doi:10.1006/jmbi.2001.4662 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2001) **309**, 657–670

JMB

Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool

Annemarie Honegger* and Andreas Plückthun

Biochemisches Institut der Universität Zürich Winterthurerstrasse 190 CH-8057 Zürich Switzerland

**Corresponding author*

A common residue numbering scheme for all immunoglobulin variable domains (immunoglobulin light chain lambda (V_λ) and kappa (V_κ) variable domains, heavy chain variable domains (V_H) and T-cell receptor alpha (V_α), beta (V_β), gamma (V_γ) and delta (V_δ) variable domains) has been devised. Based on the spatial alignment of known three-dimensional structures of immunoglobulin domains, it places the alignment gaps in a way that minimizes the average deviation from the averaged structure of the aligned domains. This residue numbering scheme was applied to the immunoglobulin variable domain structures in the PDB database to automate the extraction of information on structural variations in homologous positions of the different molecules. A number of methods are presented that allow the automated projection of information derived from individual structures or from the comparison of multi-structure alignments onto a graphical representation of the sequence alignment.

Keywords: immunoglobulin variable domains; numbering scheme; protein engineering; antibody engineering

Introduction

When analyzing an individual protein structure, sequential numbering of the amino acid residues within each chain, combined with a unique chain label for each chain in multi-chain structures, is the most convenient method to identify individual residues within the protein structure and sequence. When comparing a large number of related structures showing multiple sequence insertions and deletions, such a simple scheme becomes very cumbersome. To facilitate discussion of common sequence and structural features within a protein family, a unified residue nomenclature, in which structurally equivalent residues in different family members are identified by the same residue labels, is very useful. In an attempt to automate the comparative analysis of a large number of related sequences and structures, the unambiguous identification of structurally equivalent residues is of crucial importance. The quality of the final results of such an analysis is limited by the accuracy with which the input sequences and structures have been aligned. As Protein Data Bank (PDB) files (coordinates derived from X-ray and NMR structures) represent the major input for this type of analysis, such a numbering scheme should be compatible with the specifications of the PDB file format. This format allows a one-character chain identifier, a four-digit residue number and a one-character insertion

Abbreviations used: CDR, complementarity-determining region; FR, framework region of immunoglobulin variable domains; PDB, Protein Data Bank; Fv, heterodimeric fragment consisting of a V_L and a V_H domain; Fab, heterodimeric fragment consisting of an antibody light chain and the first two domains of the heavy chain; scFv, single-chain Fv fragment of an antibody, V_L and V_H connected by a flexible peptide linker; TCR, T-cell receptor; V_L and V_H, variable domains of the antibody light and heavy chains; C_L and C_H, constant domains antibody light and heavy chains; V_λ and V_κ, lambda and kappa subtypes of light chain variable domains; V_α, V_β, V_γ, and V_δ, variable domains of T-cell receptor alpha, beta, gamma and delta chains.

E-mail address of the corresponding author: honegger@bioc.unizh.ch

0022-2836/01/030657–14 $35.00/0

code. However, the field allocated to the insertion code according to the most recent definition of the PDB ATOM and HETATM records† is frequently used to indicate alternate atom locations or, occasionally, used for the chain identifier, depending on the computer program that generated the file.

In this analysis, we are concentrating on the immunoglobulin superfamily, and especially on the family of immunoglobulin variable domains, which consists of the immunoglobulin light chain lambda (V_λ) and kappa (V_κ) variable domains, the heavy chain variable domains (V_H) and the T-cell receptor alpha (V_α), beta (V_β), gamma (V_γ) and delta (V_δ) variable domains, which share sufficient structural and functional similarity that a comparative analysis can yield valuable insights into the rules governing the correlation between sequence, structure and domain folding behavior. The immunoglobulin variable domains represent a particularly well-suited target for such a study. The creation, selection and evolution of functional antibodies and T-cell receptors from the natural combinatorial libraries encoded in the immunoglobulin loci by genetic recombination and somatic hypermutation in the time-course of an immune response create a degree of variability within a single organism unseen in any other protein family.[1,2]

The usefulness of antibodies and of antibody-derived artificial constructs in various medical and biochemical applications has made them a prime target for protein engineering.[3] Fv fragments and single-chain Fv fragments, in which the antibody V_L and V_H domain are connected by a flexible linker, represent the minimal building blocks needed to preserve the antigen-recognition function.[4–6] ScFv or Fab fragments can be obtained from given monoclonal antibodies, from libraries derived from immunized animals, from the naïve B-cell repertoire or from germline genes.[7,8] Although some of the fragments generated in this way perform very well, many others show insufficient stability and production yields for the intended application and have to be improved by rational engineering[9–12] or *in vitro* evolution schemes.[13,14] Feedback from these experiments provides valuable input to be combined with the data derived from public domain databases.

Currently (June 2000) the PDB‡ contains the 3D structures of 281 V_H domains representing 181 non-identical sequences, 269 V_κ domains representing 183 non-identical sequences and 56 V_λ-domains representing 28 non-identical sequences. In addition, 13 (eight non-identical) TCR V_α, 16 (six non-identical) TCR V_β and 1 TCR V_δ domain structures are available. Combined with the knowledge of the sequence variability allowed by the immunoglobulin germline repertoire (IMGT,[15]§ VBase and ABG¶), and the thousands of rearranged sequences collected in the Kabat database||, a wealth of information is available, which just has to be combined and visualized in a form suitable to facilitate interpretation.

The combined information can be used to improve the structural modeling and functional predictions involved in the optimization of engineered antibody fragments.[16] The insights gained from the analysis can then be applied in the construction of synthetic antibody libraries,[17] and used to help in the interpretation of *in vitro* evolution experiments.[18]

To facilitate this analysis, the data derived from the different databases has been rearranged in such a way that it can be accessed by molecule and by residue. With the new scheme proposed here, using PDB coordinate sets that are consistently numbered and superimposed, the features can be extracted and displayed for the entire family. Queries and analyses are easily possible, such as to display the superimposed structures of all V_L kappa domains from residue 20 to 47 (all complementarity-determining region (CDR) L1 loops with adjacent beta sheets), or show the main-chain torsion angles of the V_H residue H7 color-coded by domain subtype, show the position-dependent amino acid compositions of all human and murine V_λ sequences contained in the Kabat database, just to illustrate some examples.

Current Residue Numbering Schemes

In their compilation of Sequences of Proteins of Immunological Interest, of which several printed editions had appeared, Kabat e*t al.*[19] collected and aligned the sequences of different members of the immunoglobulin superfamily. They proposed numbering schemes for many of the different protein families, which make up the immunoglobulin superfamily. The placement of sequence gaps was based on sequence variability rather than on the spatial structure, as far fewer structures were known at the time. Chothia & Lesk[20] corrected the positioning of CDR L1 and CDR H1 sequence length variability in the antibody variable domains to better fit their actual position in the three-dimensional structure. In 1989, Chothia and colleagues changed the insertion point in CDR L1,[21] but returned to the old definition in 1997.[22] The different families of immunoglobulin domains are treated separately by both Kabat and Chothia, despite the high degree of sequence and structural homology. Gelfand and colleagues.[23–25] studied in detail the structurally invariant core of antibody V_L and V_H domains, but they basically restricted their analysis to a set of residues that is almost identical with the core residues we use for least-squares

† http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html
‡ http://www.rcsb.org/
§ http://imgt.cnusc.fr:8104/
VBase, http://www.mrc-cpe.cam.ac.uk/imt-doc/
¶ http://www.ibt.unam.mx/vir/V_mice.html
|| http://immuno.bme.nwu.edu/

superpositions and ignored the less conserved positions. They identified the residues by their secondary structure position, resulting in a complex nomenclature incompatible with the PDB format, indicating the β-strand in which a residue is located and its position in this strand. Lefranc and colleagues (IMGT[15,26]†) proposed a unified numbering scheme for immunoglobulin variable domain germline sequences,[27] including antibody lambda and kappa light and heavy chain variable domains as well as T-cell receptor alpha, beta, gamma and delta chain variable domains. However, since they deal only with germ-line sequences, their numbering scheme reaches only into CDR 3 and does not address the residues in CDR 3 or framework 4. An important difference from the numbering scheme presented here (AHo) is that in the IMGT scheme insertions and deletions ''grow'' unidirectionally, as in the original Chothia definition,[20] while in the AHo scheme, insertions and deletions are placed symmetrically around the key position marked in yellow (Figure 1(a)). Furthermore, length variations in CDR 1 and CDR 2 are represented by a single gap in IMGT and by one or two gaps in AHo (Figures 1 and 3). The different numbering schemes are compared in Figure 1(a), together with the sequences (Figure 1(b) and (c)) of the different variable domain PDB structures representing the structural diversity of immunoglobulin and T-cell receptor variable domains.

In the course of our efforts to study the influence of individual sequence features on functionality, stability and folding efficiency of antibody single-chain constructs, as well as of constructs derived from T-cell receptor variable domains, we had to automate the comparison of a large number of immunoglobulin variable domain X-ray structures, including residues from the loop regions. This necessitated the introduction of a unified residue labeling scheme carefully designed to preserve the positional information derived from the comparison of the experimental structures and models. Based on structural criteria, it seemed clear that sequence insertions and deletions, indicated by alignment gaps, should be placed symmetrically, centered on turn and loop positions in the structure, as this is the only location where they can be accommodated without distorting the surrounding structure. In contrast, the existing numbering schemes all imply a unidirectional insertion.

In order to facilitate automated processing, we tried to avoid alphabetic modifiers to the residue numbers (insertion numbering, such as 30A, 30B, etc.), but chose to allow large enough gaps in the numbering scheme to accommodate the length variation of all known germline sequences as well as to give ample space for long CDR H3 loops. Sequence insertions due to somatic mutations[28] and extremly long CDR H3 loops[29] exceeding the length provided for in our numbering scheme still may nessecitate the use of insertion numbering. Its use would, however, mark rare exceptions that defy the statistical norm. Antibody light and heavy chains are identified by the chain identifier L and H, T-cell receptor alpha, beta, gamma and delta chains by A, B, C and D, respectively, placed before the residue number (e.g. H6 for position 6 in the antibody heavy chain, A6 for the structurally equivalent position in the TCR alpha chain).

Immunoglobulin variable domain sequences extracted from the Kabat‡, IMGT†, VBase§, ABG¶ and PDB|| databases were aligned using the GCG (Wisconsin Package Version 9.0, Genetics Computer Group, Madison, WI) module PILEUP. Sequence alignments were tabulated using the GCG module PRETTY and imported into EXCEL98 (Microsoft). To determine the optimal placement of the alignment gaps, the 3D structures of several representative members of the different families covering the observed length variability were aligned by a least-squares fit of C^{α} positions in the structurally most conserved core region of the domains, consisting of residues 3-7, 20-24, 41-47, 51-57, 78-82, 89-93, 102-108 and 138-144, and indicated in dark gray in Figure 1(a) and in white in Figure 2(a). The C^{α} coordinates of the aligned structures were imported into EXCEL. The sequence alignment of the residues not used for the structural alignment was then optimized in such a way that the sum of the average deviation from the mean C^{α} position for each position was minimized, resulting in the sequence alignment shown in Figure 1(b) and (c). The positions of sequence gaps in the other alignments were corrected to conform to the AHo numbering scheme using EXCEL Visual Basic macros. These alignments were color-coded according to different criteria: type of amino acid (Figure 1(b)), hydropathy, predicted secondary structure, and similarity to a reference sequence (not shown). Other macros were written to calculate consensus sequences, distance matrices and to determine the sequence variability and position-dependent amino acid composition of sequences grouped according to different criteria (see Figure 5(b) and (c)).

Description of the AHo Numbering Scheme

The first gap to be placed concerns the one-amino acid insertions in the V_L kappa framework 1 compared to V_L lambda and V_H (Figure 3(a)). Kabat placed this insertion/deletion at the tenth residue from the amino terminus. However, as a

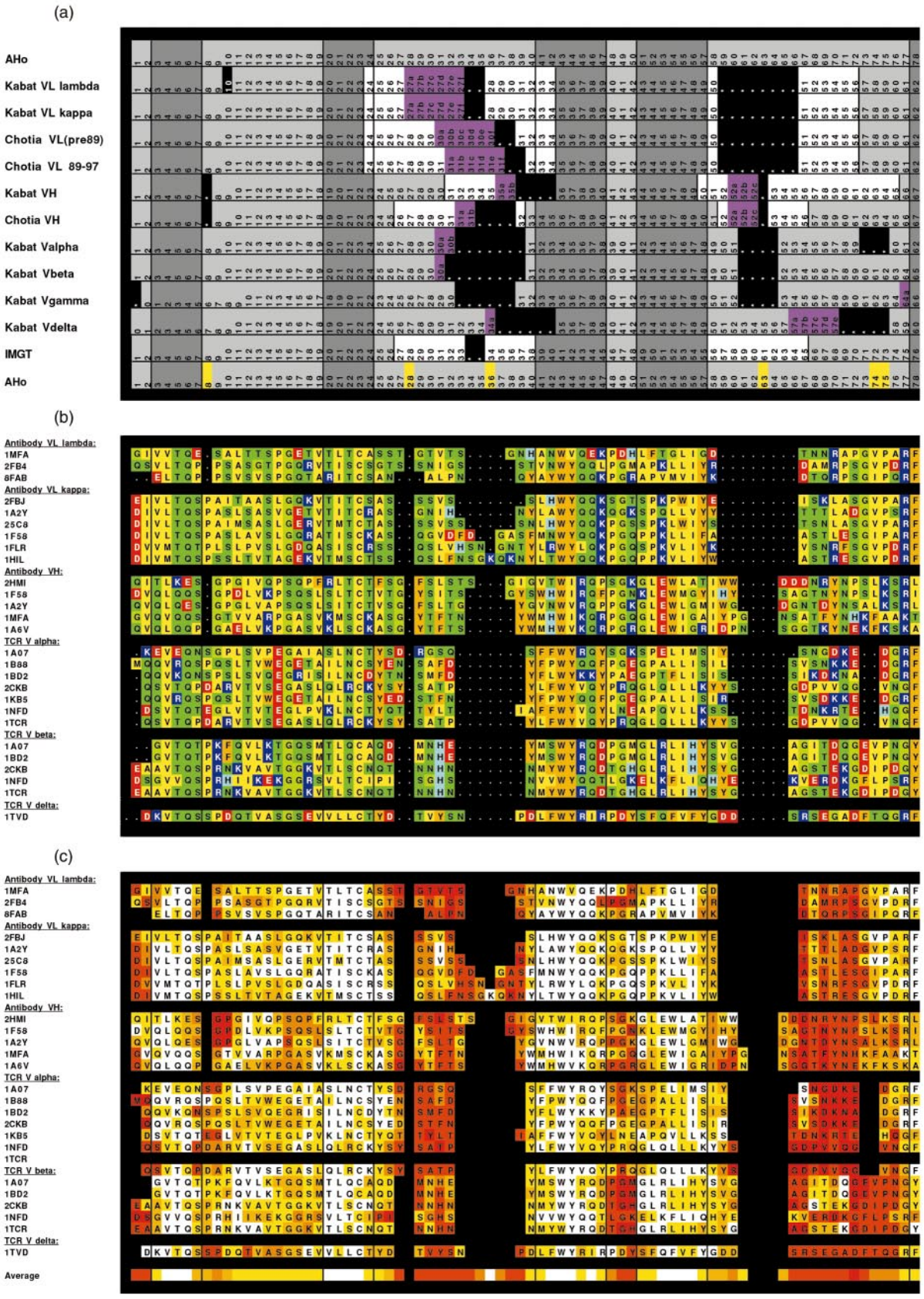

Figure 1 (*legend shown on page 672*)

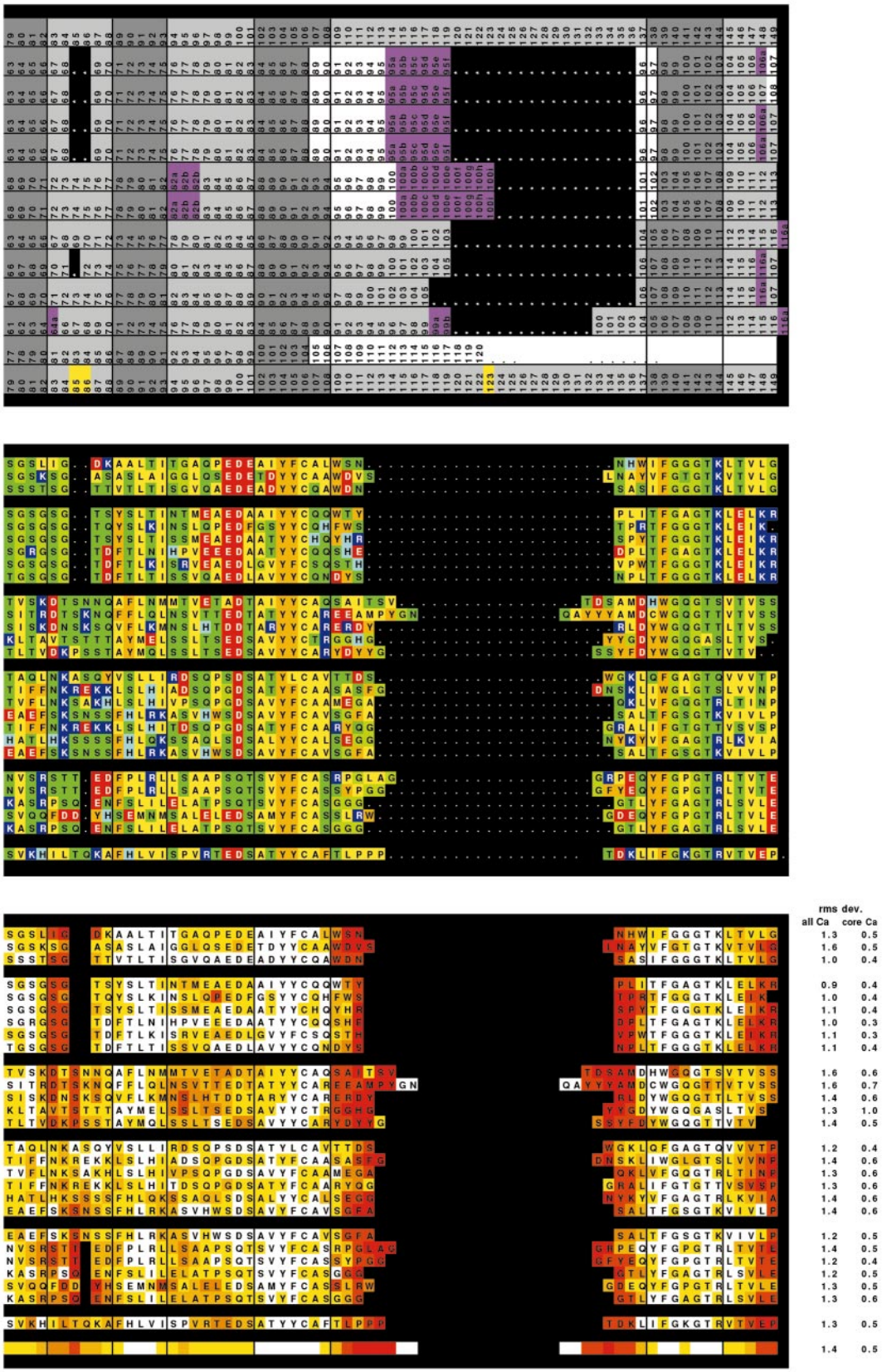

Figure 1 (*legend shown on page 672*)

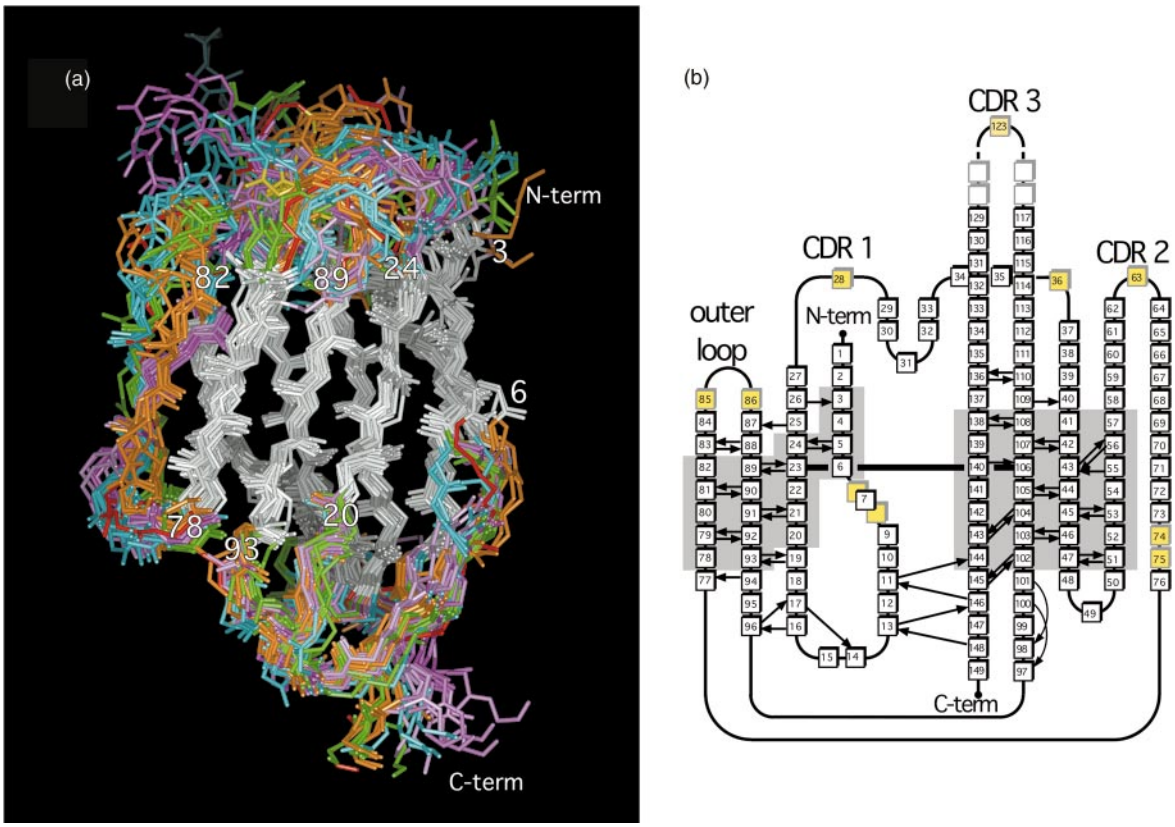

Figure 2. (a) Representative structures of V_λ (PDB entries 1MFA,[37] 2FB4,[38] and 8FAB[39] pink), V_κ (PDB entries 1A2Y,[40] 1F58,[36] 1FLR,[41] 1HIL,[42] 25C8,[43] and 2FBJ[44] magenta), V_H (PDB entries 1A2Y,[40] 1A6V,[45] 1F58,[36] 1FLR,[41] 1MFA,[37] 1MRC,[46] and 2HMI[47] cyan), TCR V_α (PDB entries 1A07,[48] 1B88,[49] 1BD2,[50] 1KB5,[51] 1NFD,[52] and 1TCR[53] orange), TCR V_β (PDB entries 1A07,[48] 1BD2,[50] 1KB5,[51] 1NFD,[52] 1TCR[53] green) and TCR V_δ (PDB entry 1TVD[31] red) were aligned by a least-squares fit of the C^α atoms of residues 3-7, 20-24, 41-47, 51-57, 78-82, 89-93, 102-108 and 138-144 (indicated in white). Numbers indicate residue positions of the aligned β-strands. (b) A representation of the consensus structure and main-chain hydrogen bonding pattern of immunoglobulin variable domains. Arrows indicate hydrogen bonds that are present in the majority of the structures in all types of immunoglobulin variable domains. The loop and turn regions that accommodate gaps are indicated in yellow. Gray areas underlie the residues whose C^α positions were used for least-squares superposition of the structures.

Figure 1. (a) The different numbering schemes for immunoglobulin variable domains (Kabat,[19] Chothia,[20,22] IMGT,[27] AHo (this study)) have been aligned. Boxes with dark gray background indicate the residues whose C^α coordinates were used for the least-squares alignment of the representative structures of each type of variable domain (Figure 2). Pink background indicates the placement of insertions/deletions implied by the numbering scheme. White dots on black background indicate additional gaps introduced to bring the different numbering schemes into structural alignment. White background indicates the position of the complementarity-determining regions. In the AHo numbering scheme, gaps are centered on the positions indicated in yellow. (b) Sequences representing the different types of immunoglobulin variable domains (immunoglobulin V_H, V_κ and V_λ, T-cell receptor V_α, V_β and V_δ) shown in Figure 2 were aligned so as to minimize the average deviation from the average structure. The amino acids are color coded according to residue type: aromatic residues (Tyr, Phe, Trp), orange; hydrophobic residues (Leu, Ile, Val, Met, Cys, Pro, Ala), yellow; uncharged hydrophilic residues (Ser, Thr, Gln, Asn, Gly), green; acidic residues (Asp, Glu), red; basic residues (Arg, Lys, His), blue. The corresponding 3D structures are shown in Figures 2 and 3. (c) Average deviation from the mean C^α position. Individual domains were excised from the corresponding PDB files and aligned by a least-squares fit of the C^α positions of the core residues (3-7, 20-24, 41-47, 51-57, 78-82, 89-93, 102-108 and 138-144) to the corresponding C^α positions of a reference structure (V_κ domain of PDB entry 1F58[36]). The mean C^α coordinates for each position in the alignment were calculated and the deviation of each residue in each structure from the average structure calculated and color coded in the corresponding sequence alignment (white, rms deviation <0.5 Å; yellow, 0.5-1 Å; yellow-orange, 1-1.5 Å; orange, 1.5-2 Å; orange-red, 2-4 Å; red, >4 Å). The C^α rms deviations of each structure from the average and the rms deviation of the C^α positions used for the alignment (dark gray boxes in (a)) are indicated in the last two columns.

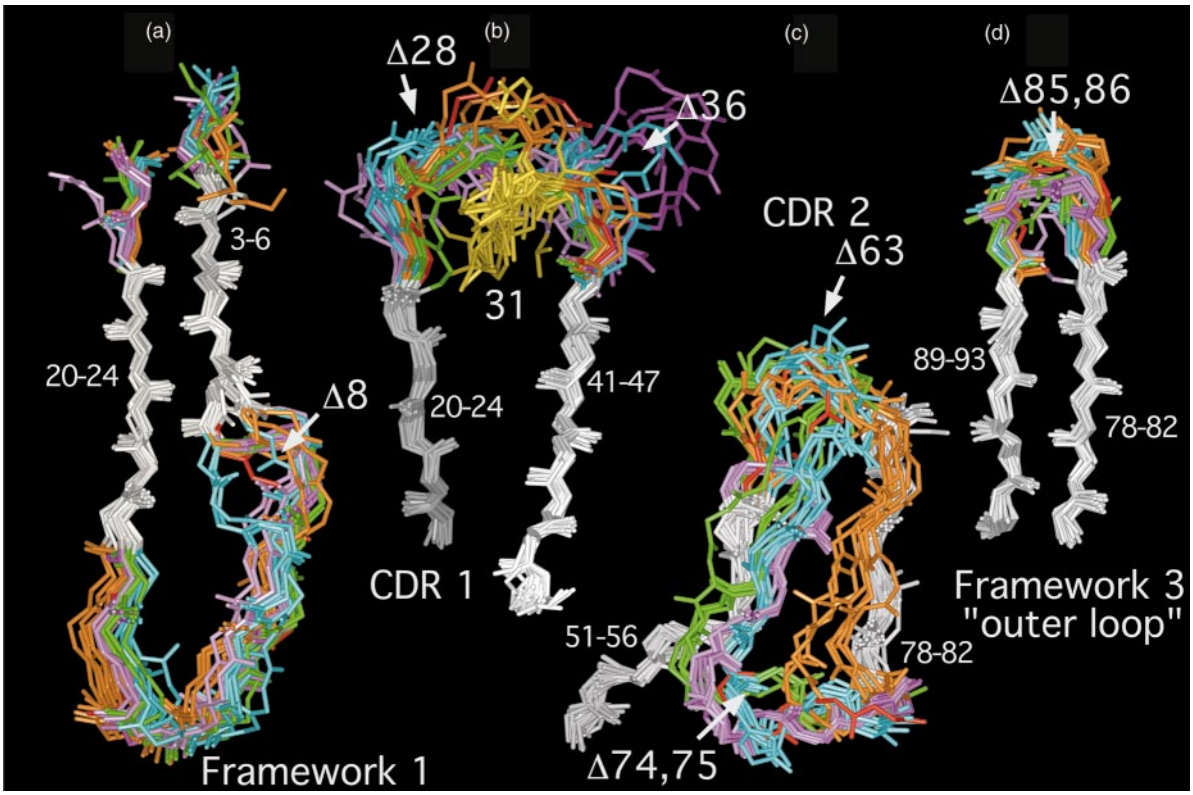

Figure 3. Placement of sequence length variability in the 3D structure of immunoglobulin variable domains. The domains were structurally aligned and color coded as described in the legend to Figure 2. The chain segments used for least-squares alignment are white, the loop regions pink (V_λ), magenta (V_κ) , cyan (V_H), green (TCR V_α) or orange (TCR V_β), depending on the domain type. The symbol Δ indicates the positions where gaps are placed to keep the numbering scheme consistent between different types of variable domains. Larger gaps are placed symmetrically, centered on this position. (a) Framework 1 region. V_λ and V_H have a one-residue gap in position 8 compared to V_κ. The TCR V_α and V_β domains whose structure have been solved so far are all V_κ-like in length, although an alignment of TCR germlines shows that V_α sequences with a V_λ-like length also exist. Most of TCR V_β and 50 % of the V_κ-like TCR V_α germlines have Pro in position 8. The available structures suggest that this Pro is a *cis*-Pro, as it is in V_κ domains. (b) Complementarity-determining region 1 (CDR 1). The segment containing the complementary-determining region 1 is divided into two loops by residue 31, which intercalates between the outer and inner (dimer interface) beta sheet of the immunoglobulin variable domain beta sandwich. The length variability between CDR 1 loops of the same class (e.g. V_κ, V_H) is mainly accommodated in the inner, C-terminal loop, centered around position 36, while the different classes of molecules differ in the length of the outer, N-terminal loop, accommodated by a gap centered around position 28. (c) CDR 2. While the length difference between V_L and V_H can be accommodated at the turn centered on positions 62 and 63, the descending branch of CDR 2 in TCR V_α is retracted from the dimer interface in such a way that the structural alignment is best described by placing an additional two-residue gap in positions 74 and 75. (d) Framework 3 region. The length variability of the framework 3 sequence is accommodated in the hairpin loop sometimes called CDR 4 in T-cell receptors. V_L domains have a two-residue gap in positions 85 and 86 compared to V_H and V_α domains, V_β domains have a one residue gap in position 86.

superposition of representative lambda and kappa chain structures shows, the amide nitrogen atom of the seventh residue assumes a structurally equivalent position in V_λ and V_κ chains, while the C═O oxygen of the seventh residue of V_λ is structurally equivalent to the C═O oxygen of the eighth residue in V_κ. Structurally, the gap in V_λ relative to V_κ should therefore be placed either in position L7 or L8.[30] The corresponding segment of TCR V_β chains is V_L kappa-like in length and structure, including the very frequent occurrence of a *cis*-proline residue in position B8, while some V_α-chains are V_κ-like and others resemble V_λ with its one-residue deletion. For V_γ and V_δ there is not enough structural information available yet, although the single delta chain for which a structure has been solved[31] (PDB entry 1TVD) is V_κ-like in length. V_H domains, while corresponding in length to V_λ, show a larger structural variation in this region,[32–34] and the gap could equally well be placed in position H7, H8 or H9. We therefore chose to place the gap in position H8. The conserved cysteine residue forming the intradomain disulfide bridge therefore always carries the label 23, as in the IMGT numbering scheme, while according to Kabat, it was labeled L23 in V_κ and V_λ, H22 in V_H, 23 in TCR V_α and V_β and 21 in TCR V_γ and V_δ.

The first complementarity-determining region, CDR L1 and CDR H1 in antibodies, is treated as a single loop in all existing numbering schemes. With CDR 1, the amino acid chain switches from the ''outer'' beta-sheet to the ''inner'' one, which forms the heterodimer interface between the antibody light and heavy chain, between the alpha and beta chains in alpha-beta TCRs and between the gamma and delta chains in gamma-delta TCRs. Structurally, in the CDR 1 segment, one usually hydrophobic residue assumes a distinguished position. It intercalates between the two beta sheets and divides the CDR 1 into two loops (Figure 3(b)). In our numbering scheme, the residue number of this amino acid is 31. The outer, N-terminal loop is longest in some TCR V_α domains and shortest in some V_λ domains, which both show some length variability in this part of CDR 1. V_H domains, as well as the majority of the V_α domains, have a one-residue gap in position 28, V_κ and V_β domains a two-residue gap in positions 27 and 28. CDR L1 of the kappa chain shows the largest degree of length variability, affecting only the loop C-terminal to the intercalating residue 31. Here, the gap is centered on position 36. The highly conserved core tryptophan residue (Kabat H36, L35, A34, B34, C35 or D35, IMGT 41) is always numbered 43 in our scheme.

Since the two gap positions, centered on residues 28 and 36, are treated separately and the intercalated residue is always kept fixed at position 31, a shorter N-terminal loop cannot compensate a longer C-terminal loop. Compared to the IMGT scheme, two positions more are needed to represent the length variability, since in the IMGT alignment, an insertion in the N-terminal loop of CDR 1 can be compensated by a short C-terminal loop (Figure 1(a)).

The gap correcting for the length variability in the CDR 2 region has been placed to be centered on position 63, resulting in an eight residue gap in the V_L domains (L59-L66, placed between L50 and L51 according to Kabat nomenclature) and a one to four residue gap in V_H domains (between H52 and H53 according to Kabat) (Figure 3(c)). In V_α domains, a second gap has been placed in positions A74 and A75 to reflect the way in which the C-terminal branch of the CDR A2 loop is retracted from the dimer interface compared to the CDR 2 of other immunoglobulin variable domains. While in all other variable domains the descending branch of CDR 2 is part of the inner beta-sheet (which forms the dimer interface), the descending branch of the V_α CDR 2 (residues 67-73) is more closely associated with the outer beta-sheet, although the hydrogen bonds to the strand formed by residues 78-82 are not very well conserved in the few V_α structures known. Again, this splitting of the CDR 2 length variability into two insertion points necessitates more positions than allowed for in the IMGT nomenclature.

The outer loop, sometimes called CDR 4 in T-cell receptors, contains a two residue gap in V_L (L74, L75, between Kabat L68 and L69). V_H and V_α have no gap, while V_β has a one residue gap in position B86 (Figure 3(d)). This places the second cysteine residue in position 106 (Kabat L88, H92, A90, B92, C94 and D88, IMGT 104). The CDR 3 residues are divided symmetrically between this cysteine residue and the first glycine residue of the framework 4 beta bulge in position 140 (Kabat L99, H104, A107, B109, C109 and D107, undefined in IMGT), leaving ample space for long CDR 3 sequences and centering the CDR 3 gap around position 123. The number of residues allowed for the CDR 3 length variability may seem excessive, but CDR3 loops close to that length can be found in the Kabat database.

Verification and Practical Applications of the New Numbering Scheme

To prove that the new numbering scheme is indeed a better representation of the structural equivalence of residues in the immunoglobulin domains than the classical schemes, the X-ray structures of the individual domains were extracted from the corresponding PDB files and structurally aligned by a least-squares fit between the least variable C^α positions (3-7, 20-24, 41-47, 51-57, 78-82, 89-93, 102-108 and 138-144) using the program Insight II (MSI/Biosym). The C^α coordinates of the aligned domains were exported to EXCEL98 (Microsoft), preserving the new relative orientations. This allowed the calculation of the mean C^α coordinates for each position and the deviation of the corresponding residue in the actual structures from this consensus position, taking into account the alignments implied by the different numbering schemes. To minimize the effects of the choice of reference structure on the quality of the final alignment, the structural alignment was performed in two steps: first, an arbitrary structure was used as a reference structure. From the alignment to this structure, the mean C^α coordinates for each position were calculated and all structures compared to this mean structure. The structure with the lowest deviation from the average was selected as new reference structure and the analysis repeated. The numerical values for the structural deviations were translated into a color code projected onto the sequence alignment as shown in Figure 1(c) and plotted against the residue number in a graphical representation (not shown). This representation also provides an indication of which positions can truly be considered structurally equivalent (deviation from average C^α position $\ll$ distance between two neighbouring C^α positions).

Such a coloring of an alignment of individual structures within a family serves well to quickly identify outliers. Larger deviations from the consensus of individual sequences or groups of sequences help to identify residues that cause a conformational change. For instance, a strong

structural deviation in the outer loop of V_κ (L83-L87) from the average V_κ conformation correlates with the presence of a non-Gly residue in position L82. This residue usually assumes a positive ϕ torsion angle disallowed for residues with bulkier side-chains, and a non-Gly residue in this position results in an outward kink of the loop (Figure 5(a)). The mean C^α deviation for each position, indicated in the header of an alignment of related sequences, serves as a quick reminder of which parts of the sequence represent structurally conserved residues, and which positions are more variable (Figure 4(a)-(c)).

The core C^α positions of the different variable domain structures shown in Figure 2 fit with a rms deviation of 0.52 Å. Using all C^α positions, with gaps positioned as defined by the AHo numbering scheme, the rms deviation obtained was 1.38 Å, with gaps needed to align the different types of variable domains positioned as implied by the IMGT numbering scheme, 1.5 Å. Since only the small fraction of the residues that align differently in the two schemes contribute to this difference, this represents quite a large difference in the residues affected. The core residues of 185 V_κ domains fit with an average rms deviation of 0.3 Å, the core residues of 23 V_λ with an average rms deviation of 0.3 Å, and the core residues of 206 V_H domains with an average rms deviation of 0.4 Å. The average structural variability for all C^α positions was 0.6 Å for the V_κ domains, 0.7 Å for the V_λ domains, and 1.2 Å for the V_H domains.

The residue and side-chain solvent-accessible surfaces of the residues were calculated from the individual V_L and V_H domain structures as well as for liganded and unliganded Fv fragments and Fab fragments using the program NACCESS†. To correct for the different absolute surface areas of the different residue types, accessibilities were expressed as a percentage of the exposed surface area of the same amino acid in the context of a poly(Ala)peptide in extended conformation. Thus, highly exposed residues, e.g. in turns or at the termini, can have relative accessibilities of more than 100%. The relative side-chain accessibilites were color-coded onto the sequence alignment to indicate buried and solvent-exposed residues (Figure 4(e)). From the relative reduction of the absolute solvent-accessible surface area of each residue in the Fv fragment upon antigen binding, the average contribution of the different residue positions to the antigen binding site could be evaluated (Figure 4(f)). This automation enabled us to expand the study by Padlan *et al.*[35] to enumerate the antigen/antibody contacts for a total of 45 antibody/protein complex structures, 30 antibody/oligomer complexes and 52 antibody/hapten complexes and investigate which residues contribute to the interface, and to link this positional information to the sequence variability observed in each position in murine and human germline sequences, and in the rearranged sequences, subdivided according to germline families (Figure 5(b) and (c)). The same method of evaluating differences in residue solvent-accessibility allowed to identify the residues contributing to the V_L/V_H C_L/C_H dimer interface, the V_L/C_L and V_H/C_H interface[11] (Figure 4(g) and (h)).

Lists of main-chain and side-chain hydrogen bonds, torsion angles, hydrogen bonds and other properties were extracted from the individual domain structures and correlated with sequence pattern and structural properties (Figure 5(b)). In Figure 5(c), preferred residue types for different antigen types were extracted. Having the different data tables in an interactive spreadsheet application and the structurally equivalent sequence positions linked by a common residue numbering scheme allowed us to sort the data according to various criteria intrinsic or extrinsic to the data tables analyzed, and thus to test for the influence of different factors on the data.

One example is given by Langedijk *et al.*,[32] where the influence of particular sequence positions on framework conformation and hydrogen bonding pattern was analyzed. In the work by Honegger & Plückthun,[34] these observations were generalized to allow the prediction of the influence of frequently observed primer-induced mutations on the conformation of the framework 1 region of V_H. As another example, the analysis of the interaction modes and antibody sequence preferences of different types of antigens allowed us to extract rules for the optimization of randomization strategies to be employed in the refinement of synthetic antibody libraries such as the human combinatorial antibody library (HuCAL).[17]

While the numbering of the sequence by itself may not seem like an important step in the analysis, these examples show that the correct structural alignment, which forms the basis for the new numbering scheme, simplifies all analyses enormously. By having a strict structural correspondence in residue numbers across all variable domains, very fast access to all comparative and structural parameters is possible, using standard molecular modeling packages and spreadsheet applications.

The aligned and renumbered PDB coordinates, sequence alignments and derived information will be collected in an appropriate database system to be made available on the Internet‡. A set of EXCEL macros will be provided, which allow the semi-automatic renumbering of multiple PDB-files: reading the files into an EXCEL workbook, extracting the sequence (from the data part, not from the header) and displaying a sequence alignment ungapped or gapped according to the residue labels present in the PDB file. The alignment can be changed manually or exported to be altered by external programs (e.g. GCG module PILEUP, or for a better automated fit to the proposed

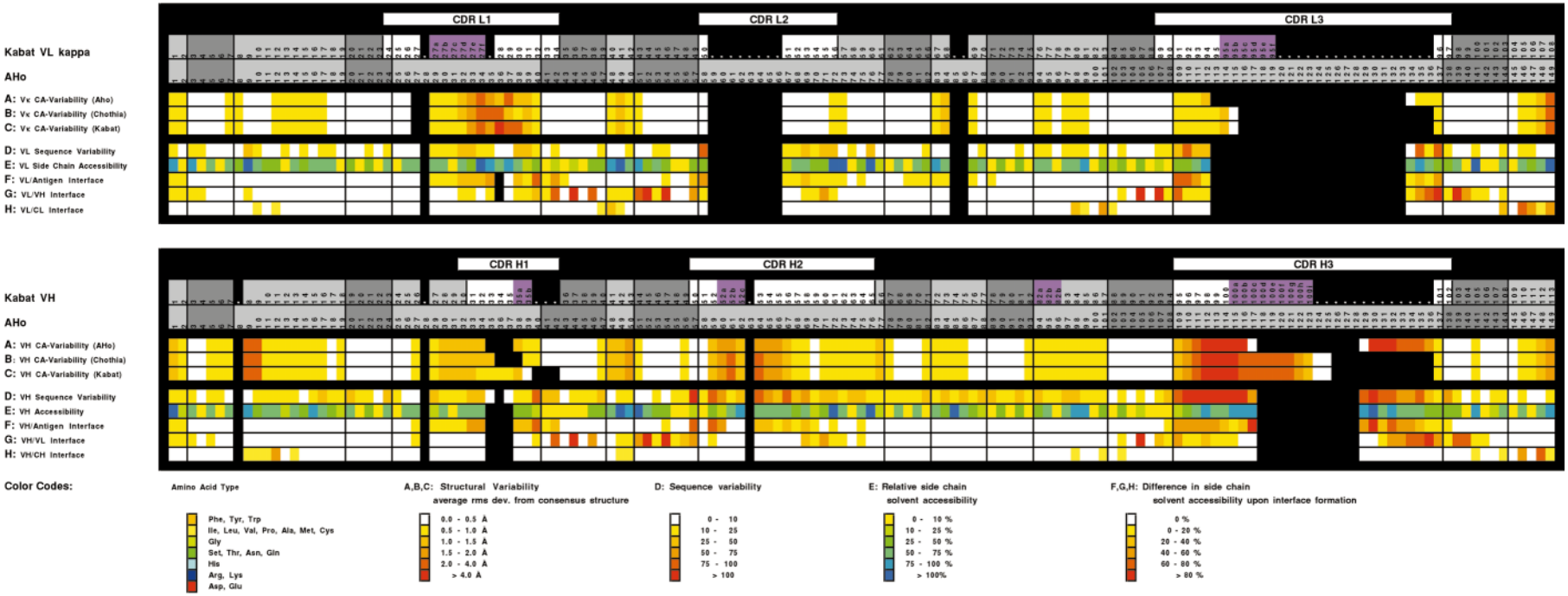

Figure 4. Standard header for V_L and V_H sequence alignments, summarizing the consensus structural properties and interaction residues of the domains, easily extractable from the aligned domains with the new numbering scheme. (a)-(c) Structural variability: average rms deviation from mean C^α position (average of 185 V_κ and 206 V_H structures representing >100 non-identical sequences, all the experimental Fv and Fab structures with a resolution better than 3.0 Å available in the PDB database at the time of the analysis) for structures aligned according to the (a) AHo, (b) Chotia and (c) Kabat alignment. Individual domains were excised from the corresponding PDB files and aligned by a least-squares fit of the C^α positions of the core residues (3-7, 20-24, 41-47, 51-57, 78-82, 89-93, 102-108 and 138-144) to the corresponding C^α-positions of a reference structure 1YEH[54] for V_κ and 1MFD[55] for V_H. The mean C^α positions for each residue were calculated and the average deviation for each residue position in the alignment is indicated by a color code (white, rms deviation <0.5 Å, yellow; 0.5-1 Å, yellow-orange; 1-1.5 Å, orange; 1.5-2 Å, orange-red, 2-4 Å; red, >4 Å). (d) Sequence variability. From the position-dependent amino acid composition, the sequence variability was calculated according to Kabat *et al.*:[19] Variability = 100 × (number of different amino acids at position *n*)/(frequency of the most common amino acid at that position). Color code: white, <10; yellow, 10-25; yellow-orange, 25-50; orange, 50-75; red-orange, 75-100; red, >100. (e) Average relative side-chain accessibility. The side-chain solvent-accessible surface of each residue was calculated as a percentage of the solvent-accessible surface that the same residue would have in the context of a poly(Ala) peptide in extended conformation (program NACCESS (http://wolf.bms.umist.ac.uk/naccess)) (yellow, 0-10 %; yellow-green, 10-25 %; buried; green, 25-50 %; green-blue, 50-75 %, semi-buried; blue, 75-100 %; dark blue, >100 %, exposed). (f) Reduction of the side-chain accessible surface upon formation of the complex of the Fv fragment with a protein antigen. Relative reduction of the side-chain accessible surface of each residue in the complex of antigen-Fv fragment compared to the same residue in the Fv fragment without antigen (white, 0 % reduction; yellow, 0-20 %; yellow-orange, 20-40 %; orange, 40-60 %; red-orange, 60-80 %; red, 80-100 %). (g) Reduction of the side-chain accessible surface upon formation of the dimer interface between V_L and V_H. Average relative reduction of the side-chain accessible surface of each residue in the Fv fragment compared to its accessible surface in the isolated V_L or V_H domain (white, 0 % reduction; yellow, 0-20 %; yellow-orange, 20-40 %; orange, 40-60 %; red-orange, 60-80 %; red, 80-100 %). (h) Reduction of the side-chain accessible surface upon formation of the interface between V_L and C_L or between V_H and C_H. Average relative reduction of the side-chain accessible surface of each residue in the Fab fragment compared to its accessible surface in the Fv fragment (white, 0 % reduction; yellow, 0-20 %; yellow-orange, 20-40 %; orange, 40-60 %; red-orange, 60-80 %; red, 80-100 %).

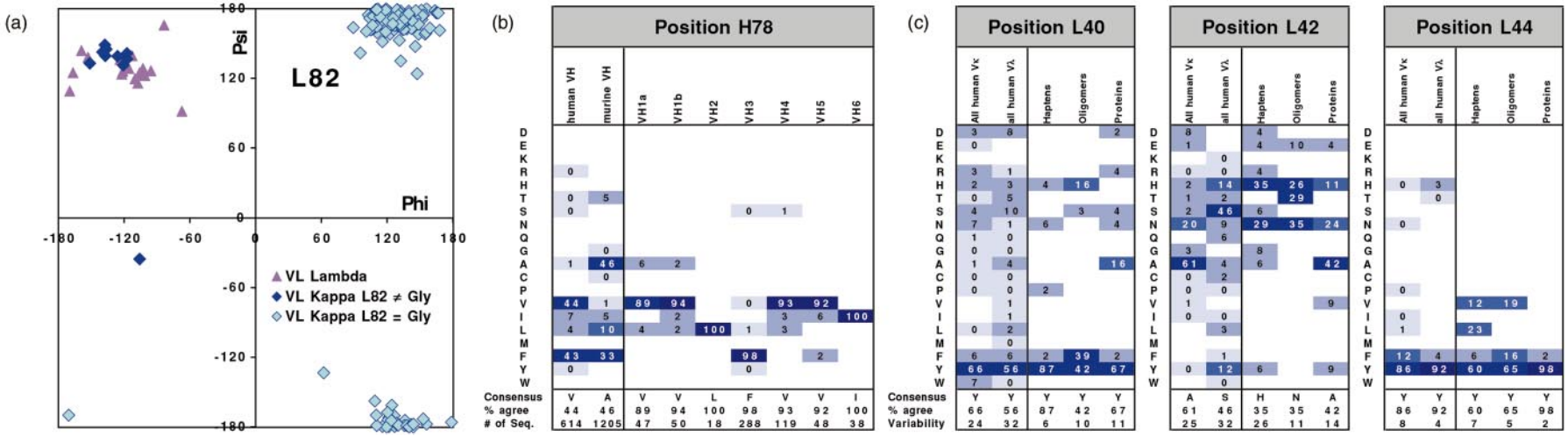

Figure 5. Examples of analyses greatly facilitated by the common residue numbering scheme. (a) Ramachandran plot of torsion angles V_L lambda and kappa residue L82. Magenta triangles, V_λ; blue diamonds, V_κ with a Gly residue in position L82; dark blue diamonds, V_κ with a non-Gly residue in position L82 (usually Arg). Such residue-by-residue correlation between sequence and structural features can be very valuable in avoiding errors in homology modeling and assessing the effects of point mutations acquired as somatic point mutations or during *in vitro* evolution on structure and function. (b) Position-dependent amino acid composition of human and murine antibody variable domain sequences collected in the Kabat database (complete sequences only). The rearranged human sequences were sorted according to the major germline subtypes to analyze the proportional representation, consensus sequence and sequence variability of the different classes as a base for the construction of synthetic consensus frameworks.[17] (Sequences in the Kabat database were considered to be rearranged if the sequence information extended beyond CDR 3 into FR 4; shorter sequences were omitted from the analysis). Fields containing values of exactly zero were colored white and the number was supressed if the amino acid was not observed in any sequence. Those containing higher values were color-coded from light blue (values 0 %-1 %) to very dark blue (90 %-100 %), and the amino acids were sorted according to their properties (charged, uncharged hydrophilic, aliphatic, aromatic side chains) to facilitate the perception of pattern of position-dependent residue properties and sequence conservation for each position in the domain. (c) Position-dependent amino acid composition of structures represented in the PDB database. The sequences were sorted according to the type and size of the antigen (hapten, oligomer (peptide or oligosaccharide) and protein), in order to analyze the correlation between sequence, antibody structure and type of antigen.

numbering scheme, module PROFILEGAP) and re-imported. All PDB files in the workbook can then be renumbered taking into account the gap positions in the sequence alignment and the residue labels indicated in the header line of the alignment. The renumbered coordinates can be re-exported as PDB files. Other macros allow color sequence alignment according to properties intrinsic (e.g. residue type, hydropathy, similarity to a reference sequence) or extrinsic (according to data provided in a separate table) to the information contained in the sequence alignment and to calculate consensus sequences and residue statistics.

Acknowledgements

This work was supported by a grant from the Schweizerische Nationalfonds 3100-046624.

References

1. Frippiat, J. P., Williams, S. C., Tomlinson, I. M., Cook, G. P., Cherif, D., Le Paslier, D., Collins, J. E., Dunham, I., Winter, G. & Lefranc, M. P. (1995). Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11. 2. *Hum. Mol. Genet.* **4**, 983-991.
2. Barbie, V. & Lefranc, M. P. (1998). The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Expt. Clin. Immunogenet.* **15**, 171-183.
3. Huston, J. S., Tai, M. S., McCartney, J., Keck, P. & Oppermann, H. (1993). Antigen recognition and targeted delivery by the single-chain Fv. *Cell. Biophys.* **22**, 189-224.
4. Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S. M., Lee, T., Pope, S. H., Riordan, G. S. & Whitlow, M. (1988). Single-chain antigen-binding proteins. *Science,* **242**, 423-426.
5. Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M. S., Novotny, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E. & Crea, R. *et al*. (1988). Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Escherichia coli*. *Proc. Natl Acad. Sci. USA,* **85**, 5879-5883.
6. Glockshuber, R., Malia, M., Pfitzinger, I. & Plückthun, A. (1990). A comparison of strategies to stabilize immunoglobulin Fv-fragments. *Biochemistry,* **29**, 1362-1367.
7. Winter, G., Griffiths, A. D., Hawkins, R. E. & Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433-455.
8. Vaughan, T. J., Williams, A. J., Pritchard, K., Osbourn, J. K., Pope, A. R., Earnshaw, J. C., McCafferty, J., Hodits, R. A., Wilton, J. & Johnson, K. S. (1996). Human antibodies with sub-nanomolar affinities isolated from a large non-immunized phage display library. *Nature Biotechnol.* **14**, 309-314.
9. Pantoliano, M. W., Bird, R. E., Johnson, S., Asel, E. D., Dodd, S. W., Wood, J. F. & Hardman, K. D. (1991). Conformational stability, folding, and ligand-binding affinity of single-chain Fv immunoglobulin fragments expressed in *Escherichia coli*. *Biochemistry,* **30**, 10117-10125.
10. Young, N. M., MacKenzie, C. R., Narang, S. A., Oomen, R. P. & Baenziger, J. E. (1995). Thermal stabilization of a single-chain Fv antibody fragment by introduction of a disulphide bond. *FEBS Letters,* **377**, 135-139.
11. Nieba, L., Honegger, A., Krebber, C. & Plückthun, A. (1997). Disrupting the hydrophobic patches at the antibody variable/constant domain interface. *Protein Eng.* **10**, 435-444.
12. Jung, S. & Plückthun, A. (1997). Improving *in vivo* folding and stability of a single-chain Fv antibody fragment by loop grafting. *Protein Eng.* **10**, 959-966.
13. Hanes, J., Jermutus, L., Weber-Bornhauser, S., Bosshard, H. R. & Plückthun, A. (1998). Ribosome display efficiently selects and evolves high-affinity antibodies *in vitro* from immune libraries. *Proc. Natl Acad. Sci. USA,* **95**, 14130-14135.
14. Jung, S., Honegger, A. & Plückthun, A. (1999). Selection for improved protein stability by phage display. *J. Mol. Biol.* **294**, 163-180.
15. Lefranc, M. P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbie, V. & Chaume, D. (1999). IMGT, the international ImMunoGeneTics database. *Nucl. Acids Res.* **27**, 209-212.
16. Wörn, A. & Plückthun, A. (2001). Stability engineering of antibody single-chain Fv fragments. *J. Mol. Biol.* **305**, 989-1010.
17. Knappik, A., Ge, L., Honegger, A., Pack, P., Fischer, M., Wellnhofer, G., Hoess, A., Wölle, J., Plückthun, A. & Virnekäs, B. (2000). Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol.* **296**, 57-86.
18. Proba, K., Wörn, A., Honegger, A. & Plückthun, A. (1998). Antibody scFv fragments without disulfide bonds made by molecular evolution. *J. Mol. Biol.* **275**, 245-253.
19. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesmann, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, 5th edit., NIH Publication no. 91-3242 U.S. Department of Health and Human Services.
20. Chothia, C. & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901-917.
21. Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D. & Tulip, W. R., *et al*. (1989). Conformations of immunoglobulin hypervariable regions. *Nature,* **342**, 877-883.
22. Al-Lazikani, B., Lesk, A. M. & Chothia, C. (1997). Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **273**, 927-948.
23. Gelfand, I., Kister, A., Kulikowski, C. & Stoyanov, O. (1998). Algorithmic determination of core positions in the V_L and V_H domains of immunoglobulin molecules. *J. Comput. Biol.* **5**, 467-477.
24. Gelfand, I., Kister, A., Kulikowski, C. & Stoyanov, O. (1998). Geometric invariant core for the V_L and V_H domains of immunoglobulin molecules. *Protein Eng.* **11**, 1015-1025.
25. Gelfand, I. M., Kister, A. E. & Leshchiner, D. (1996). The invariant system of coordinates of antibody molecules: prediction of the ''standard'' C alpha

framework of V_L and V_H domains. *Proc. Natl Acad. Sci. USA,* **93**, 3675-3678.
26. Ruiz, M., Giudicelli, V., Ginestoux, C., Stoehr, P., Robinson, J., Bodmer, J., Marsh, S. G., Bontrop, R., Lemaitre, M., Lefranc, G., Chaume, D. & Lefranc, M. P. (2000). IMGT, the international ImMunoGeneTics database. *Nucl. Acids Res.* **28**, 219-221.
27. Lefranc, M. P. (1997). Unique database numbering system for immunogenetic analysis. *Immunol. Today,* **18**, 509.
28. De Wildt, R. M., van Venrooij, W. J., Winter, G., Hoet, R. M. & Tomlinson, I. M. (1999). Somatic insertions and deletions shape the human antibody repertoire. *J. Mol. Biol.* **285**, 701-710.
29. Saini, S. S., Allore, B., Jacobs, R. M. & Kaushik, A. (1999). Exceptionally long CDR3H region with multiple cysteine residues in functional bovine IgM antibodies. *Eur. J. Immunol.* **29**, 2420-2426.
30. Spada, S., Honegger, A. & Plückthun, A. (1998). Reproducing the natural evolution of protein structural features with the selectively infective phage (SIP) technology. The kink in the first strand of antibody kappa domains. *J. Mol. Biol.* **283**, 395-407.
31. Li, H., Lebedeva, M. I., Llera, A. S., Fields, B. A., Brenner, M. B. & Mariuzza, R. A. (1998). Structure of the Vdelta domain of a human gamma-delta T-cell antigen receptor. *Nature,* **391**, 502-506.
32. Langedijk, A. C., Honegger, A., Maat, J., Planta, R. J., van Schaik, R. C. & Plückthun, A. (1998). The nature of antibody heavy chain residue H6 strongly influences the stability of a V_H domain lacking the disulfide bridge. *J. Mol. Biol.* **283**, 95-110.
33. Jung, S., Spinelli, S., Schimmele, B., Honegger, A., Pugliese, L., Cambillau, C. & Plückthun, A. (2001). The importance of framework residues H6, H7 and H10 in antibody heavy chains: experimental evidence for a new structural subclassification of antibody V_H domain. *J. Mol. Biol.* **309**, 701-716.
34. Honegger, A. & Plückthun, A. (2001). The influence of the buried glutamine or glutamate residue in position 6 on the structure of immunoglobin variable domains. *J. Mol. Biol.* **309**, 687-699.
35. Padlan, E. A., Abergel, C. & Tipper, J. P. (1995). Identification of specificity-determining residues in antibodies. *FASEB J.* **9**, 133-139.
36. Stanfield, R. L., Cabezas, E., Satterthwait, A. C., Stura, E. A., Profy, A. T. & Wilson, I. A. (1999). Dual conformations for the HIV-1 Gp120 V3 loop in complexes with different neutralizing Fabs. *Struct. Fold. Des.* **7**, 131-142.
37. Zdanov, A., Li, Y., Bundle, D. R., Deng, S. J., MacKenzie, C. R., Narang, S. A., Young, N. M. & Cygler, M. (1994). Structure of a single-chain antibody variable domain (Fv) fragment complexed with a carbohydrate antigen at 1.7-Å resolution. *Proc. Natl Acad. Sci. USA,* **91**, 6423-6427.
38. Kratzin, H. D., Palm, W., Stangel, M., Schmidt, W. E., Friedrich, J. & Hilschmann, N. (1989). Die Primaerstruktur des kristallisierbaren monoklonalen Immunoglobulins IgG1 Kol. II. Aminosaeuresequenz der L-Kette, Lambda-Typ, Subgruppe I. *Biol. Chem. Hoppe-Seyler,* **370**, 263-272.
39. Strong, R. K., Campbell, R., Rose, D. R., Petsko, G. A., Sharon, J. & Margolies, M. N. (1991). Three-dimensional structure of murine anti-p-azophenylarsonate Fab 36-71.1. X-ray crystallography, site-directed mutagenesis, and modeling of the complex with hapten. *Biochemistry,* **30**, 3739-3748.
40. Dall'Acqua, W., Goldman, E. R., Lin, W., Teng, C., Tsuchiya, D., Li, H., Ysern, X., Braden, B. C., Li, Y., Smith-Gill, S. J. & Mariuzza, R. A. (1998). A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry,* **37**, 7981-7991.
41. Whitlow, M., Howard, A. J., Wood, J. F., Voss, E. W., Jr & Hardman, K. D. (1995). 1.85 Å structure of anti-fluorescein 4-4-20 Fab. *Protein Eng.* **8**, 749-761.
42. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. (1992). Structural evidence for induced fit as a mechanism for antigen-antibody recognition. *Science,* **255**, 959-965.
43. Gruber, K., Zhou, B., Houk, K. N., Lerner, R. A., Shevlin, C. G. & Wilson, I. A. (1999). Structural basis for antibody catalysis of a disfavored ring closure reaction. *Biochemistry,* **38**, 7062-7074.
44. Navia, M. A., Segal, D. M., Padlan, E. A., Davies, D. R., Rao, N., Rudikoff, S. & Potter, M. (1979). Crystal structure of galactan-binding mouse immunoglobulin J539 Fab at 4.5 Å resolution. *Proc. Natl Acad. Sci. USA,* **76**, 4071-4074.
45. Simon, T. & Rajewsky, K. (1992). A functional antibody mutant with an insertion in the framework region 3 loop of the V_H domain: implications for antibody engineering. *Protein Eng.* **5**, 229-234.
46. Pokkuluri, P. R., Bouthillier, F., Li, Y., Kuderova, A., Lee, J. & Cygler, M. (1994). Preparation, characterization and crystallization of an antibody Fab fragment that recognizes RNA. Crystal structures of native Fab and three Fab- mononucleotide complexes. *J. Mol. Biol.* **243**, 283-297.
47. Ding, J., Das, K., Yu, H., Sarafianos, S. G., Clark, A. D., Jr, Jacobo-Molina, A., Tantillo, C., Hughes, S. H. & Arnold, E. (1998). Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA and an Antibody Fab fragment at 2.8 Ångstrom resolution. *J. Mol. Biol.* **284**, 1095-1111.
48. Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. & Wiley, D. C. (1996). Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature,* **384**, 134-141.
49. Plaksin, D., Chacko, S., McPhie, P., Bax, A., Padlan, E. & Margulies, D. (1996). A T cell receptor V alpha domain expressed in bacteria: does it dimerize in solution? *J. Exp. Med.* **184**, 1251-1258.
50. Ding, Y. H., Smith, K. J., Garboczi, D. N., Utz, U., Biddison, W. E. & Wiley, D. C. (1998). Two human T Cell receptors bind in a similar diagonal mode to the HLA-A2/Tax peptide complex using different TCR amino acids. *Immunity,* **8**, 403-411.
51. Housset, D., Mazza, G., Gregoire, C., Piras, C., Malissen, B. & Fontecilla-Caps, J. C. (1997). The three-dimensional structure of a T-cell antigen receptor V alpha/V beta heterodimer reveals a novel arrangement of the V beta domain. *EMBO J.* **16**, 4205-4216.
52. Wang, J. H., Lim, K., Smolyar, A., Teng, M. K., Liu, J. H., Tse, A. G. D., Liu, J., Hussey, R. E., Chishti, Y., Thomson, C. T., Sweet, R. M., Nathenson, S. G., Chang, H.-C., Sacchettini, J. C. & Reinherz, E. L. (1998). Atomic structure of an alpha/beta T-cell receptor (TCR) heterodimer in complex with an anti-TCR Fab fragment derived from a mitogenic antibody. *EMBO J.* **17**, 10-26.
53. Garcia, K. C., Degano, M., Stanfield, R. L., Brunmark, A., Jackson, M. R., Peterson, P. A., Teyton, L. & Wilson, I. A. (1996). An alpha/beta T

cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science,* **274**, 209-219.
54. Gigant, B., Charbonnier, J. B., Eshhar, Z., Green, B. S. & Knossow, M. (1997). X-ray structures of a hydrolytic antibody and of complexes elucidate catalytic pathway from substrate binding and transition state stabilization through water attack and product release. *Proc. Natl Acad. Sci. USA,* **94**, 7857-7861.
55. Bundle, D. R., Baumann, H., Brisson, J. R., Gagne, S. M., Zdanov, A. & Cygler, M. (1994). Solution structure of a trisaccharide-antibody complex: comparison of NMR measurements with a crystal structure. *Biochemistry*, **33**, 5183-5192.

Edited by I. Wilson

(Received 4 December 2000; received in revised form 28 March 2001; accepted 29 March 2001)