JOURNAL OF MOLECULAR RECOGNITION
J. Mol. Recognit. 2004; **17**: 17–32
DOI:10.1002/jmr.647

IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties

Christelle Pommié[1], Séverine Levadoux[1], Robert Sabatier[2], Gérard Lefranc[1] and Marie-Paule Lefranc[1,3]*

[1]IMGT, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier II, UPR CNRS 1142 Institut de Génétique Humaine IGH, 141 rue de la Cardonille, F-34396 Montpellier Cedex 5, France
[2]Laboratoire de Physique Moléculaire et Structurale UMR 5094, Faculté de Pharmacie, 15 Avenue Charles Flahault, F-34060 Montpellier, France
[3]Institut Universitaire de France, 103 Boulevard Saint-Michel, 75005 Paris, France

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) is a high-quality integrated information system specializing in immunoglobulins (IG), T cell receptors (TR) and major histocompatibility complex (MHC) of human and other vertebrates. IMGT comprises IMGT/LIGM-DB, the comprehensive database of IG and TR sequences from human and other vertebrates (76 846 sequences in September 2003). In order to define the IMGT criteria necessary for standardized statistical analyses, the sequences of the IG variable regions (V-REGIONs) from productively rearranged human IG heavy (IGH) and IG light kappa (IGK) and lambda (IGL) chains were extracted from IMGT/LIGM-DB. The framework amino acid positions of 2474 V-REGIONs (1360 IGHV, 585 IGKV, 529 IGLV) were numbered according to the IMGT unique numbering. Two statistical methods (correspondence analysis and hierarchic classification) were used to analyze the 237 framework positions (80 for IGHV, 79 for IGKV, 78 for IGLV), for three properties (hydropathy, volume and chemical characteristics) of the 20 common amino acids. Results of the analyses are shown as standardized two-dimensional representations, designated as IMGT Colliers de Perles statistical profiles. They provide a characterization of the amino acid properties at each framework position of the expressed IG V-REGIONs, and a visualization of the resemblances and differences between heavy and light, and between kappa and lambda sequences. The standardized criteria defined in this paper, amino acid positions and property classes, will be useful to study the mutations and allele polymorphisms, to establish correlations between amino acids in the IG and TR protein three-dimensional structures and to extract new knowledge from V-like domains of chains, other than IG and TR, belonging to the immunoglobulin superfamily. Copyright © 2004 John Wiley & Sons, Ltd.

Keywords: **statistics; IMGT; immunoglobulin; variable gene; amino acid property; V-REGION; heavy; kappa; lambda; Colliers de Perles**

Received 23 July 2003; revised 1 August 2003; accepted 5 August 2003

INTRODUCTION

Owing to their fundamental role in the immune system, the immunoglobulin (IG) and T cell receptor (TR) variable domains have been extensively studied. IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr (Lefranc, 2003, 2001a; Lefranc *et al.*, 1998, 1999; Ruiz *et al.*, 2000) created in 1989 by Marie-Paule Lefranc (Université Montpellier II, CNRS) at Montpellier, France, provides a standardized and integrated access to immunogenetics data of IG, TR, major histocompatibility complex (MHC) and related proteins of the immune system (RPI) from human and other vertebrate species (150 species in September 2003). As a result of the recent years sequencing effort, all the human IG and TR genes are now characterized (for review see Lefranc and Lefranc, 2001a,b). The IG and TR variable domains (corresponding to the V-J-REGION and V-D-J-REGION labels in IMGT) represent a privileged situation by the conservation of their three-dimensional structure despite divergent amino acid sequences, and by the considerable amount of available annotated data (Artero and Lefranc, 2000a,b; Barbié and Lefranc, 1998; Bosc and Lefranc, 2000, 2003; Bosc *et al.*, 2001; Folch and Lefranc, 2000a,b; Folch *et al.*, 2000; Lefranc, 2000a,b, 2001b,c,d; Lefranc and Lefranc, 2001a,b; Martinez and Lefranc, 1998; Martinez-Jean *et al.*, 2001; Pallarès *et al.*, 1998, 1999; Ruiz and Lefranc, 2002; Ruiz *et al.*, 1999; Scaviner *et al.*, 1999; Scaviner and Lefranc, 2000a,b; IMGT Repertoire http://imgt.cines.fr).

Before any antigen-bound immunoglobulin structure had been determined, Kabat (1970) defined regions of high amino acid diversity or complementarity determining

---

*Correspondence to: M.-P. Lefranc, IMGT, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Université Montpellier II, UPR CNRS 1142 Institut de Génétique Humaine, 141 rue de la Cardonille, F-34396 Montpellier Cedex, France. E-mail: lefranc@ligm.igh.cnrs.fr

Abbreviations used: CDR, complementarity determining regions; FR, framework regions; IG, immunoglobulin; MHC, major histocompatibility complex; RPI, related proteins of the immune system; TR, T cell receptor.

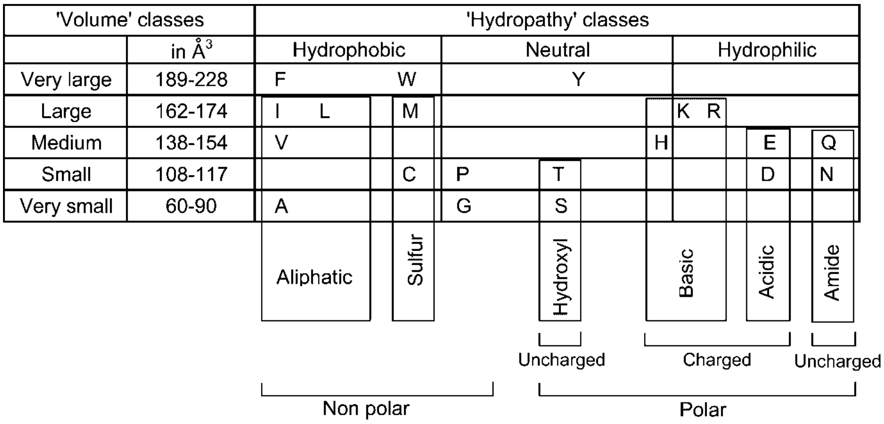
regions (CDR) and conserved regions of low amino acid variability or Framework Regions (FR), with the CDRs being the antigen binding sites of the immunoglobulins (or antibodies). These predictions have since been verified by IG crystal structures (Padlan *et al.*, 1995, for review see Ruiz and Lefranc, 2002). Even now, sequence variability analysis can fill in the gaps in our structural knowledge of the IG and TR variable domains (Ruiz and Lefranc, 2000a,b). However, data on the conserved amino acids, which determine the structural specifications of the variable domains, are not readily available in the publications and databases, owing to a lack of standardization. One of the goals of IMGT has been to develop a formal specification of the terms to be used in the domain of immunogenetics and immunoinformatics. This has been the basis of IMGT-ONTOLOGY (Giudicelli and Lefranc, 1999), the first ontology in the domain, which allows the management of the immunogenetics knowledge for all vertebrate species (Giudicelli and Lefranc, 2003). Control of coherence in IMGT combines data integrity control and biological data evaluation (Giudicelli *et al.*, 1998a,b). More particularly, the IMGT unique numbering for IG and TR V-REGION sequences of all vertebrate species was established to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type (heavy or light chains for IG; alpha, beta, gamma or delta chains for TR), or the species (Lefranc, 1997, 1999; Lefranc and Lefranc, 2001a,b; Lefranc *et al.*, 2003). In the IMGT unique numbering, conserved amino acids from FR always have the same number whatever the IG or TR variable sequence, and whatever the species they come from, e.g. Cysteine 23 (1st-CYS in FR1-IMGT), Tryptophan 41 (CONSERVED-TRP in FR2-IMGT), Cysteine 104 (2nd-CYS in FR3-IMGT) (Lefranc *et al.*, 2003). The IMGT unique numbering has allowed redefinition of the limits of the FR and CDR regions (Ruiz and Lefranc, 2002; Lefranc *et al.*, 2003). The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterizes variable regions belonging to a group, a subgroup and/or a gene (Lefranc and Lefranc, 2001a,b; Scaviner *et al.*, 1999; Folch *et al.*, 2000; Ruiz and Lefranc, 2002), and this can be applied to all vertebrate species (Artero and Lefranc, 2000a,b). FR amino acids located at the same position in different sequences can be compared without requiring sequence alignments. This also holds for amino acids belonging to CDR-IMGT of the same length (Lefranc *et al.*, 2003). The IMGT unique numbering allows to obtain standardized multi-sequence alignments and to set up statistical approaches of the amino acid physico-chemical properties, position by position in the FR-IMGT. These analyses are not only useful to study mutations and allele polymorphisms, but are also needed to establish correlations between amino acids in the protein sequences and three-dimensional (3D) structures of IG and TR V-REGIONs and V-DOMAINs as shown in IMGT/3Dstructure-DB, the IMGT 3D structure database, http://imgt.cines.fr (Ruiz and Lefranc, 2002). In this paper, we define the most appropriate amino acid property classes to analyse the amino acid resemblances and differences between IG and TR chains. We then describe the statistical analysis of the hydropathy, volume and chemical side chain properties of the amino acids, found at the IMGT standardized positions of the three FR (FR1-, FR2- and FR3-IMGT) of the V-REGIONs of human expressed IG heavy (IGH) and IG light kappa (IGK) and lambda (IGL) chains, extracted from IMGT/LIGM-DB.

MATERIALS AND METHODS

Data matrix

Amino acid property classes. The amino acid property classes were defined for the 'hydropathy', 'volume' and 'chemical characteristics' of the 20 common amino acids. The amino acid hydropathy, which derives from the physico-chemical properties of the amino acid side chains, determines in part, at a given position, the side chain orientation of the amino acid in the 3D structure (inside a protein, on its surface or neutral). For example, the Arginine $HN{=}C(NH_2){-}NH{-}(CH_2)_3{-}$ side chain is usually on the surface of a protein, whereas the Valine $CH_3{-}CH(CH_3){-}$ side chain is usually orientated inside of a protein, and the Serine $HO{-}CH_2{-}$ side chain is neutral. The amino acid 'hydropathy' classes were defined based on the Kyte and Doolittle (1982) amino acid hydropathy index (IMGT Aide-mémoire; http://imgt.cines.fr). The amino acids with a hydropathy index equal to or more than 1.8 were defined as hydrophobic. The amino acids with a hydropathy index equal to or less than −3.3 were defined as hydrophilic. The amino acids with a hydropathy index less than 1.8 and more than − 3.3 were defined as neutral. Three classes were thus defined (amino acid hydropathy index is decreasing between parentheses): hydrophobic (I, V, L, F, C, M, A, W), neutral (G, T, S, Y, P, H) and hydrophilic (D, N, E, Q, K, R) (Table 1). Tryptophan (W) was included in the hydrophobic class, its hydropathy index varying, depending from the studies, from −0.9 (Kyte and Doolittle, 1982) to 1.9 (Engelman *et al.*, 1986). The amino acid 'volume' classes were defined based on the amino acid volumes in angstrom^3 (Å^3) (Zamyatnin, 1972). Five classes were defined (the amino acid volume is increasing between parentheses): very small (60–90 Å^3; G, A, S); small (108–117 Å^3; C, D, P, N, T); medium (138–154 Å^3; E, V, Q, H); large (162–174 Å^3; M, I, L, K, R); and very large (189–228 Å^3; F, Y, W; Table 1). The amino acid 'chemical characteristics' classes were defined based on the principal chemical property of the amino acid side chain. Eleven classes were defined (Table 1), six of them contain several amino acids—aliphatic (A, V, I, L); sulfur (C, M); hydroxyl (S, T); acidic (D, E); amide (N, Q); basic (H, K, R)—and five classes correspond to a single amino acid (F, W, Y, G, P) owing to their particular characteristics (Rawn, 1989). F, W and Y have aromatic side chains. The bicyclic structure of tryptophan side chain is indole. The side chains of F and W are usually buried in the hydrophobic interior of proteins. Tyrosine (Y) differs from phenylalanine by a *para*-hydroxyl group which results in a polar side chain. The hydroxyl group of Y is weakly acidic. G has the simplest structure of any amino acid; its side chain is merely a hydrogen atom. The small size of this side chain gives glycine a unique function in the structure of many proteins since it will fit in niches that can accommodate no other amino acid. G is classified as nonpolar because the bond linking the α-carbon and the hydrogen atom side chain

Table 1. IMGT classes of the 20 common amino acids for the 'hydropathy', 'volume', 'chemical characteristics' properties. The three 'hydropathy' classes (hydrophobic, neutral, hydrophilic), five 'volume' classes in angstrom^3 (A^3) ([60–90], [108–117], [138–154], [162–174], [189–228]) and eleven 'chemical characteristics' (aliphatic, sulfur, hydroxyl, acidic, amide, basic, F, W, Y, G, P) classes as defined in this study



'Volume' classes		'Hydropathy' classes		
	in Å^3	Hydrophobic	Neutral	Hydrophilic
Very large	189-228	F W	Y	
Large	162-174	I L M		K R
Medium	138-154	V	H	E Q
Small	108-117	C	P T	D N
Very small	60-90	A	G S	

Amino acid side chains are also described in the literature as polar or nonpolar, charged or uncharged. Correspondence with the classes defined in this study are shown. Nonpolar (aliphatic, sulfur) amino acid side chains are hydrophobic. Uncharged polar side chains are neutral (hydroxyl) or hydrophilic (amide). Tyrosine is in the neutral class, although its OH is weakly acidic and polar. Charged polar side chains (basic or acidic) are hydrophilic. Histidine is in the neutral class, although it is weakly basic.

Amino acid one-letter abbreviation: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; Y (Tyr), tyrosine.

Detailed information on the physico-chemical characteristics of the amino acids are available in IMGT Aide-Mémoire, http://imgt.cines.fr.

is nonpolar. The structure of P differs sharply from that of the other amino acids in that its side chain is bonded to the nitrogen as well as to the α-carbon in a pyrrolidine ring which restricts the geometry of the backbone chain of the protein that contains it, and introduces abrupt changes in the direction of the chain. P is not chemically reactive. Because of the bond to nitrogen, P is an imino (—NH—) rather than an amino (—NH$_2$—) acid.

Sequence data. The statistical analysis was performed on the amino acid sequences of the V-REGIONs from productively rearranged (in-frame) IG sequences which may therefore be expressed in human IGH, IGK and IGL chains. Nucleotide sequences of the V-REGIONs to analyse were extracted from IMGT/LIGM-DB (http://imgt.cines.fr), the comprehensive nucleotide sequence database of IG and TR from human and other vertebrates (76 846 sequences in September 2003; Lefranc, 2003). The sequences were selected on the following criteria: 'human'; 'rearranged'; 'productive'; 'Ig-heavy'; 'Ig-Light-Lambda' or 'Ig-Light-Kappa' (in the IMGT/LIGM-DB Taxonomy and characteristics query module), and 'V-REGION' (in the IMGT/LIGM-DB Annotation labels query module). A total of 2474 V-REGIONs from human productively rearranged (or in-frame) IG sequences was obtained: 1360 IGHV, 585 IGKV and 529 IGLV. The V-REGION nucleotide sequences were translated into amino acid sequences. FR-IMGT and CDR-IMGT were delimited and gaps were created according to the IMGT unique numbering (Lefranc and Lefranc, 2001a,b; Lefranc *et al.*, 2003). Sets of subsequences were created which correspond to FR1-IMGT (amino acid positions 1–26), FR2-IMGT (amino acid positions 39–55) and FR3-IMGT (amino acid positions 66–104), respectively (Lefranc and Lefranc, 2001a,b; Lefranc *et al.*, 2003). The position 73 in FR3-IMGT, not occupied by amino acids in the analysed set of sequences, was not included in the statistical analysis. On a total number of 82 positions (26 for FR1-IMGT, 17 for FR2-IMGT and 39 for FR3-IMGT), 81 were therefore analysed which correspond to 80 positions for IGHV, 79 for IGKV and 78 for IGLV. Indeed, position 10 (FR1-IMGT) is not occupied in IGHV, positions 81 and 82 (FR3-IMGT) are not occupied in IGKV, and the three positions 10, 81 and 82 are not occupied in IGLV (Lefranc and Lefranc, 2001a).

Contingency tables. Three contingency tables (one for IGHV, one for IGKV and one for IGLV) were established in order to perform the statistical analysis. These tables comprise 20 columns for the 20 different amino acids, and 78–80 rows (78 for IGLV, 79 for IGKV, 80 for IGHV) for the FR-IMGT amino acid positions. A Perl program (not shown) was developed to fill in the table with the number of sequences which have a given amino acid at a given position. At the same time, the frequency of each amino acid at each sequence position was calculated.

Amino acids were grouped according the 'hydropathy', 'volume' and 'chemical characteristics' classes, which resulted into tables of three, five and 11 columns, respectively, corresponding to the number of classes for each property (Table 1). The numbers of sequences having an amino acid in a given class at a given position were summed.

Statistical analysis methods

IGHV, IGKV and IGLV FR-IMGT positions were compared for the 'hydropathy', 'volume' and 'chemical characteristics' amino acid properties, by two different but complementary multivariate descriptive statistical analysis (MDSA) methods: the correspondence (or factor) analysis and the hierarchic classification (Ward's method) methods (Lebart *et al.*, 1984), using the ADE-4 software (Thioulouse *et al.*, 1997). For both methods, two types of analysis were performed: 'IGHV' vs 'IGKV+IGLV', in order to compare the heavy and light chain V-REGIONs, and 'IGKV' vs 'IGLV', in order to compare the kappa and lambda V-REGIONs.

The correspondence analysis (COA in ADE-4) was done on a 159-row contingency table (80 rows for 'IGHV' and 79 for 'IGKV+IGLV') for the heavy/light comparison, and on a 157-row contingency table (79 rows for 'IGKV' and 78 for 'IGLV') for the kappa/lambda comparison. The results of the correspondence analysis are shown on graphs that represent the configurations of points in projection planes formed by the first principal axes taken two at a time, and that visualize, for a given property, the statistical resemblances (displayed as packets) or differences (displayed as isolated points) between the amino acid positions. Only the COA results for the hydropathy analysis are shown in this paper. Indeed, when the number of classes is limited to three such as in the hydropathy analysis, the COA allows to single out 'isolated' positions in a 2D plot of maximal information and therefore provides an easy and reliable characterization of the data which do not fall in the defined classes. When the number of classes increases, for instance, in the case of volume property with five classes, the COA graph results become less interpretable. For the chemical characteristics property with 11 classes, only limited information could eventually be retrieved from the COA graphs.

The hierarchic classification (CAH in ADE-4) was carried out on each set of sequences separately: 'IGHV', 'IGKV+IGLV', 'IGKV', 'IGLV'. The CAH was done by projecting separately the 'IGHV', 'IGKV+IGLV', 'IGKV' and 'IGLV' contingency tables, as supplementary rows (80, 79, 79 and 78, respectively), on the first principal axes of the precedent COA, and then by computing a hierarchy (Clusters module) from an Euclidian distance matrix. The number of first principal axes to be used in the CAH was determined from the eigenvalues diagrams and the cumulated inertia for these axes, two for hydropathy, four for volume and 10 for chemical characteristics. In Ward's method, each group is replaced by its center of gravity. This hierarchy is used for computing the algorithm of hierarchic classification (Dendograms module). The results of the Ward method are presented as hierarchic classification dendograms in which the IGHV, IGKV+IGLV, IGKV and IGLV FR-IMGT amino acid positions are classed.

Colliers de Perles

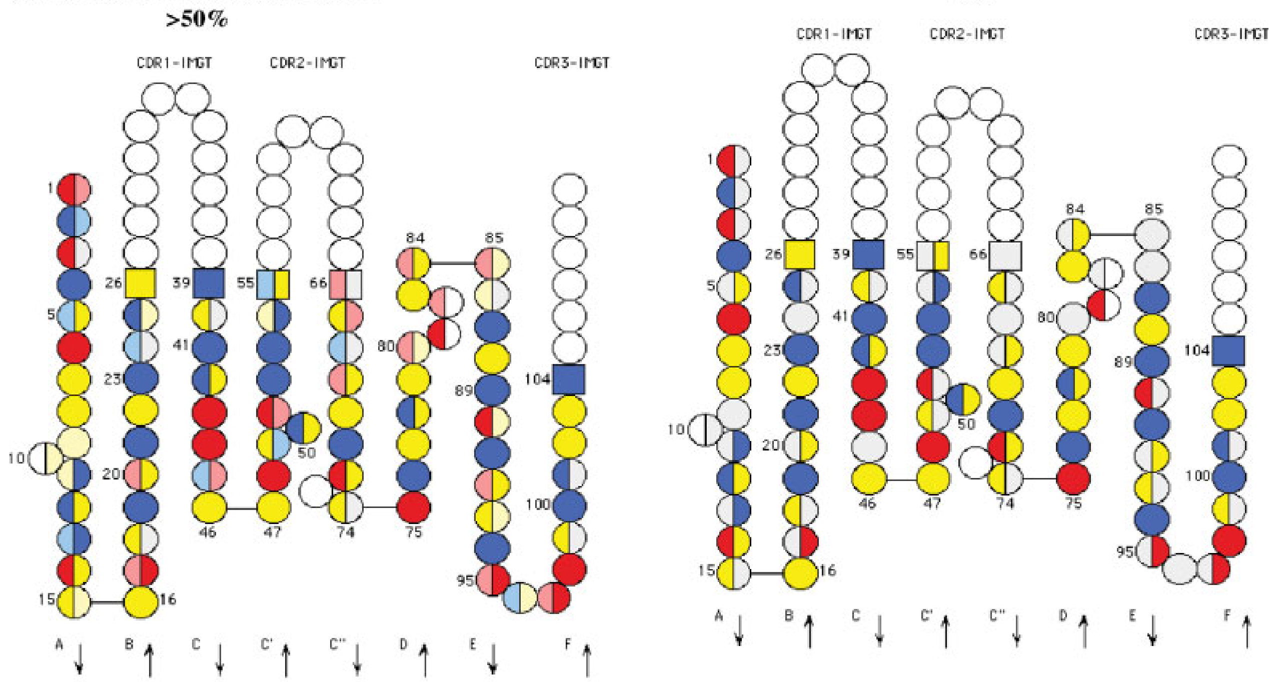
V-REGION statistical results from the CAH are displayed as graphical 2D representations or Colliers de Perles (cover of *Nucleic Acids Research* 27(1), Database issue, January 1999; Lefranc, 1999; Lefranc and Lefranc, 2001a; Lefranc *et al.*, 2003; Ruiz and Lefranc, 2002; Ruiz *et al.*, 2000) with FR-IMGT and CDR-IMGT delimitations, based on the IMGT unique numbering and with the standardized IMGT color menus for the hydropathy, volume and chemical characteristics classes (IMGT Scientific chart Representation rules http://imgt.cines.fr). These standardized 2D representations are designated as IMGT Colliers de Perles statistical profiles. The IMGT color menu for the chemical characteristics is derived from the IMGT standardized amino acid sequence profiles (Ruiz and Lefranc, 2000a,b).

RESULTS

The statistical analysis of three amino acid properties (hydropathy, volume, chemical characteristics) was done for 2474 V-REGIONs (1360 IGHV, 585 IGKV and 529 IGLV), extracted from human productively rearranged sequences. FR-IMGT amino acid positions (80 for IGHV, 79 for IGKV+IGLV, 79 for IGKV and 78 for IGLV) were compared by the correspondence analysis (COA) and the hierarchic classification (CAH; Ward's method), using the ADE-4 software (Thioulouse *et al.*, 1997). Each IGHV, IGKV+IGLV, IGKV and IGLV FR-IMGT position was checked for the amino acid statistical appartenance to one or the other property class by combining the COA and CAH results. However, except for the hydropathy property, for which the number of classes is limited and the COA graph easily interpretable, results for the volume and chemical characteristics are only derived from the CAH, which identifies the FR-IMGT positions on the hierarchic classification dendograms.

Hydropathy property. The COA graphs in Figure 1 represent the projection of the 159 FR-IMGT positions for the heavy/light V-REGION comparison [80 for IGHV and 79 for IGKV+IGLV; Fig. 1(A)], and that of the 157 FR-IMGT positions for the kappa/lambda V-REGION comparison [79 for IGKV and 78 for IGLV; Fig. 1(B)] for the three variable classes (hydrophobic, neutral and hydrophilic) of the hydropathy property. Proximity between positions means statistically similar amino acid properties at these positions. Three distinct packets, designated 1, 2 and 3, can be identified (Fig. 1). They correspond to positions which statistically are characterized by the property 'hydrophobic', 'neutral' and 'hydrophilic', respectively. The intermediary positions correspond to 'unclassified' positions. The CAH dendrograms in Fig. 2 allow to more easily classify the FR-IMGT positions inside the three 'hydrophobic', 'neutral' and 'hydrophilic' classes. However, the FR-IMGT positions which fall in a given class show percentages which vary in a large range. For example, positions 89 and 55 of IGHV [Fig. 2(A)] fall in the hydrophobic class, whereas the percentage of sequences with an hydrophobic amino acid is 99.7% for position 89 but only 40.8% for position 55, as calculated from the contingency tables (not shown). Therefore, in order to discriminate between such positions in the CAH dendrograms, a threshold was applied to the contingency tables. A threshold corresponding to a percentage equal or superior to 80% for a given class was determined as being the more appropriate one for comparison of IG V-REGIONs. A second threshold ($\geq 50\%$) was determined for analysis where it is necessary to identify the trend of the amino acid property. The CAH analysis results, taking into account these thresholds, are reported in graphical 2D representations or IMGT Colliers de Perles statistical profiles (Plate 1). Comparison of the heavy/light

A-Human IGHV and IGKV+IGLV



B- Human IGKV and IGLV

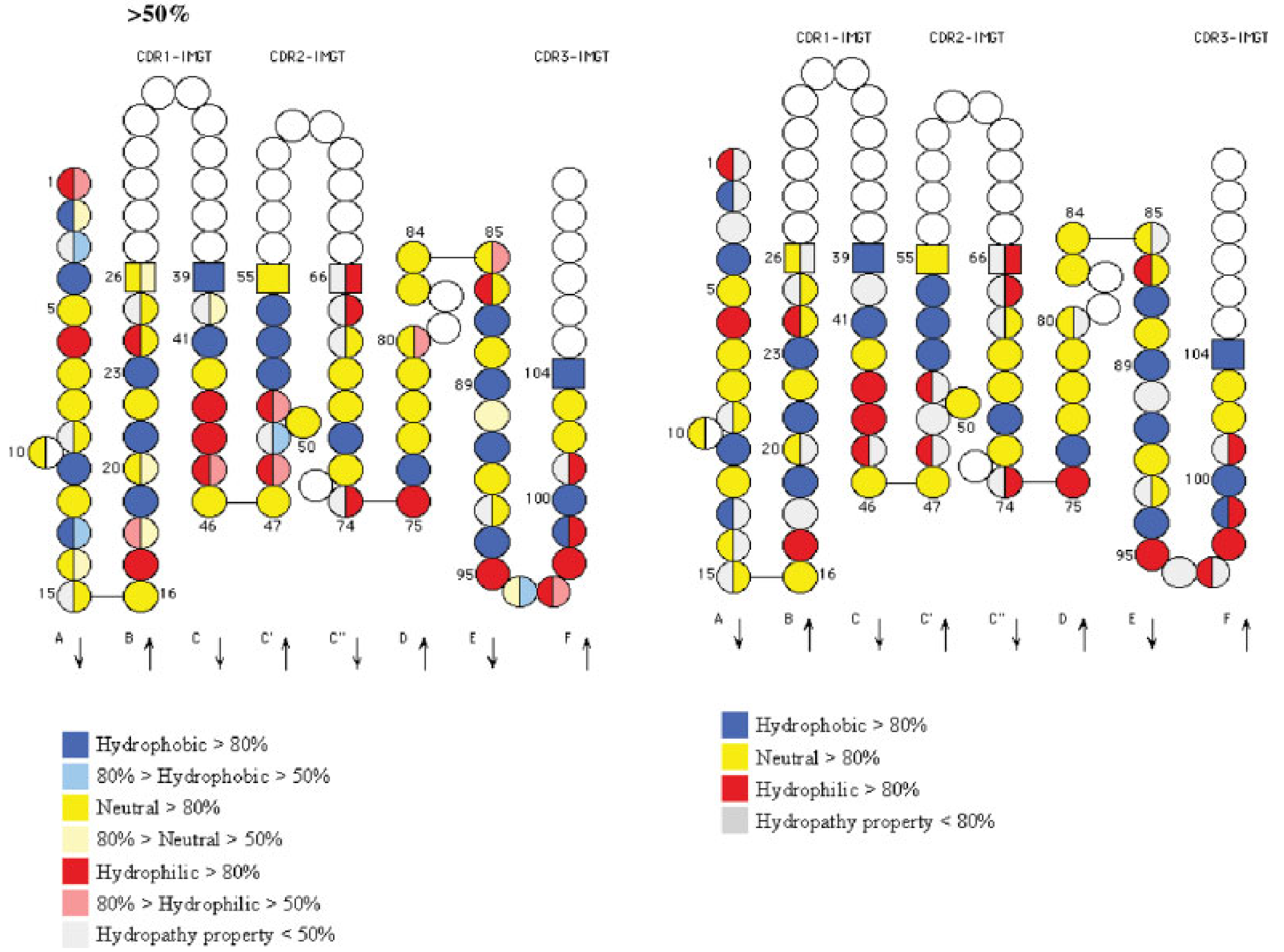
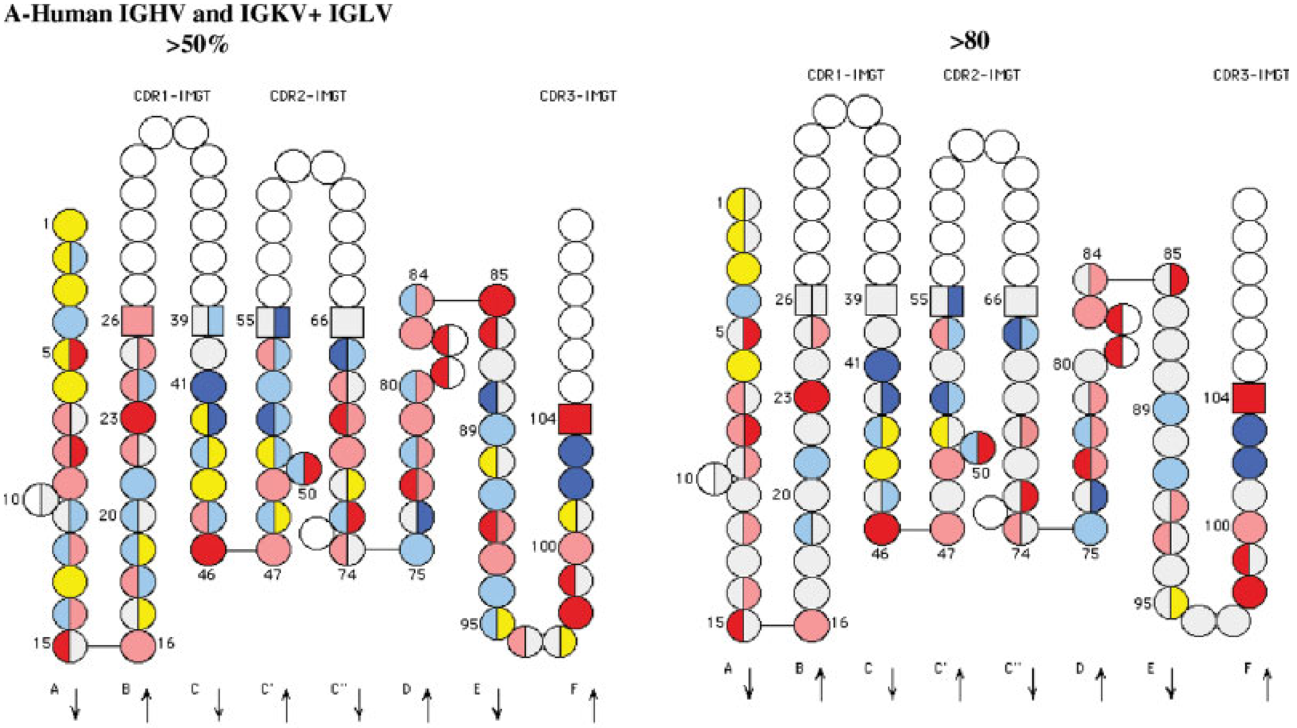


Plate 1. IMGT Collier de Perles statistical profile for hydropathy. (A) Human IGHV and IGKV+IGLV. (B) Human IGKV and IGLV. The positions are shown with a hydropathy profile defined at ≥50% and ≥80% threshold (see text). In (A), half circles correspond to IGHV (left) and IGKV+IGLV (right), and in (B), to IGKV (left) and IGLV (right).



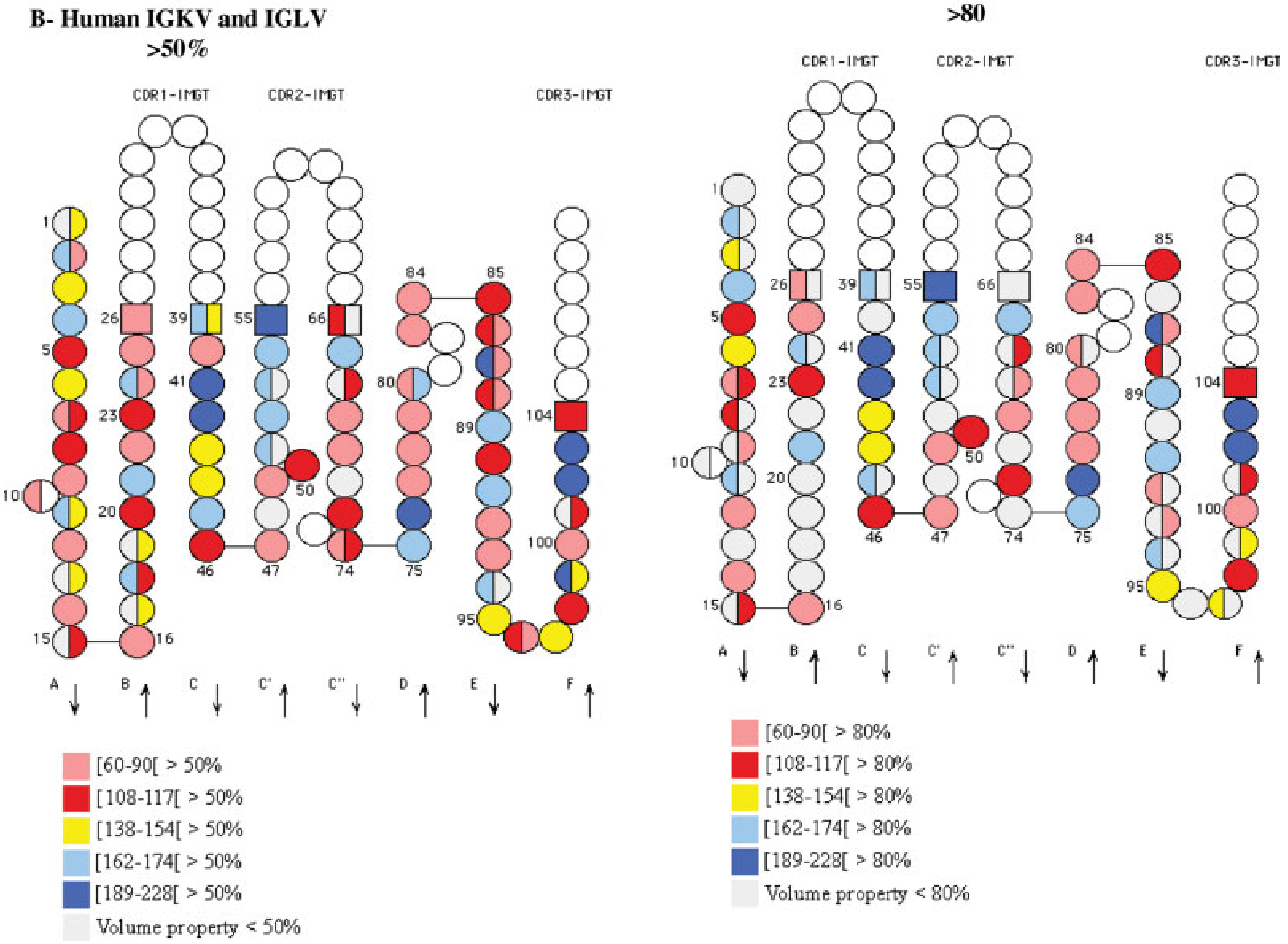
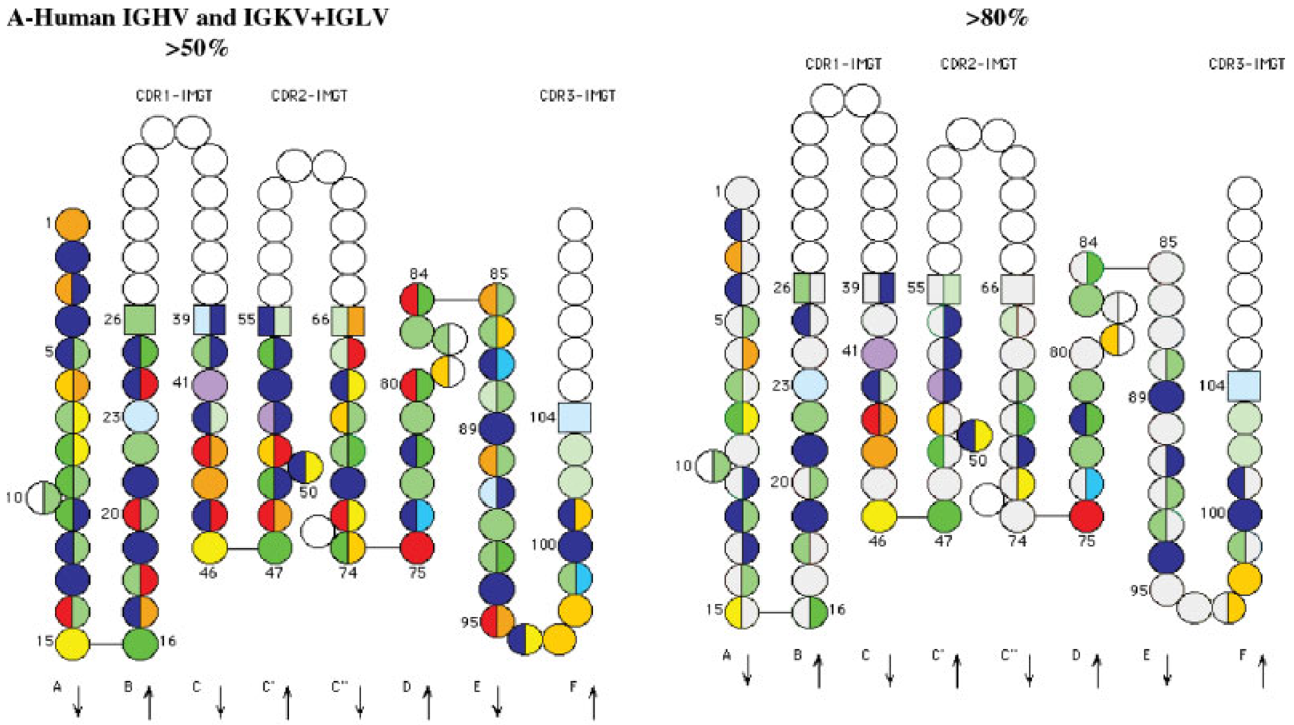


Plate 2. IMGT Collier de Perles statistical profile for volume. (A) Human IGHV and IGKV+IGLV. (B) Human IGKV and IGLV. The positions are shown with a volume profile defined at ≥50% and ≥80% threshold (see text). In (A), half circles correspond to IGHV (left) and IGKV + IGLV (right), and in (B), to IGKV (left) and IGLV (right).


A-Human IGHV and IGKV+IGLV

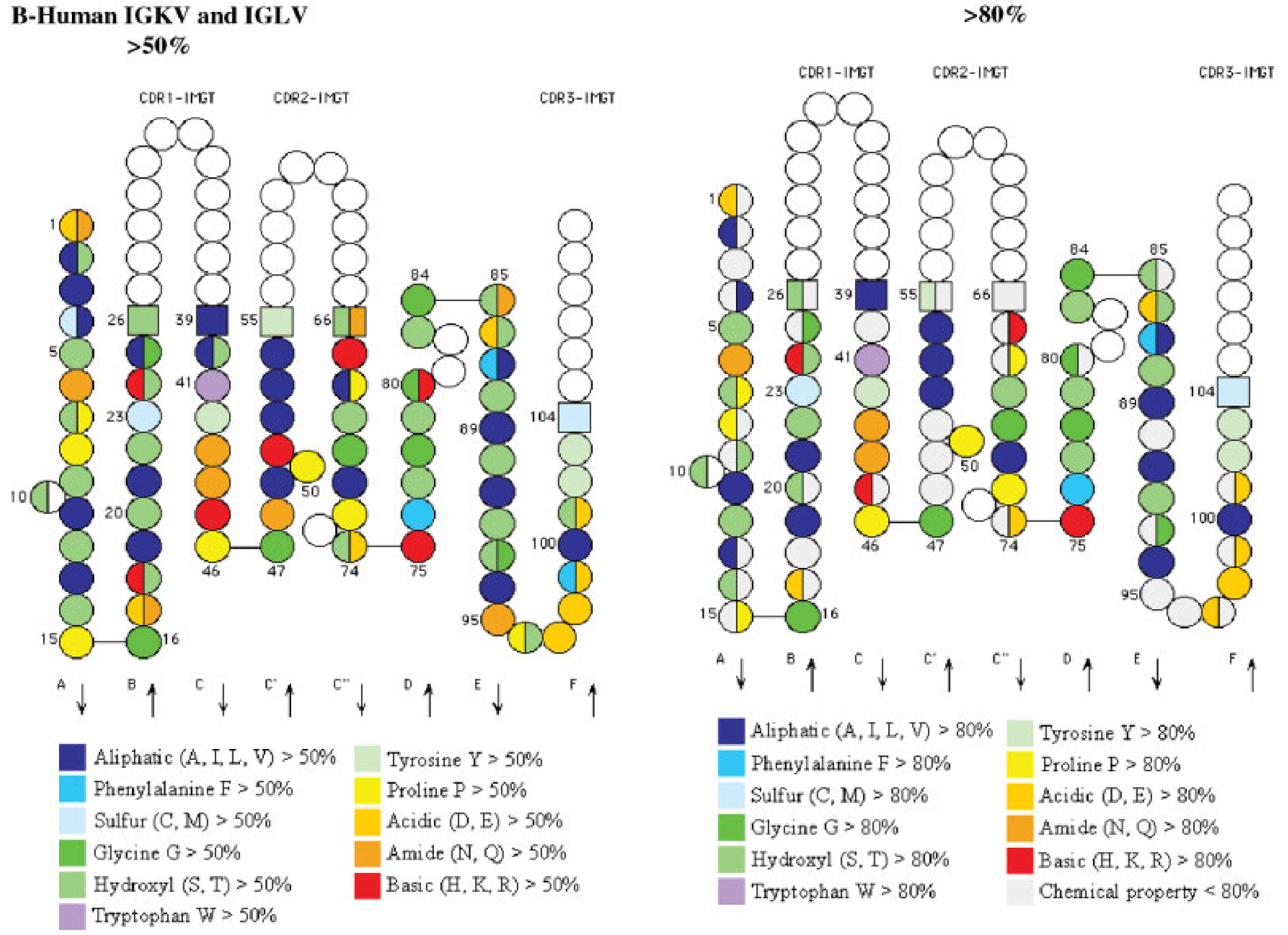

B-Human IGKV and IGLV

Plate 3. IMGT Collier de Perles statistical profile for chemical characteristics. (A) Human IGHV and IGKV+IGLV. (B) Human IGKV and IGLV. The positions are shown with a chemical profile defined at ≥50% and ≥80% threshold (see text). In (A), half circles correspond to IGHV (left) and IGKV+IGLV (right), and in (B), to IGKV (left) and IGLV (right).

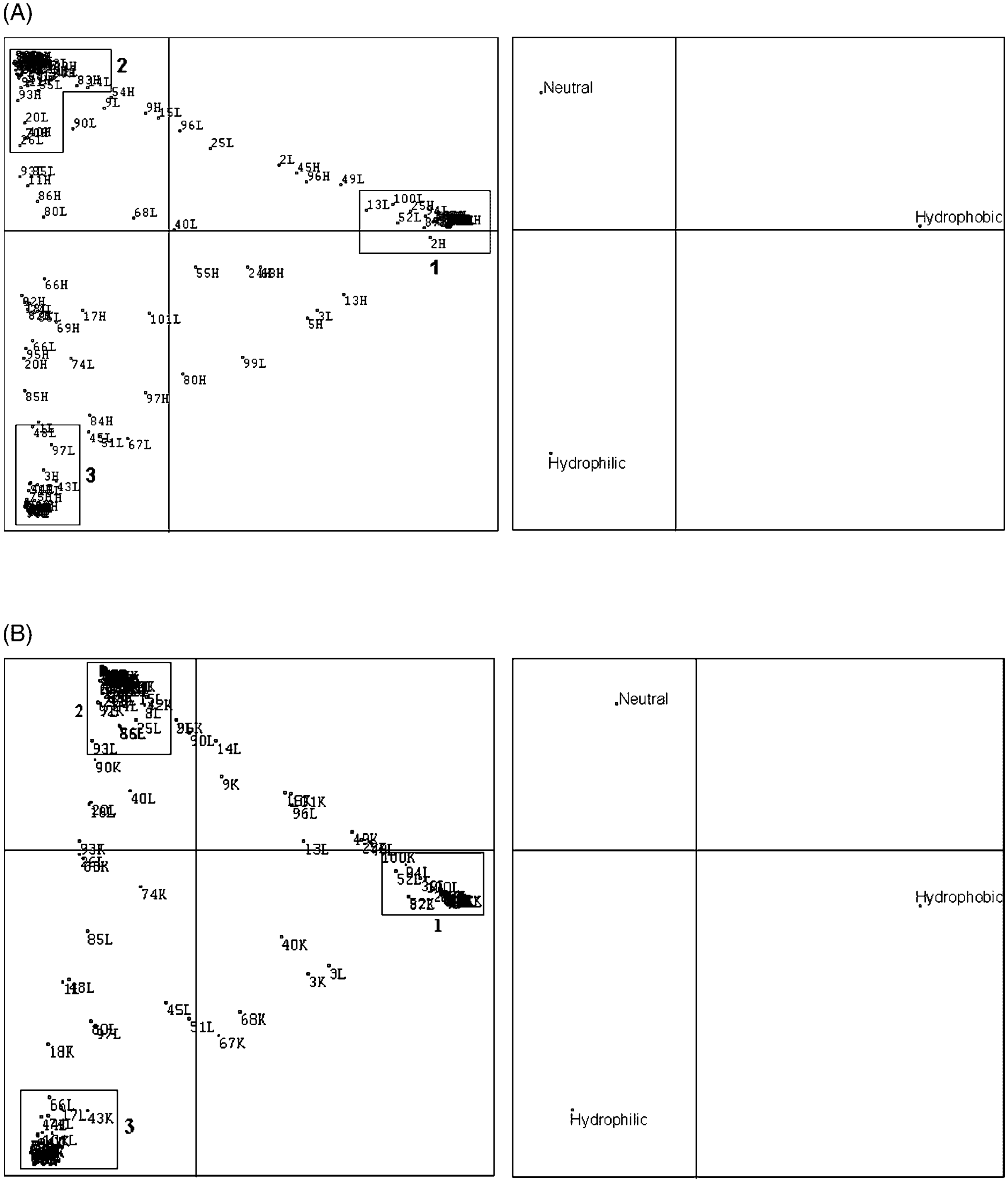


Figure 1. Correspondence analysis (COA) graphs for the hydropathy property of the human IG FR-IMGT amino acid positions. (A) Human IGHV and IGKV+IGLV. A total of 159 FR-IMGT positions were analysed, 80 from IGHV ('H') and 79 from IGKV+IGLV ('L'). (B) Human IGKV and IGLV. A total of 157 FR-IMGT positions were analysed, 79 from IGKV ('K') and 78 from IGLV ('L'). The packets 1, 2, 3 comprise positions which fall in the hydrophobic, neutral and hydrophilic classes, respectively, with a percentage equal or superior to 80% (see text). The advantage of the COA graph is to visualize positions which are situated along the diagonal between two packets or are isolated and therefore do not fall clearly in one or two classes.

V-REGIONs [IGHV vs IGKV+IGLV; Fig. 2(A) and (B)] shows that 36 FR-IMGT positions (out of 81) have conserved hydropathy properties, with a percentage threshold of 80%, whereas six FR-IMGT positions (12, 14, 42, 50, 72 and 78) have different hydropathy properties. The 36 conserved positions between IGHV and IGKV+IGLV comprise: 16 hydrophobic, 14 neutral and 6 hydrophilic positions (Table 2). The six positions which have different properties are qualified as 'specific'. A position is qualified as 'specific' if the amino acid property, in the two analysed sets of sequences, belongs to different classes with a percentage $\geq$80%. Four positions (12, 42, 50 and 78) are hydrophobic in IGHV and neutral in IGKV+IGLV. Two positions (14 and 72) are hydrophilic in IGHV and neutral in IGKV+IGLV [Plate 1(A)].

Comparison of IGKV and IGLV (Fig. 2(C) and (D)] shows that 48 FR-IMGT positions (out of 79) have conserved hydropathy properties, with a percentage threshold of 80% whereas three FR-IMGT positions (24, 86 and 99) have specific hydropathy properties. The 48 conserved positions between the IGKV and IGLV comprise: 18 hydrophobic, 23 neutral and seven hydrophilic positions (Table 2). Positions 24 and 86 are hydrophilic in IGKV and neutral in IGLV. Position 99 is hydrophobic in IGKV and hydrophilic in IGLV [Plate 1 (B)].

Volume property. The CAH dendrograms display the five classes '(60–90)', '(108–117)', '(138–154)', '(162–174)' and '(189–228)' for the volume property (Figure 3). Comparison of IGHV and IGKV+IGLV [Fig. 3(A) and (B)]

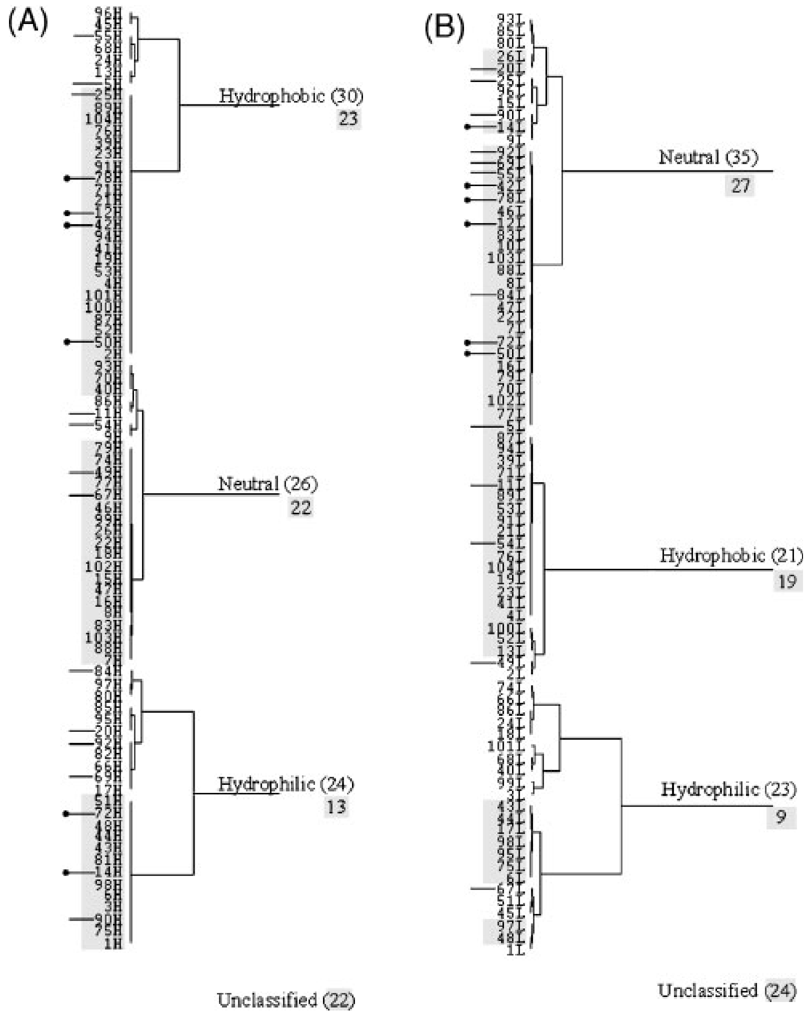


Figure 2. CAH dendrograms for the hydropathy property of the human IG FR-IMGT amino acid positions. (A) Human IGHV. (B) Human IGKV+IGLV. (C) Human IGKV. (D) Human IGLV. Eighty, 79, 79 and 78 FR-IMGT positions were analysed, respectively. In order to compare side by side the dendrograms, the right part of the graphs is not shown. Positions with an hydropathy property equal or superior to a percentage threshold of 80% are in gray. (●—) Positions where the amino acid property, in the two analysed sets of sequences, belongs to different classes, with a percentage $\geq$80%. (—) Positions where the amino acid property has a percentage $\geq$80% in only one of the two analysed sets. Numbers of positions between parentheses are from the CAH dendrogram. Numbers in gray, for the classes and for the unclassified, are those obtained with a $\geq$80% threshold.

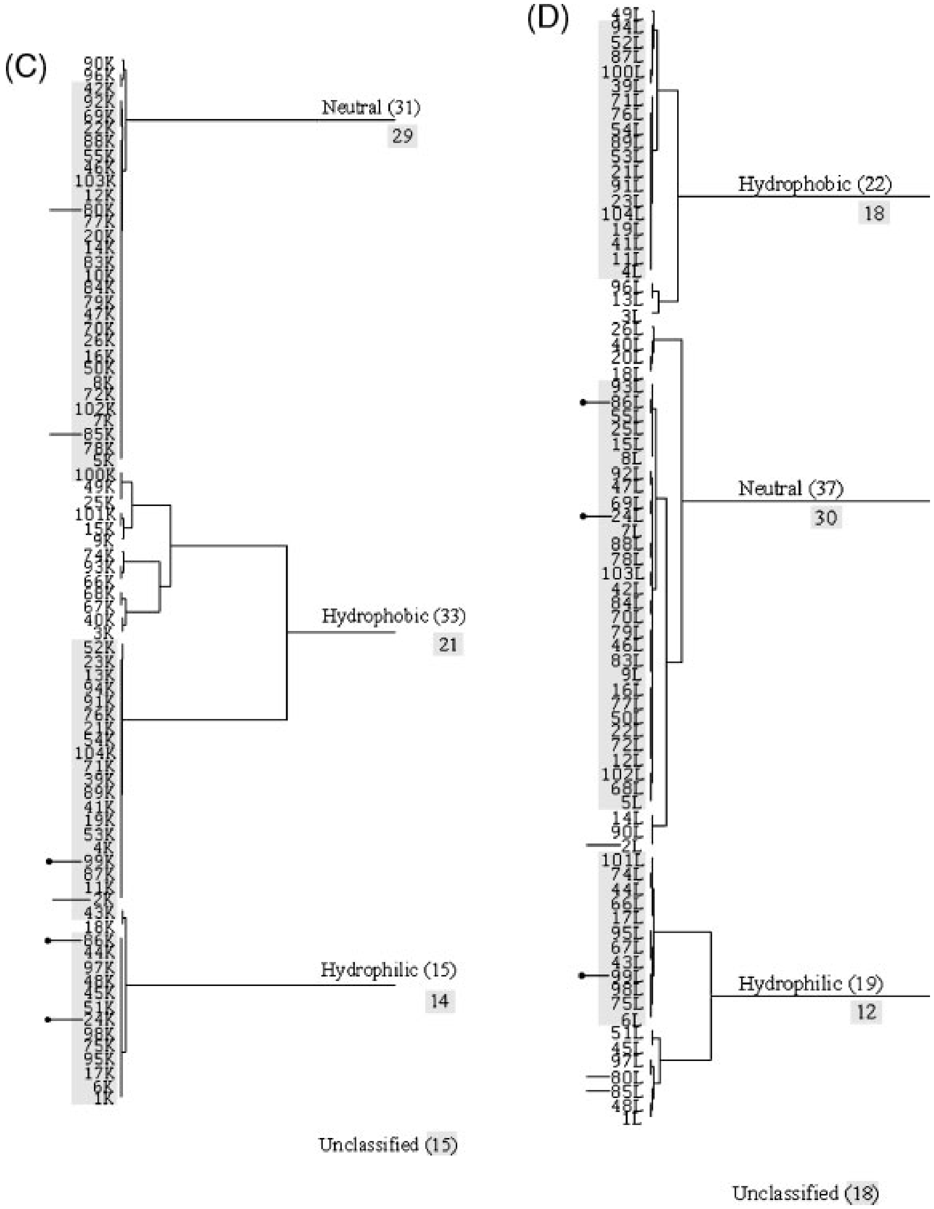


Figure 2. Continued

shows that, with a percentage threshold of 80%, 20 positions (out of 81) have conserved amino acid volume properties whereas eight positions (8, 43, 50, 52, 54, 67, 77 and 78) have specific amino acid volume properties. The 20 conserved positions between IGHV and IGKV+IGLV fall in the five volume classes with five, four, three, five and three positions, respectively, per class ranged by increased volumes (Table 2). Volume properties of the eight specific positions are reported in Table 2 and displayed in Plate 2(A). Comparison of IGKV and IGLV [Fig. 3(C) and (D)] shows that 38 positions (out of 79) have conserved volume properties whereas two positions (7 and 87) have specific amino acid volume property. The 38 conserved positions between IGKV and IGLV fall in the five volume classes with 13, eight, four, seven and six positions, respectively, per class ranged by increased volumes (Table 2). Volume properties of the two specific positions 7 and 87 are reported in Table 2 and displayed in Plate 2(B).

Chemical characteristics property. The CAH dendrograms display the eleven classes ‘aliphatic’, ‘sulfur’, ‘hydroxyl’, ‘acidic’, ‘amide’, ‘basic’, ‘F’, ‘W’, ‘G’, ‘Y’ and ‘P’ for the chemical characteristics property (Fig. 4). Comparison of IGHV and IGKV+IGLV (Figs 4(A) and (B)] shows that, with a percentage threshold of 80%, 19 FR-IMGT positions (out of 81) have conserved amino acid chemical characteristics properties whereas seven positions (8, 12, 42, 43, 50, 52 and 78) have specific amino acid chemical characteristics properties (Table 2). The 19 conserved positions between IGHV and IGKV+IGLV comprise five aliphatic (positions 19, 21, 89, 94 and 100), two sulfur (positions 23 and 104), four hydroxyl (positions 22, 77, 79 and 83), one acidic (position 98), one amide (position 44), one basic (position 75), one Tryptophan 41, two Tyrosine (positions 102 and 103), one Glycine 47 and one Proline 46. Chemical characteristics properties of the seven specific positions are reported in Table 2 and displayed in Plate 3(A).

Table 2. Conserved and specific amino acid properties at the FR-IMGT positions of human IGHV and IGKV+IGLV (heavy/light chain comparison) and human IGKV and IGLV (kappa/lambda chain comparison) as deduced from the statistical analyses: (A) hydropathy, (B) volume and (C) chemical characteristics. *N* = Number of positions with a threshold ≥80%. FR-IMGT positions are according to the IMGT unique numbering (Lefranc *et al.*, 2002). 81 and 79 positions were analysed in the IGHV/IGKV+IGLV and in the IGKV/IGLV comparisons, respectively

	N	Human IGHV and IGKV+IGLV FR-IMGT positions with conserved properties	*N*	Human IGKV and IGLV FR-IMGT positions with conserved properties
(A) Hydropathy				
Hydropathy classes				
Hydrophobic	16	4, 19, 21, 23, 39, 41, 52, 53, 71, 76, 87, 89, 91, 94, 100, 104	18	4, 11, 19, 21, 23, 39, 41, 52, 53, 54, 71, 76, 87, 89, 91, 94, 100, 104
Neutral	14	7, 8, 16, 22, 26, 46, 47, 70, 77, 79, 83, 88, 102, 103	23	5, 7, 8, 12, 16, 22, 42, 46, 47, 50, 55, 69, 70, 72, 77, 78, 79, 83, 84, 88, 92, 102, 103
Hydrophilic	6	6, 43, 44, 48, 75, 98	7	6, 17, 43, 44, 75, 95, 98
Total	36		48	

	N	FR-IMGT positions with specific properties	IGHV	IGKV+IGLV	*N*	FR-IMGT positions with specific properties	IGKV	IGLV
	4	12, 42, 50, 78	Hydrophobic	Neutral	2	24, 86	Hydrophilic	Neutral
	2	14, 72	Hydrophilic	Neutral	1	99	Hydrophobic	Hydrophilic

	N	Human IGHV and IGKV+IGLV FR-IMGT positions with conserved properties	*N*	Human IGKV and IGLV FR-IMGT positions with conserved properties
(B) Volume				
Volume classes				
(60–90)	5	16, 47, 49, 83, 100	13	12, 14, 16, 25, 47, 49, 70, 77, 78, 79, 83, 84, 100
(108–117)	4	23, 46, 98, 104	8	5, 23, 46, 50, 72, 85, 98, 104
(138–154)	3	3, 6, 44	4	6, 43, 44, 95
(162–174)	5	4, 21, 75, 89, 91	7	4, 21, 54, 67, 75, 89, 91
(189–228)	3	41, 102, 103	6	41, 42, 55, 76, 102, 103
Total	20		38	

	N	FR-IMGT positions with specific properties	IGHV	IGKV+IGLV	*N*	FR-IMGT positions with specific properties	IGKV	IGLV
	1	8	(60–90)	(108–117)	1	7	(60–90)	(108–117)
	1	54		(162–174)	1	87	(189–228)	(60–90)
	1	77	(108–117)	(60–90)				
	1	43	(162–174)	(138–154)				
	1	50		(108–117)				
	1	78		(60–90)				
	2	52, 67	(189–228)	(162–174)				

	N	Human IGHV and IGKV+IGLV FR-IMGT positions with conserved properties	*N*	Human IGKV and IGLV FR-IMGT positions with conserved properties
(C) Chemical characteristics[a]				
Chemical characteristics classes				
Aliphatic	5	19, **21**, **89**, 94, **100**	12	11, 19, **21**, 39, 52, 53, **54**, 71, **89**, **91**, 94, **100**
Sulfur	2	**23**, **104**	2	**23**, **104**
Hydroxyl	4	22, 77, 79, **83**	9	**5**, **12**, 22, 69, **77**, **79**, **83**, 88, 92

Continues

Table 2. Continued

Acidic	1	**98**	1	**98**
Amide	1	**44**	3	**6**, **43**, **44**
Basic	1	**75**	1	**75**
F	0		1	**76**
W	1	**41**	1	**41**
Y	2	**102**, **103**	3	**42**, **102**, **103**
G	1	**47**	5	**16**, **47**, **70**, **78**, **84**
P	1	**46**	3	**46**, **50**, **72**
Total	19		41	

N	FR-IMGT positions with specific properties[b]	IGHV	IGKV+IGLV	*N*	FR-IMGT positions with specific properties[c]	IGKV	IGLV
1	12	Aliphatic	Hydroxyl	1	7	Hydroxyl	P
1	42		Y	1	24	Basic	Hydroxyl
1	50		P	1	86	Acidic	
1	78		G	1	87	F	Aliphatic
1	8	G	P				
1	43	Basic	Amide				
1	52	W	Aliphatic				

[a] In bold are shown the positions conserved for the three properties hydropathy, volume, chemical characteristics, fourteen positions are conserved in both IGHV and IGKV+IGLV for the three properties. The five other positions are conserved for two properties (hydropathy and chemical characteristics) but show either differences (position 77) or a greater variability for the volume (19, 22, 77, 79, 94). These 19 positions are in yellow in Table 3A. 30 positions are conserved in both IGKV and IGLV for the three properties. The 11 other positions are conserved for two properties (hydropathy and chemical characteristics). These 41 positions are in yellow (19), light red (6) and green (16) in Table 3B.

[b] These seven positions (in red in Table 3A) identify the specific differences between the human immunoglobulin heavy (IGHV) and light (IGKV+IGLV) variable regions.

[c] These four positions (in orange in Table 3B) identify the specific differences between the human immunoglobulin light kappa (IGKV) and lambda (IGLV) variable regions.

Comparison of IGKV and IGLV [Fig. 4(C) and (D)] shows that up to 41 FR-IMGT positions (out of the 79) have conserved chemical properties whereas only four positions (7, 24, 86, 87) have specific amino acid chemical properties. The 41 conserved positions between IGKV and IGLV comprise 12 aliphatic (11, 19, 21, 39, 52, 53, 54, 71, 89, 91, 94 and 100), two sulfur (23 and 104), nine hydroxyl (5, 12, 22, 69, 77, 79, 83, 88, 92), one acidic (position 98), three amide (positions 6, 43 and 44), one basic (position 75), one Phenylalanine (position 76), one Tryptophan 41, three Tyrosine (positions 42, 102 and 103), five Glycine (positions 16, 47, 70, 78, 84), three Proline (positions 46, 50 and 72). Chemical characteristics properties of the four specific positions are reported in Table 2 and shown in Plate 3(B).

Comparison of statistical data with 3D structural data. A V-DOMAIN is formed from nine β-strands (labeled A, B, C, C', C'', D, E, F, G from N-terminal to C-terminal) in two β-sheets, one with four strands (A, B, E and D) and the other with five strands (G, F, C, C' and C''). They pack to form a sandwich which encloses an hydrophobic core constituted by side-chains of amino acids from both sheets (IMGT Education http://imgt.cines.fr) (Chothia *et al.*, 1998). The VL and VH domains associate to form a dimer with contacts between the GFCC’ β-sheet strands. Note that, owing to the combinatorial rearrangements between V and (D)-J to form the FG loop (CDR3-IMGT), the G strand (FR4-IMGT, $3'$ part of the J-REGION) was not included in the statistical analysis.

Positions with conserved properties for both the IGHV and IGKV+IGLV are expected to be important for the conserved structure of the immunoglobulin fold. With a percentage threshold of 80%, 14 FR-IMGT positions (out of 81) are conserved in both IGHV and IGKV+IGLV, for the three properties (Table 3). They include two positions (21, 23) in the B strand, two (41, 44) in the C strand, two (46, 47) at the CC' turn, one (75) at the $C''D$ turn, one (83) in the D strand, one (89) in the E strand and five (98, 100, 102, 103, 104) in the F strand. They correspond to the two conserved Cysteine at positions 23 (1st-CYS) and 104 (2nd-CYS) involved in the disulfide bridge (Rudikoff and Pumphrey, 1986), the conserved Tryptophan at position 41 (CONSERVED-TRP), the aliphatic (21, 89 and 100) with an inner side-chain orientation which belong to the hydrophobic core, the amide (44), the Proline (46) and Glycine (47) involved at the CC’ turn, the basic (75) at the $C''D$ turn, the hydroxyl (83) with an outer side-chain orientation, the acidic (98) with an inner side-chain orientation, and the Tyrosine (102) (with an inner side-chain orientation) and Y 103 (interface VH-VL; Ruiz and Lefranc, 2000a,b, 2002; IMGT/3Dstructure-DB, http://imgt.cines.fr). Five additional positions [19, 22 (strand A), 77, 79 (strand D), 94 (strand F)] are conserved for the chemical characteristics and the hydropathy but show differences for the volume properties (Table 3). Position 77

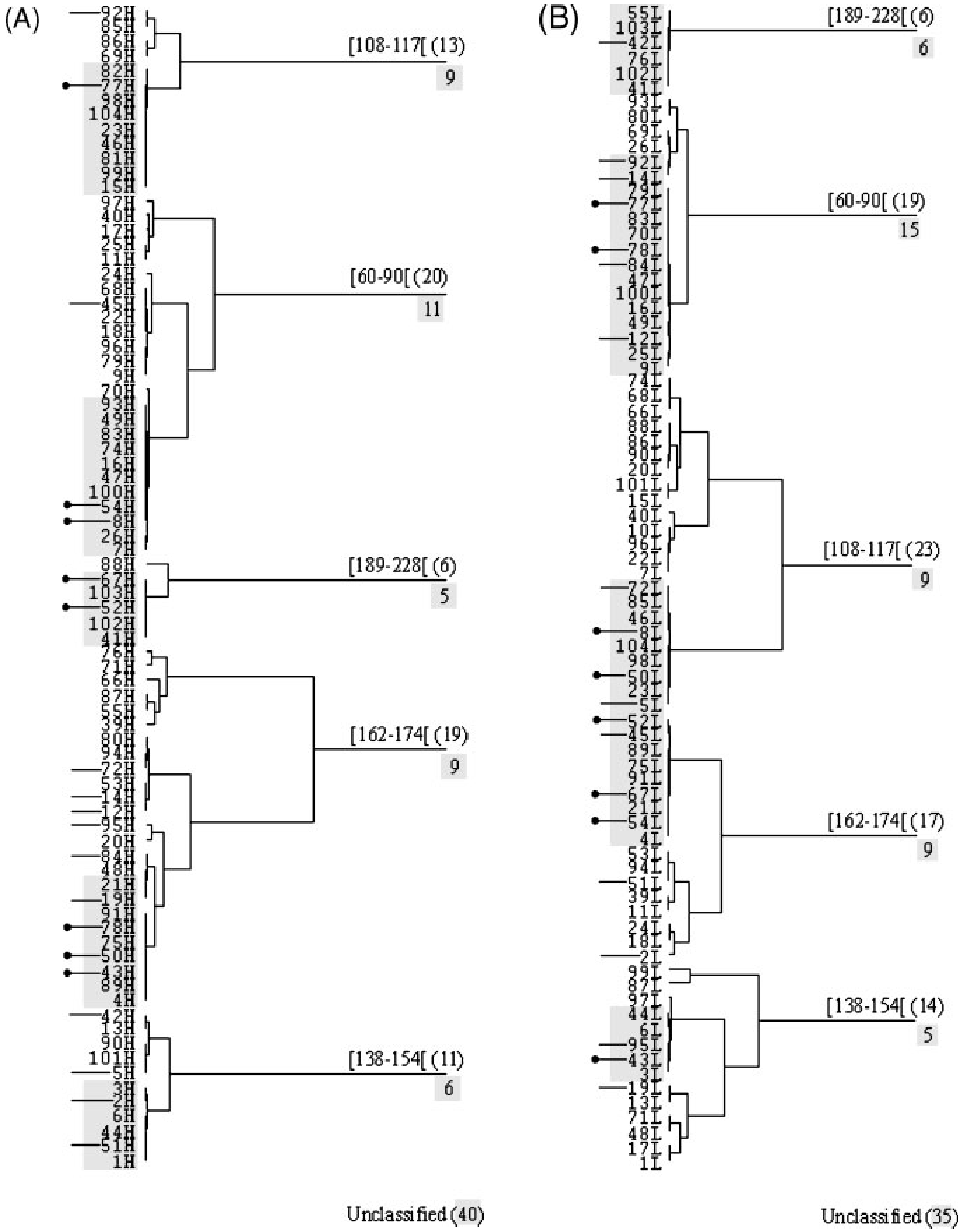


Figure 3. CAH dendrograms for the volume property of the human IG FR-IMGT amino acid positions. (A) Human IGHV. (B) Human IGKV+IGLV. (C) Human IGKV. (D) Human IGLV. Eighty, 79, 79 and 78 FR-IMGT positions were analysed, respectively. In order to compare side by side the dendrograms, the right part of the graphs is not shown. Positions with a volume property equal or superior to a percentage threshold of 80% are in gray. (●—) Positions where the amino acid property, in the two analysed sets of sequences, belongs to different classes, with a percentage ≥80%. (—) Positions where the amino acid property has a percentage ≥80% in only one of the two analysed sets. Numbers of positions between parentheses are from the CAH dendrogram. Numbers in gray, for the classes and for the unclassified, are those obtained with a ≥80% threshold.

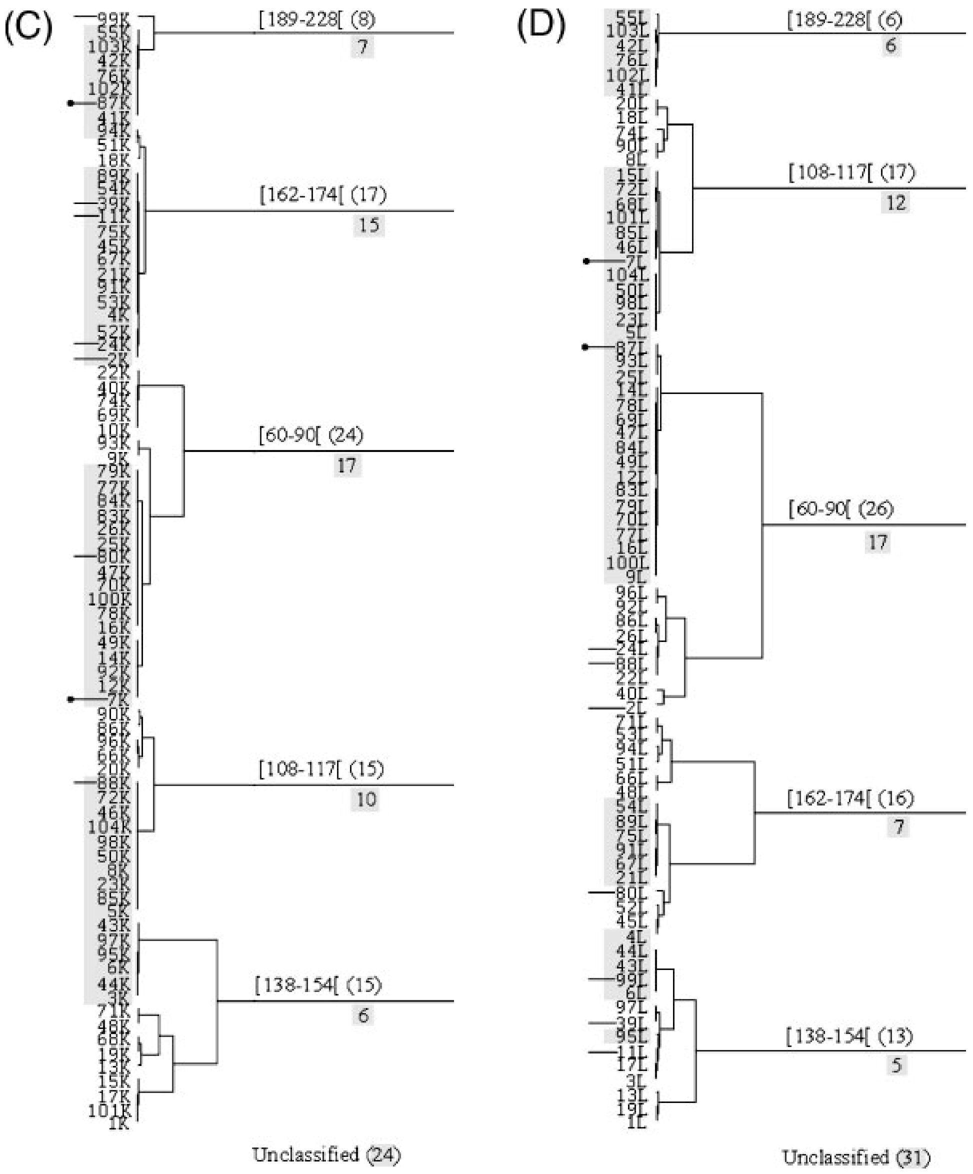


Figure 3. Continued

(hydroxyl with an outer orientation) corresponds to two different volume classes at a threshold of 80% (108–117) in IGHV, (60–90) in IGKV+IGLV, whereas the positions 22 and 79 (hydroxyl with an outer orientation), 19 and 94 (aliphatic with an inner orientation) display a greater variability for the volume properties (Table 3). Note that positions 4 and 91 are conserved for the hydropathy and volume, the threshold of 80% for aliphatic is not reached due to the frequency of Methionine at position 4 in IGKV+IGLV and at position 91 in IGHV.

In the comparison of IGHV and IGKV+IGLV, seven positions have statistical differences at a threshold of 80% for two or three properties (Table 3). These positions identify the specific differences between the human heavy (IGHV) and light (IGKV + IGLV) variable regions (Table 2). They include two positions (8, 12) in the A strand, two (42, 43) in the C strand, two (50, 52) in the C′ strand and one (78) in the D strand. The gap at position 10 in IGHV and the gaps at positions 81 and 82 should also be considered. The two positions which differ for the three properties are positions 50 (aliphatic in IGHV/Proline in IGKV+IGLV) and 78 (aliphatic/Glycine). The five positions which differ for two properties comprise position 8 (Glycine in IGHV/Proline in IGKV + IGLV), position 12 (aliphatic/hydroxyl) involved in IGHV at the VH-CH1 interface in a ball-and-socket joint (Lesk and Chothia, 1988), position 42 (aliphatic/Tyrosine) at the VH–VL interface, position 43 (basic/amide) with an inner orientation and position 52 (Tryptophan/aliphatic) at the VH–VL interface.

In the comparison of IGHV and IGKV+IGLV, 12 out of the 15 conserved hydrophobic positions (4, 19, 21, 23, 39, 41, 53, 76, 87, 89, 91, 104; Table 3) participate to the hydrophobic core and have amino acids with an inner side-chain orientation, whereas one (position 52) is involved in the VH–VL domain interaction (Ruiz and Lefranc, 2002).

The IGKV and IGLV comparison shows that the human light chains have 30 positions conserved for the three

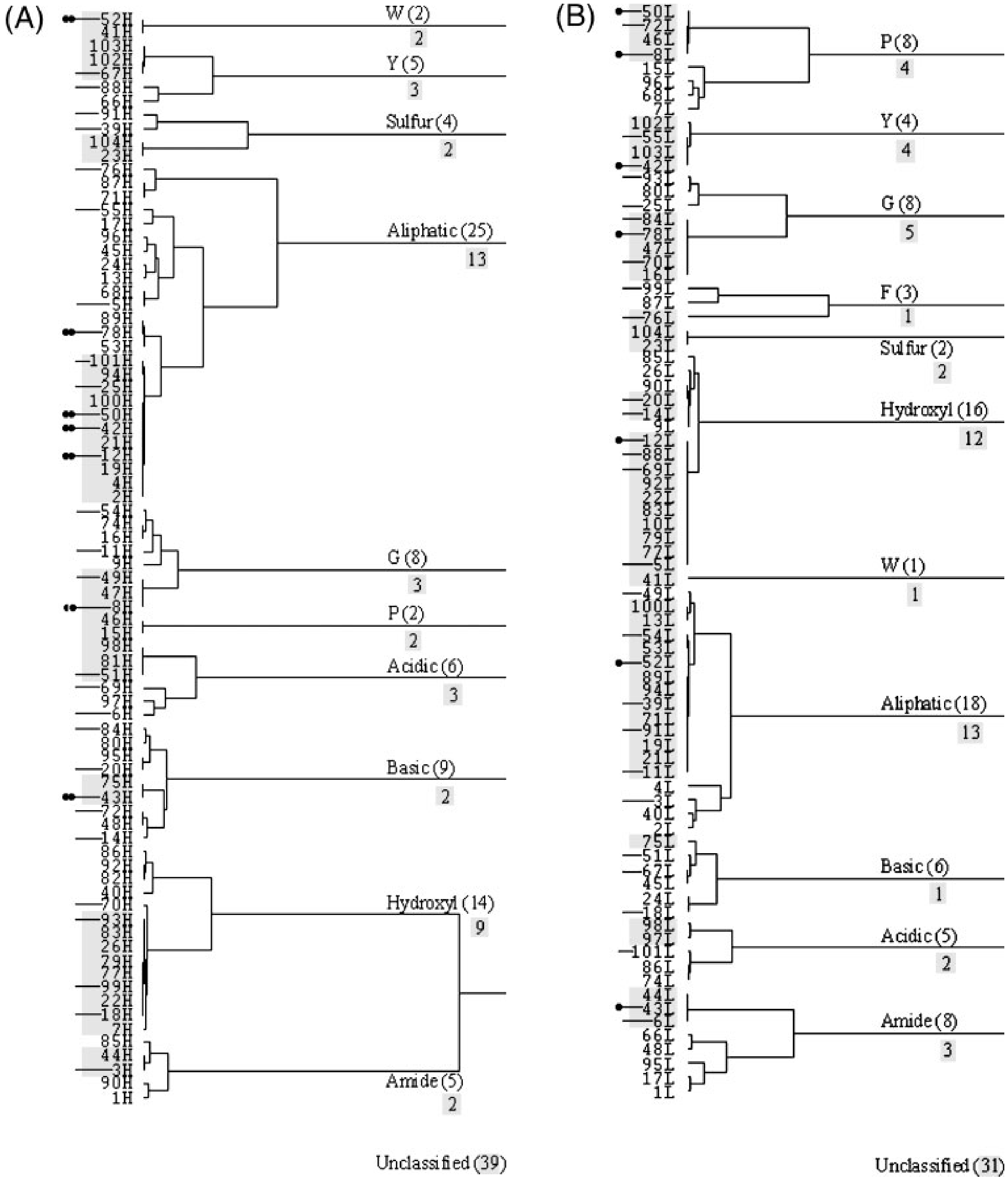


Figure 4. CAH dendrograms for the chemical property of the human IG FR-IMGT amino acid positions. (A) Human IGHV. (B) Human IGKV+IGLV. (C) Human IGKV. (D) Human IGLV. Eighty, 79, 79 and 78 FR-IMGT positions were analysed, respectively. In order to compare side by side the dendrograms, the right part of the graphs is not shown. Positions with a chemical property equal or superior to a percentage threshold of 80% are in gray: (●—) Positions where the amino acid property, in the two analysed sets of sequences, belongs to different classes, with a percentage ≥80%. (—) Positions where the amino acid property has a percentage ≥80% in only one of the two analysed sets. Numbers of positions between parentheses are from the CAH dendrogram. Numbers in gray, for the classes and for the unclassified, are those obtained with a ≥80% threshold.

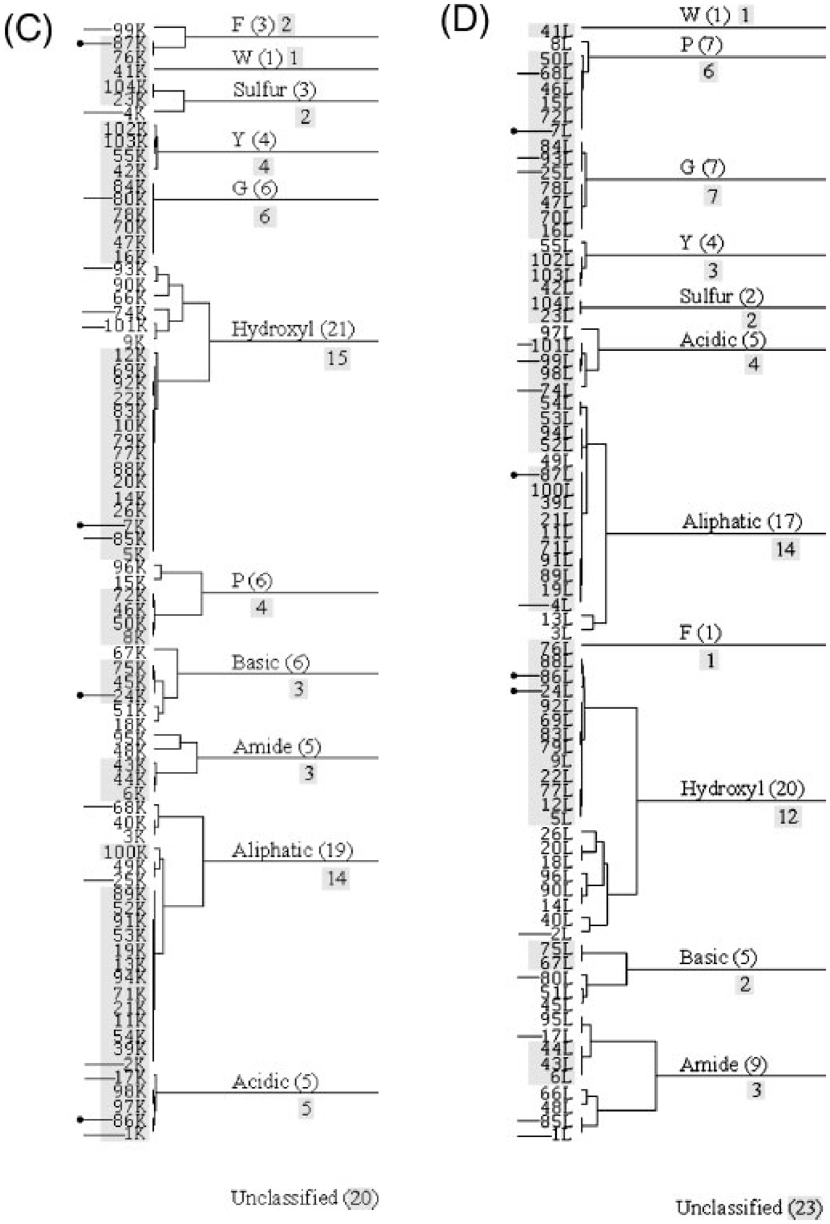


Figure 4. Continued

properties. They include nine new positions (5, 6, 16, 54, 70, 72, 76, 84 and 91) in addition to the 21 positions already described in the heavy/light comparison: 16 identified as conserved [the 14 positions conserved for the three properties (21, 23, 41, 44, 46, 47, 75, 83, 89, 98, 100, 102, 103, 104) and two of the five conserved for two properties in the IGHV and IGKV+IGLV comparison (77 and 79); Table 2C], and five positions of the seven identified as specific for the light (12, 42, 43, 50, 78). Eleven additional positions are conserved between the IGKV and IGLV for the chemical characteristics and the hydropathy but show differences in the volume properties. They include three of the positions conserved for two properties in the IGHV and IGKV+IGLV comparison (19, 22, 94), one specific for the light (position 52) and the following seven positions: 11, 39, 53, 69, 71, 88, 92. The position 8, although specific for the light in the IGHV and IGKV+IGLV comparison, does not reach the threshold of 80% for two properties of the IGLV in the IGKV and IGLV comparison and therefore was not included in Table 2. Four positions (14, 26, 45, 97) share the same hydropathy, volume and chemical properties in IGKV and in IGLV but with a higher degree of conservation ($\geq$80%) in IGKV compared with IGLV. The presence of five Glycine should be noted at positions 16 (AB turn), 47 (CC′ turn, conserved also in IGHV), 70 (C″ strand), 78 (D strand) and 84 (DE turn).

In the IGKV and IGLV comparison, no positions have statistical differences at a threshold of 80%, and only four positions (7, 24, 86 and 87) have differences for two properties, of which the chemical properties. These four positions identify the specific differences between the human IGKV and IGLV regions (Table 2): positions 7 (hydroxyl in IGKV/Proline in IGLV), 24 (basic in IGKV/hydroxyl in IGLV), 86 (acidic in IGKV/hydroxyl in IGLV) and

Table 3. Hydropathy, volume and chemical characteristics of the amino acids at the FR-IMGT positions of (A) human IGHV and IGHKV+IGLV (heavy/light chain comparison), (B) human IGKV and IGLV (kappa/lambda chain comparison) as deduced from CAH statistical analyses. A total of 2474 V-REGIONs from human productively rearranged (in-frame) sequences extracted from IMGT/LIGM-DB (http://imgt.cines.fr) were analysed: 1360 IGHV and 1114 IGKV+IGLV (A), 585 IGKV and 529 IGLV (B). The 82 FR-IMGT positions were analysed. However, position 73, not occupied by amino acids in the analysed set of sequences, was not included in the statistical analysis. Conserved property ($\geq$80%) at a given FR-IMGT position is shown in bold. Position 10 is only present in the 585 IGKV sequences of the IGKV+IGLV set and therefore its property appears with a percentage less than 80%. A dash indicates the absence of amino acid at a given position

87 (Phenylalanine in IGKV/aliphatic in IGLV; Table 2). The gap at position 10 in IGLV should also be considered. A fifth position (99) may be added, the IGKV hydrophobic property being mostly contributed by Phenylalanine, whereas IGLV position 99 is acidic at a threshold of 80% (Table 2).

Side chains of the five conserved aliphatic (19, 21, 89, 94, 100), the two Cysteine (first-CYS 23, second-CYS 104) and Tryptophan (CONSERVED-TRP 41) form the hydrophobic region which fills the interior of the sandwich between the beta sheets. In addition to these eight strongly conserved positions, six other positions (4, 39, 53, 76, 87, 91) participate in the hydrophobic core. Tyrosine 102, strongly conserved, has an inner side-chain orientation. Seven positions are at the VH-VL domain interface: positions 40, 42, 44 (in the C strand), 49, 50, 52 (in the C' strand) and 103 (in the F strand). The strong conservation of position 44 (amide) in the C strand and 103 (Tyrosine) in the F strand should be noted since the camel VH domains of the heavy chains, expressed without light chains, have mutated amino acids at positions 42, 49, 50 and 58 but still have the conserved 44 amide and 103 Tyrosine (IMGT Repertoire, http://imgt.cines.fr; Conrath *et al.*, 2003; Nguyen *et al.*, 1998, 2000).

CONCLUSION

Standardized criteria, amino acid positions and property classes, necessary for statistical analyses of the immunoglobulin V-REGIONs were defined. They were applied to the comparison of the human immunoglobulin, IGHV vs IGKV+IGLV (heavy/light chain comparison), and those of IGKV vs IGLV (kappa/lambda chain comparison), for the hydropathy, volume and chemical characteristics properties, for each FR-IMGT position. Eighty-one FR-IMGT positions from 2474 human productively rearranged (in-frame) IG V-REGION sequences from IGH, IGK and IGL were analysed by correspondence analysis and hierarchical classification for the hydropathy, volume and chemical characteristics amino acid properties. Such an approach was feasible owing to the standardization of the FR positions in IMGT sequences according to the IMGT unique numbering (Lefranc, 1997, 1999; Lefranc *et al.*, 2003).

Our standardized approach based on careful class definition and on the IMGT unique numbering will be extended to other properties [amino acid solvent accessibility (Bordo and Argos, 1991), hydrogen and Van der Waals bondings] and to other sets of sequences. This will be applied to immunoglobulins from the different vertebrate species, and more specifically to those found in lower vertebrates, such as the immunoglobulin light iota chains from Teleostei (IMGT Repertoire, http://imgt.cines.fr), to the T cell receptors and to proteins with V-like and/or C-like domains (Williams and Barclay, 1988). This will be particularly useful to establish correlations between amino acid positions of the IG fold.

Acknowledgements

We thank Elodie Duprat, Olivier Elemento, Quentin Kaas and Manuel Ruiz for helpful discussion. IMGT is funded by the 5th PCRDT (QLQ2-2000-01287) programme, the Centre National de la Recherche Scientifique (CNRS), the Ministère de la Recherche et de l'Education Nationale.

REFERENCES

Artero S, Lefranc M-P. 2000a. The Teleostei immunoglobulin light IGL1 and IGL2 V, J and C genes. *Exp. Clin. Immunogenet.* **17**: 162–172.

Artero S, Lefranc M-P. 2000b. The Teleostei immunoglobulin heavy IGH genes. *Exp. Clin. Immunogenet.* **17**: 148–161.

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

Bordo D, Argos P. 1991. Solvent accessibility of amino acids in known protein structures. *J. Mol. Biol.* **217**: 721–729.

Bosc N, Lefranc M-P. 2000. The mouse (*Mus musculus*) T cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. *Exp. Clin. Immunogenet.* **17**: 216–228.

Bosc N, Lefranc M-P. 2003. IMGT Locus in Focus: the mouse (*Mus musculus*) T cell receptor alpha (TRA) and delta (TRD) variables genes. *Dev. Comp. Immunol.* **27**: 465–497.

Bosc N, Contet V, Lefranc M-P. 2001. The mouse (*Mus musculus*) T cell receptor delta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. *Exp. Clin. Immunogenet.* **18**: 51–58.

Chothia C, Gelfand I, Kister A. 1998. Structural determinants in the sequences of the immunoglobulin variable domain. *J. Mol. Biol.* **278**: 457–479.

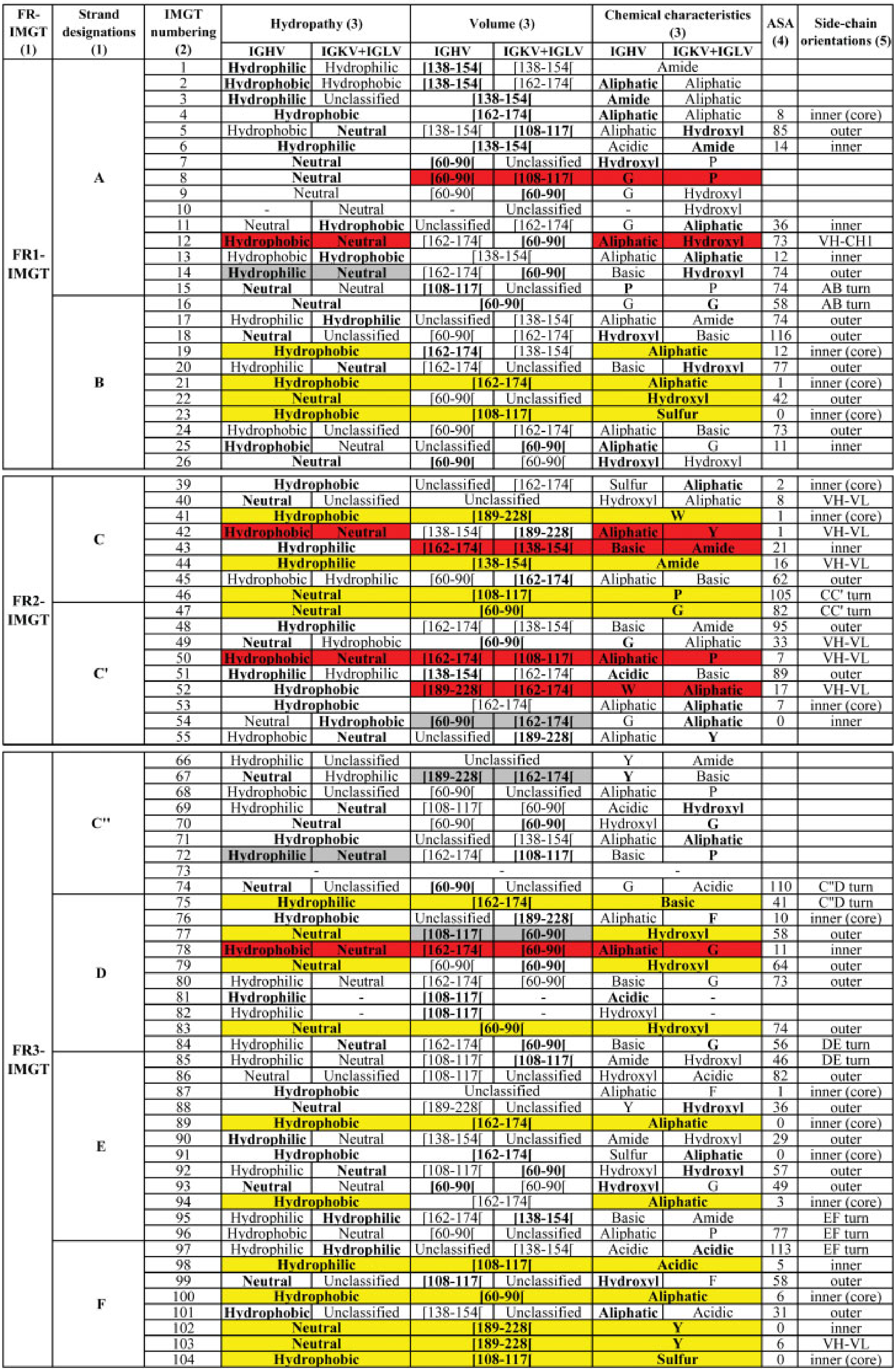
Conrath KE, Wernery U, Muyldermans S, Nguyen VK. 2003. Emergence and evolution of functional heavy-chain antibodies in Camelidae. *Dev. Comp. Immunol.* **27**: 87–103.

Engelman DM, Steitz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Ann. Rev. Biophys. Chem.* **15**: 321–353.

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

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

A- Human IGHV and IGKV+IGLV



FR-IMGT (1)	Strand designations (1)	IMGT numbering (2)	Hydropathy (3)		Volume (3)		Chemical characteristics (3)		ASA (4)	Side-chain orientations (5)
			IGHV	IGKV+IGLV	IGHV	IGKV+IGLV	IGHV	IGKV+IGLV		
FR1-IMGT	**A**	1	**Hydrophilic**	Hydrophilic	**[138-154[**	[138-154[	Amide			
		2	**Hydrophobic**	Hydrophobic	**[138-154[**	[162-174[	**Aliphatic**	Aliphatic		
		3	**Hydrophilic**	Unclassified	**[138-154[**		**Amide**	Aliphatic		
		4	**Hydrophobic**		**[162-174[**		**Aliphatic**	Aliphatic	8	inner (core)
		5	Hydrophobic	**Neutral**	[138-154[	**[108-117[**	Aliphatic	**Hydroxyl**	85	outer
		6	**Hydrophilic**		**[138-154[**		Acidic	**Amide**	14	inner
		7	**Neutral**		**[60-90[**	Unclassified	**Hydroxyl**	P		
		8	**Neutral**		**[60-90[**	**[108-117[**	**G**	**P**		
		9	Neutral		[60-90[	**[60-90[**	G	Hydroxyl		
		10	-	Neutral	-	Unclassified	-	Hydroxyl		
		11	Neutral	**Hydrophobic**	Unclassified	[162-174[	G	**Aliphatic**	36	inner
		12	**Hydrophobic**	**Neutral**	[162-174[	**[60-90[**	**Aliphatic**	**Hydroxyl**	73	VH-CH1
		13	Hydrophobic	**Hydrophobic**	[138-154[		Aliphatic	**Aliphatic**	12	inner
		14	**Hydrophilic**	**Neutral**	[162-174[	**[60-90[**	Basic	**Hydroxyl**	74	outer
		15	**Neutral**	Neutral	**[108-117[**	Unclassified	**P**	P	74	AB turn
	B	16	**Neutral**		**[60-90[**		G	**G**	58	AB turn
		17	Hydrophilic	**Hydrophilic**	Unclassified	[138-154[	Aliphatic	Amide	74	outer
		18	**Neutral**	Unclassified	[60-90[	[162-174[	**Hydroxyl**	Basic	116	outer
		19	**Hydrophobic**		**[162-174[**	[138-154[	**Aliphatic**		12	inner (core)
		20	Hydrophilic	**Neutral**	[162-174[	Unclassified	Basic	**Hydroxyl**	77	outer
		21	**Hydrophobic**		**[162-174[**		**Aliphatic**		1	inner (core)
		22	**Neutral**		[60-90[	Unclassified	**Hydroxyl**		42	outer
		23	**Hydrophobic**		**[108-117[**		**Sulfur**		0	inner (core)
		24	Hydrophobic	Unclassified	[60-90[	[162-174[	Aliphatic	Basic	73	outer
		25	**Hydrophobic**	Neutral	Unclassified	**[60-90[**	**Aliphatic**	G	11	inner
		26	**Neutral**		**[60-90[**	[60-90[	**Hydroxyl**	Hydroxyl		
FR2-IMGT	**C**	39	**Hydrophobic**		Unclassified	[162-174[	Sulfur	**Aliphatic**	2	inner (core)
		40	**Neutral**	Unclassified	Unclassified		Hydroxyl	Aliphatic	8	VH-VL
		41	**Hydrophobic**		**[189-228[**		**W**		1	inner (core)
		42	**Hydrophobic**	**Neutral**	[138-154[	**[189-228[**	**Aliphatic**	**Y**	1	VH-VL
		43	**Hydrophilic**		**[162-174[**	**[138-154[**	**Basic**	**Amide**	21	inner
		44	**Hydrophilic**		**[138-154[**		**Amide**		16	VH-VL
		45	Hydrophobic	Hydrophilic	[60-90[	**[162-174[**	Aliphatic	Basic	62	outer
		46	**Neutral**		**[108-117[**		**P**		105	CC' turn
	C'	47	**Neutral**		**[60-90[**		**G**		82	CC' turn
		48	**Hydrophilic**		[162-174[	[138-154[	Basic	Amide	95	outer
		49	**Neutral**	Hydrophobic	**[60-90[**		**G**	Aliphatic	33	VH-VL
		50	**Hydrophobic**	**Neutral**	**[162-174[**	**[108-117[**	**Aliphatic**	**P**	7	VH-VL
		51	**Hydrophilic**	Hydrophilic	**[138-154[**	[162-174[	**Acidic**	Basic	89	outer
		52	**Hydrophobic**		**[189-228[**	**[162-174[**	**W**	**Aliphatic**	17	VH-VL
		53	**Hydrophobic**		[162-174[		Aliphatic	**Aliphatic**	7	inner (core)
		54	Neutral	**Hydrophobic**	**[60-90[**	**[162-174[**	G	**Aliphatic**	0	inner
		55	Hydrophobic	**Neutral**	Unclassified	**[189-228[**	Aliphatic	**Y**		
FR3-IMGT	**C"**	66	Hydrophilic	Unclassified	Unclassified		Y	Amide		
		67	**Neutral**	Hydrophilic	**[189-228[**	**[162-174[**	**Y**	Basic		
		68	Hydrophobic	Unclassified	[60-90[	Unclassified	Aliphatic	P		
		69	Hydrophilic	**Neutral**	[108-117[	[60-90[	Acidic	**Hydroxyl**		
		70	**Neutral**		[60-90[	**[60-90[**	Hydroxyl	**G**		
		71	**Hydrophobic**		Unclassified	[138-154[	Aliphatic	**Aliphatic**		
		72	**Hydrophilic**	**Neutral**	[162-174[	**[108-117[**	Basic	**P**		
		73	-		-		-			
		74	**Neutral**	Unclassified	**[60-90[**	Unclassified	G	Acidic	110	C"D turn
	D	75	**Hydrophilic**		**[162-174[**		**Basic**		41	C"D turn
		76	**Hydrophobic**		Unclassified	**[189-228[**	Aliphatic	**F**	10	inner (core)
		77	**Neutral**		**[108-117[**	**[60-90[**	**Hydroxyl**		58	outer
		78	**Hydrophobic**	**Neutral**	**[162-174[**	**[60-90[**	**Aliphatic**	**G**	11	inner
		79	**Neutral**		[60-90[	**[60-90[**	**Hydroxyl**		64	outer
		80	Hydrophilic	Neutral	[162-174[	[60-90[	Basic	G	73	outer
		81	**Hydrophilic**	**-**	**[108-117[**	-	**Acidic**	-		
		82	Hydrophilic	-	**[108-117[**	-	Hydroxyl	-		
		83	**Neutral**		**[60-90[**		**Hydroxyl**		74	outer
		84	Hydrophilic	**Neutral**	[162-174[	**[60-90[**	Basic	**G**	56	DE turn
	E	85	Hydrophilic	Neutral	[108-117[	**[108-117[**	Amide	Hydroxyl	46	DE turn
		86	Neutral	Unclassified	[108-117[	Unclassified	Hydroxyl	Acidic	82	outer
		87	**Hydrophobic**		Unclassified		Aliphatic	F	1	inner (core)
		88	**Neutral**		[189-228[	Unclassified	Y	**Hydroxyl**	36	outer
		89	**Hydrophobic**		**[162-174[**		**Aliphatic**		0	inner (core)
		90	**Hydrophilic**	Neutral	[138-154[	Unclassified	Amide	Hydroxyl	29	outer
		91	**Hydrophobic**		**[162-174[**		Sulfur	**Aliphatic**	0	inner (core)
		92	Hydrophilic	**Neutral**	[108-117[	**[60-90[**	Hydroxyl	**Hydroxyl**	57	outer
		93	**Neutral**	Neutral	**[60-90[**	[60-90[	**Hydroxyl**	G	49	outer
		94	**Hydrophobic**		[162-174[		**Aliphatic**		3	inner (core)
		95	Hydrophilic	**Hydrophilic**	[162-174[	**[138-154[**	Basic	Amide		EF turn
		96	Hydrophobic	Neutral	[60-90[	Unclassified	Aliphatic	P	77	EF turn
	F	97	Hydrophilic	**Hydrophilic**	Unclassified	[138-154[	Acidic	**Acidic**	113	EF turn
		98	**Hydrophilic**		**[108-117[**		**Acidic**		5	inner
		99	**Neutral**	Unclassified	**[108-117[**	Unclassified	**Hydroxyl**	F	58	outer
		100	**Hydrophobic**		**[60-90[**		**Aliphatic**		6	inner (core)
		101	**Hydrophobic**	Unclassified	[138-154[	Unclassified	**Aliphatic**	Acidic	31	outer
		102	**Neutral**		**[189-228[**		**Y**		0	inner
		103	**Neutral**		**[189-228[**		**Y**		6	VH-VL
		104	**Hydrophobic**		**[108-117[**		**Sulfur**		0	inner (core)

B- Human IGKV and IGLV

FR-IMGT (1)	**Strand designations (1)**	**IMGT numbering (2)**	**Hydropathy (3)**		**Volume (3)**		**Chemical characteristics (3)**		**ASA (4)**	**Side-chain orientations (5)**
			IGKV	**IGLV**	**IGKV**	**IGLV**	**IGKV**	**IGLV**		
FR1-IMGT	**A**	1	**Hydrophilic**	Hydrophilic	Unclassified	[138-154[	**Acidic**	Amide		
		2	**Hydrophobic**	Neutral	**[162-174[**	[60-90[	**Aliphatic**	Hydroxyl		
		3	Unclassified	Hydrophobic	**[138-154[**	[138-154[	Aliphatic			
		4	**Hydrophobic**		**[162-174[**		Sulfur	**Aliphatic**	8	inner (core)
		5	**Neutral**		**[108-117[**		**Hydroxyl**		85	outer
		6	**Hydrophilic**		**[138-154[**		**Amide**		14	inner
		7	**Neutral**		**[60-90[**	**[108-117[**	**Hydroxyl**	**P**		
		8	**Neutral**		**[108-117[**	[108-117[	**P**	P		
		9	Unclassified	**Neutral**	[60-90[	**[60-90[**	Hydroxyl	**Hydroxyl**		
		10	**Neutral**	-	[60-90[	-	**Hydroxyl**	-		
		11	**Hydrophobic**		**[162-174[**	[138-154[	**Aliphatic**		36	inner
		12	**Neutral**		**[60-90[**		**Hydroxyl**		73	
		13	**Hydrophobic**	Hydrophobic	Unclassified	[138-154[	**Aliphatic**	Aliphatic	12	inner
		14	**Neutral**	Neutral	**[60-90[**		**Hydroxyl**	Hydroxyl	74	outer
		15	Unclassified	**Neutral**	Unclassified	**[108-117[**	P	**P**	74	AB turn
	B	16	**Neutral**		**[60-90[**		**G**		58	AB turn
		17	**Hydrophilic**		Unclassified	[138-154[	**Acidic**	Amide	74	outer
		18	Hydrophilic	Neutral	[162-174[	[108-117[	Basic	Hydroxyl	116	outer
		19	**Hydrophobic**		Unclassified	[138-154[	**Aliphatic**		12	inner (core)
		20	**Neutral**	Neutral	[108-117[		**Hydroxyl**	Hydroxyl	77	outer
		21	**Hydrophobic**		**[162-174[**		**Aliphatic**		1	inner (core)
		22	**Neutral**		[60-90[		**Hydroxyl**		42	outer
		23	**Hydrophobic**		**[108-117[**		**Sulfur**		0	inner (core)
		24	**Hydrophilic**	**Neutral**	**[162-174[**	[60-90[	**Basic**	**Hydroxyl**	73	outer
		25	Unclassified	**Neutral**	**[60-90[**		Aliphatic	**G**	11	inner
		26	**Neutral**	Neutral	**[60-90[**	[60-90[	**Hydroxyl**	Hydroxyl		
FR2-IMGT	**C**	39	**Hydrophobic**		**[162-174[**	[138-154[	**Aliphatic**		2	inner (core)
		40	Unclassified	Neutral	[60-90[		Aliphatic	Hydroxyl	8	VH-VL
		41	**Hydrophobic**		**[189-228[**		**W**		1	inner (core)
		42	**Neutral**		**[189-228[**		**Y**		1	VH-VL
		43	**Hydrophilic**		**[138-154[**		**Amide**		21	inner
		44	**Hydrophilic**		**[138-154[**		**Amide**		16	VH-VL
		45	**Hydrophilic**	Hydrophilic	**[162-174[**	[162-174[	**Basic**	Basic	62	outer
		46	**Neutral**		**[108-117[**		**P**		105	CC' turn
	C'	47	**Neutral**		**[60-90[**		**G**		82	CC' turn
		48	**Hydrophilic**	Hydrophilic	Unclassified		Amide		95	outer
		49	Unclassified	Hydrophobic	**[60-90[**		Aliphatic		33	VH-VL
		50	**Neutral**		**[108-117[**		**P**		7	VH-VL
		51	**Hydrophilic**	Hydrophilic	[162-174[	Unclassified	Basic		89	outer
		52	**Hydrophobic**		**[162-174[**	[162-174[	**Aliphatic**		17	VH-VL
		53	**Hydrophobic**		**[162-174[**	Unclassified	**Aliphatic**		7	inner (core)
		54	**Hydrophobic**		**[162-174[**		**Aliphatic**		0	inner
		55	**Neutral**		**[189-228[**		**Y**	Y		
FR3-IMGT	**C''**	66	Unclassified	**Hydrophilic**	[108-117[	Unclassified	Hydroxyl	Amide		
		67	Unclassified	**Hydrophilic**	**[162-174[**		Basic	**Basic**		
		68	Unclassified	**Neutral**	Unclassified	**[108-117[**	Aliphatic	**P**		
		69	**Neutral**		[60-90[	**[60-90[**	**Hydroxyl**			
		70	**Neutral**		**[60-90[**		**G**			
		71	**Hydrophobic**		Unclassified		**Aliphatic**			
		72	**Neutral**		**[108-117[**		**P**			
		73	-		-		-			
		74	Unclassified	**Hydrophilic**	[60-90[	[108-117[	Hydroxyl	**Acidic**	110	C"D turn
	D	75	**Hydrophilic**		**[162-174[**		**Basic**		41	C"D turn
		76	**Hydrophobic**		**[189-228[**		**F**		10	inner (core)
		77	**Neutral**		**[60-90[**		**Hydroxyl**		58	outer
		78	**Neutral**		**[60-90[**		**G**		11	inner
		79	**Neutral**		**[60-90[**		**Hydroxyl**		64	outer
		80	**Neutral**	Hydrophilic	**[60-90[**	[162-174[	**G**	Basic	73	outer
		81	-		-		-			
		82	-		-		-			
		83	**Neutral**		**[60-90[**		**Hydroxyl**		74	outer
		84	**Neutral**		**[60-90[**		**G**		56	DE turn
	E	85	**Neutral**	Hydrophilic	**[108-117[**		**Hydroxyl**	Amide	46	DE turn
		86	**Hydrophilic**	**Neutral**	[108-117[	[60-90[	**Acidic**	**Hydroxyl**	82	outer
		87	**Hydrophobic**		**[189-228[**	**[60-90[**	**F**	**Aliphatic**	1	inner (core)
		88	**Neutral**		**[108-117[**	[60-90[	**Hydroxyl**		36	outer
		89	**Hydrophobic**		**[162-174[**		**Aliphatic**		0	inner (core)
		90	Neutral		[108-117[		Hydroxyl		29	outer
		91	**Hydrophobic**		**[162-174[**		**Aliphatic**		0	inner (core)
		92	**Neutral**		**[60-90[**	[60-90[	**Hydroxyl**		57	outer
		93	Unclassified	**Neutral**	[60-90[	**[60-90[**	Hydroxyl	**G**	49	outer
		94	**Hydrophobic**		**[162-174[**	Unclassified	**Aliphatic**		3	inner (core)
		95	**Hydrophilic**		**[138-154[**		Amide			EF turn
		96	Neutral	Hydrophobic	[108-117[	[60-90[	P	Hydroxyl	77	EF turn
	F	97	**Hydrophilic**	Hydrophilic	**[138-154[**	[138-154[	**Acidic**	Acidic	113	EF turn
		98	**Hydrophilic**		**[108-117[**		**Acidic**		5	inner
		99	**Hydrophobic**	**Hydrophilic**	[189-228[	**[138-154[**	F	**Acidic**	58	outer
		100	**Hydrophobic**		**[60-90[**		**Aliphatic**		6	inner (core)
		101	Unclassified	**Hydrophilic**	Unclassified	**[108-117[**	Hydroxyl	**Acidic**	31	outer
		102	**Neutral**		**[189-228[**		**Y**		0	inner
		103	**Neutral**		**[189-228[**		**Y**		6	VH-VL
		104	**Hydrophobic**		**[108-117[**		**Sulfur**		0	inner (core)

Folch G, Scaviner D, Contet V, Lefranc M-P. 2000. Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. *Exp. Clin. Immunogenet.* **17**: 205–215.

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

Giudicelli V, Lefranc M-P. 2003. IMGT-ONTOLOGY: gestion et découverte de connaissances au sein d'IMGT. In *Extraction et gestion des connaissances EGC 2003, Extraction des connaissances et apprentissage*, Revue des Sciences et Technologie de l'Information RSTI, Series RIA-ECA 17; Lavoisier: Paris; 13–23.

Giudicelli V, Chaume D, Mennessier G, Althaus HH, Müller W, Bodmer J, Malik A, Lefranc M-P. 1998a. IMGT, the international ImMunoGeneTics database: a new design for immunogenetics data access. In *MEDINFO'1998*, Cesnik B, *et al.* (eds). IOS Press: Amsterdam; 351–355.

Giudicelli V, Chaume D, Lefranc M-P. 1998b. IMGT/LIGM-DB: a systematized approach for ImMunoGeneTics database coherence and data distribution improvement. International Conference on Intelligent Systems ISBM 98; 59–68.

Kabat EA. 1970. Heterogeneity and structure of antibody-combining sites. *Ann. NY Acad. Sci.* **169**: 43–54.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105–132.

Lebart L, Morineau A, Warwick KM. 1984. *Multivariate descriptive statistical analysis*. Wiley: New York.

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

Lefranc M-P. 1999. The IMGT unique numbering for Immunoglobulins, T cell receptors and IG-like domains. *Immunologist* **7**: 132–136.

Lefranc M-P. 2000a. *Nomenclature of the Human Immunoglobulin Genes*. Current Protocols in Immunology. Wiley: New York; A.1P.1–A.1P.37.

Lefranc M-P. 2000b. *Nomenclature of the Human T cell Receptor Genes*. Current Protocols in Immunology. Wiley: New York; A.1O.1–A.1O.23.

Lefranc M-P. 2001a. IMGT, the international ImMunoGeneTics database. *Nucl. 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. 2001c. Nomenclature of the human immunoglobulin kappa (IGK) genes. *Exp. Clin. Immunogenet.* **18**: 161–174.

Lefranc M-P. 2001d. Nomenclature of the human immunoglobulin lambda (IGL) genes. *Exp. Clin. Immunogenet.* **18**: 242–254.

Lefranc M-P. 2003. IMGT, the international ImMunoGeneTics database®, http://imgt.cines.fr. *Nucl. Acids Res.* **31**: 307–310.

Lefranc M-P, Lefranc G. 2001a. *The Immunoglobulin FactsBook*. Academic Press: London; 448 pages.

Lefranc M-P, Lefranc G. 2001b. *The T cell receptor FactsBook*. Academic Press: London; 384 pages.

Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Chaume D. 1998. IMGT, the international ImMunoGeneTics database. *Nucl Acids Res* **26**: 297–303.

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

Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* **27**: 55–77.

Lesk AM, Chothia C. 1988. Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature* **335**: 188–190.

Martinez C, Lefranc M-P. 1998. The Mouse (*Mus musculus*) immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp. Clin. Immunogenet.* **15**: 184–193.

Martinez-Jean C, Folch G, Lefranc M-P. 2001. Nomenclature and overview of the mouse (*Mus musculus and Mus sp.*) immunoglobulin kappa (IGK) genes. *Exp. Clin. Immunogenet.* **15**: 255–279.

Nguyen VK, Muyldermans S, Hamers R. 1998. The specific variable domain of camel heavy-chain antibodies is encoded in the germline. *J. Mol. Biol.* **23**: 413–418.

Nguyen VK, Hamers R, Wyns L, Muyldermans S. 2000. Camel heavy-chain antibodies: diverse germline V(H)H and specific mechanisms enlarge the antigen-binding repertoire. *EMBO J.* **19**: 921–930.

Padlan EA, Abergel C, Tipper JP. 1995. Identification of specificity-determining residues in antibodies. *FASEB J.* **9**: 133–139.

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 (IGHV) genes. *Exp. Clin. Immunogenet.* **16**: 36–60.

Rawn JD. 1989. *Biochemistry.* Carolina Biological Supply Company: Burlington.

Rudikoff S, Pumphrey JG. 1986. Functional antibody lacking a variable-region disulfure bridge. *Proc. Natl. Acad. Sci. USA* **83**: 7875–7878.

Ruiz M, Lefranc M-P. 2000a. IMGT sequence profile: a standardized visualization for the immunoglobulin and T cell receptor V-REGIONs applicable to other protein alignments. In *Currents in Computational Molecular Biology*, Miyano S, *et al.* (eds). Frontiers Science Series no. 30. Universal Academy Press: Tokyo; 126–127.

Ruiz M, Lefranc M-P. 2000b. Sequence profiles of immunoglobulin and T cell receptor V-REGIONs according to IMGT data. In *JOBIM 2000 Recueil des Actes*, Caraux G, *et al.* (eds). Montpellier; 293–300.

Ruiz M, Lefranc M-P. 2002. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics* **53**: 857–883.

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 SGE, Bontrop R, Lemaître M, Lefranc G, Chaume D, Lefranc M-P. 2000. IMGT, the international ImMunoGeneTics database. *Nucl. Acids Res.* **28**: 219–221.

Scaviner D, Barbié V, Ruiz M, Lefranc M-P. 1999. Protein displays of the human immunoglobulin heavy, kappa and lambda

---

[a] FR-IMGT and strands designations are as described in IMGT Colliers de Perles (IMGT Repertoire, http://imgt.cines.fr). CDR1-IMGT, CDR2-IMGT and CDR3-IMGT positions are not shown.

[b] IMGT numbering is according to the IMGT unique numbering for V–REGION (Lefranc, 1997, 1999; Lefranc *et al*., 2003). For simplification, amino acids from the beta turns (Ruiz and Lefranc, 2000a, b) are included in the adjacent strands.

[c] Classes are as in Table 1.

[d] ASA Mean accessible surface areas (in $Å^2$) are from Chothia *et al*. (1998). Amino acids of VH-VL are orientated towards the outside of the sandwich but being buried in the domain interface. They have limited accessible surface area in 3D structure.

[e] The side-chain orientation and the localization in the 3D structure are described based on the 3D structures available in IMGT/3Dstructure-DB, http://imgt.cines.fr (Ruiz and Lefranc, 2002). Outer: outer surface of the ABED beta sheet and of the F strand (exposed to solvent); inner: inner surface of the ABED and GFCC′C″ beta sheets (inside of the 'sandwich'); VH-VL: interface between the VH-DOMAIN and the VL-DOMAIN (Ruiz and Lefranc, 2000a,b).

variable and joining regions. *Exp. Clin. Immunogenet.* **16**: 234–240.

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

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

Thioulouse J, Chessel D, Dolédec S. 1997. A multivariate analysis and graphical display software. *Stat. Comput.* **7**: 75–83.

Williams AF, Barclay AN. 1988. The immunoglobulin superfamily-domains for cell surface recognition. *A. Rev. Immunol.* **6**: 381–405.

Zamyatnin AA. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* **24**: 107–123.