Geometric Invariant Core for the $C_L$ and $C_{H1}$ Domains of Immunoglobulin Molecules

OGNYAN STOYANOV, 1 ALEXANDER KISTER, 1 ISRAEL GELFAND, 1
CASIMIR KULIKOWSKI, 2 and CYRUS CHOTHIA 3

ABSTRACT

A previously developed algorithmic method for identifying a geometric invariant of protein structures, termed geometrical core, is extended to the $C_L$ and $C_{H1}$ domains of immunoglobulin molecules. The method uses the matrix of $C_\alpha-C_\alpha$ distances and does not require the usual superposition of structures. The result of applying the algorithm to 53 Immunoglobulin structures led to the identification of two geometrical core sets of $C_\alpha$ atom positions for the $C_L$ and $C_{H1}$ domains.

Key words: atom coordinate prediction, immunoglobulin geometry, preferred coordinate system, protein core.

1. INTRODUCTION

This paper extends the study of geometric and sequence invariants of the immunoglobulin (Ig) family previously reported by us (Gelfand and Kister, 1995; Gelfand et al., 1996, 1998b) to analytically extract a geometrical core for the constant domain of this family. We define a geometrical core as a set of residues whose $C_\alpha$ atoms occupy the same relative positions in space, which is consistent with other prior definitions (Chothia and Lesk, 1986, 1987; Altman and Gerstein, 1994, 1995).

Two main approaches are usually used in discovering a common core. In the first, all structures are described in one preferred system of coordinates. Several methods based mainly on the superposition of structures have been suggested. We have proposed an analytical method for introducing an invariant system of coordinates based on secondary structure alignment and inherent geometrical properties of the variable domains common to all members of the Ig family, thereby avoiding the need for a graphical superposition of structures (Gelfand et al., 1996, 1998b).

The main idea of the second approach is to identify a geometrical core as a subset of all aminoacid residues such that the distances between pairs $C_\alpha(I)-C_\alpha(J)$ of $C_\alpha$ atoms in the $I^{th}$ and $J^{th}$ secondary structure positions are (almost) equal in all members of the family. We developed an algorithm for finding such a “minimal dispersion” subset of residues and applied it to identify a geometrical core for the VL and VH domains (Gelfand et al., 1998a, 1998b). The advantage of this approach is that it does not require

1Department of Mathematics, Rutgers University, New Brunswick, NJ 08903.
2Department of Computer Science, Rutgers University, New Brunswick, NJ 08903.
3MRC Laboratory of Molecular Biology, Cambridge, CB2 2QH, U.K.
a superposition procedure and does not depend on any system of coordinates. For the variable domains of IgS, both methods produced almost identical subsets for the geometrical cores.

In the present work, we extend our study of the immunoglobulin family to the $C_L$ and $C_{H1}$ domains, applying the minimal dispersion algorithm to identify a common geometrical core for each of these two types of domains. We then compare the residue conservation at core positions between the variable and constant domains.

2. MATERIALS AND METHODS

As described previously (Gelfand et al., 1998a, 1998b), the method has two major components: (1) secondary structure alignment and (2) comparison of distances between $C_\alpha$ atoms at pairs of identical secondary structure positions across the Ig family.

2.1. Secondary structure alignment

Any method of sequence and structural classification of proteins presupposes a sequence and structural alignment. In this work we use secondary structure alignment based on both sequence and structural information. The sequences were divided into fragments which approximately correspond to strands and loops. Alignment of these fragments gives us structural units of a protein—which we denote as words—with a uniquely defined indexing of residues to the positions within words (Gelfand and Kister, 1995).

In our previous research, 53 Ig structures were selected to include no identical sequences and to contain only data at the highest resolution (Gelfand et al., 1998a, 1998b). Using the same 53 Ig structures reported for the above study, we fragmented the light constant domain of the structures into 16 fragments, or words, and the heavy constant domain into 15 words. Using the standard notation for immunoglobulin secondary structures, these are: A, AB, B, BC, C, $C'$, $C'$D, D, $D'$, $D'$E, E, EF, F, FG, G, and $G'$ for the heavy constant domain, and all the above with the exception of $C'$ for the light constant domain (which also results in the subsequent fragment being labeled CD). Once a sequence has been fragmented into these words, it is possible to assign a position to every residue. Tables 1 and 2 show a sample of the immunoglobulin sequences and their secondary structure alignment.

2.2. Comparison of distances between $C_\alpha$ atoms

The core finding algorithm has been defined by Gelfand et al. (1998b) and can be summarized as follows:

1. We start by aligning the Ig structures according to the secondary structure assignments described in Section 2.1 (Gelfand and Kister, 1995). Further, we selected 53 Ig structures with $C_L$ (that have a $\nu_L$ $\kappa$ chain) and $C_{H1}$ domains with the highest resolution and nonidentical sequences. All $C_L$ (with $\nu_L$ $\kappa$ chains) domains were aligned, and the set of positions that were common to all of them (which we dubbed “candidate” positions) was selected. (The set of (100) common $C_L$ $\kappa$ positions includes: A1–A9, AB1–AB10, B1–B11, BC1–BC2, C1–C9, $C'$–$C'$4, $C'$D1–$C'$D2, D1–D6, D1′–D3, D'E1–D'E5, E1–E10, EF1–EF10, F1–F7, FG1–FG2, G1–G4, G'1–G'6.) This was the starting set of positions of the $C_L$ $\kappa$ domain from which a geometrical core set was later obtained. Similarly, all $C_{H1}$ domains were aligned and a set of positions common to all of them (the set of (94) common $C_{H1}$ positions includes: A1–A9, AB1–AB9, B1–B11, BC1–BC2, C1–C7, CD1–CD6, D1–D7, D1′–D'3, D'E1–D'E2, E1–E10, EF1–EF9, F1–F7, FG1–FG3, G1–G3, G'1–G'5) was taken as a starting point for obtaining a set of geometrical core positions for the $C_{H1}$ domain.

2. The set of fully aligned candidate positions is pruned to yield a subset of positions having minimal dispersion of distances between their backbone $C_\alpha$ atoms. These can be construed as the most stable

---

1These are: lacy, lbaf, lbbd, lbbq, lcbw, lcs, ldby, leap, lfai, lfdl, lfgy, lffr, lftr, lfg, lflc, lgb, lhi, lgbq, liff, lgg, lqm, lkkf, lje1, lje1k, lkn0, lma, lmc, lmc, lmd, lbv, lmb, lpg, lpg, lte, lvia, lvge, 2cgr, 2dib, 2f19, 2fby, 2fgw, 2ff, 2gff, 3hfl, 3hfm, 6fab, 1igj, 1ncc, 4fab.
Table 1. Sample of Ig Sequences and the Alignment of Their $V_L\kappa$ and $C_L$ Domains. The Switch between the Two Domains is Indicated by the Segment SW

<table>
<thead>
<tr>
<th>Alignment of the $V_L\kappa$ domains:</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA $A$ $A\prime$ $A\prime\prime$ $AB$ $BC$ $CB$ $C$ $C\prime$ $C\prime\prime$ $C\prime\prime\prime$ $C\prime\prime\prime\prime$</td>
</tr>
<tr>
<td>DIV-MTQ-SPAS-LVLS-RATISCR-SES-VDSYKSF-----MHWVQQ--KPGQP---KVLVY--1ACY</td>
</tr>
<tr>
<td>QIV-LTQ-SPAI-MSAS-PGE-KVMTMCSA-SSS-VYY--------MYWQQ--KPGSS--RLLYI--1BAF</td>
</tr>
<tr>
<td>DIV-MTQ-SPSS-LLVT-TGE-KVMTMCSK-SQS-LNSRTQKNY----LTVQQ--KPGQS--KLLLYI--1BBD</td>
</tr>
<tr>
<td>DIV-LTQ-SPAI-MSAS-PGE-KVMTMCSA-SSS-VNY--------MYWQQ--KSGTSP--KRWYI--1BQL</td>
</tr>
<tr>
<td>DVV-MTQ-PLPS-LLGD-QASISCRS-SQS-LVHSNGNTY----LHWLQ--KPGQSP--KLLLYI--1CBV</td>
</tr>
<tr>
<td>ELV-MTQ-SPIL-PLVG-LGD-QASISCRP-SQS-LVHSNGNTY----LHWLQ--KPGQSP--KLLLYI--1CGS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alignment of the $C_L$ domains:</th>
</tr>
</thead>
<tbody>
<tr>
<td>C $C\prime$ $C\prime\prime$ $C\prime\prime\prime$ $D$ $DE$ $E$ $EF$</td>
</tr>
<tr>
<td>IAS-----NLE---SGVPA----RFSGGS--RT-----DFTLTID----PVEADDA--1ACY</td>
</tr>
<tr>
<td>DTS-----NLA--SGPVV--RFSGGS--GT-----SYSLTI-------RMEADA--1BAF</td>
</tr>
<tr>
<td>WAS-----TRE---SGVPD--RFSGGS--GT-----DFTLSIS----GQVDEDL--1BBD</td>
</tr>
<tr>
<td>DTS-----KLA--SGPVE--RFSGGS--GT-----SYSLTI-------SMEADA--1BQL</td>
</tr>
<tr>
<td>KVS-----NRF---SGVPD--RFSGGS--GT-----DFTLKIS----RVEADEL--1CBV</td>
</tr>
<tr>
<td>RVS-----NRF---SGVPD--RFSGGS--GT-----AFTLKIS----RVEADEL--1CGS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alignment of the $C_{H1}$ domains:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A $AB$ $B$ $BC$ $C$ $C\prime$ $C\prime\prime$ $C\prime\prime\prime$ $D$ $D\prime$</td>
</tr>
<tr>
<td>AAPTVISIFP--PSEQLTSGG--ASVVFNLNNFY--PK--DINVWKID--GSER--QN--GVLNSW--TDQ--1ACY</td>
</tr>
<tr>
<td>AAPTVISIFP--PSEQLTSGG--ASVVFNLNNFY--PK--DINVWKID--GSER--QN--GVLNSW--TDQ--1BAF</td>
</tr>
<tr>
<td>AAPTVISIFP--PSEQLTSGG--ASVVFNLNNFY--PK--DINVWKID--GSER--QN--GVLNSW--TDQ--1BBD</td>
</tr>
<tr>
<td>AAPTVISIFP--PSEQLTSGG--ASVVFNLNNFY--PK--DINVWKID--GSER--QN--GVLNSW--TDQ--1CBV</td>
</tr>
<tr>
<td>AAPTVISIFP--PSEQLTSGG--ASVVFNLNNFY--PK--DINVWKID--GSER--QN--GVLNSW--TDQ--1CGS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alignment of the $C_{L}$ domains:</th>
</tr>
</thead>
<tbody>
<tr>
<td>D $E$ $EF$ $F$ $FG$ $G$ $G\prime$</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---R------1ACY</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---RNEC--1BAF</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---RNEC--1BBD</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---RNEC--1BQL</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---RNEC--1CBV</td>
</tr>
<tr>
<td>DSKDS---TYSMSSSSL--TKDEYERHNS--YTECAH--KT--STSP----IVKSNF---RNEC--1CGS</td>
</tr>
</tbody>
</table>

representative positions covering different parts of a fold. We used a threshold $T_1 = 0.25 \text{Å}$ on the dispersion of distances for the $C_{H1}$ domain ($T_1 = 0.20 \text{Å}$ for the $C_L$ domain) and checked all positions for self-consistency, as described below, yielding the subset denoted as FOLDREP.

(3) We next expand the FOLDREP subset to include all positions with low distance dispersion with respect to the elements of FOLDREP, as defined by another threshold ($T_2 \leq 0.30 \text{Å}$ for $C_{H1}$ and $T_2 \leq 0.25 \text{Å}$ for $C_L$ domains) and, after again checking for internal consistency, denote it as the CENTRAL subset.

(4) Next, we further expand this subset to include all positions with a slightly larger dispersion of their distances (less than or equal to $T_3 = 0.40 \text{Å}$ for the $C_{H1}$ and $T_3 = 0.35 \text{Å}$ for the $C_L$ domain positions) to
Table 2. Sample of Ig Sequences and the Alignment of Their $V_H$ and $C_H1$ Domains

Alignment of the $V_H$ domains:

<table>
<thead>
<tr>
<th>OA</th>
<th>AA'</th>
<th>A'</th>
<th>AB</th>
<th>B</th>
<th>BC</th>
<th>CB</th>
<th>C</th>
<th>CC'</th>
<th>C'</th>
</tr>
</thead>
<tbody>
<tr>
<td>QVK-LQE-SGP-AVIK-PSQ---SLSLTCIV---SGFS-ITRTNYC---WHWIRG--APGKGL---EWMGRI--1ACY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DVG-LQE-SGP-GLVK--PSQ---SQSSLCTIV---TYS-ITSDYA---WNNWIRG--FPGNKL--EWGMYG--1BAF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVQ-LQQ-SGA-ELVR-PGA-SVKLSCCTT---SGFN--1KDIY---IHWKQR--REPQGL--EWIGRL--1BBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>?VQ-LQQ-SGA-ELMK-PA--SVKISKCA---SGYT-FSDYW---IEWKQR--RPQHGL--EWIGEI--1BQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EVQ-PVE-TGG-GLLQ--PKG---SLKLSCCA---SGFS-FNTNA---MNWIRG--APGKGL--EWVAR--1CBV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RVQ-LE-LGA-ELMK-PA--SVQISCKA---TGYT-FSEYW---IEWKE--RPQHGL--EWIGEI--1CGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alignment of the $C_H1$ domains:

<table>
<thead>
<tr>
<th>F</th>
<th>FG</th>
<th>G</th>
<th>SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMYYCSR---ENHYETYFD---------WQQGTTTVTS----SAKT----1ACY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATYFCAR---GWPLA---------WQQGTTWSYS----EAKT----1BAF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVYYCDG---YSSYYDM---------WQPGRSTTVS----SAKT----1BD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GVYYCLH---GYYDFD---------WQQGTVTLS----SAKT----1QL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMYYCVR---DQTGTAFA---------WQQGTVTVS----AATK----1DV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVYYCTR---GYSSMD---------WQQGTVTVS----AATK----1GS</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

elements of PRE_CORE and, after filtering for self-consistency (with an internal threshold on dispersion of distances of 0.35Å for $C_H1$ and 0.40Å for the $C_L$ domain), take the resulting positions to be members of the geometrical CORE subset. Each of the above subsets defines different levels of conservation of the positions for the particular domain. The thresholds are different for each of the domain types, determined empirically based on the distribution of distances and their dispersions. Here is how it is done in more detail. We take the set of 100 positions, denoted CANDIDATES, from the $C_L$ domain common to all structures under consideration. As mentioned above, we did computations...
using 53 Igs with $V_L \kappa$ chains only. For each pair $(I, J)$, (e.g., A1–G3) of positions from this set we compute the distance $D_i(I, J)$ between these two positions for each of the $i = 1, 2, \ldots, N(= 53)$ Igs. Then we take the average

$$
AVER(I, J) = \frac{1}{N} \sum_{i=1}^{N} D_i(I, J)
$$

over these $N = 53$ Igs and compute the dispersion

$$
DISP(I, J) = \left[ \frac{1}{N} \sum_{i=1}^{N} (D_i(I, J) - AVER(I, J))^2 \right]^{1/2}.
$$

Thus, with each pair $(I, J)$ of positions from our CANDIDATES list we associate a number $DISP(I, J)$. The set of all these numbers form a symmetric matrix $DISP$ with zeros along the main diagonal. Figures 1 and 2 show the density distribution of number of pairs versus their dispersion for the $C_L$ and $C_H$ domains respectively.

To determine a stable set of core positions for the $C_L$ domain we need only know the matrix $DISP$ and the following natural characteristic one can ascribe to each of the positions of the starting CANDIDATES set. Namely, for each position $I$ we define

$$
AVERDISP(I) := \frac{1}{M} \sum_{J=1}^{M} DISP(J, I),
$$

FIG. 1. Density distribution of number of pairs with respect to their dispersion for the CL $\kappa$ chains.
FIG. 2. Density distribution of number of pairs with respect to their dispersion for the CH1 chains.

where $M$ is the number of positions in CANDIDATES (the size of the matrix $DISP$). $AVERDISP(I)$ gives a rough characterization of the overall “poor-fitness” of the position $I$.

3. RESULTS: CORE EXTRACTION FOR IG CONSTANT DOMAINS

The set of geometric core positions is determined in the following algorithmic way.

- Using the matrix $DISP$, we obtained the set $FOLDREP$ of positions, each representing a distinct part of the protein fold, having low $AVERDISP$ and being self-consistent in the following sense. For any element $I$ of $FOLDREP$ we must have

$$AVERDISP(I) = \frac{1}{L} \sum_{J \in FOLDREP} DISP(J, I) \leq 0.20 \text{Å}.$$ 

Here $L$ is the number of elements in $FOLDREP$. One subset of CANDIDATES satisfying these conditions turned out to be $FOLDREP = \{A5 (Val), B5 (Cys), C6 (Trp), D5 (Ser), E3 (Ser), F2 (Thr, Ala)\}$. (Replacing $A5$ with $A4$ (Ser, Thr) and $B5$ with $B3$ (Val) also yields a set satisfying the above conditions.) Note that $F3$ (Cys) is what one would expect more likely to enter the above set, for it is a part of the Cys–Cys disulfide bridge. However, our method indicated that including $F3$ in $FOLDREP$ destroys self-consistency with the above constraints. We can interpret this result by inferring that the proximity of $F2$ to the bridge makes its $C_\alpha$ atom even more structurally stable than that of $F3$. One could also attribute to this the stability of $B3$, in view of the allowable substitution with $B5$. The submatrix of dispersions for these “super conservative” positions is
GEOMETRIC INVARIANT CORE

\[
\begin{array}{cccccc}
A5 & B5 & C6 & D5 & E3 & F2 \\
A5 & 0.00 & 0.16 & 0.24 & 0.19 & 0.20 & 0.19 \\
B5 & 0.16 & 0.00 & 0.23 & 0.20 & 0.17 & 0.19 \\
C6 & 0.24 & 0.23 & 0.00 & 0.19 & 0.19 & 0.19 \\
D5 & 0.19 & 0.20 & 0.19 & 0.00 & 0.18 & 0.17 \\
E3 & 0.20 & 0.17 & 0.19 & 0.18 & 0.00 & 0.19 \\
F2 & 0.19 & 0.19 & 0.19 & 0.17 & 0.19 & 0.00 \\
Av & 0.17 & 0.16 & 0.17 & 0.16 & 0.15 & 0.16.
\end{array}
\]

In the last row is shown the average of the dispersions over the members of FOLDREP which are all \( T_1 = 0.20\text{Å}. \)

- Next we take the \( 6 \times 100 \) submatrix \( DISP_{TO\_FOLDREP} \) of \( DISP \)

\[
\begin{array}{cccccccc}
A1 & \ldots & AB3 & \ldots & B7 & \ldots & F6 & \ldots & G'5 & G'6 \\
A5 & 0.22 & \ldots & 0.26 & \ldots & 0.19 & \ldots & 0.36 & \ldots & 0.30 & 0.30 \\
B5 & 0.20 & \ldots & 0.28 & \ldots & 0.15 & \ldots & 0.31 & \ldots & 0.30 & 0.29 \\
C6 & 0.23 & \ldots & 0.30 & \ldots & 0.27 & \ldots & 0.19 & \ldots & 0.28 & 0.30 \\
D5 & 0.27 & \ldots & 0.31 & \ldots & 0.14 & \ldots & 0.28 & \ldots & 0.30 & 0.31 \\
E3 & 0.17 & \ldots & 0.25 & \ldots & 0.15 & \ldots & 0.21 & \ldots & 0.28 & 0.27 \\
F2 & 0.25 & \ldots & 0.32 & \ldots & 0.24 & \ldots & 0.19 & \ldots & 0.26 & 0.34 \\
Av & 0.22 & \ldots & 0.29 & \ldots & 0.19 & \ldots & 0.26 & \ldots & 0.29 & 0.30.
\end{array}
\]

The last row presents the average dispersions relative to the 6 members of FOLDREP. The distribution of the 100 positions of CANDIDATES with respect to their average dispersions relative to the 6 members of FOLDREP is shown in Figure 3.

**FIG. 3.** Density distribution of positions in CANDIDATES with respect to the average dispersion relative to the members of FOLDREP for the CL \( \kappa \) domain.
The second level of conservation is determined as follows. Based on the data contained in the submatrix \( \text{DISP}_{\text{TO}} \), we select a subset \( \text{PRE}_{\text{CENTRAL}} \) of positions of \( \text{CANDIDATES} \) satisfying the following criterion: every position \( I \) in \( \text{PRE}_{\text{CENTRAL}} \) has \( \text{Av} \leq T_1 = 0.25 \text{Å} \), where \( \text{Av} \) is the average dispersion of the distances between the members of \( \text{FOLDREP} \) and all members of \( \text{CANDIDATES} \). There are 59 positions that satisfied this criterion: A1–A9, AB1, AB4–AB5, AB10, B1–B11, C2, C4–C7, D3–D6, D’1, D’3, D’E5, E1–E10, EF1–EF3, EF5, EF8, EF10, F1–F5, G’3–G’4.

The set \( \text{PRE}_{\text{CENTRAL}} \) is checked for self-consistency. Namely, we take the \( 59 \times 59 \) submatrix \( \text{DISP}_{\text{PRE}_{\text{CENTRAL}}} \) of \( \text{DISP} \), the 59 rows and 59 columns of which correspond to the elements of \( \text{PRE}_{\text{CENTRAL}} \).

Then all positions in \( \text{PRE}_{\text{CENTRAL}} \) with \( \text{Av} > T_1 = 0.25 \text{Å} \) are excluded from \( \text{PRE}_{\text{CENTRAL}} \) yielding a new self-consistent set \( \text{CENTRAL} \). The excluded positions are AB10, C2, C7, D3, EF3, EF8. Thus, after the requirement for self-consistency has been met, the set \( \text{CENTRAL} \) then consists of the following 53 positions: A1–A9, AB1, AB4–AB5, B1–B11, C4–C6, D4–D6, D’1, D’3, D’E5, E1–E10, EF1–EF2, EF5, EF10, F1–F5, G’3–G’4.

To identify the third level of conservation (which we termed the geometrical core) we take the \( 53 \times 100 \) submatrix of \( \text{DISP} \), the rows of which are the ones labeled by the positions of the set \( \text{CENTRAL} \) and the columns by all positions in \( \text{CANDIDATES} \).

A new set \( \text{PRE}_{\text{CORE}} \) of positions is selected the elements of which satisfy the requirement \( \text{Av} \leq T_2 = 0.35 \text{Å} \). It consists of 87 positions.

Finally, the set \( \text{PRE}_{\text{CORE}} \) is checked for self-consistency by taking the submatrix of \( \text{DISP} \) with rows and columns labeled by the members of the set \( \text{PRE}_{\text{CORE}} \):
and then excluding from \( \text{PRE\_CORE} \) all positions with \( A^v > T_2 = 0.35\text{Å} \). In this case only two positions, \( D'\text{E}2 \) and \( D'\text{E}4 \), were excluded. Thus, this yields a set of 85 positions: \( A1–A9, \text{AB}1–\text{AB}10, \text{B}1–\text{B}11, \text{BC}1–\text{BC}2, \text{C}1–\text{C}7, \text{D}2–\text{D}6, \text{D}'1–\text{D}'3, \text{D}’\text{E}5, \text{E}1–\text{E}10, \text{EF}1–\text{EF}10, \text{F}1–\text{F}7, \text{FG}1–\text{FG}2, \text{G}1–\text{G}4, \text{G}'1–\text{G}'6 \), which we term geometrical core, \( \text{CORE}_{T_L} \), for the \( C_L \) domain of the immunoglobulin family at variability threshold \( T_L = 0.35\text{Å} \).

For the \( C_{H1} \) domain, our method determined the initial set to be \( \text{FOLDREP} = \{ A5 \text{ (Val), B5 (Cys), C6 (Trp), D5 (Phe), E3 (Thr), F3 (Cys)} \} \), to which corresponds a dispersion submatrix:

\[
\begin{array}{cccccc}
A5 & B5 & C6 & D5 & E3 & F3 \\
A5 & 0.00 & 0.26 & 0.24 & 0.21 & 0.14 & 0.26 \\
B5 & 0.26 & 0.00 & 0.25 & 0.20 & 0.17 & 0.18 \\
C6 & 0.24 & 0.25 & 0.00 & 0.34 & 0.31 & 0.21 \\
D5 & 0.21 & 0.20 & 0.34 & 0.00 & 0.19 & 0.24 \\
E3 & 0.14 & 0.17 & 0.31 & 0.19 & 0.00 & 0.23 \\
F3 & 0.26 & 0.18 & 0.21 & 0.24 & 0.23 & 0.00 \\
\text{Av} & 0.18 & 0.18 & 0.22 & 0.20 & 0.17 & 0.10 \\
\end{array}
\]

All of these positions have \( A^v \mu T_1 = 0.25\text{Å} \). Further, analogously to how it was done for the \( C_L \) domain, we selected a subset \( \text{PRE\_CENTRAL} \) of positions having average dispersion (average taken over the number of positions in \( \text{FOLDREP} \) \( A^v \leq T_2 = 0.30\text{Å} \). After the self-consistency requirement was met, this yielded the set \( \text{CENTRAL} \), consisting of the following 51 positions: \( A1–A9, \text{B}1–\text{B}11, \text{C}4–\text{C}6, \text{D}'1–\text{D}'2, \text{E}1–\text{E}10, \text{EF}1, \text{F}1–\text{F}7 \). Similarly, a subset \( \text{PRE\_CORE} \) is further selected consisting of positions whose average dispersion satisfy \( A^v \leq T_3 = 0.40\text{Å} \) (the average being over the members of \( \text{CENTRAL} \)). Finally, the refinement of \( \text{PRE\_CORE} \) yields the geometrical core \( \text{CORE}_{T_H} \) for the \( C_{H1} \) domain at variability threshold \( T_H = 0.40\text{Å} \). This core consists of the following 59 positions: \( A1–A9, \text{AB}10, \text{B}1–\text{B}11, \text{BC}1–\text{BC}2, \text{C}1–\text{C}6, \text{D}1–\text{D}7, \text{D}'1–\text{D}'2, \text{E}1–\text{E}10, \text{EF}1–\text{EF}2, \text{F}1–\text{F}7, \text{FG}1, \text{FG}3 \).

**FIG. 4.** Density distribution of positions in \( \text{CANDIDATES} \) with respect to the average dispersion relative to the members of \( \text{FOLDREP} \) for the \( CH1 \) domain.
Here we observed the unexpected fact that the $C_{H1}$ domain shows greater variability than the $V_H$ domain previously reported by Gelfand et al. (1998b). This can be seen from the comparison of the corresponding density distributions of number of pairs versus the dispersion of the distance between each of them (taken across the family of 53 Ig structures) for the $C_{H1}$ and $V_H$ domains given in Figures 4 and 5 respectively. Similar differences can be seen when comparing the corresponding density distributions of positions in CANDIDATES with respect to the average dispersion relative to the members of FOLDREP for the $C_{H1}$ (Figure 4) and $V_H$ (Figure 6).

4. DISCUSSION

(1) The simple geometrical technique described above leads to the identification of two sets, $CORE_{T_L}$ and $CORE_{T_H}$, of $C_\alpha$ atom secondary structure positions for the variable $C_L$ $\kappa$ and $C_{H1}$ domains of the immunoglobulin family. It is based on the analysis of the X-ray crystallographic data for 53 immunoglobulin structures. The sets $CORE_{T_L}$ and $CORE_{T_H}$ consist of secondary structure positions with low structural variation. The positions of the $C_\alpha$ atoms in $CORE_{T_L}$ can be thought of as vertices of a skeleton that forms a geometrical “core” of the $C_L$ domain and that is the same across the Ig family. Analogously, the secondary structure positions from $CORE_{T_H}$ are vertices of a skeleton for the $C_{H1}$ domain that is conserved across the family.

(2) To determine invariant characteristics and to compare different individual structures we need to have the coordinates of atoms of each structure in a single coordinate system (Gelfand et al., 1998b). We transformed the X-ray crystallography coordinates of the $C_L$ and $C_{H1}$ core positions (as well as those of all heavy atoms) of all 53 Ig structures to the coordinate system introduced by Gelfand et al. (1998b), using the 1CGS as a reference structure.

(3) Just as in the case for the $V_L$ and $V_H$ in Gelfand et al. (1998b) we could have sought to construct “an axis of symmetry” between the $C_L$ and $C_{H1}$ domains. Such an axis may exist and this would provide
FIG. 6. Density distribution of positions in CANDIDATES with respect to the average dispersion relative to the members of FOLDREP for the VH domain.

a useful geometric characterization of the mutual orientation between the variable and constant domains. This matter needs further investigation.

(4) The classification of residue frequencies at positions in the geometrical core of the $C_L$ and $C_{H1}$ domains was also addressed. For the classification of residue conservation, we suggested dividing the positions into three kinds (Chothia et al., 1998). At the first kind of position, only one residue is found in almost all sequences (these kinds of positions are dubbed invariant residue (IR) positions). At the second kind of position, two or several chemically similar residues are found (dubbed similar residue (SR) positions). At the third kind, a wider range of residues is found and the conservation is best described in terms of the chemical nature of the residues that occur at the position. We refer to these positions as residue class (RC) positions.

The frequency of residues at the core positions of constant domains was calculated from sequences in the Kabat database (Chothia, Gelfand, and Kister, unpublished data). It includes the sequences with kappa and heavy chain constant domains from different species. About 130 heavy chains and 85 kappa chains were aligned, and each residue was assigned a double label that includes a name of a word and a position of the given residue within the word. It is remarkable that the extent of residue conservation at different positions in the constant domain is less than that in the variable domain. The inspection of residue frequencies in the constant domain shows five invariant positions in the chains of $C_L$ and $C_{H1}$ domains (Pro at A3, Trp at C4, Ser at E6, and Cys at B5 and F3 positions). Note that our analysis of residues in the variable domains revealed 9 positions that are occupied by a single specific residue in almost all 5500 sequences (Chothia et al., 1998).

Further inspection of the constant domain residue frequencies shows six SR positions (B3, B7, C2, E8, E10, F3) that are occupied by hydrophobic residues. This is in contrast to the eight hydrophobic positions found in the variable domain. A total of thirteen positions with invariant residues or positions that contain closely related residues were found in the constant domains, while twenty-six such conserved positions were found in the variable domain.
ACKNOWLEDGMENTS

The authors are grateful to Dr. O. Pitsyn for useful discussions. AK and OS would like to thank the Gabriella and Paul Rosenbaum Foundation for support. We wish to thank Mrs. M. Goldman for continuous encouragement. We would also like to thank SROA at Rutgers University for partial support.

REFERENCES


Address correspondence to:
Israel Gelfand
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903

E-mail: igelfand@math.rutgers.edu
This article has been cited by: