffraction	2.70	10 Aug 2001	52
FcE-1	1ige	Fc	1		theoretical model	2.70	15 Oct 1994	5
	2ige	Fc	1		theoretical model	2.70	15 Oct 1994	9
FcE-2	1f6a	CH3-CH4	1	Fc epsilon RI alpha (high-affinity receptor)	X-ray diffraction	3.50	20 Jul 2000	44
	1fp5	CH3-CH4	1		X-ray diffraction	2.30	27 Sep 2000	50
FcG1–1	1e4k	Fc	1	Fc gamma III low affinity receptor (CD16)	X-ray diffraction	3.20	06 Aug 2000	49
	1fc1	Fc	1		X-ray diffraction	2.90	15 Jul 1992	3
	1fc2	Fc	1	Fragment B of protein A [*Staphylococcus aureus*]	X-ray diffraction	2.80	15 Jul 1992	3
	1fcc	Fc	1	Fragment C2 of protein G [streptococcal]	X-ray diffraction	3.50	20 Jul 1995	17
FcG1–2	1dn2	Fc	1	Asp-Cys-Ala-Trp-His-Leu-Gly-Glu-Leu-Val-Trp-Cys-Thr-NH2	X-ray diffraction	2.70	17 May 2000	43
FcG1–3	1iis	Fc	1	Fc gamma III low affinity receptor	X-ray diffraction	3.00	09 May 2001	54
	1iix	Fc	1	Fc gamma III low affinity receptor	X-ray diffraction	3.50	09 May 2001	54
Fv-1	1hou	Fv	1		theoretical model		23 Dec 1999	35
HULYS11	1bvk	Fv	2	Lysozyme [hen egg white]	X-ray diffraction	2.70	16 Feb 1999	29
	1bvl	Fv	2		X-ray diffraction	2.87	16 Feb 1999	24
Hil	8fab	Fab	2		X-ray diffraction	1.80	31 Oct 1993	10
IgA1	1iga	IgA1	1		theoretical model; scattering fitting		15 Jun 1999	33
IgmRf2A2	1dee	Fab	3		X-ray diffraction	2.70	14 Jun 2000	45
Jto	1cd0	V-LAMBDA	2		X-ray diffraction	1.90	06 Mar 2000	37
Kau	1dn0	Fab	2		X-ray diffraction	2.28	24 Jan 2001	42
	1qlr	Fab	2		X-ray diffraction	2.83	14 Sep 2000	42
Kol	2fb4	Fab	1		X-ray diffraction	1.90	12 Jul 1989	7
	2ig2	Fab	1		X-ray diffraction	3.00	12 Jul 1989	7
Len K36>T	4lve	V-KAPPA	2		X-ray diffraction	2.30	18 May 1999	31

Table 1 (continued)

IMGT protein name[a]	PDB code	Fragment[b]	Number of fragments	Ligands	Experimental technique	Resolution	PDB release date	References[c]
Len M4>L, Y30>D, Q105>D, T114>H	1eeu	V-KAPPA	2		X-ray diffraction	1.60	03 Feb 2001	57
Len M4>L, Y30>D, T114>H	1eeq	V-KAPPA	2		X-ray diffraction	1.50	01 Feb 2001	57
Len Q105>A	5lve	V-KAPPA	1		X-ray diffraction	2.00	18 Feb 2000	31
Len Q105>L	1qac	V-KAPPA	2		X-ray diffraction	1.80	23 Feb 2000	47
Len Q44>D	1efq	V-KAPPA	1		X-ray diffraction	1.60	09 Feb 2001	57
Len Q44>E	3lve	V-KAPPA	1		X-ray diffraction	2.00	18 May 1999	31
Len	1lve	V-KAPPA	1		X-ray diffraction	1.95	21 Jan 1998	25
	2lve	V-KAPPA	1		X-ray diffraction	2.70	18 May 1999	31
Loc	1bjm	L-LAMBDA	2		X-ray diffraction	2.20	07 Dec 1995	21
	3bjl	L-LAMBDA	2		X-ray diffraction	2.30	07 Dec 1995	21
	4bjl	L-LAMBDA	2		X-ray diffraction	2.40	07 Dec 1995	21
Loi	2loi	V-LAMBDA	2		theoretical model		29 Dec 1999	36
M3C65	1dl7	Fv	1		X-ray diffraction	2.35	13 Dec 2000	41
Mak33	1fh5	Fab	1		X-ray diffraction	2.90	13 Sep 2000	51
Mcg-Weir hybrid	1mcw	L-LAMBDA	2		X-ray diffraction	3.50	15 Oct 1990	8
Mcg	1a8j	L-LAMBDA	2	Aspartame	X-ray diffraction	2.70	17 Jun 1998	27
	1dcl	L-LAMBDA	2		X-ray diffraction	2.30	15 May 1997	6
	1mcb	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mcc	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mcd	L-LAMBDA	2	*N*-Acetyl-D-Phe-B-Ala-L-His-D-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mce	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-B-Ala-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mcf	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-B-Ala-B-Ala-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mch	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-B-Ala-B-Ala-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mci	L-LAMBDA	2	*N*-Acetyl-D-Phe-L-His-D-Pro-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mcj	L-LAMBDA	2	*N*-Acetyl-D-Phe-L-His-D-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mck	L-LAMBDA	2	*N*-Acetyl-D-Glu-L-His-D-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mcl	L-LAMBDA	2	*N*-Acetyl-D-His-L-Pro-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mcn	L-LAMBDA	2	*N*-Acetyl-D-His-L-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mcq	L-LAMBDA	2	*N*-Acetyl-L-His-D-Pro-NH2	X-ray diffraction	2.70	31 Jan 1994	14
	1mcr	L-LAMBDA	2	*N*-Acetyl-L-His-D-Pro-OH	X-ray diffraction	2.70	31 Jan 1994	14
	1mcs	L-LAMBDA	2	*N*-Acetyl-L-Gln-D-Phe-L-His-D-Pro-OH	X-ray diffraction	2.70	31 Jan 1994	14
	2mcg	L-LAMBDA	2		X-ray diffraction	2.00	15 Jul 1992	2
	3mcg	L-LAMBDA	2		X-ray diffraction	2.00	15 Oct 1990	6
McgHL	1mco	H-GAMMA1 (with a hinge deletion) and L-LAMBDA	1		X-ray diffraction	3.20	31 Jan 1994	15

Table 1 (continued)

IMGT protein name[a]	PDB code	Fragment[b]	Number of fragments	Ligands	Experimental technique	Resolution	PDB release date	References[c]
Mez	1dql	Fv	1		X-ray diffraction	2.60	04 Oct 2000	48
Newm	7fab	Fab	1		X-ray diffraction	2.00	31 Jan 1994	13
Pot	1igm	Fv	1		X-ray diffraction	2.30	31 Oct 1993	11
Rea	1adq	Fc	1	Rf-An	X-ray diffraction	3.15	16 Sep 1998	23
Rec	1ek3	V-KAPPA	2		X-ray diffraction	1.90	06 Mar 2001	57
Rei C23>V, Y32>H	1ar2	V-KAPPA	1		X-ray diffraction	2.80	12 Nov 1997	26
Rei T45>K	1bww	V-KAPPA	2		X-ray diffraction	1.70	29 Dec 1999	26
Rei	1rei	V-KAPPA	2		X-ray diffraction	2.00	17 Feb 1984	1
Rf-An	1adq	Fab	1	Autoantigen IgG4 Fc Rea	X-ray diffraction	3.15	16 Sep 1998	23
Rhe	2rhe	V-LAMBDA	1		X-ray diffraction	1.60	09 Oct 1988	4
Tr1.9	1vge	Fab	1		X-ray diffraction	2.00	10 Jun 1996	20
Wat	1wtl	V-KAPPA	2		X-ray diffraction	1.90	01 Nov 1994	16
Wil	2cd0	V-LAMBDA	2		X-ray diffraction	1.80	08 Mar 2000	37

[a] When the protein names were undefined in the PDB files, standardized names were created corresponding to the fragment type, chain type, and eventually a number separated by hyphens (CH2E-1, FcE-1, FcE-2, FcG1-1, FcG1-2, FcG1-3, Fv-1, FabM). The designation *Newm* is used in IMGT (this protein is designated as New in PDB) for the human IgG1 myeloma protein Fab (PDB code: 7fab) [58], to avoid confusion with the human New Bence Jones protein [59], which was sequenced before Newm. Detailed comparison of the different nomenclatures for Newm, in publications and in databases, is available in IMGT Repertoire (Particularities in protein designations, http://imgt.cines.fr:8104/textes/IMGTrepertoire/IMGTrepProteins.html#5). For the protein mutants (mutants of Len and Rei), the protein names are redefined according to the description of mutations in the IMGT Scientific chart

[b] *CH1*, *CH2*, *CH3*, *CH4* C-DOMAIN encoded by a IG heavy CH1, CH2, CH3, CH4 exon, respectively. *Fab* Fragment antigen-binding, consists of L-KAPPA or L-LAMBDA disulfide-linked with a IG heavy chain fragment consisting of VH and CH1. *Fc* Fragment crystallizable, two identical disulfide-linked fragments each comprising the CH2 and CH3 domains for alpha, gamma, and delta heavy chains, and CH2, CH3 and CH4 domains for mu and epsilon heavy chains. *Fv* two V-DOMAINs of different chains non covalently linked. For the IG, a VH with a V-KAPPA or a V-LAMBDA. *IgA1* complete immunoglobulin A1. *IgG1* complete immunoglobulin G1. *L-KAPPA* Complete light kappa chain. Kappa chains are usually found as non-covalently linked light chain dimers in the crystals. *L-LAMBDA* Complete light lambda chain. Lambda chains are usually found as non-covalently linked light chain dimers in the crystals. *VH* IG heavy V-DOMAIN, encoded by the IGH V-D-J-REGION. *V-KAPPA* IG light kappa V-DOMAIN, encoded by the IGK V-J-REGION. *V-LAMBDA* IG light lambda V-DOMAIN, encoded by the IGL V-J-REGION

[c] List of primary bibliographic references of the structure determination: [1] Epp O et al. (1975) Biochemistry 14:4943–4945; [2] Ely KR et al. (1978) Biochemistry 17:158–167; [3] Deisenhofer J (1981) Biochemistry 20:2361–2370; [4] Furey W Jr et al. (1983) J Mol Biol 167:661–692; [5] Padlan EA, Davies DR (1986) Mol Immunol 23:1063–1075; [6] Ely KR et al. (1989) J Mol Biol 210:601–615; [7] Kratzin HD et al. (1989) Biol Chem Hoppe Seyler 370:263–267; [8] Ely KR et al. (1990) Mol Immunol 27:101–114; [9] Helm BA et al. (1991) Eur J Immunol 21: 1543–1548; [10] Strong RK et al. (1991) Biochemistry 30:3739–3748; [11] Fan ZC et al. (1992) J Mol Biol 228:188–207; [12] He XM et al. (1992) Proc Natl Acad Sci USA 89:7154–7158; [13] Saul FA et al. (1992) Proteins 14:363–367; [14] Edmundson AB et al. (1993) Proteins 16:246–267; [15] Guddat LW et al. (1993) Proc Natl Acad Sci U S A 90:4271–4275; [16] Huang DB et al. (1994) Biochemistry 33:14848–14857; [17] Sauer-Eriksson AE et al. (1995) Structure (Lond) 3:265–278; [18] Schormann N et al. (1995) Proc Natl Acad Sci U S A 92:9490–9494; [19] Stigler RD et al. (1995) Protein Eng 8:471–479; [20] Chacko S et al. (1996) J Biol Chem 271:12191–12198; [21] Huang DB et al. (1996) Proc Natl Acad Sci U S A 93:7017–7021; [22] Huang DB et al. (1996) Acta Crystallogr D Biol Crystallogr 52:1058; [23] Corper AL et al. (1997) Nat Struct Biol 4:374–378; [24] Holmes MA et al. (1997) J Immunol 158:2192–2201; [25] Huang DB et al. (1997) Mol Immunol 34:1291–1301; [26] Uson I et al. (1997) Fold Des 2:357–361; [27] Edmundson AB, Manion CV (1998) Clin Pharmacol Ther 63:580–593; [28] Faber C et al. (1998) Immunotechnology 3:253–270; [29] Holmes MA et al. (1998) J Exp Med 187:479–485; [30] Kwong PD et al. (1998) Nature 393:648–659; [31] Pokkuluri PR et al. (1998) Structure 6:1067–1073; [32] Schormann N et al. (1998) Amyloid 5:175–187; [33] Boehm MK et al. (1999) J Mol Biol 286:1421–1447; [34] Chen Y et al. (1999) J Mol Biol 293:865–868; [35] Hougs L et al. (1999) Infect Immun 67:2503–2514; [36] Jokiranta TS et al. (1999) J Immunol 163: 4590–4596; [37] Pokkuluri PR et al. (1999) Amyloid 6:165–167; [38] Roussel A et al. (1999) Eur J Biochem 260:192–199; [39] Steinrauf LK et al. (1999) J Biochem (Tokyo) 125:422–429; [40] Beiboer SH et al. (2000) J Mol Biol 296:833–849; [41] Brown M et al. (2000) J Exp Med 191:2101–2112; [42] Cauerhff A et al. (2000) J Immunol 165:6422–6428; [43] Delano WL et al. (2000) Science 287:1279–1283; [44] Garman SC et al. (2000) Nature 406:259–266; [45] Graille M et al. (2000) Proc Natl Acad Sci U S A 97:5399–5404; [46] Kwong PD et al. (2000) Structure Fold Des 8:1329–1339; [47] Pokkuluri PR et al. (2000) Protein Sci 9:1852–1855; [48] Ramsland PA et al. (2000) Mol Immunol 37:295–310; [49] Sondermann P et al. (2000) Nature 406:267–273; [50] Wurzburg BA et al. (2000) Immunity 13:375–385; [51] Augustine JG et al. (2001) J Biol Chem 276:3287–3294; [52] Graille M et al. (2001) Structure (Lond) 9:679–687; [53] McDonnell JM et al. (2001) Nat Struct Biol 8:437–441; [54] Radaev S et al. (2001) J Biol Chem 276:16469–16477; [55] Saphire EO et al. (2001) Science 239:1155–1159; [56] Spiegel PC Jr et al. (2001) Blood 98:13–19; [57] Pokkuluri PR et al. to be published; [58] Chen BL, Poljak RJ (1974) Biochemistry 13:1295–1302; [59] Langer B et al. (1968) Hoppe-Seyler's Z Physiol Chem 349:945–951

Table 2 Identification of the genes and alleles expressed in the human IGH V-DOMAINs (V-D-J-REGIONs) and IGK or IGL V-DOMAINs (V-J REGIONs) of associated heavy and light chains. For each protein, the IMGT protein name, the protein fragment and the PDB code are indicated. For each protein, the genes' identification for the heavy chain and for its associated light (kappa or lambda) chain are displayed on the same table row. As discussed in the text IGHD genes are not shown. For each chain, the PDB chain name, the identified IG V and IG J gene and allele(s), the CDR-IMGT lengths, the identified IG C gene or, for the heavy chain, CH exons are indicated

Protein name	Fragment	PDB code	Heavy chain						Light chain				
			PDB chain name	IGH V-D-J-REGION			IGH C-REGION		PDB chain name	IGK or IGL V-J-REGION			IGK or IGL C-REGION
				IGHV gene and allele name	*IGHJ* gene and allele name	CDR-IMGT lengths	*IGHC* gene name	Exons		*IGKV* or *IGLV* gene and allele name	*IGKJ* or *IGLJ* gene and allele name	CDR-IMGT lengths	*IGKC* or *IGLC* gene name
17B	Fab	1g9m 1g9n 1gc1	H H H	*IGHV1-69*02* or *IGHV1-69*04*	*IGHJ1*01*	[8.8.21]	*IGHG1*	CH1	L L L	*IGKV3-15*01*	*IGKJ2*01*	[6.3.11]	*IGKC*
3D6	Fab	1dfb 1obe	H H	*IGHV3-9*01*	*IGHJ3*01* or *IGHJ3*02*	[8.8.19]	*IGHG1*	CH1	L L	*IGKV1-5*03*	*IGKJ3*01*	[6.3.7]	*IGKC*
9E	Fv	1dx3	H	*IGHV7-4-1*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.9]			L	*IGKV3-11*01*	*IGKJ1*01*	[6.3.10]	–
B12	IgG1	1hzh 1hzh	H K	*IGHV1-3*01*	*IGHJ6*03*	[8.8.20]	*IGHG1*	CH1 CH2 CH3	L M	*IGKV3-20*01*	*IGKJ2*01*	[7.3.9]	*IGKC*
B7-15A2	Fab	1aqk	H	*IGHV3-30*01* or *IGHV3-30*04* or *IGHV3-30*07* or *IGHV3-30*11* or *IGHV3-30*14* or *IGHV3-30*16* or *IGHV3-30*17* or *IGHV3-30-3*0*	*IGHJ3*02*	[8.8.16]	*IGHG1*	CH1	L	*IGLV1-40*01*	*IGLJ3*01* or *IGLJ3*02*	[9.3.10]	*IGLC3*
BO2C11	Fab	1iqd	B	*IGHV1-24*01*	*IGHJ3*02*	[8.8.10]	*IGHG4*	CH1	A	*IGKV3-20*01*	*IGKJ5*01*	[7.3.9]	*IGKC*
Fab-12	Fab	1cz8 1cz8	H Y	*IGHV7-4-1*02*	*IGHJ2*01*	[8.8.16]	*IGHG1*	CH1	L X	*IGKV1-33*01*	*IGKJ1*01*	[6.3.9]	*IGKC*
FabM	Fab	1hez 1hez	B D	*IGHV3-30*18*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.14]	*IGHM*	CH1	A C	*IGKV1-39*01*	*IGKJ1*01*	[6.3.9]	*IGKC*
Fv-1	Fv	1hou	H	*IGHV3-23*01*	*IGHJ6*01* or *IGHJ6*02*	[8.8.8]			L	*IGKV2-29*01*	*IGKJ1*01*	[11.3.10]	
HULYS11	Fv	1bvk 1bvk 1bvl 1bvl	B E A C	*IGHV4-59*01* or *IGHV4-59*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.7.10]			A D B D	*IGKV1-27*01*	*IGKJ1*01*	[6.3.9]	
Hil	Fab	8fab 8fab	B D	*IGHV3-33*01* or *IGHV3-33*04*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.14]	*IGHG1*	CH1	A C	*IGLV3-25*02*	*IGLJ2*01*	[6.3.9]	*IGLC2*
IgA1	IgA1	1iga 1iga	A B	*IGHV1-3*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.14]	*IGHA1*	CH1 CH2 CH3	C D	*IGKV1-13*02*	*IGKJ4*01*	[6.3.9]	*IGKC*

Table 2 (continued)

Protein name	Fragment	PDB code	Heavy chain						Light chain				
			PDB chain name	IGH V-D-J-REGION			IGH C-REGION		PDB chain name	IGK or IGL V-J-REGION			IGK or IGL C-REGION
				IGHV gene and allele name	*IGHJ* gene and allele name	CDR-IMGT lengths	*IGHC* gene name	Exons		*IGKV* or *IGLV* gene and allele name	*IGKJ* or *IGLJ* gene and allele name	CDR-IMGT lengths	*IGKC* or *IGLC* gene name
IgmRf2A2	Fab	1dee 1dee 1dee	B D F	*IGHV3-30*18*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.14]	*IGHM*	CH1	A C E	*IGKV1-39*01*	*IGKJ1*01*	[6.3.9]	*IGKC*
Kau	Fab	1dn0 1dn0 1qlr 1qlr	B D B D	*IGHV4-34*01* or *IGHV4-34*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.7.14]	*IGHM*	CH1	A C A C	*IGKV3-20*01*	*IGKJ4*01*	[7.3.9]	*IGKC*
Kol	Fab	2fb4 2ig2	H H	*IGHV3-33*01* or *IGHV3-33*04*	*IGHJ6*01* or *IGHJ6*02*	[8.8.19]	*IGHG1*	CH1	L L	*IGLV1-44*01*	*IGLJ1*01*	[8.3.11]	*IGLC1*
M3C65	Fv	1dl7	H	*IGHV4-59*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.7.7]			L	*IGLV7-46*01*	*IGLJ3*02*	[9.3.9]	
Mak33	Fab	1fh5	H	*IGHV3-21*01* or *IGHV3-21*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.7]	*IGHG1*	CH1	L	*IGKV3-15*01*	*IGKJ4*01*	[6.3.9]	*IGKC*
McgHL	H-GAMMA1 (with a hinge deletion) and L-LAMBDA	1mco	H	*IGHV4-39*01* or *IGHV4-39*06*	*IGHJ5*02*	[10.7.9]	*IGHG1*	CH1 CH2 CH3	L	*IGLV2-8*01*	*IGLJ1*01*	[9.3.10]	*IGLC1*
Mez	Fv	1dql	H	*IGHV3-30*10*	*IGHJ3*01* or *IGHJ3*02*	[8.8.16]			L	*IGKV1-17*01*	*IGKJ1*01*	[6.3.8]	
Newm	Fab	7fab	H	*IGHV4-59*04*	*IGHJ6*01* or *IGHJ6*02*	[8.7.11]	*IGHG1*	CH1	L	*IGLV1-40*01*	*IGLJ3*01* or *IGLJ3*02*	[9.3.9]	*IGLC3*
Pot	Fv	1igm	H	*IGHV3-23*01*	*IGHJ5*01*	[8.8.14]			L	*IGKV1-33*01*	*IGKJ3*01*	[6.3.9]	
Rf-An	Fab	1adq	H	*IGHV3-9*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.16]	*IGHM*	CH1	L	*IGLV3-21*02*	*IGLJ3*01* or *IGLJ3*02*	[6.3.11]	*IGLC3*
Tr1.9	Fab	1vge	H	*IGHV1-3*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	[8.8.14]	*IGHG1*	CH1	L	*IGKV1-13*02*	*IGKJ4*01*	[6.3.9]	*IGKC*

Table 3 Identification of the genes and alleles expressed in the human IGK V-DOMAINs (V-J-REGIONs) of kappa chains not associated with a heavy chain. For each protein, the IMGT protein name, the protein fragment and the PDB code are indicated. For each chain, the PDB chain name, the identified *IGKV* and *IGKJ* gene and allele(s), the CDR-IMGT lengths and the *IGKC* gene, if present, are indicated

IMGT protein name	Fragment	PDB code	Light chain				
			PDB chain name	IGK V-J-REGION			IGK C-REGION *IGKC* gene name
				IGKV gene and allele name	*IGKJ* gene and allele name	CDR-IMGT lengths	
Bre	V-KAPPA	1b0w 1b0w 1b0w 1bre 1bre 1bre 1bre 1bre 1bre 1qp1 1qp1 1qp1	A B C A B C D E F A B C	*IGKV1-33*01*	*IGKJ2*01*	[6.3.9]	
Del	L-KAPPA	1b6d 1b6d	A B	*IGKV1-33*01*	*IGKJ4*01*	[6.3.9]	*IGKC*
Len	V-KAPPA	1lve 2lve	– –	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len K36>T	V-KAPPA	4lve 4lve	A B	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len M4>L, Y30>D, Q105>D, T114>H	V-KAPPA	1eeu 1eeu	A B	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len M4>L, Y30>D, T114>H	V-KAPPA	1eeq 1eeq	A B	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len Q105>A	V-KAPPA	5lve	A	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len Q105>L	V-KAPPA	1qac 1qac	A B	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len Q44>D	V-KAPPA	1efq	A	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Len Q44>E	V-KAPPA	3lve	–	*IGKV4-1*01*	*IGKJ2*01*	[12.3.9]	
Rec	V-KAPPA	1ek3 1ek3	A B	*IGKV4-1*01*	*IGKJ4*01*	[12.3.9]	
Rei	V-KAPPA	1rei 1rei	A B	*IGKV1-33*01*	*IGKJ2*01*	[6.3.9]	
Rei C23>V, Y32>H	V-KAPPA	1ar2	–	*IGKV1-33*01*	*IGKJ2*01*	[6.3.9]	
Rei T45>K	V-KAPPA	1bww 1bww	A B	*IGKV1-33*01*	*IGKJ2*01*	[6.3.9]	
Wat	V-KAPPA	1wtl 1wtl	A B	*IGKV1-33*01*	*IGKJ4*01*	[6.3.9]	

gions, like framework regions (FR) and complementarity-determining regions (CDR), according to the IMGT standardized nomenclature (Giudicelli and Lefranc 1999) and IMGT unique numbering (Lefranc 1997, 1998, 1999). Two-dimensional graphical representations of the V-DOMAINs designated as Colliers de Perles (Lefranc et al. 1999) are automatically produced. A query Web interface allows interactive search of the IMGT/3Dstructure-DB database. In this article, IMGT gene identification and Colliers de Perles of human IGs with known 3D structures in PDB are presented.

Materials and methods

IG and TR structural data are stored in a relational database managed by the Mysql (http://www.mysql.com) RDBMS (relational database management system). Different Perl (http://www.perl.com) programs were implemented to collect and extract IG and

Table 4 Identification of the genes and alleles expressed in the human IGL V-DOMAINs (V-J-REGIONs) of lambda chains not associated with a heavy chain. For each protein, the IMGT protein name, the protein fragment and the PDB code are indicated. For each chain, the PDB chain name, the identified *IGLV* and *IGLJ* gene and allele(s), the CDR-IMGT lengths, and the *IGLC* gene, if present, are indicated

IMGT protein name	Fragment	PDB code	Light chain				
			PDB chain name	IGK V-J-REGION			IGK C-REGION *IGKC* gene name
				IGKV gene and allele name	*IGKJ* gene and allele name	CDR-IMGT lengths	
Cle	L-LAMBDA	1lil	A	*IGLV3-1*01*	*IGLJ2*01*	[6.3.10]	*IGLC2*
		1lil	B				
Jto	V-LAMBDA	1cd0	A	*IGLV6-57*01*	*IGLJ2*01* or *IGLJ3*01*	[8.3.9]	
		1cd0	B				
Loc	L-LAMBDA	1bjm	A	*IGLV1-44*01*	*IGLJ1*01*	[8.3.11]	*IGLC1*
		1bjm	B				
		3bjl	A				
		3bjl	B				
		4bjl	A				
		4bjl	B				
Loi	V-LAMBDA	2loi	A	*IGLV3-21*01*	*IGLJ2*01* or *IGLJ3*01*	[6.3.11]	
		2loi	B				
Mcg	L-LAMBDA	1a8j	H	*IGLV2-8*01*	*IGLJ1*01*	[9.3.10]	*IGLC1*
		1a8j	L				
		1dcl	A				
		1dcl	B				
		1mcb	A				
		1mcb	B				
		1mcc	A				
		1mcc	B				
		1mcd	A				
		1mcd	B				
		1mce	A				
		1mce	B				
		1mcf	A				
		1mcf	B				
		1mch	A				
		1mch	B				
		1mci	A				
		1mci	B				
		1mcj	A				
		1mcj	B				
		1mck	A				
		1mck	B				
		1mcl	A				
		1mcl	B				
		1mcn	A				
		1mcn	B				
		1mcq	A				
		1mcq	B				
		1mcr	A				
		1mcr	B				
		1mcs	A				
		1mcs	B				
		2mcg	1				
		2mcg	2				
		3mcg	1				
		3mcg	2				
Mcg-Weir hybrid	L-LAMBDA	1mcw	M	*IGLV2-8*01*	*IGLJ1*01*	[9.3.10]	*IGLC1*
		1mcw	W	*IGLV2-23*02*	*IGLJ1*01*	[9.3.10]	*IGLC1*
Rhe	V-LAMBDA	2rhe	–	*IGLV1-36*01*	*IGLJ2*01* or *IGLJ3*01* or *IGLJ3*02*	[8.3.11	
Wil	V-LAMBDA	2cd0	A	*IGLV6-57*01*	*IGLJ2*01* or *IGLJ3*01* or *IGLJ3*02*	[8.3.9]	
			B				

Table 5 Classification by V gene names of the human IG proteins with known 3D structures: *IGHV*. For each IG V gene name, the number of PDB entries, the PDB codes, the number of proteins and the IMGT protein names associated with the gene are displayed. Only functional and mapped human *IGHV* genes are shown (Lefranc and Lefranc 2001a). *IGHV7-4-1* is, so far, the only *IGHV* polymorphic gene by insertion/deletion found in the 3D structures. Functional *IGHV* polymorphic genes by insertion/deletion (two *IGHV3* and three *IGHV4* subgroup genes) are not shown

IGHV subgroup	IMGT *IGHV* gene name	Number of PDB entries	PDB codes	Number of different heavy chains	IMGT protein names
IGHV1	*IGHV1-2*				
	IGHV1-3	3	1hzh; 1iga; 1vge	3	B12; IgA1; Tr1.9
	IGHV1-8				
	IGHV1-18				
	IGHV1-24	1	1iqd	1	BO2C11
	IGHV1-45				
	IGHV1-46				
	IGHV1-58				
	IGHV1-69	3	1g9m, 1g9n, 1gc1	1	17B
IGHV2	*IGHV2-5*				
	IGHV2-26				
	IGHV2-70				
IGHV3	*IGHV3-7*				
	IGHV3-9	3	1dfb, 1obe; 1adq	2	3D6; Rf-An
	IGHV3-11				
	IGHV3-13				
	IGHV3-15				
	IGHV3-20				
	IGHV3-21	1	1fh5	1	Mak33
	IGHV3-23	2	1hou; 1igm	2	Fv-1; Pot
	IGHV3-30	4	1aqk; 1hez; 1dee; 1dal	4	B7-15A2; FabM; IgmRf2A2; Mez
	IGHV3-33	3	8fab; 2fb4, 2ig2	2	Hil; Kol
	IGHV3-43				
	IGHV3-48				
	IGHV3-49				
	IGHV3-53				
	IGHV3-64				
	IGHV3-66				
	IGHV3-72				
	IGHV3-73				
	IGHV3-74				
IGHV4	*IGHV4-4*				
	IGHV4-28				
	IGHV4-31				
	IGHV4-34	2	1dn0, 1qlr	1	Kau
	IGHV4-39	1	1mco	1	McgHL
	IGHV4-59	4	1bvk, 1bvl; 1dl7; 7fab	3	HULYS11; M3C65; Newm
	IGHV4-61				
IGHV5	*IGHV5-51*				
IGHV6	*IGHV6-1*				
IGHV7	*IGHV7-4-1*	2	1dx3; 1cz8	2	9E; Fab-12
Total		29		23	

TR PDB data, to analyze PDB sequences, to update data, and to display structural data overviews in the IMGT Web server.

A first Perl module collects the PDB format files containing IG and TR structural data from the PDB database, weekly on the basis of keyword presence in the PDB file text. The keywords used are ANTIBODY, IMMUNOGLOBULIN, IG, FAB, FV, FC (corresponding to the Fab, Fv and Fc IG fragments, respectively), T CELL RECEPTOR, and TCR (for T cell receptors). An additional manual control allows elimination of non-IG and non-TR data, and the opportunity to check the protein names, associated ligands, protein fragment definition, and species determination. Experimental technique, X-ray diffraction resolution, bibliographic references, PDB release date, and PDB chain names are automatically extracted from the PDB files.

A Perl module extracts amino acid sequences from the PDB file ATOM records, corresponding to structure atomic coordinates. It is worth noting that there may be conflicts between sequence data from the PDB file ATOM records and those from the PDB file SEQRES records.

The IMGT gene identification is automatically determined by comparing PDB amino acid sequences with the IMGT reference directory translated sequences (Lefranc et al. 1999), using a standalone implementation version of the BLAST2 program (Altschul et al. 1990). The IMGT reference directory consists of sets of IG or TR sequences isolated from the functional and ORF allele IMGT reference sequences (Lefranc et al. 1999). By definition, the IMGT reference directory sets contain one sequence for each allele. The identification of the V gene, J gene, C gene, and, for

Table 6 Classification by V gene names of the human IG proteins with known 3D structures: *IGKV*. For each IG V gene name, the number of PDB entries, the PDB codes, the number of proteins and the IMGT protein names associated with the gene are displayed. Only functional and mapped human *IGKV* genes are shown (Lefranc and Lefranc 2001a). Functional *IGKV* genes of the distal locus (Lefranc and Lefranc 2001a) not found yet in the 3D structures are not shown

IGKV subgroup	IMGT *IGKV* gene name	Number of PDB entries	PDB codes	Number of different kappa chains	IMGT protein names
IGKV1	*IGKV1-5* *IGKV1-6* *IGKV1-9* *IGKV1-12*	2	1dfb, 1obe	1	3D6
	IGKV1-13 *IGKV1-16*	2	1iga; 1vge	2	IgA1; Tr1.9
	IGKV1-17	1	1dql	1	Mez
	IGKV1-27	2	1bvk, 1bvl	1	HULYS11
	IGKV1-33	10	1b0w, 1bre, 1qp1; 1b6d; 1cz8; 1igm; 1rei; 1ar2; 1bww; 1wtl	8	Bre; Del; Fab-12; Pot; Rei; Rei C23>V, Y32>H; Rei T45>K; Wat
	IGKV1-39	2	1hez; 1dee	2	FabM; IgmRf2A2
IGKV2	*IGKV2-24* *IGKV2-28*				
	IGKV2-29 *IGKV2-30* *IGKV2-40*	1	1hou	1	Fv-1
IGKV3	*IGKV3-11*	1	1dx3	1	9E
	IGKV3-15	4	1g9m, 1g9n, 1gc1; 1fh5	2	17B; Mak33
	IGKV3-20	4	1hzh; 1iqd; 1dn0, 1qlr	3	B12; BO2C11; Kau
IGKV4	*IGKV4-1*	10	1lve, 2lve; 4lve; 1eeu; 1eeq; 5lve; 1qac; 1efq; 3lve; 1ek3	9	Len; Len K36>T; Len M4>L, Y30>D, Q105>D, T114>H; Len M4>L, Y30>D, T114>H; Len Q105>A; Len Q105>L; Len Q44>D; Len Q44>E; Rec
IGKV5	*IGKV5-2*				
Total		39		31	

the IG heavy chains, CH exons is done by the best alignment scores in the corresponding BLAST2 output. IMGT unique numbering for V-DOMAIN and FR-IMGT and CDR-IMGT delimitations are automatically applied to the PDB sequences.

Associated heavy and light chains (designated as "partners"), which belong to the same receptor, are automatically determined on the basis of the chain type identification (heavy or light), and of the PDB chain identifier (almost always a one-letter code that follows a logical alphabetic order). A Perl module renumbers PDB atomic coordinates according to the IMGT unique numbering.

Two-dimensional Colliers de Perles representations of the V-DOMAINs and HTML pages displaying data overviews are automatically created and updated. Before public display, HTML pages are checked and additional standardized information added manually if necessary (protein mutant description in Protein name, ligand description).

Results and discussion

As of August 2001 53 different human IG proteins and 86 different PDB entries containing these proteins have been retrieved. Chimeric and humanized immunoglobulins are not included in this analysis. An overview of the human IG protein 3D structures is shown in Table 1. Seven of these proteins correspond to different mutants of the same protein Len, and two proteins are mutants of the protein Rei (Table 1). Two IG proteins are found in the same PDB structure (PDB code: 1adq; Rea, chain A, in complex with Rf-An, chains H and L). Most of the IG proteins were crystallized as IG fragments. The total different fragments found are 14 Fab, 6 Fv, 1 L-KAPPA, 4 L-LAMBDA, 14 V-KAPPA, 4 V-LAMBDA, 5 Fc, 1 CH2-EPSILON, and a CH3-CH4-EPSILON (Table 1). Only one complete IG is available, a recently crystallized complete human IgG1 (B12, PDB code: 1hzh), although there is also a complete human IgA1 theoretical model (IgA1, PDB code: 1iga, containing only alpha carbon coordinates) and one heavy chain gamma 1, characterized by a hinge deletion, associated to a light chain (McgHL, PDB code: 1mco). Some of the PDB protein names have been modified or created (if undefined) by IMGT in a standardized way (Table 1). Table 1 gives, for each IG protein, the IMGT protein name, the corresponding PDB code, the IG fragment type whose 3D structure has been determined, the number of identical IG fragments found in the structure, the name of the ligand associated with the IG protein in this structure, the experimental technique, the resolution for the X-ray diffraction experiments, the last PDB release date of the PDB file, and the primary bibliographic reference of the structure determination.

Table 7 Classification by V gene names of the human IG proteins with known 3D structures: *IGLV*. For each IG V gene name, the number of PDB entries, the PDB codes, the number of proteins and the IMGT protein names associated with the gene are displayed. Only functional and mapped human *IGLV* genes are shown (Lefranc and Lefranc 2001a)

IGLV subgroup	IMGT *IGLV* gene name	Number of PDB entries	PDB codes	Number of different lambda chains	IMGT protein names
IGLV1	*IGLV1-36*	1	2rhe	1	Rhe
	IGLV1-40	2	1aqk; 7fab	2	B7-15A2; Newm
	IGLV1-44	5	2fb4, 2ig2; 1bjm, 3bjl, 4bjl	2	Kol; Loc
	IGLV1-47				
	IGLV1-51				
IGLV2	*IGLV2-8*	19[a]	1a8j, 1dcl, 1mcb, 1mcc, 1mcd, 1mce, 1mcf, 1mch, 1mci, 1mcj, 1mck, 1mcl, 1mcn, 1mcq, 1mcr, 1mcs, 2mcg, 3mcg; 1mcw[a]; 1mco	1[a]	Mcg; Mcg-Weir hybrid[a]; McgHL
	IGLV2-11				
	IGLV2-14				
	IGLV2-18				
	IGLV2-23	1[a]	1mcw[a]	1[a]	Mcg-Weir hybrid[a]
IGLV3	*IGLV3-1*	1	1lil	1	Cle
	IGLV3-9				
	IGLV3-10				
	IGLV3-12				
	IGLV3-16				
	IGLV3-19				
	IGLV3-21	2	2loi; 1adq	2	Loi; Rf-An
	IGLV3-22				
	IGLV3-25	1	8fab	1	Hil
	IGLV3-27				
IGLV4	*IGLV4-3*				
	IGLV4-60				
	IGLV4-69				
IGLV5	*IGLV5-37*				
	IGLV5-39				
	IGLV5-45				
	IGLV5-52				
IGLV6	*IGLV6-57*	2	1cd0; 2cd0	2	Jto; Wil
IGLV7	*IGLV7-43*				
	IGLV7-46	1	1dl7	1	M3C65
IGLV8	*IGLV8-61*				
IGLV9	*IGLV9-49*				
IGLV10	*IGLV10-54*				
Total		35		14	

[a] Mcg-Weir is a hybrid lambda chain dimer consisting of one lambda Mcg chain non-covalently linked to one lambda Weir chain. The Mcg-Weir hybrid protein and the 1mcw PDB code are only counted once (on the *IGLV2-23* line for the Weir chain)

IMGT gene identification

An important increment value added by the IMGT automatic expertise is the identification of the genes and alleles expressed in the IG V-DOMAIN and C-DOMAIN of the PDB sequences, according to the standardized IMGT gene nomenclature (Tables 2, 3, 4). This identification allows automatic direct links between gene and structural data for the first time (Tables 5, 6, 7).

Twenty-three different associations of heavy and light chains (16 with kappa and 7 with lambda), 15 different kappa light chains not associated to heavy chains, and 8 different lambda light chains not associated to heavy chains are found (Tables 2, 3, 4; Figs. 1, 2, 3). Twelve of the 38–46 functional human IGHV genes (Table 5), 11 of the 17–20 functional human IGKV genes of the proximal cluster (Table 6), and 10 of the 32–33 human IGLV genes (Table 7) are expressed in the IG V-DOMAIN of the PDB sequences. One-third of the total number of human functional IG V genes have a corresponding IG protein structure. It seems that many more human IG 3D structures will be necessary to obtain a more complete overview of the global structural repertoire of the human IG proteins.

The IMGT/JunctionAnalysis program available in the IMGT Web interface (http://imgt.cines.fr) allows D gene identification from nucleotide sequences, but not from protein sequences. Indeed, D gene identification from PDB amino acid sequences is too uncertain owing to the shortness of the D sequences in the V-D-J rearrange-

Fig. 1 Protein display of the human IGH V-DOMAINs (V-D-J-REGIONs). Numbering is according to the IMGT unique numbering for V-DOMAIN. CDR-IMGT regions are colored as follows: CDR1-IMGT (*red*), CDR2-IMGT (*orange*), CDR3-IMGT (*purple*). The *asterisk* indicates that there is a partner light chain. The amino acid difference found in the CDR3-IMGT of the protein 17B, at position 112.2, between 1g9m_H and 1g9n_H (amino acid R) and 1gc1_H (amino acid D) is shown by a *red vertical bar*. For the CDR3-IMGT, if all positions are occupied, this numbering corresponds to a rearranged CDR3-IMGT of 13 amino acids (positions 105–117). This numbering is convenient to use since 80% of the IMGT/LIGM-DB immunoglobulin and T cell receptor rearranged sequences have a CDR3-IMGT length less than or equal to 13 amino acids. If the CDR3-IMGT length is less than 13 amino acids, gaps are created from the top of the loop, in the following order 111, 112, 110, 113, 109, 114, etc. If the CDR3-IMGT length is more than 13 amino acids, additional positions are created between positions 111 and 112 at the top of the CDR3-IMGT loop in the following order 112.1,111.1, 112.2, 111.2, 112.3, 111.3, etc. The four first *underlined amino acids* QVKL of the Tr1.9 IGH V-REGION in PDB : 1vge_H and in the corresponding L12098 EMBL/GenBank/DDBJ/IMGT accession number (not shown) are introduced by the primer. The IgA1 theoretical model (PDB: 1iga) uses the 1vge_H sequence and structure for the heavy chain V-DOMAIN (M.-P.L., IMGT, http://imgt.cines.fr, 28/09/2001)

ment, to the N-region diversity and to the somatic hypermutations. This identification was therefore not included in the tables.

Concerning the different C-DOMAINs found in the IG fragments, 9 IGHG1 CH1, 1 IGHG4 CH1, 4 IGHM CH1, 6 IGHE CH2, 6 IGHE CH3, and 1 IGHE CH4 were found in the heavy chains; 12 IGKC, 6 IGLC1, 2 IGLC2, and 3 IGLC3 were found in the light chains.

Protein displays and Colliers de Perles

Another important area of expertise is the description of the IG V-DOMAIN of the PDB sequences according to the IMGT unique numbering. This numbering represents a unified amino acid nomenclature, in which structurally equivalent amino acids in the different antigen receptors (IG and TR), different chain types (heavy or light chains for IG; alpha, beta, gamma or delta for TR), and different species are identified by the same number (Lefranc 1997, 1998). Corresponding protein displays of the PDB sequences with FR-IMGT and CDR-IMGT delimitations are shown (Figs. 1, 2, 3).

This standardization is useful for describing mutations and allelic polymorphisms and for establishing correlations between amino acid positions, in the sequences

Fig. 2 Protein display of the human IGK V-DOMAINs (V-J-REGIONs). CDR-IMGT regions are colored as follows: CDR1-IMGT (*blue*), CDR2-IMGT (*green*), CDR3-IMGT (*green-blue*). The *asterisk* indicates that there is a partner heavy chain. For the mutants of the proteins Len and Rei, mutation positions are indicated by *red vertical bars*. The four first *underlined amino acids* ELVM of the Tr1.9 IGK V-REGION in PDB: 1vge_L and in the corresponding L12099 EMBL/GenBank/DDBJ/IMGT accession number (not shown) are introduced by the primer. The IgA1 theoretical model (PDB: 1iga) uses the 1vge_L sequence and structure for the kappa chain V-DOMAIN (M.-P.L., IMGT, http://imgt.cines.fr, 28/09/2001)

and in the protein 3D structures. Data summary on the CDR-IMGT lengths, known to be important for the CDR conformations, can be automatically extracted (Tables 8, 9, 10). Two-dimensional Colliers de Perles representations of the different V-DOMAINs are provided (Fig. 4). A crucial advantage is the renumbering of the PDB atomic coordinates according to the IMGT unique numbering, allowing large and automatic sequence–structure relationship analysis. The corresponding files will be available in IMGT.

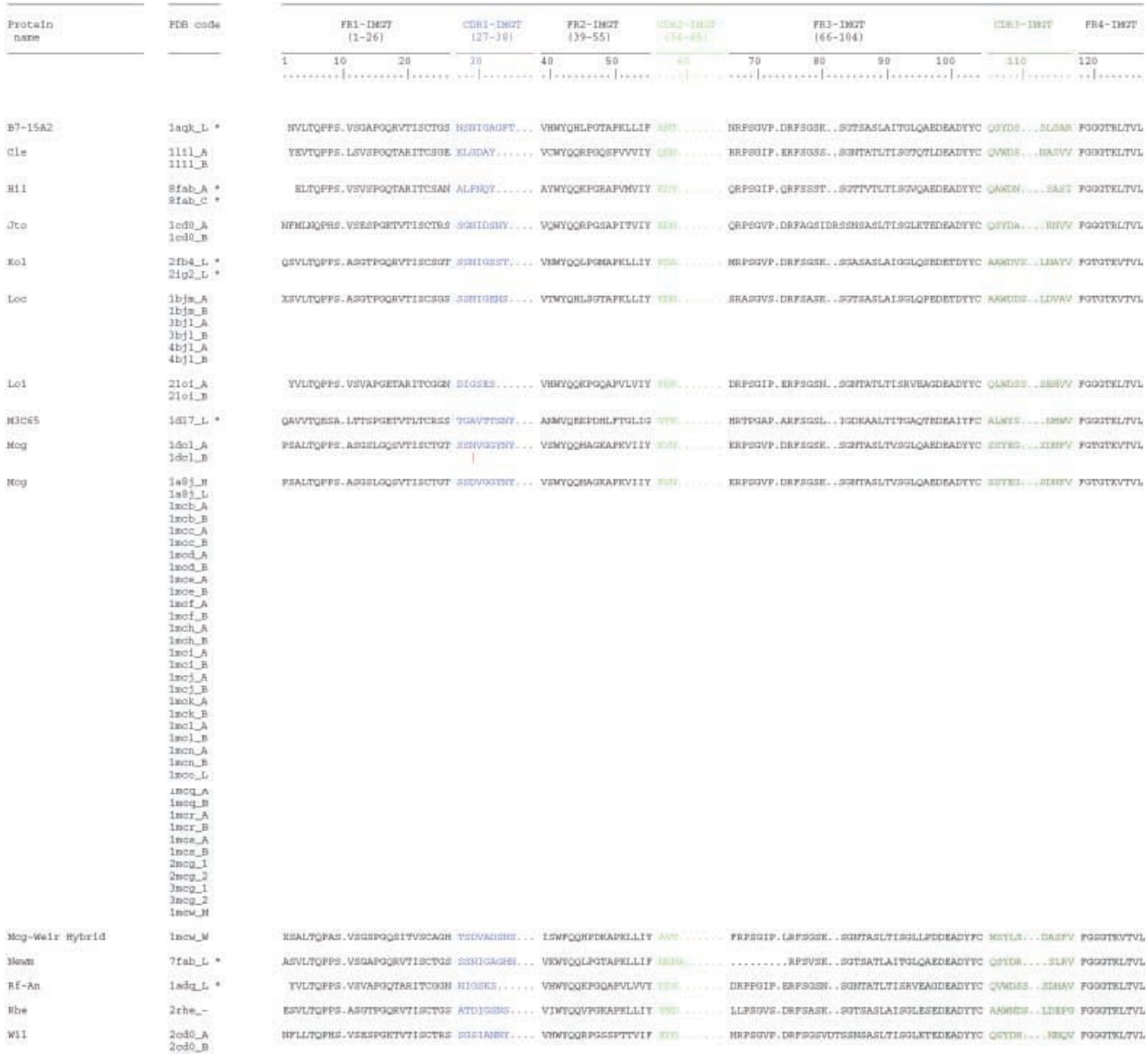
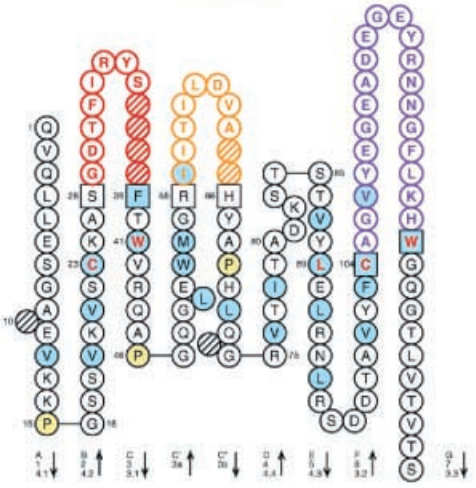


Fig. 3 Protein display of the human IGL V-DOMAINs (V-J-REGIONs). CDR-IMGT regions are colored as follows: CDR1-IMGT (*blue*), CDR2-IMGT (*green*), CDR3-IMGT (*green-blue*). The *asterisk* indicates that there is a partner heavy chain. For the protein Mcg an amino acid difference (*red vertical bar*) was found at position 29, in the 1dcl PDB entry

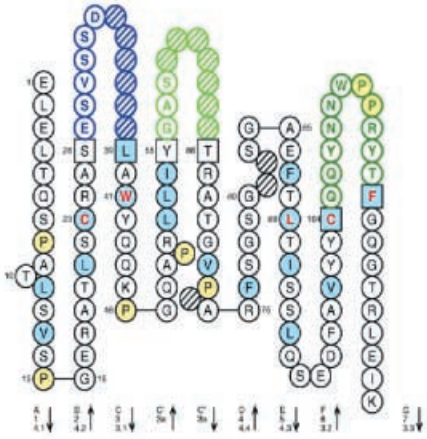
Fig. 4a–c IMGT V-DOMAINs Collier de Perles of **a** the human IGH V-DOMAINs (*left*) and IGK or IGL V-DOMAINs (*right*) for the 23 proteins with associated heavy and light chains (see Table 2), **b** the human IGK V-DOMAINs of the six kappa light chains not associated to heavy chains in the crystals (mutants not shown) (see Table 3), **c** the human IGL V-DOMAINs of the eight lambda light chains not associated to heavy chains in the crystals (see Table 4). The IgA1 theoretical model (PDB: 1iga) uses the Tr1.9 (PDB: 1vge) sequences and structures for the heavy and light chain V-DOMAINs. The four first amino acids QVKL and ELVM of the IGH and IGK V-DOMAINs, respectively are introduced by the primers (M.-P.L., IMGT, http://imgt.cines.fr, 28/09/2001). Amino acids are shown in the *one-letter abbreviation*. Hydrophobic amino acids (hydropathy index with positive value) and Tryptophan (W) found at a given position in more than 50% of analyzed IG sequences are shown in *blue*. All Proline (P) are shown in *yellow*. The CDR-IMGT are limited by amino acids shown in *squares*, which belong to the neighboring FR-IMGT. The CDR3-IMGT extend from position 105 to position 117 preceding the 118 J-PHE or J-TRP. Numbering of the CDR3-IMGT amino acids is shown in protein displays. *Hatched circles or squares* correspond to missing positions according to the IMGT unique numbering. *Arrows* indicate the direction of the beta sheets and their different designations in 3D structure. CDR-IMGT regions are colored as follows: for IGK and IGL V-DOMAIN : CDR1-IMGT (*blue*), CDR2-IMGT (*green*), and CDR3-IMGT (*green-blue*), and for IGH V-DOMAIN: CDR1-IMGT (*red*), CDR2-IMGT (*orange*), and CDR3-IMGT (purple) ▶

A

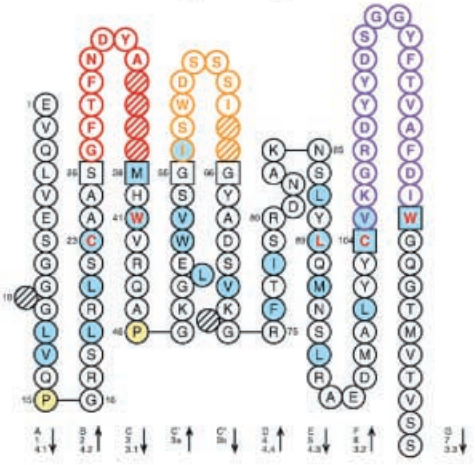
17B (1g9m_H)
[8.8.21]



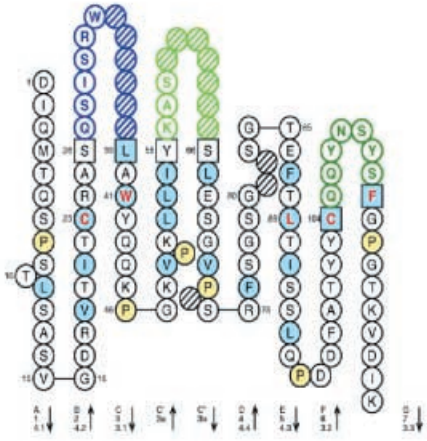
17B (1g9m_L)
[6.3.11]



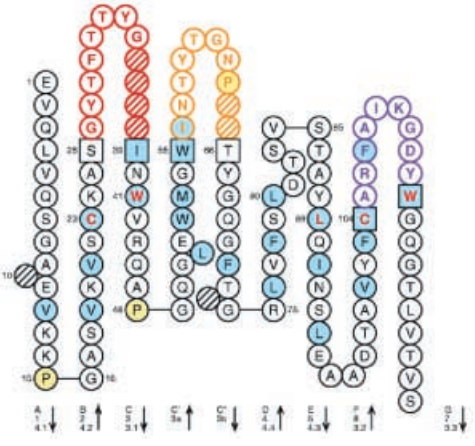
3D6 (1dfb_H)
[8.8.19]



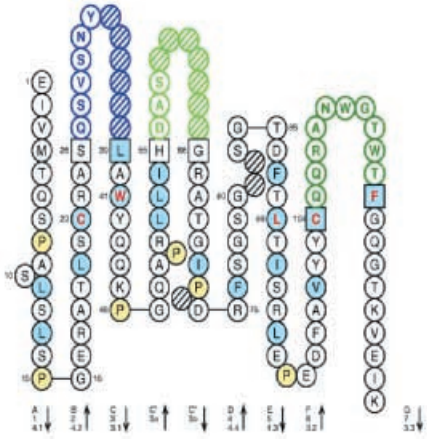
3D6 (1dfb_L)
[6.3.7]



9E (1dx3_H)
[8.8.9]



9E (1dx3_L)
[6.3.10]



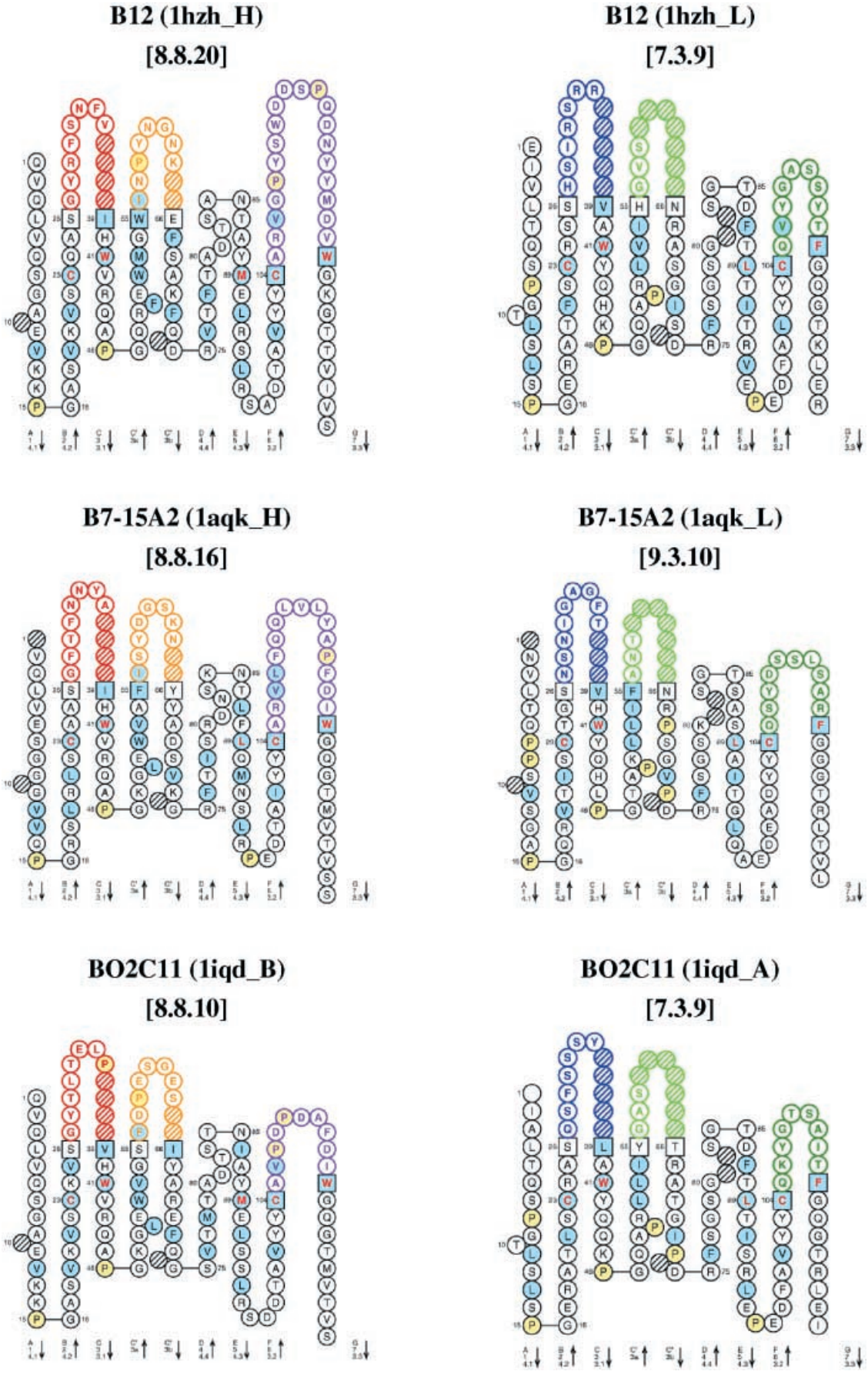


Fig. 4 Legend see page 870

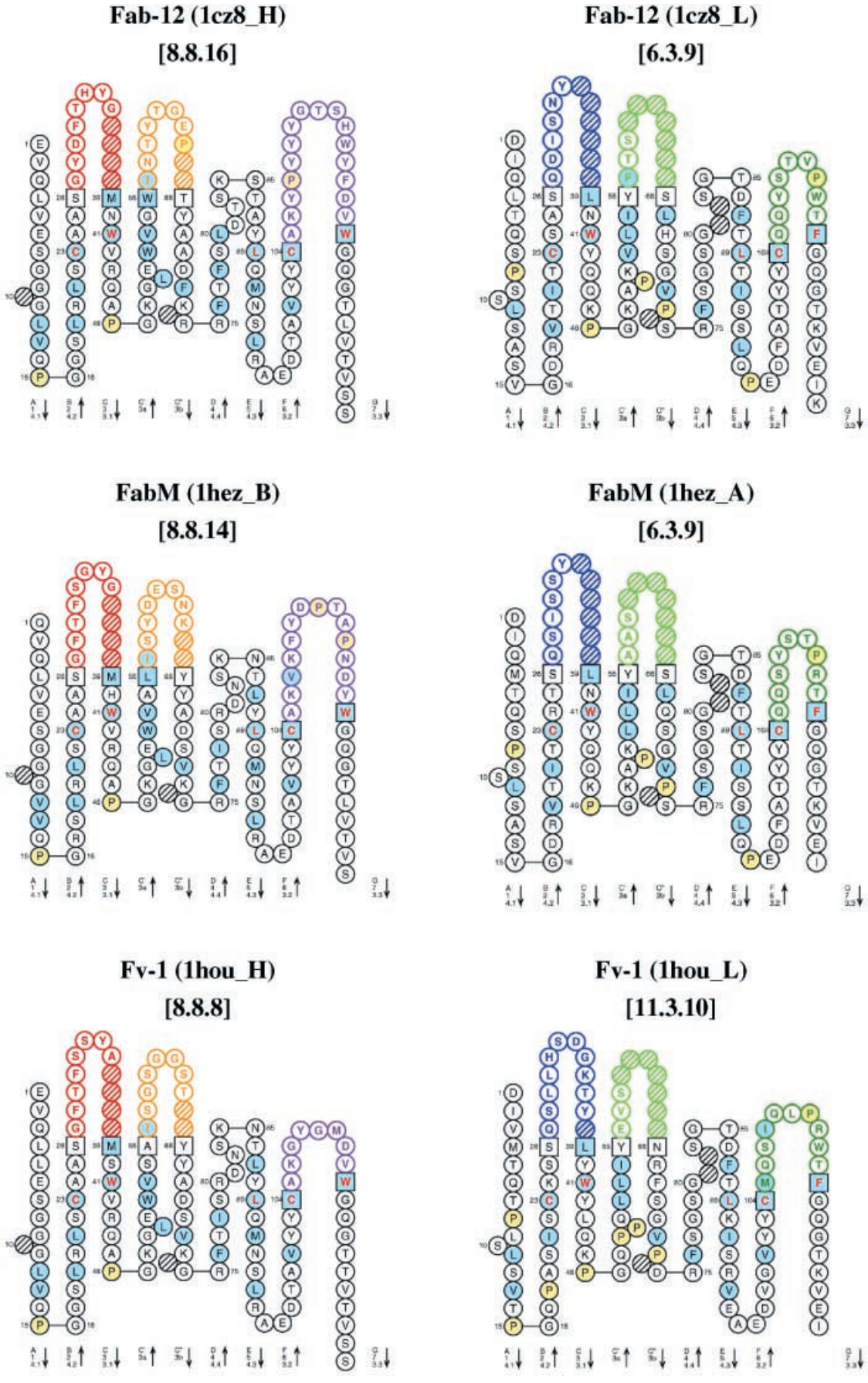


Fig. 4 Legend see page 870

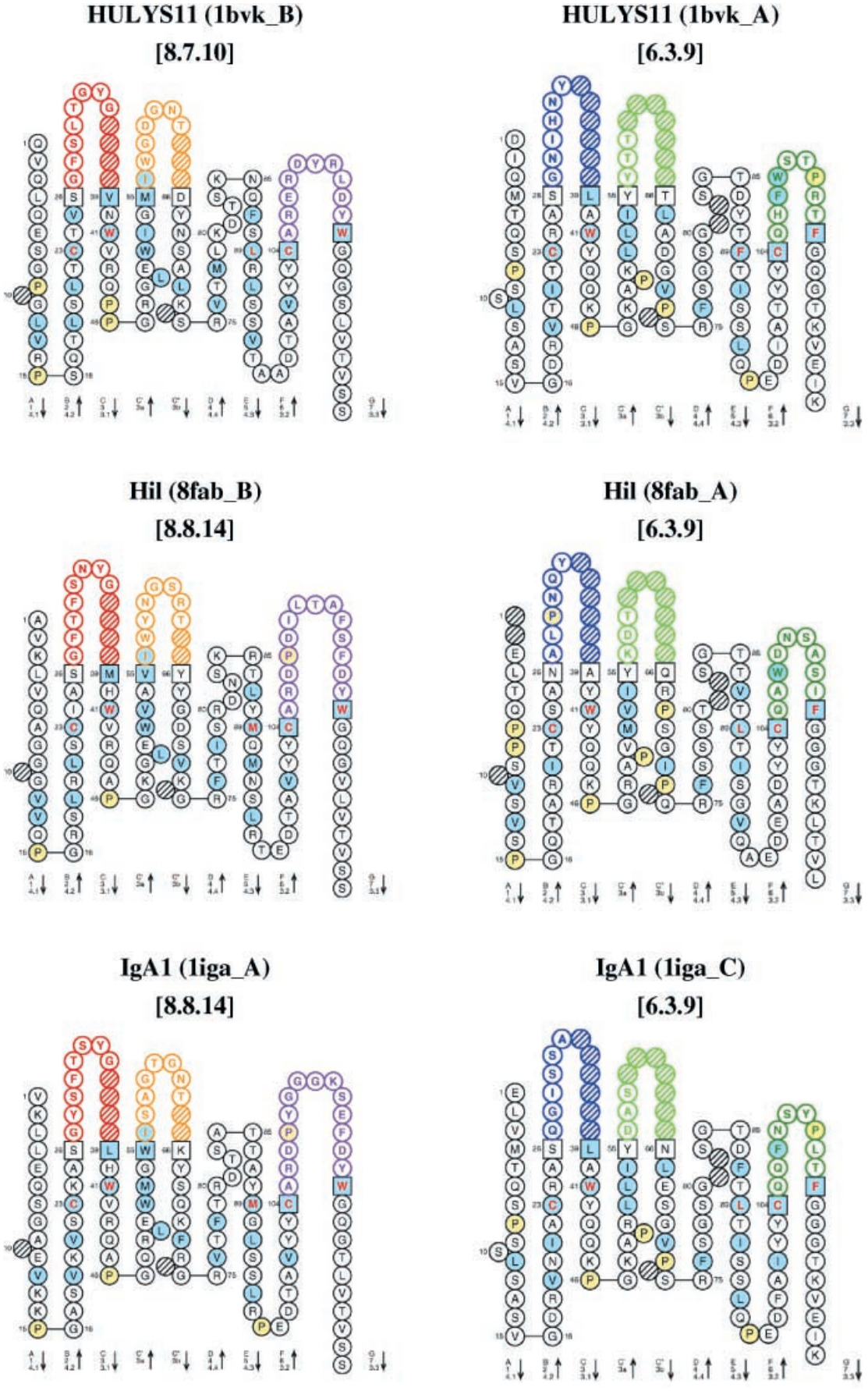


Fig. 4 Legend see page 870

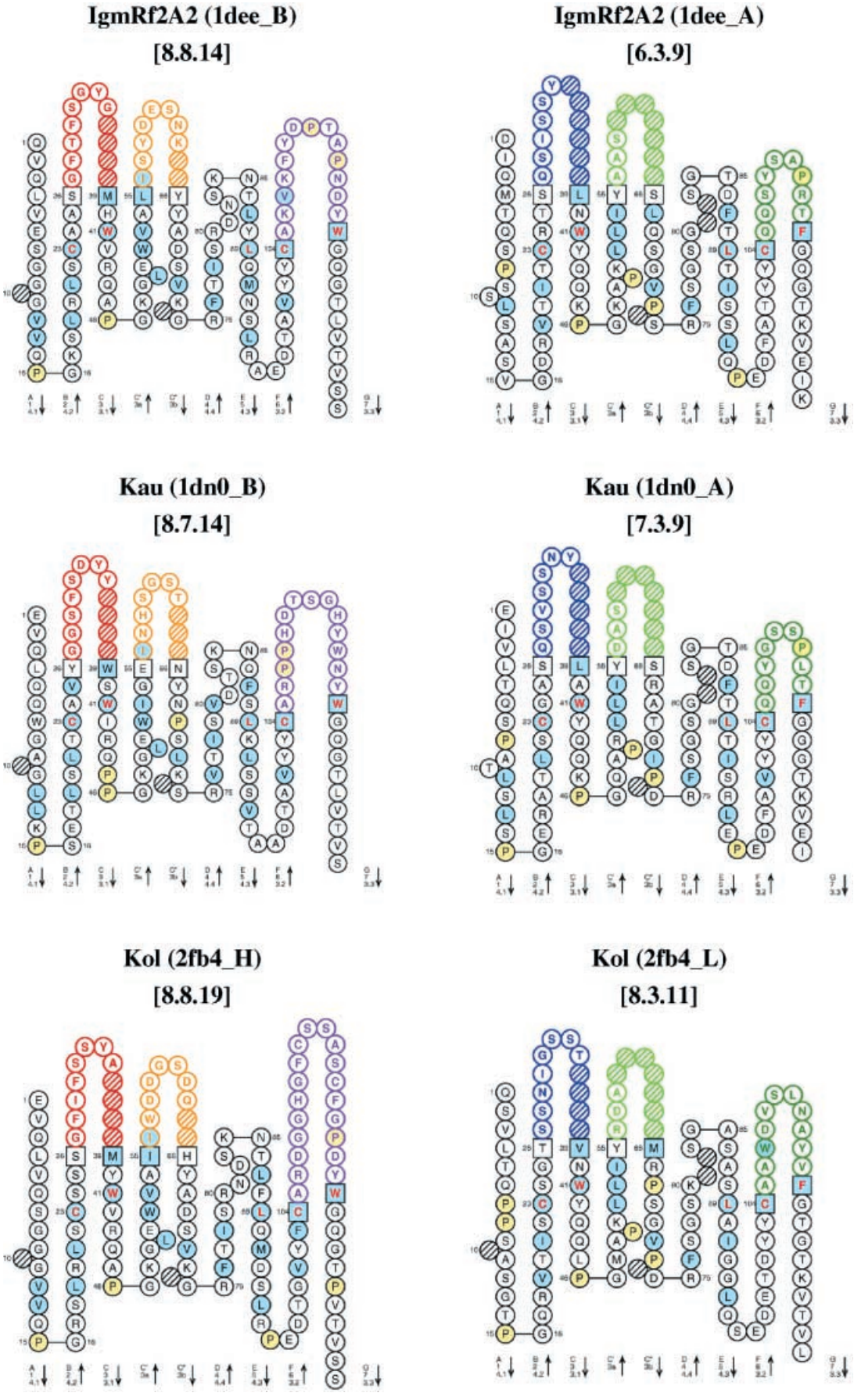


Fig. 4 Legend see page 870

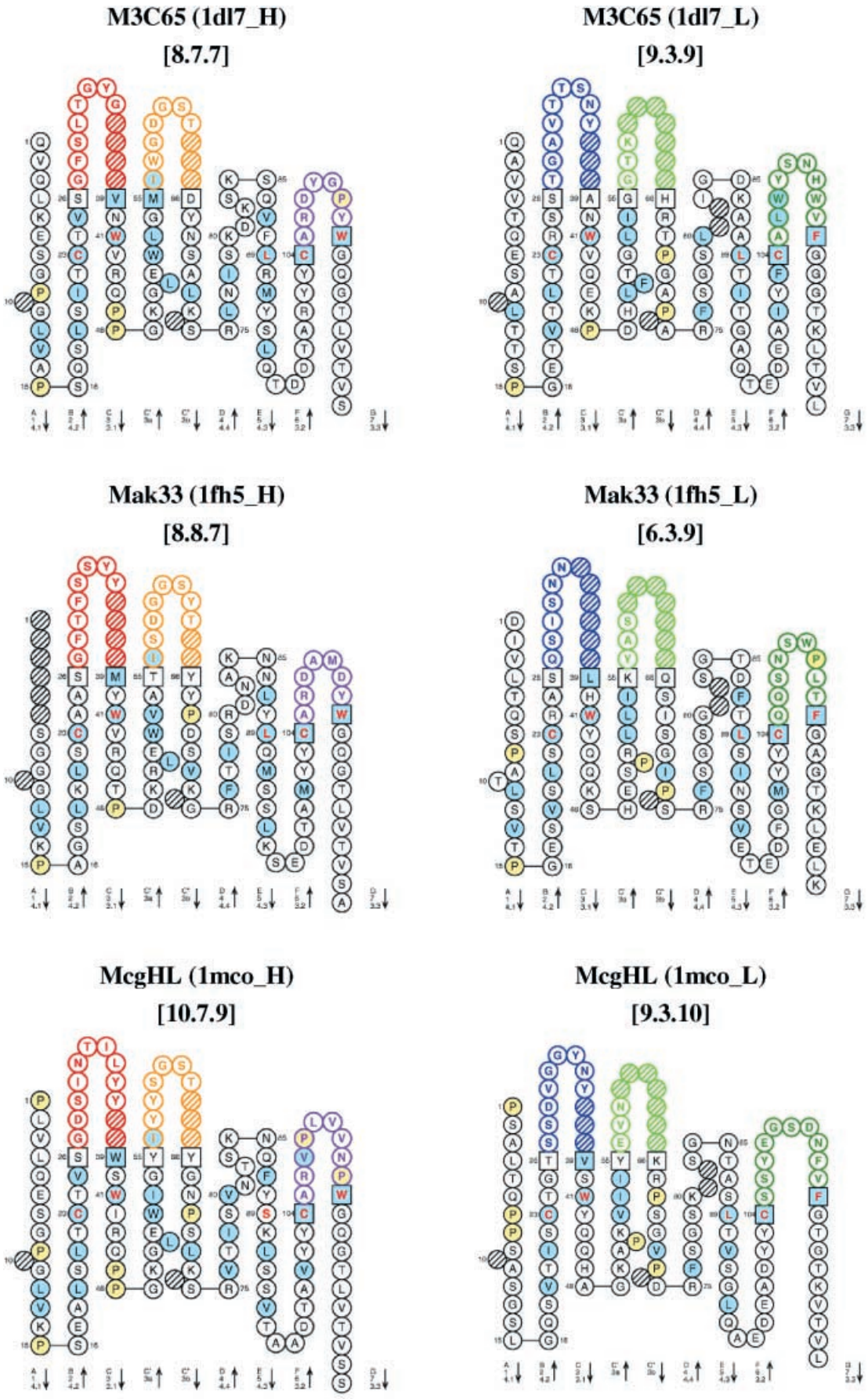


Fig. 4 Legend see page 870

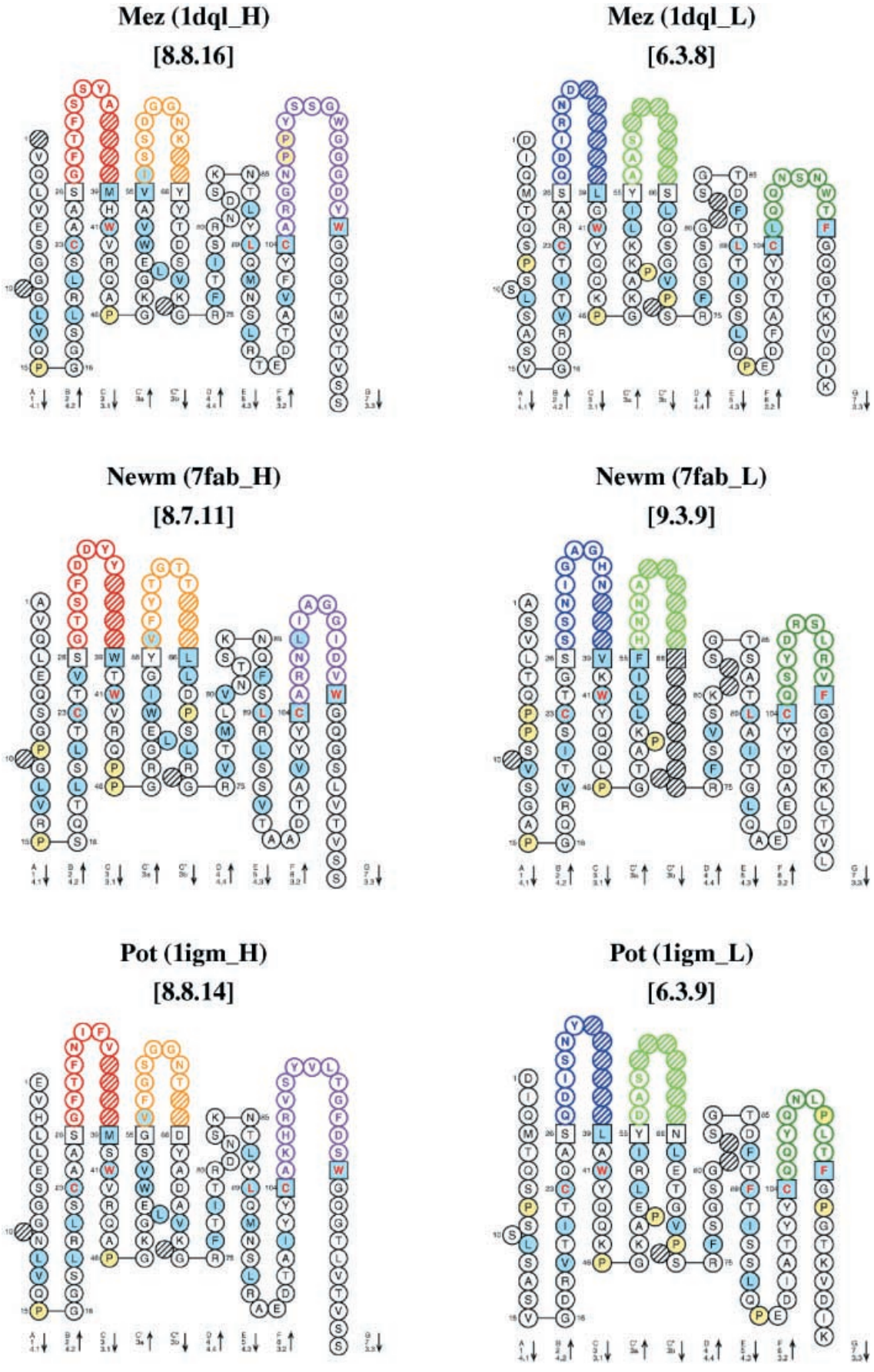


Fig. 4 Legend see page 870

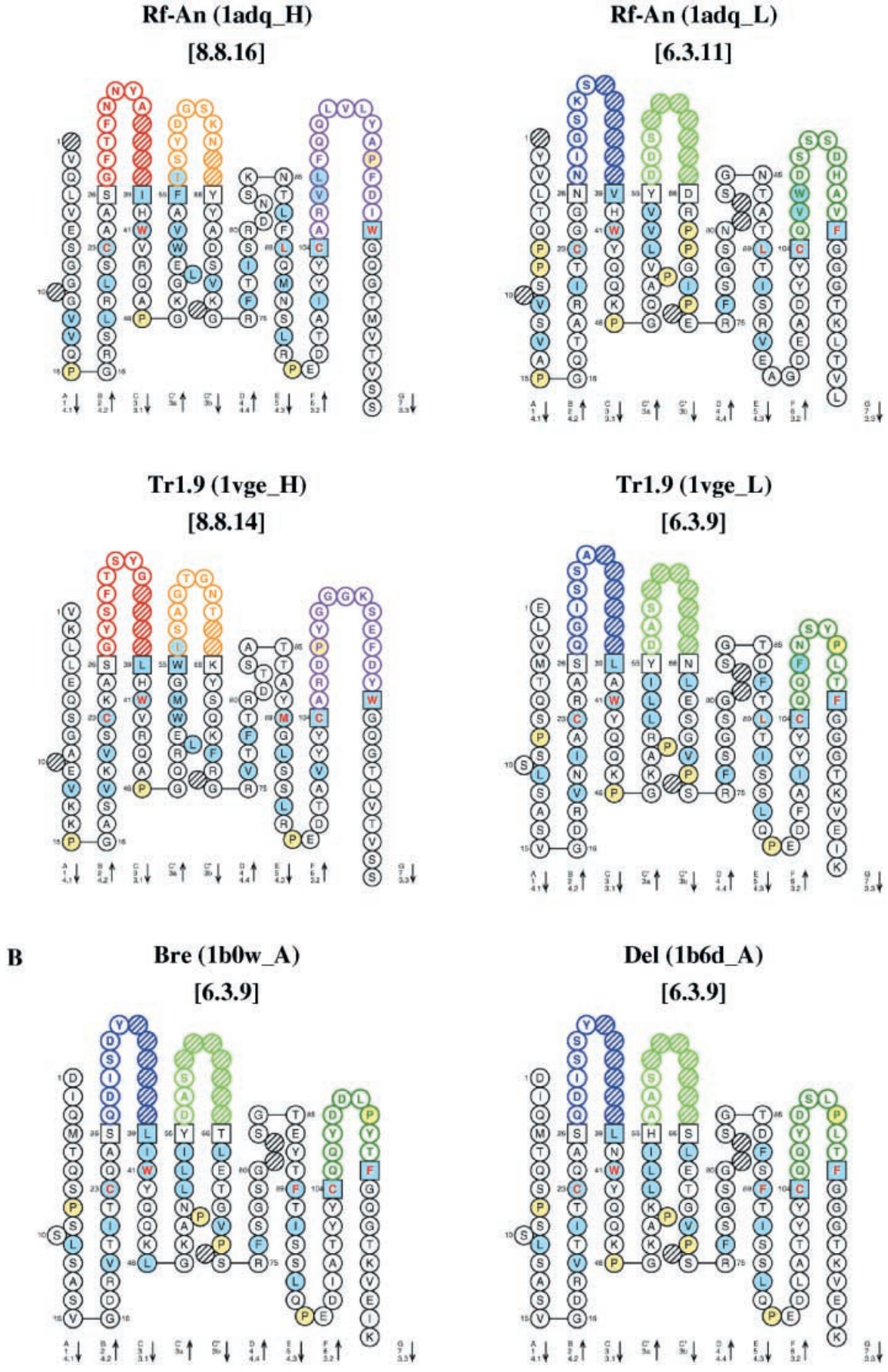


Fig. 4 Legend see page 870

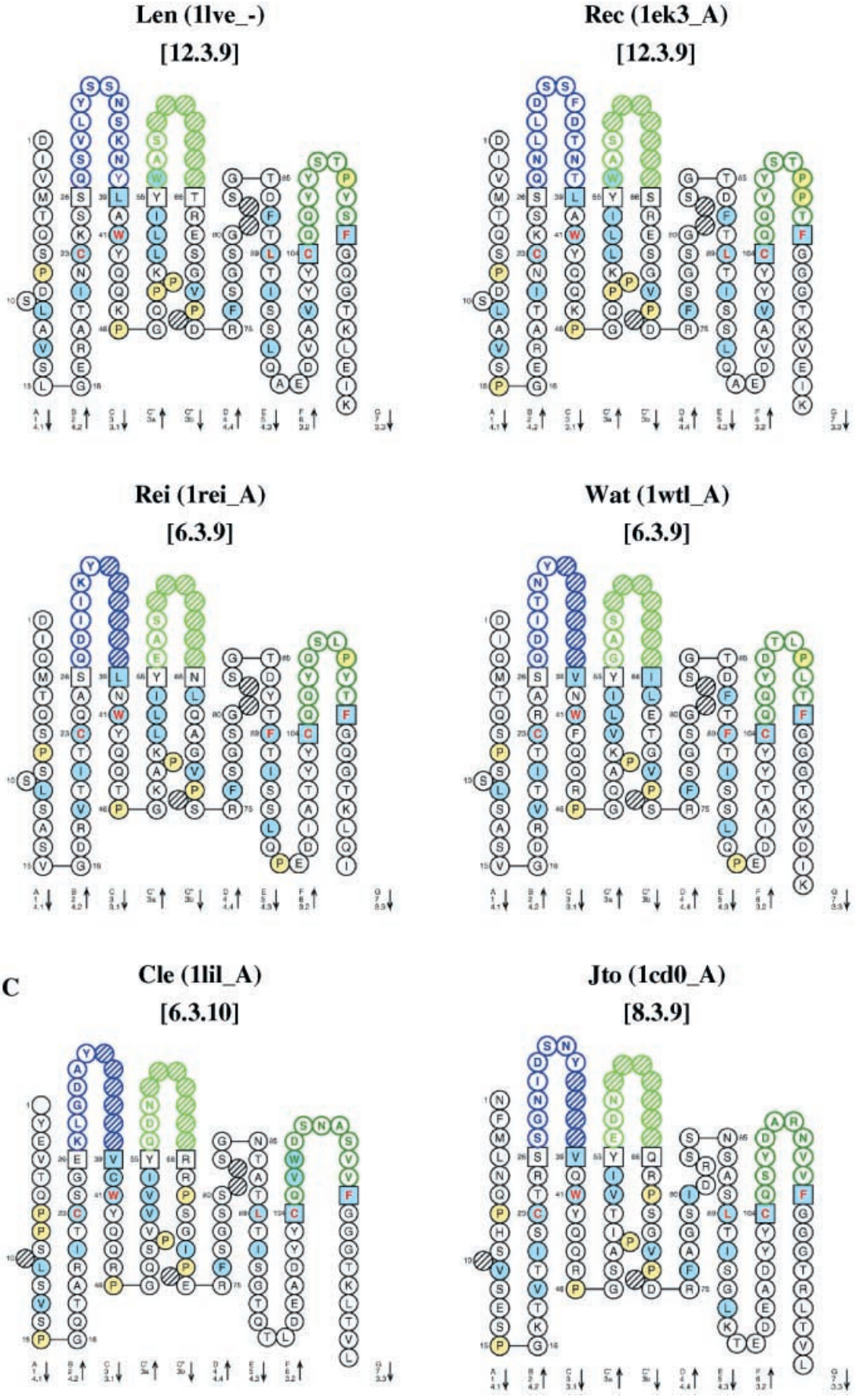


Fig. 4 Legend see page 870

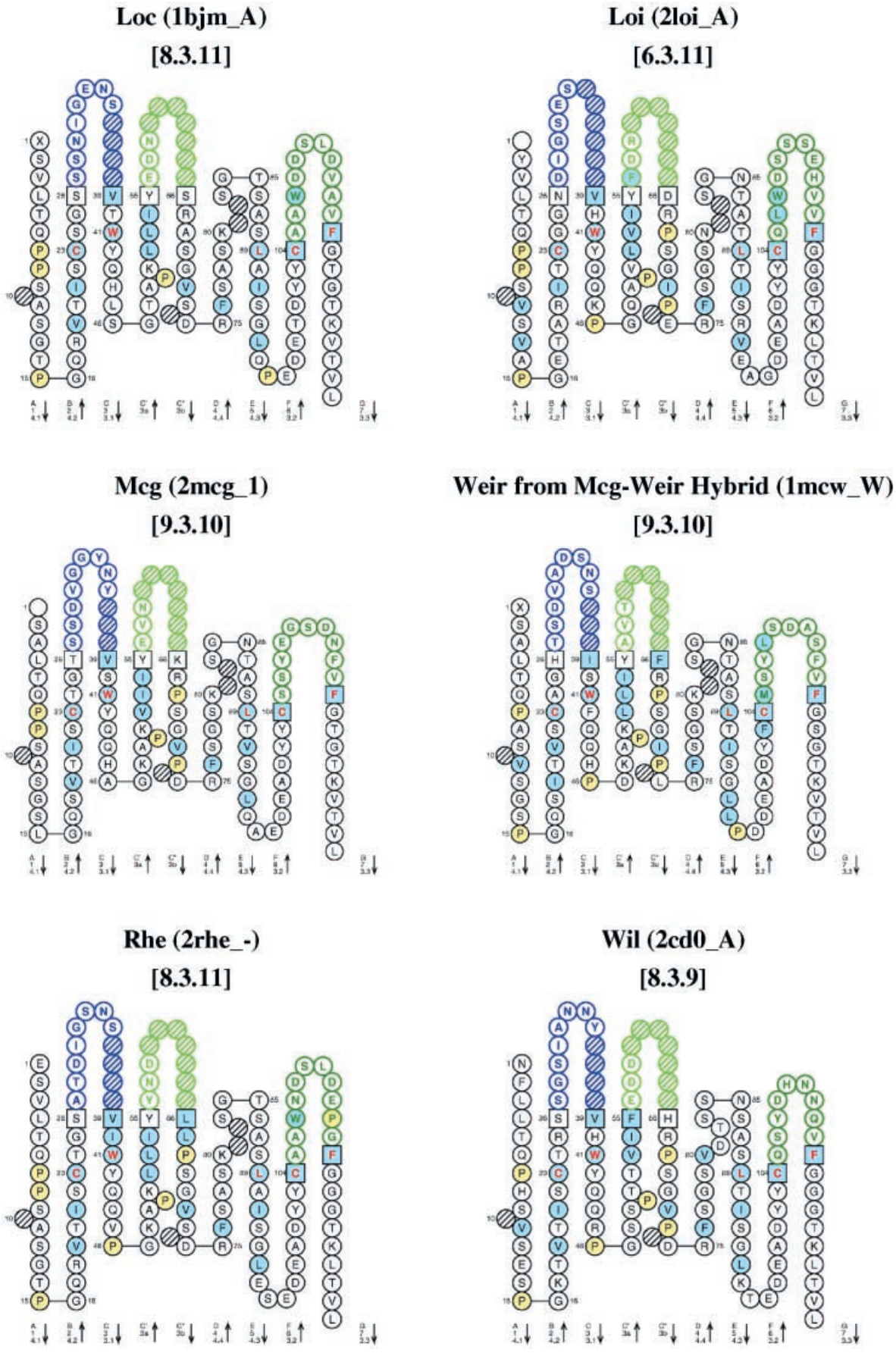


Fig. 4 Legend see page 870

Table 8 Classification by CDR-IMGT lengths of the human IG proteins with known 3D structures: *IGHV*. For each CDR-IMGT length, the corresponding identified V and J gene and allele(s), the IMGT protein name, and the PDB code are shown

CDR-IMGT lengths	IMGT *IGHV* gene and allele name	IMGT *IGHJ* gene and allele name	IMGT protein name	PDB codes
[8.7.7]	*IGHV4-59*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	M3C65	1dl7
[8.7.10]	*IGHV4-59*01* or *IGHV4-59*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	HULYS11	1bvk,1bvl, 1bvl
[8.7.11]	*IGHV4-59*04*	*IGHJ6*01* or *IGHJ6*02*	Newm	7fab
[8.7.14]	*IGHV4-34*01* or *IGHV4-34*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	Kau	1dn0,1qlr, 1qlr
[8.8.7]	*IGHV3-21*01* or *IGHV3-21*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	Mak33	1fh5
[8.8.8]	*IGHV3-23*01*	*IGHJ6*01* or *IGHJ6*02*	Fv-1	1hou
[8.8.9]	*IGHV7-4-1*02*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	9E	1dx3
[8.8.10]	*IGHV1-24*01*	*IGHJ3*02*	BO2C11	1iqd
[8.8.14]	*IGHV1-3*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	IgA1	1iga
			Tr1.9	1vge
	*IGHV3-23*01*	*IGHJ5*01*	Pot	1igm
	*IGHV3-30*18*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	IgmRf2A2	1dee
			FabM	1hez
	*IGHV3-33*01* or *IGHV3-33*04*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	Hil	8fab
[8.8.16]	*IGHV3-9*01*	*IGHJ4*01* or *IGHJ4*02* or *IGHJ4*03*	Rf-An	1adq
	*IGHV3-30*01* or *IGHV3-30*04* or *IGHV3-30*07* or *IGHV3-30*11* or *IGHV3-30*14* or *IGHV3-30*16* or *IGHV3-30*17* or *IGHV3-30-3*01*	*IGHJ3*02*	B7-15A2	1aqk
	*IGHV3-30*10*	*IGHJ3*01* or *IGHJ3*02*	Mez	1dql
	*IGHV7-4-1*02*	*IGHJ2*01*	Fab-12	1cz8
[8.8.19]	*IGHV3-9*01*	*IGHJ3*01* or *IGHJ3*02*	3D6	1dfb,1obe
	*IGHV3-33*01* or *IGHV3-33*04*	*IGHJ6*01* or *IGHJ6*02*	Kol	2fb4,2ig2
[8.8.20]	*IGHV1-3*01*	*IGHJ6*03*	B12	1hzh
[8.8.21]	*IGHV1-69*02* or *IGHV1-69*04*	*IGHJ1*01*	17B	1g9m,1g9n, 1gc1
[10.7.9]	*IGHV4-39*01* or *IGHV4-39*06*	*IHGJ5*02*	McgHL	1mco

Table 9 Classification by CDR-IMGT lengths of the human IG proteins with known 3D structures: *IGKV*. For each CDR-IMGT length, the corresponding identified V and J gene and allele(s), the IMGT protein name, and the PDB code are shown

CDR-IMGT lengths	IMGT *IGKV* gene and allele name	IMGT *IGKJ* gene	IMGT protein name	PDB codes and allele name
[6.3.7]	*IGKV1-5*03*	*IGKJ3*01*	3D6	1dfb,1obe
[6.3.8]	*IGKV1-17*01*	*IGKJ1*01*	Mez	1dql
[6.3.9]	*IGKV1-13*02*	*IGKJ4*01*	IgA1	1iga
			Tr1.9	1vge
	*IGKV1-27*01*	*IGKJ1*01*	HULYS11	1bvk,1bvl, 1bvl
	*IGKV1-33*01*	*IGKJ1*01*	Fab-12	1cz8
		*IGKJ2*01*	Rei C23>V, Y32>H	1ar2
			Bre	1b0w,1bre, 1bre, 1bre, 1bre, 1bre, 1bre, 1qp11qp1, 1qp1,
			Rei T45>K	1bww
			Rei	1rei
		*IGKJ3*01*	Pot	1igm
		*IGKJ4*01*	Del	1b6d
			Wat	1wtl
	*IGKV1-39*01*	*IGKJ1*01*	IgmRf2A2	1dee
			FabM	1hez
	*IGKV3-15*01*	*IGKJ4*01*	Mak33	1fh5
[6.3.10]	*IGKV3-11*01*	*IGKJ1*01*	9E	1dx3
[6.3.11]	*IGKV3-15*01*	*IGKJ2*01*	17B	1g9m,1g9n, 1gc1
[7.3.9]	*IGKV3-20*01*	*IGKJ2*01*	B12	1hzh
		*IGKJ4*01*	Kau	1dn0,1qlr, 1qlr
		*IGKJ5*01*	BO2C11	1iqd
[11.3.10]	*IGKV2-29*01*	*IGKJ1*01*	Fv-1	1hou
[12.3.9]	*IGKV4-1*01*	*IGKJ2*01*	Len M4>L, Y30>D, T114>H	1eeq
			Len M4>L, Y30>D, Q105>D, T114>H	1eeu
			Len Q44>D	1efq
			Len	1lve,2lve
			Len Q105>L	1qac
			Len Q44>E	3lve
			Len K36>T	4lve
			Len Q105>A	5lve
		*IGKJ4*01*	Rec	1ek3

Table 10 Classification by CDR-IMGT lengths of the human IG proteins with known 3D structures: *IGLV*. For each CDR-IMGT length, the corresponding identified V and J gene and allele(s), the IMGT protein name, and the PDB code are shown

CDR-IMGT lengths	IMGT *IGLV* gene and allele name	IMGT *IGLJ* gene and allele name	IMGT protein name	PDB codes
[6.3.9]	*IGLV3-25*02*	*IGLJ2*01*	Hil	8fab
[6.3.10]	*IGLV3-1*01*	*IGLJ2*01*	Cle	1lil
[6.3.11]	*IGLV3-21*01*	*IGLJ2*01* or *IGLJ3*01*	Loi	2loi
	*IGLV3-21*02*	*IGLJ3*01* or *IGLJ3*02*	Rf-An	1adq
[8.3.9]	*IGLV6-57*01*	*IGLJ2*01* or *IGLJ3*01*	Jto	1cd0
		*IGLJ2*01* or *IGLJ3*01* or *IGLJ3*02*	Wil	2cd0
[8.3.11]	*IGLV1-36*01*	*IGLJ2*01* or *IGLJ3*01* or *IGLJ3*02*	Rhe	2rhe
	*IGLV1-44*01*	*IGLJ1*01*	Loc	1bjm,3bjl, 3bjl, 4bjl, 4bjl
			Kol	2fb4,2ig2
[9.3.9]	*IGLV1-40*01*	*IGLJ3*01* or *IGLJ3*02*	Newm	7fab
	*IGLV7-46*01*	*IGLJ3*02*	M3C65	1dl7
[9.3.10]	*IGLV1-40*01*	*IGLJ3*01* or *IGLJ3*02*	B7-15A2	1aqk
	*IGLV2-8*01*	*IGLJ1*01*	Mcg	1a8j,1dcl, 1dcl, 1mcb, 1mcb, 1mcc, 1mcc, 1mcd, 1mcd, 1mce, 1mce, 1mcf, 1mcf, 1mch, 1mch, 1mci, 1mci, 1mcj, 1mcj, 1mck, 1mck, 1mcl, 1mcl, 1mcn, 1mcn, 1mcq, 1mcq, 1mcr, 1mcr, 1mcs, 1mcs, 2mcg, 2mcg, 3mcg, 3mcg
			McgHL	1mco
			Mcg-Weir hybrid	1mcw
	*IGLV2-23*02*	*IGLJ1*01*	Mcg-Weir hybrid	1mcw

Conclusion

By providing the precise identification of the genes expressed in the proteins with known 3D structures, the IMGT/3Dstructure-DB database realizes, for the first time, the interoperability between a sequence database and the PDB 3D structure database. Since IMGT nomenclature has been approved by HUGO and has reciprocal links to GDB and LocusLink, this interoperability can now be extended to the genome databases. Protein displays and Colliers de Perles representations of the human IG with known 3D structures are described according to the IMGT unique numbering and can therefore be easily compared with corresponding germline data (Lefranc and Lefranc 2001a). A user-friendly query Web interface allows interactive search of the IMGT/3Dstructure-DB data. This unique expertised resource will be extended to comprise immunoglobulins and T cell receptors from other species for which 3D structures are available.

Acknowledgements We are grateful to Gérard Lefranc for helpful discussion. We thank Olga Posukh, Oksana Kravchuk, Gunilla Norhagen, and Per-Erik Engstrom for their contribution to the gene identification and ligand standardized description. IMGT is funded by the European Union's 5th PCRDT (QLG2-2000-01287) program, the Centre National de la Recherche Scientifique, the Ministère de la Recherche and the Ministère de l'Education Nationale. M.R. was supported by the Association pour la Recherche sur le Cancer and the Région Languedoc-Roussillon.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barbié V, Lefranc M-P (1998) The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp Clin Immunogenet 15:171–183

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J, Berman HM (2001) The PDB data uniformity project. Nucleic Acids Res 29:214–218

Folch G, Lefranc M-P (2000a) The human T cell receptor beta diversity (TRBD) and beta joining (TRBJ) genes. Exp Clin Immunogenet 17:107–114

Folch G, Lefranc M-P (2000b) The human T cell receptor beta variable (TRBV) genes. Exp Clin Immunogenet 17:42–54

Giudicelli V, Lefranc M-P (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 15:1047–1054

Lefranc M-P (1997) Unique database numbering system for immunogenetic analysis. Immunol Today 18:509

Lefranc M-P (1998) IMGT (ImMunoGeneTics) locus on focus. A new section of experimental and clinical immunogenetics. Exp Clin Immunogenet 15:1–7

Lefranc M-P (1999) The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. Immunologist 7:132–136

Lefranc M-P (2001a) IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 29:207–209

Lefranc M-P (2001b) Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp Clin Immunogenet 18:100–116

Lefranc M-P, Lefranc G (2001a) The immunoglobulin FactsBook. Harcourt Academic Press, London

Lefranc M-P, Lefranc G (2001b) The T cell receptor FactsBook. Harcourt Academic Press, London

Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Muller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D (1999) IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 27:209–212

Pallarès N, Frippiat JP, Giudicelli V, Lefranc M-P (1998) The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp Clin Immunogenet 15:8–18

Pallarès N, Lefebvre S, Contet V, Matsuda F, Lefranc M-P (1999) The human immunoglobulin heavy variable genes. Exp Clin Immunogenet 16:36–60

Ruiz M, Pallarès N, Contet V, Barbié V, Lefranc M-P (1999) The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Exp Clin Immunogenet 16:173–184

Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G, Chaume D, Lefranc M-P (2000) IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 28:219–221

Scaviner D, Lefranc M-P (2000a) The human T cell receptor alpha joining (TRAJ) genes. Exp Clin Immunogenet 17:97–106

Scaviner D, Lefranc M-P (2000b) The human T cell receptor alpha variable (TRAV) genes. Exp Clin Immunogenet 17:83–96

Scaviner D, Barbié V, Ruiz M, Lefranc M-P (1999) Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 16:234–240