REVISED manuscript (PEW 32/95)
for *J. Mol. Biol.* (sent 25 Apr 1995)

Average Core Structures and Variability Measures for Protein Families: Application to the Immunoglobulins

Mark Gerstein [1†] & Russ B Altman [2†]

[1] Department of Structural Biology, Fairchild D109
and
[2] Section on Medical Informatics, MSOB X215
Stanford University, Stanford, CA 94305

† e-mail: mbg@hyper.stanford.edu, altman@camis.stanford.edu

MG: FAX (415) 725-8464; Phone 725-0754
RBA: FAX (415) 725-7944; Phone 725-3394

Keywords: alignment, structural; classification; protein structure, substructures; statistical analysis; structure motif; Immunoglobulin; protein cores; RMS

Running Title: Average Core Structures

Subject classification: Proteins

Abstract

A variety of methods are currently available for creating multiple alignments, and these can be used to define and characterize families of related proteins, such as the globins or the immunoglobulins. We have developed a method for using a multiple alignment to identify an average structural "core," a subset of atoms with low structural variation. We show how the means and variances of core-atom positions summarize the commonalities and differences within a family, making them particularly useful in compiling libraries of protein folds. We show further how it is possible to describe the rotation and translation relating two core structures, as in two domains of a multi-domain protein, in a consistent fashion in terms of a "mean" transformation and a deviation about this mean. Once determined, our average core structures (with their implicit measure of structural variation) allow us to define a measure of structural similarity more informative than the usual RMS deviation in atomic position, i.e. a "better RMS." Our average structures also permit straightforward comparisons between variation in structure and sequence at each position in a family.

We have applied our core finding methodology in detail to the immunoglobulin family. We find that the structural variability we observe *just within* the VL and VH domains anticipates the variability that others have observed throughout the whole immunoglobulin superfamily; that a core definition based on sequence conservation, somewhat surprisingly, does not agree with one based on structural similarity; and that the cores of the VL and VH domains vary about 5° in relative orientation across the known structures.

1. Introduction

The number of protein structures in the Protein Data Bank is now quite large and is rapidly increasing. A recent estimate puts the total number of chains within the databank at 3000, increasing by about one a day (Orengo, 1994). One way to deal with this huge amount of structural information is by grouping related proteins into families, such as such as the globins or the immunoglobulins (Levitt & Chothia, 1976; Chothia & Finkelstein, 1990; Richardson, 1981). Members of protein families have similar overall folds but differences in their detailed structure. The classification of the entire databank using protein families has recently been attempted by a number of groups (Johnson *et al*., 1990; Sander & Schneider, 1991; Murzin *et al*., 1994; Holm *et al*., 1993; Orengo *et al*., 1993; Pascarella & Argos, 1992; Orengo *et al.,* 1994), and it has been found that some protein families are quite large. The immunoglobulin family is a case in point. In the databank, there are currently over 25 structures of different antibody molecules, and when the whole immunoglobulin superfamily is considered (Bork *et al.,* 1994), there are at least 20 additional immunoglobulin-like structures, which include such disparate proteins as the enzyme myosin light chain kinase and the cell-surface receptor CD2.

Thus, because of both the great numbers of structures and of families, it has become desirable (even necessary) to summarize the common features within a family, whilst separating out the variable ones. One of the most basic commonalities shared by each member of a family is a set of atoms which occupy the same relative positions in space. Our focus here is in identifying these atoms, and then in characterizing them statistically. We show how to construct an average core structure for a protein family in such a way that the average is unbiased and the resulting structure has acceptable stereochemistry. This core structure can then be used to characterize the structural variability within a family, to define the average relative orientation of domains in multi-domain complexes, and to develop new measures of similarity between members of the same structural family. We illustrate our ideas here through application to the archetypal protein family: the all β-sheet immunoglobulins. Previously, we had demonstrated some preliminary aspects of the core calculation on the all α-helical globin family (Altman & Gerstein, 1994). Our method for defining regions of low structural variation is also useful for the analysis of structures solved NMR spectroscopy and generated by molecular dynamics since both techniques produce an ensemble of structures — in a sense, a family of very similar structures.

For the purposes of this discussion, we define “core atoms” as those having the same local conformation (i.e. secondary structure) and the same global conformation relative to their non-bonded neighbors across a family. This is similar to core-structure concept developed extensively by Chothia & Lesk, who have used it for such applications as analyzing protein motions (Lesk & Chothia, 1984; Chothia & Lesk, 1986; Lesk, 1991). However, Chothia & Lesk confine their core structure calculations to comparisons between pairs of proteins, while we aim to generalize the calculations so they are applicable to the many proteins in a family. Many other investigators have also used the term “core structure” but in a different sense from that used here. For instance, Bryant & Lawrence (1993) define a core in terms of conserved secondary structure elements, and others define a core structure based on measures of sequence conservation or hydrophobicity (Swindells, 1995). As will be discussed later, our results indicate that a core based purely on structural considerations is not the same as one based on sequence considerations, so, clearly, these definitions of “core” do not always coincide.

Practically, our method directly builds upon the large amount of recent work on superposition of families of structures. This work has either focused on finding the optimum superposition of a series of structures in which the corresponding atoms already have been defined (Diamond, 1992; Gerber & Müller, 1987; Kearsley, 1990; Shapiro *et al.*, 1992) or on finding a structural alignment between a pair of structures for which the corresponding atoms have not been defined (Taylor & Orengo, 1989; Sali & Blundell, 1990; Holm & Sander, 1993; Subbiah *et al*., 1993; Yee & Dill, 1993). Our core finding builds on both of these foundations: structural alignments are refined iteratively to include only low variance “core” atoms using the techniques of series superposition.

Because the average cores we define have reasonable stereochemistry, coupled with a consensus sequence, profile, or hidden Markov model (Gribskov *et al*., 1990 ; Bowie *et al*., 1991; Overington *et al*., 1992; Krogh *et al*., 1994), they may also be useful as starting points for a variety of homology modeling tasks. Furthermore, a library of carefully constructed core structures could eventually prove useful for summarizing the roughly 1000 folds that are thought to occur in nature (Chothia, 1992). This is particularly true with regard to the nine superfold structures (Orengo *et al.,* 1994). A library of core structures could also help speed up threading calculations, which match a sequence against a collection of structures (Ponder & Richards, 1987; Sippl, 1990; Jones *et al*., 1992; Bryant & Lawrence, 1993; Madej & Mossing, 1994), and make them less sensitive to the non-essential details of individual structures.

Once a core for a family of structures has been calculated, it is possible to use it to assess the similarity of two structures in the family in a better way than the usual RMS deviation in atom positions after doing a fit. The problem with the usual RMS value is that it weights each atom position equally and it gives equal weight to deviations in any direction. That is, a poorly fitting atom in the core of a structure is given equal weight to a poorly fitting atom in a highly variable surface loop. We use the variances calculated as part of the average core structure to weight the deviations and then use these “calibrated” deviations to calculate a better RMS.

For a multi-domain protein, such as the antibody molecule, one could compute an average core structure for all the domains taken together. However, this average structure would have the effect of any rigid-body motion between the domains spread throughout it in a correlated and highly redundant fashion. Consequently, we describe the average core structure of a multi-part protein in terms of the individual core structures of its component parts and then the average rigid-body positioning of one part relative to the another. We, furthermore, show how this average positioning can be described in terms of a mean and a variance in a manner that is completely consistent with the formalism we use to construct the average core structures.

2. Methods

The methods developed and applied in this paper fall into five categories: finding an average core, characterizing structural variation using this core, calculating sequence variation in a way that can be related to structure variation, using our calculation of structural variation to define a better RMS, and defining the average positioning of two cores.

a. Core finding algorithm

The core finding algorithm has been described in detail previously (Altman & Gerstein, 1994). It can be summarized as a five step process: (1) We start with an ensemble of aligned structures (e.g. all the immunoglobulin structures after they have aligned according to the Kabat numbering). Initially, we consider each atom position that occurs in every aligned structure to be a member the of the "putative core." (2) We construct an unbiased average structure from all these putative core positions, using one of the available methods for superimposing a series of corresponding structures (Altman & Gerstein, 1994; Diamond, 1992; Gerber & Müller, 1987; Kearsley, 1990; Shapiro *et al.*, 1992). (3) We determine the spatial variation of each group of aligned atoms about the average structure in terms of the volume of an "error ellipsoid" (as described in the next section). (4) We remove from the putative core the position with the largest structural variation. We then return to step 2 with a smaller core. We repeat this process until all aligned positions have been removed from the core. (This is a generalization of the "sieve-fit" procedure described in Gerstein & Chothia, 1991). The result is an ordered list of atomic positions, ranked according to structural variability, which we call the "throw-out" order. (5) Based on a variety of different criteria, discussed in detail in the caption to Figure 3, we pick one cycle in this overall core-finding procedure as best representing the separation between core and non-core atoms. The average structure calculated in this cycle is what we call the "core structure."

b. Calculating the Structural Variation of a Given Position

When all the structures in the ensemble are fit to the average core structure, it is possible to quantify the structural variation at each aligned position using an "error ellipsoid." We summarize the variability of each aligned position i over all structures in the ensemble of structures by using the variance/covariance matrix **C**. Each element in this 3×3 matrix, represented by cov(m,n), is the covariance between two coordinates, m and n, where m and n can be 1, 2, or 3 (representing the x, y, and z coordinates), over the ensemble at position i. Thus, for instance, cov(1,3) represents the covariance between the x and z coordinates, and the diagonal elements of the matrix, cov(m,m), contain the variance in coordinate m at position i over the ensemble. The variance/covariance matrix can be translated into an "ellipsoid of errors" centered at the mean position of the atom $\bar{\mathbf{x}}_i$. To find the orientation and axes lengths of the ellipsoid, the matrix is diagonalized in standard fashion to give:

$$\mathbf{C}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{R}_i^{-1} \qquad [1]$$

where $\mathbf{R}_i$ is a rotation matrix that specifies the orientation of the principal axes of the ellipsoid, and $\mathbf{S}_i$ is a diagonal matrix of eigenvalues. The square roots of the eigenvalues (denoted by σ_x, σ_y, and σ_z) give the standard deviation of the distribution in its three principal directions and are conventionally taken as the lengths of the semi-axes of the ellipsoid. Consequently, if the distribution of atomic positions is a three-dimensional normal distribution (for which we provide evidence for in Section 3.c), then an ellipsoid drawn at one standard deviation should contain approximately two thirds of the atoms in the sample. In order to get a single scalar estimate of the amount of variation in atomic position, we calculate the volume of the ellipsoid from the eigenvalues:

$$V = \tfrac{4}{3}\pi\sigma_x\sigma_y\sigma_z = \tfrac{4}{3}\pi\sqrt{\det \mathbf{C}} \ . \qquad [2]$$

c. Calculating the Sequence Variation of a Given Position

One of the advantages of our "volumetric" measure of structural variability is that it can be directly compared with measures of sequence variability at each aligned position, to see whether they are correlated. We calculate sequence variability through the computation of an information-theoretic entropy (Schneider & Stephens, 1991; Schneider *et al*., 1991; Shenkin *et al*., 1991). In particular, we measure variability of a given position i in the alignment in terms of its entropy relative to that if the sequences were aligned randomly:

$$R_{seq}(i) = \sum_{t=1}^{20} \bar{f}(t)\log_2 \bar{f}(t) - \sum_{t=1}^{20} f(i,t)\log_2 f(i,t) \ . \qquad [3]$$

where the first and second terms represent the standard Shannon entropy H for the random and actual alignments, f(i,t) is frequency of amino acid t at position i, and $\bar{f}(t)$ is the average frequency of residue type t over the whole the alignment.

d. Using the Core to Calculate a “Better RMS”

Once a core for a family of structures has been calculated, it can be used to assess the similarity of two structures in the family in a better way than the usual RMS value in atom positions. The basic idea is that one *uses the amount of structural variation observed at each position in the core to scale the interatomic distances between two structures*. If an atom position has low structural variation in the core structure and if the difference in the position of the corresponding atoms in two structures is large relative to this variation, then this difference should contribute more to the overall of difference between the two structure. Conversely, if the difference in position of corresponding points is small relative to the variation, then it should contribute less, regardless of the absolute value of the difference.

In the usual RMS measure of similarity, D_{RMS}, at each residue position i, one takes the vector difference between the coordinate positions in the first structure and the second structure

$$\mathbf{d}_i = \mathbf{x}_{1i} - \mathbf{x}_{2i} \qquad [4]$$

Then one averages the squares of these differences (i.e. the Euclidean distance) and takes the square root to get the RMS value,

$$D_{RMS} = \sqrt{\left\langle \mathbf{d}_i^2 \right\rangle} \qquad [5]$$

where the brackets denote averaging over the residues position index i. It is obviously harder to fit more residues well, and the overall RMS value, from fitting two arbitrarily selected segments of protein structure of length M, increases proportionately with the square root of M (McLachlan, 1984; Remington & Matthews, 1980).

The usual RMS measure weights the difference $\mathbf{d}_i$ at each residue position equally. However, if a given position is highly variable in all the structures used in the core finding procedure, it might be more reasonable to down-weight the coordinate difference at this position. Furthermore, the structural variability in a family may be oriented preferentially along a certain direction, so one would want to weight down the variation in this direction more than in the other directions. Using the “ellipsoid of errors” representation for structural variability developed in the previous sections, it is possible to make such a compensation. The weighted coordinate difference at each position $\mathbf{w}_i$

is the normal distance expressed in the units of standard deviation along the principal axes of the errors ellipsoid (discussed above):

$$\mathbf{w}_i = \mathbf{S}_i^{-\frac{1}{2}}\mathbf{R}_i^{-1}\mathbf{d}_i \qquad [6]$$

The above formula expresses the operation of translating and rotating the atoms into the coordinate system of the errors ellipsoid and then scaling their separation by an amount inversely proportional to the lengths of its principal axes. The root mean square (i.e. RMS) of this "standard-deviation" distance can then simply be computed to give, what we call an SD-RMS: $D_{SD} = \sqrt{\langle \mathbf{w}_i^2 \rangle}$.

Because $\mathbf{w}_i$ and the SD-RMS are expressed in units of standard deviation, they can not simply be related back to normal RMS values, which are usually expressed in units of Ångstroms. Consequently, it is helpful to introduce a conversion of the standard deviation "unit" back into Ångstroms. For each position, we define a "calibrated Ångstroms" distance $\mathbf{a}_i$ to be

$$\mathbf{a}_i = \frac{D_{RMS}}{D_{SD}}\mathbf{w}_i \qquad [7]$$

where the ratio $D_{RMS/}D_{SD}$ is the "conversion factor" between the average standard deviation unit for atom *i* and Ångstroms. Thus, the calibrated-Ångstroms distance expresses the coordinate difference between two atoms in Ångstrom units that have be inflated or deflated according to the structural variability observed at position i the family of structures. Note that the RMS value of the calibrated Ångstrom distances (which is taken over all atoms) is necessarily *the same as that of the standard RMS* value. However, at particular positions the calibrated Ångstrom distance will be larger than the normal Cartesian distance if there is little variability in the family — i.e. these are expensive "Ångstroms." Conversely, if there is much variability in the family, the calibrated distance will be less than the normal distance.

Finally, it is possible to do this whole analysis with one of the two structures as the average structure — i.e. $\mathbf{x}_{2i} = \bar{\mathbf{x}}_i$. This allows one to decide which of an ensemble of structures is most like the mean structure overall or at a given residue position.

e. Finding the Average Orientation of Two Rigid Cores

For a multi-domain protein, such as the immunoglobulins, one could compute an average core structure for all the domains taken together. However, this core structure would have the effect of any rigid-body motion between the domains spread throughout all of its error ellipsoids, in a correlated and highly redundant fashion. This would obviously create problems in the core-finding procedure, as one would tend to throw out all the atoms of mobile domain together, and it would make detecting the subtle structural variations within a mobile domain more difficult. Consequently, it is useful to describe the average core structure of a multi-part protein in terms of individual core structures of its component parts and then the average rigid-body positioning of one part relative to the another. The variation about this mean positioning would summarize the range of orientations that are adopted throughout the ensemble of structures.

Describing the average relative position of two rigid bodies involves calculating an average translation and an average rotation. The translation is straightforward. We treat all the translation vectors exactly the same way as the atom coordinates at one aligned position in a family. That is, to get a mean translation we just average the vectors. Likewise, to describe the variation, we can form a variance/covariance matrix **C** from the translation vectors and then compute an “errors ellipsoid” volume from the determinant this matrix. This allows us to describe a rigid-body translation in exactly the same the language as we use to describe the individual atom positions.

Describing an average rotation is not so simple. As rotations do not commute, one can not simply compute a normal (commutative) average all the components in a list of rotation matrices. However, for small rotations, is possible to express each rotation as a vector and then average these in a straightforward fashion. By a small rotation we mean one with a small rotation angle θ, such that $\sin\theta \approx \theta$. This is usually about 10°. The vector representation we use for rotations are the four quaternion parameters, $\mathbf{q}^T = (\lambda\ \mu\ \ \nu\ \ \sigma)$ (Altmann, 1986). These are simply related to the rotation angle θ and the direction cosines of the rotation axis l, m, and n:

$$\mathbf{q}^T = \quad (ls \quad ms \quad ns \quad c) \qquad [8]$$

where

$$s = \sin\frac{\theta}{2} \text{ and } c = \cos\frac{\theta}{2}.$$

As opposed to the direction cosines, for which the rotation angle is treated distinctly differently from the rotation axis, quaternions have the advantage that all four components enter on essentially equal footing. Consequently, in averaging rotations it is reasonable to average each component of series of quaternions in an equivalent and symmetric fashion. Since each quaternion has unit length (i.e. $\mathbf{q}^2 = \lambda^2 + \mu^2 + \nu^2 + \sigma^2 = 1$), we treat the average quaternion as a vector constrained to be on the unit sphere (in four dimensions), and we average a number of orientations by vector summing their quaternion vectors and then normalizing the result. That is,

$$\bar{\mathbf{q}} = \frac{\sum \mathbf{q}_j}{\left\| \sum \mathbf{q}_j \right\|}, \qquad [9]$$

where the $\mathbf{q}_j$ are individual rotations and $\bar{\mathbf{q}}$ is the average rotation.[1]

The beauty of the quaternion representation is that since it uses vectors to describe orientation, it is possible to describe the variation in orientations using the same formalism we used above to describe the variation of atom positions in core structures and of the translations. Since both the 4-component quaternion and the direction cosines (l m n) are normalized to have a length of one, we find that, for small rotations, the variation in orientation can be described completely in terms of RMS variation in the rotation angle, $\sqrt{\langle \theta^2 \rangle}$.

In averaging rotations, one additional point must be considered. Often the rotation that one wants to average is conventionally described as large. For instance, the average orientation of one immunoglobulin domain relative to the another (i.e. VL to VH) is usually described in terms of a pseudo two-fold axis or a 180° rotation. This is hardly a small rotation and so would not appear amenable to the averaging scheme discussed above. However, in comparing a series of Fv fragments, the *difference* in orientation of one VL relative to another VL is small. Consequently, as shown in Figure 8, it possible to do the averaging calculation in “bootstrap” fashion. One finds

[1] Note that for small θ, $2\mathbf{q}^T \approx (\ \theta l\ \ \theta m\ \ \theta n\ 2)$, so averaging quaternions is very similar to averaging the vector representation for infinitesimal rotations (Goldstein, 1980).

the difference in orientation of all VL domains as compared to an arbitrarily chosen first one and then averages these small differences. Then one combines the large rotation of the first VL relative to VH with average of the difference rotations to get the correct average rotation of VL relative to VH. Differences relative this unbiased average can then be computed to get the variation in orientations.

Fig. 8 her

3. Results

a. Immunoglobulin core

Fig. 1 her

We started with 12 VL (κ) domains and 12 VH domains (listed in Table 1A). As shown in Figure 1, we used the the Kabat numbering scheme (Kabat *et al.*, 1983) to align the VL and VH domains individually, and we used the structural alignment in Chothia and Lesk (1987) to align the VL and VH frameworks to each other. After removing the three difficult-to-align hypervariable loops, we started our core finding with 90 aligned positions for VL, 99, for VH, and 89 for the combined VL-VH. We found cores with 70 Cα atoms for VL, 87, for VH, and 52 for VL-VH combined. The error ellipsoids for cores are shown in Figure 2.

Table 1 h
Fig. 2 h

Because of the way we chose the core vs non-core threshold, all the core structures had acceptable Cα stereochemistry. As discussed in Figure 3, the Cα-Cα bond lengths were all nearly 3.8 Å and the Cα bond and torsion angles were within normal ranges.

Fig. 3 h

The throw out order was very similar for all three runs: from the original positions we first threw out loop regions — in particular, the long loop connecting the C” strand to the D strand — and then A’ and C” strands and the ends of G and C’ strands. The second group of strands to be thrown out included A and D; the third group included E and the rest of G and C’; and the last group included B, C, and F.

The core structures included all of strands A, B, C, D, E, and F and most of the strands C’ and G. Only one of the strand residues sandwiched between the two sheets and none of the strand residues forming the VL-VH interface were thrown out in the VL or VH cores. (The sandwich and interface residues were identified by Chothia *et al*. (1986) and indicated by “•” and “x” in Figure 1).

The last residues to be thrown out include the absolutely conserved disulfide bridge, the buried Trp (C4), and the residues immediately surrounding the disulfide (B5, F3, C4, B3, B4, E2, E3, E4, and F2, indicated by a ‘p’ in Figure 1). These residues constitute the “pin” holding together the two immunoglobulin sheets (Lesk & Chothia, 1982).

We performed similar core calculations to the ones described above starting with a larger subset of aligned atoms: i.e., we started with all backbone atoms or with all possible backbone atoms plus all alignable sidechain atoms. These calculations resulted in essentially the same core structure and throw-out order, indicating that Cα atoms alone were enough to define the essential features of the core.

b. No correlation between sequence and structure variation

Figure 4 shows the relationship between sequence variation, measured by Shannon entropy, and structural variation, measured by ellipsoid volume. We find there is no significant correlation between them (discussed more fully in the figure caption). This is true whether we consider all the aligned positions (89), the core positions (52), or the non-core positions (27).

Fig. 4 h

c. SD-RMS Calculations

After calculating the VL-VH combined core, we fit each of the 24 immunoglobulin variable domains to it and then computed RMS and SD-RMS distances between all pairs. We then used these distances to cluster the immunoglobulins into two different trees, which are shown in Figure 5. The trees from the normal RMS and SD-RMS calculations are distinctly different, having a rank correlation of only ~0.8. This means that the SD-RMS is not a redundant calculation and does indeed provide new and potentially useful information to use in comparing structures. Furthermore, the comparison of the trees well illustrates how SD-RMS de-emphasizes the usual sources of variation between between structures and accentuates the unique differences of each individual structure.

Figure 6 illustrates the use of the calibrated Ångstrom measure: for two representative variable domains, we compare at each aligned positions the standard Cα-Cα distance with the calibrated

Ångstroms distance. In regions of little structural variability, such as in the B strand the calibrated distance is greater than the normal distance. This is because even small differences between structures are very significant in this highly conserved region. The contrasting situation is observed in highly variable regions, such as the C”D turn, where the calibrated distance is usually less than the normal distance.

Fig. 6 her

In order to evaluate the degree to which a three-dimensional normal distribution accurately describes the distribution of actual atom positions around the average core structure, we plotted the displacement of each atom in each variable domain structures from the average core structure (after fitting all structures to the calculated core). To put all the displacements on the same scale, we expressed them in S.D. units as weighted coordinate differences $\mathbf{w}_i$. Figure 7 shows the distribution of these weighted coordinate differences for three arbitrarily selected α-carbons and the aggregate distribution of weighted differences derived from all α-carbons. The distributions are unimodal, peaked at zero, and nearly symmetric. Thus, to a good degree they can be considered normal.

Fig. 7 her

d. Average orientation

The average positioning of the VH domains relative to a VL domains is described in Table 2. We express this positioning as the rotation and translation necessary to superimpose a VL domain onto a VH domain. As described in the caption to the table, we chose our coordinate system so that the *average* transformation could simply be described as a 173° rotation around the z axis followed by a 24.4 Å translation along the x-axis.

Table 2 h

Once we found the mean transformation we could compute deviations about this mean for each structure. These deviations are the incremental rotation and translation that need to be applied after the mean transformation to correctly position the VL and VH domains in each Fv fragment. We find that on average the incremental rotation is 5.4° and the incremental translation is 0.94 Å. The incremental rotations appear to be roughly equally spread among the three axis directions (x, y, and z). However, most of the incremental translation is in the z direction (parallel to the pseudo-twofold), while the least is in the x-direction.

The amount of rotation and the amount of translation required are fairly well correlated (with a correlation coefficient of 0.79) so that structures that require significant additional translation also require further rotation. Structures that are close to the mean orientation are HyHEL-10, which has the smallest incremental translation (0.26 Å) and a small incremental rotation (4.5°), and B13I2, which has a small translation (0.43 Å) and the smallest rotation (3.6°). The structure farthest from the mean is NC41, which has both the largest rotation and largest translation (7.6° and 1.8 Å). Note that even this maximum rotation is rather small, so the assumptions underlying the rotational averaging procedure are fully satisfied.

4. Discussion

a. Methodology

Our core finding procedure requires an initial alignment between the structures in a family. It then iteratively refines this alignment, throwing out atoms with high spatial variation. It is designed to work on families of relatively similar structures, such as the globins or immunoglobulins, where it is possible to get an initial structural alignment by eye or by sequence alignment. We, clearly, depend on a high-quality alignment as a starting point, and the availibility of automatic structural alignment procedures (Taylor & Orengo, 1989; Sali & Blundell, 1990; Holm & Sander, 1993; Subbiah *et al*., 1993; Yee & Dill, 1993) increases the applicability of our method. Considering the analogy with methods for multiple sequence alignment (Subbiah, 1989), we believe our method for finding and refining an unbiased average could extend these pairwise structure alignment procedures to allow them to perform multiple structural alignment.

The basic core finding procedure we present is also only applicable to monomeric proteins. However, we show how it is possible to extend it for use on multi-domain proteins such as the immunoglobulins in an intelligent and consistent fashion. Since the orientation averaging calculations are not that computationally expensive, we see no limitations in applying our calculations to even larger assemblies than immunoglobulin Fv fragments. Furthermore, our formalism for describing average positioning of two rigid cores has obvious applications for describing the rigid-body motion of domains (Gerstein *et al*., 1993, 1994) and the rigid-body docking of macromolecular complexes, such a protein binding to DNA (Suzuki *et al.*, 1994, 1995).

The core structures generated by our procedure exhibit acceptable stereochemistry. This is a natural consequence of the way we discard the most variable positions and only average α-carbons (and thus never have to worry about averaging over a flipped peptide). Moreover, because of their acceptable stereochemistry and their variability measures at each site, our core structures might provide good starting points for model-building and threading, and this, in turn, suggests that they may provide a useful representation to use in building up a compact library of folds.

Another application for our core structures is that they allow one to compare structures using the SD-RMS and calibrated Ångstroms. These measures of structural similarity emphasize differences between structures in regions of low variability and discount them in regions of high variability. Consequently, they are useful in highlighting structures with particularly distorted core geometry and in assessing whether a particular structure differs from the other members of the family in a typical or unique fashion.

Unlike other aspects of our procedure, the way we represent the structural variation of a particular site with an “error ellipsoid” is somewhat arbitrary. Underlying such a representation is a normal (i.e. symmetric and unimodal) model for the distribution of actual atom coordinates around the average structure. Our results show that at least for the immunoglobulins this assumption of normality is reasonable. However, there are clearly cases (e.g. particular surface sidechains) in which the distributions may be multimodal and asymmetric. In these cases, neither our core-finding algorithm nor our SD-RMS calculation would lose its applicability or validity since it is still completely valid to perform calculations based on the first two moments, i.e. mean and variance, of an asymmetric, multi-modal distribution. However, certain common *interpretations* of our average core structures would not be valid (e.g. that the average structure is the most probable location for atoms in a family or that a one standard deviation contour contains two-thirds of the atoms in a family).

Another aspect of procedure that is somewhat arbitrary is the way we choose a threshold between core and non-core. We used the variance of the non-core atoms as a criteria because it was straightforward to calculate and produced consistent results (as discussed in the caption to Figure 3). For specific applications one might want to choose a different way of drawing the line between core and non-core atoms. For instance, one could use a particular maximum ellipsoid volume (e.g. 1 Å^3) as a cutoff for core atoms. Such a cutoff would have the advantage of being even more directly related to the stereochemistry of the resulting core than is our present criteria. Alternatively, one could choose a core cutoff based on minimizing the overlap between the ellipsoid volume distributions of core and non-core atoms.

While the particular dividing line between core and non-core is arbitrary, the throw-out order generated by our procedure is not. This ordering is essentially a *ranking of the atoms by their structural variability* in the family. The throw-out order does not depend on whether one starts core finding with all possible aligned atoms, just mainchain atoms, or just Cα atoms. Residues appear to be thrown out as units. This suggests that for finding an average core structure for a family one gets most of the relevant information by just using α-carbons.

b. Immunoglobulin Core

Our analysis of the immunoglobulins shows that the throw-out order can be quite biologically illuminating. The immunoglobulin superfamily has recently been divided into 4 groupings on the basis of sequences and structures: the V-set, the C1-set, the C2-set, and the I-set (Harpaz & Chothia, 1994; Williams & Barclay, 1988). The V-set includes the immunoglobulin variable domains (i.e. VL and VH) as well as parts of the T-cell receptor (e.g. domain 1 of CD2). Each molecule in the V-set contains the 10 β-strands found in VL and VH (A, A’, B, C, C’, C”, D, E, F, and G) with the exception of CD2 and CD4 which are missing strand A. The C1-set includes the constant domains of the immunoglobulins and various parts of Major Histocompatibility Complex (MHC) molecules. In comparison with the V-set, it is missing all of strands A’ and C” and the ends of strands G and C’. The C2-set is similar to the C1-set but is also missing strand D. The I-set, which includes cell-adhesion molecules and surface receptors, is intermediate between the V-set and the C1-set in that it contains A’ and the end of G but does not contain C” and the end of C’. All immunoglobulin molecules contain a highly conserved “pin” holding together the two sheets (Lesk & Chothia, 1982; Bork *et al.*, 1994): this consists of two pairs of interlocking strands (B-C and E-F), which contain the conserved disulfide and Trp and the residues contacting them.

The throw-out order we found during the immunoglobulin core finding is very consistent with the division of the immunoglobulin superfamily. The first strands thrown-out (C”, A’, and the ends of C’ and G) were the most variable strands in the superfamily, determining whether a molecule is in the V-set, I-set, and so forth. The next grouping of strands thrown-out included strands A and D. The presence or absence of these strands separates VL and VH from other members of the immunoglobulin family in a less fundamental way than C” or A’. Finally, the last strands to be thrown-out were B, C, and F, which contain the conserved disulfide and make up the bulk of the pin.

Thus, a ranking of structural variability *within* the variable domains is consistent with structural variability within the whole immunoglobulin superfamily. This is quite a striking finding since no where in our immunoglobulin core-finding did we incorporate any information about variability in other members of the superfamily beside VL and VH, yet our procedure identified as variable those parts of immunoglobulin structure that vary greatly throughout the superfamily.

Structural variability is clearly correlated with sequence variability at the level of the overall fold — i.e. similar sequences have the same fold. However, with regard to the immunoglobulins, we find that sequence variation is not correlated with structural variation in terms of the detailed positioning of atoms (as measured by our error ellipsoids). This is true whether we consider just the core atoms or both core and non-core atoms. Our results are consistent with the idea that structural accommodation is global: in responding to mutations helices and sheets shift slightly, more or less as rigid bodies (reflecting their fixed hydrogen-bonded geometry), and spread the effect of the mutation throughout whole core. Such global accomodation has been found in the structures of T4 lysozyme mutants (Eriksson et al., 1992; Baldwin et al., 1993).

Previously, we demonstrated similar results regarding sequence and structure variability for the globins (Altman & Gerstein, 1994). Our work with the globins also manifest the importance of the throw-out order. In particular, we found that purely on the basis of throw-out order the globins could be partitioned into a more variable region (the F helix) and a conserved core, which turned out to have the essentially the same structural elements as the repressor protein.

For the immunoglobulins, as well as the globins, it appears that the parts of the protein thrown away first had fewer tertiary interactions than those thrown away later. For the immunoglobulins this is obvious in comparing, say, the first strands thrown-out (A' and C", on the edge of the protein) with those thrown out last (B-C and E-F, in the center of the domain). Thus, the core-finding procedure appears to start at the outside and successively peel of away layers of protein until it gets to the center.

5. Conclusion

We have presented a method for finding an unbiased, average structural core and for finding the average orientation between two rigid cores. An integral part of our method is assigning a measure of variability (i.e. the error ellipsoid), to each position in a family and ranking all the positions according to their structural variability (i.e. the throw-out order). Once calculated this measure of variability can be used to calculate a more informative measure of structural similarity than the normal RMS difference atom positions — i.e. a better RMS — and can be easily compared with sequence variability. Furthermore, when applied to specific protein families, our average core structures and measures of variability yield a number of biologically significant results. For instance, by looking at variability just within the variable domains of immunoglobulin family, we are able to see patterns of variability that reoccur throughout the whole superfamily.

As we have defined them, the core atoms represent the structurally invariant components of protein families. Since they adopt the same position in all members of the family, the core atoms are probably not responsible for functional differences within these families. Instead, the atoms which are classified as non-core are logically the ones to which functional differences can be assigned.

Availability of Results on the Internet

We make available C and lisp source code for performing the core finding and calculating the SD-RMS; alignments of the immunoglobulins; the actual coordinates of the immunoglobulin cores; ProteanD, a program for displaying error ellipsoids on a Silicon Graphics workstation; and further documentation in hypertext form. These items can be retrieved by sending e-mail to mbg@hyper.stanford.edu or altman@camis.stanford.edu or through anonymous ftp to the following URL:

ftp://camis.stanford.edu/pub/AvgCore/

Acknowledgments

RBA is a Culpeper Medical Scholar and is supported by LM-05652. Computing environment provided by the CAMIS resource under NIH grant LM-05305. MG is supported by a Damon-Runyon Walter-Winchell fellowship (DRG-1272). We thank M. Levitt and A. Lesk for useful discussions and suggestions and R Diamond for providing his structure superposition program.

References

Altman, R. and Gerstein, M. (1994). Finding an Average Core Structure: Application to the Globins. *Proc. Second Int. Conf. Intell. Sys. Mol. Biol*., 19-27 (Menlo Park, CA, AAAI Press).

Altmann, S. L. (1986). *Rotations, Quaternions and Double Groups*. New York, Oxford UP.

Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. and Poljak, R. J. (1986). Three dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science.* **233**: 747-753.

Arevalo, J. H., Stura, E. A., Taussig, M. J. and Wilson, I. A. (1993). Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex. *J. Mol. Biol.* **231**: 103.

Bashford, D., Chothia, C. and Lesk, A. M. (1987). Determinants of a Protein Fold: Unique Features of the Globin Amino Acid Sequences. *J. Mol. Biol.* **196**: 199-216.

Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A. and Matthews, B. W. (1993). The Role of Backbone Flexibility in Accomodation of Variants that Repack the Core of T4 Lysozyme. *Science* **262**: 1715-1718.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**: 535-542.

Bork, P., Holm, L. and Sander, C. (1994). The Immunoglobulin Fold: Structural Classification, Sequence Patterns and Common Core. *J. Mol. Biol.* **242**: 309-320.

Bowie, J. U., Lüthy, R. and Eisenberg, D. (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science.* **253**: 164-170.

Bryant, S. H. and Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16**: 92-112.

Chothia, C. (1992). Proteins — 1000 families for the molecular biologist. *Nature.* **357**: 543-544.

Chothia, C., Boswell, D. R. and Lesk, A. M. (1988). The outline structure of the T-cell αβ– receptor. *EMBO J.* **7**: 3745-3755.

Chothia, C. and Finkelstein, A. V. (1990). The classification and origins of protein folding patterns. *Ann. Rev. Biochem.* **59**: 1007-39.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**: 823-826.

Chothia, C. and Lesk, A. M. (1987). Canonical structures for the hypervariable regions of the immunoglobulins. *J. Mol. Biol.* **196**: 901-917.

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., Colman, P. M., Spinelli, S., Alzari, P. M. and Poljak, R. J. (1989). Conformations of the immunoglobulin hypervariable regions. *Nature.* **342**: 877-883.

Chothia, C., Novotny, J., Bruccoleri, R. and Karplus, M. (1985). Domain association in immunoglobulin molecules: The packing of variable domains. *J. Mol. Biol.* **186**: 651-663.

Colman, P. M., Tulip, W. R., Varghese, J. N., Baker, A. T., and Tuloch, P. A. (1987). NC41 Complex structure determination. *Nature.* **326**: 358-363.

Epp, O., Latham, E., Schiffer, M., Huber, R. and Palm, W. (1975). *Biochemistry.* **14**: 4943-52.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. and Matthews, B. W. (1992). Response of a protein structure to cavity creating mutations and its relation to the hydrophobic effect. *Science.* **255**: 178-183.

Felsenstein, J. (1989). PHYLIP — Phylogeny Inference Package (Verstion 3.2). *Cladistics.* **5**: 164-166.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Seattle, Department of Genetics, University of Washington.

Fermi, G., Perutz, M. F., Shaanan, B. and Fourme, R. (1984). the crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* **175**: 159.

Finkelstein, A. V. and Reva, B. A. (1993). A search for the most stable folds of protein chains. *Nature.* **351**: 497-500.

Gerstein, M. and Chothia, C. H. (1991). Analysis of Protein Loop Closure: Two Types of Hinges Produce One Motion in Lactate Dehydrogenase. *J. Mol. Biol.* **220**: 133-149.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G. and Chothia, C. (1993). Domain Closure in Lactoferrin: Two Hinges produce a See-saw Motion between Alternative Close-Packed Interfaces. *J. Mol. Biol.* **234**: 357-372.

Gerstein, M., Lesk, A. M. and Chothia, C. (1994). Structural Mechanisms for Domain Movments. *Biochemistry.* **33**: 6739-6749.

Goldstein, H. (1980). *Classical Mechanics.* New York, Addison-Wesley.

Gribskov, M., Lüthy, R. and Eisenberg, D. (1990). Profile Analysis. *Meth. Enz.* **183**: 146-159.

Harpaz, Y. and Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **238**: 528-539.

Herron, J. N., He, X., Mason, M. L., Voss, E. W. and Edmundson, A. B. (1989). Three-dimensional structure of a fluorescein-Fab complex crystallized in 2-methyl-2,4-pentandiol. *Proteins.* **5**: 271-280.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G. (1993). A Database of Protein Structure Families with Common Folding Motifs. *Prot. Sci.* **1**: 1691-1698.

Holm, L. and Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **233**: 123-128.

Johnson, M. S., Sali, A. and Blundell, T. L. (1990). Phylogenetic Relationships from Three-dimensional Protein Structures. *Meth. Enz.* **183**: 670-691.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature.* **358**: 86-89.

Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Milner, M. and Perry, H. (1983). *Sequences of Proteins of Immunological interest.* Washington, DC, NIH.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modelling. *J. Mol. Biol.* **235**: 1501-1531.

Lesk, A. M. (1991). *Protein Architecture: A Practical Approach.* Oxford, IRL Press.

Lesk, A. M. and Chothia, C. (1982). Evolution of Proteins Formed by β-Sheets II. The Core of the Immunoglobulin Domains. *J. Mol. Biol.* **160**: 325-342.

Lesk, A. M. and Chothia, C. (1984). Mechanisms of Domain Closure in Proteins. *J. Mol. Biol.* **174**: 175-91.

Lesk, A. M. and Chothia, C. H. (1980). How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins. *J. Mol. Biol.* **136**: 225-270.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**: 59-107.

Levitt, M. (1992). Accurate Modeling of Protein Conformation by Automatic Segment Matching. *J. Mol. Biol.* **226**: 507-533.

Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature.* **261**: 552-558.

Madej, T. and Mossing, M. C. (1994). Hamiltonians for protein tertiary structure prediction based on three-dimensional envioronment principles. *J. Mol. Biol.* **233**: 480-487.

Marquart, M., Deisenhofer, J., Huber, R. and Palm, W. (1980). Crystallographic Refinement and atomic models of the intact immunoglobulin molecule KOL and its antigen-binding fragment at 3.0 Å and 1.9 Å resolution. *J. Mol. Biol.* **141**: 369.

McLachlan, A. D. (1984). How Alike are the Shapes of Two Random Chains? *Biopolymers.* **23**: 1325-1331.

Murzin, A., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: Structural Classification of Proteins. *J. Mol. Biol.* **247**: 536-540.

Orengo, C. A. (1994). Classification of protein folds. *Curr. Opin. Struc. Biol.* **4**: 429-440.

Orengo, C. A., Flores, T. P., Taylor, W. R. and Thornton, J. M. (1993). Identifying and Classifying Protein Fold Families. *Prot. Eng.* **6**: 485-500.

Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature.* **372**: 631-634.

Overington, J., Donnelly, D., Johnson, M. S., Sali, A. and Blundell, T. L. (1992). Environment-specific amino acid substitution tables: Tertiary templates and predition of protein folds. *Protein Sci.* **1**: 216-226.

Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smith-Gill, G. S. and Davies, D. R. (1989). Structure of an antibody-antigen complex: Crystal Structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci. USA.* **86**: 5938-5942.

Pascarella, S. and Argos, P. (1992). A Databank Merging Related Protein Structures and Sequences. *Prot. Eng.* **5**: 121-137.

Ponder, J. W. and Richards, F. M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**: 775-791.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992). *Numerical Recipes in C.* Cambridge, Cambridge University Press.

Remington, S. J. and Matthews, B. W. (1980). A systematic approach to the comparison of protein structures. *J. Mol. Biol.* **140**: 77-99.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**: 167-334.

Sali, A. and Blundell, T. L. (1990). The definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **212**: 403-428.

Sander, C. and Schneider, R. (1991). Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins: Struc. Func. Genet.* **9**: 56-68.

Satow, Y., Cohen, G. H., Padlan, E. A. and Davies, D. R. (1987). Phosphocholine binding immunoglobulin Fab McPC603: An X-ray diffraction study at 2.7 Å. *J. Mol. Biol.* **190**: 593-604.

Schneider, T. D. and Stephens, R. M. (1991). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**: 6097-6100.

Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986). Information Content of Binding Sites on Nucleotide Sequences. *J. Mol. Biol.* **188**: 415-431.

Schulze-Gahmen, U., Rini, J. M., Arevalo, J., Stura, E. A., Kenten , J. H. and Wilson, I. A. (1988). Preliminary crystallographic data, primary sequence, and binding data for an anti-peptide Fab and its complex with a synthetic peptide from influenza virus hemagglutinin. *J. Biol. Chem.* **263**: 17100-17105.

Shapiro, A., Botha, J. D., Pastore, A. and Lesk, A. M. (1992). A Method for the Multiple Superposition of Structures. *Acta. Cryst.* **A48**: 11-14.

Shenkin, P. S., Erman, B. and Mastrandrea, L. D. (1991). Information-Theoretical Entropy as a Measure of Sequence Variability. *Proteins: Struc. Func. Genet.* **11**: 297-313.

Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. and Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. USA.* **84**: 8075-8079.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J.Mol. Biol.* **213**: 859-884.

Stanfield, R. L., Fieser, T. M., Lerner, R. A. and Wilson, I. A. (1990). Crystal Structures of an Antibody to a Peptide and Its Complex with Peptide Antigen at 2.8 Å. *Science.* **248**: 712-719.

Strong, R. K., Campbell, R., Rose, D. R., Petsko, G. A., Sharon, J. and 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**(15): 3739-3748.

Subbiah, S. and Harrison, S. C. (1989). A Method for Multiple Sequence Alignment With Gaps. *J. Mol. Biol.* **209**: 539-548.

Subbiah, S., Laurents, D. V. and Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* **3**: 141-148.

Suh, S. W., Bhat, T. N., Navia, M. A., Cohen, G. H., Rao, D. N., Rudikoff, S. and Davies, D. R. (1986). The galactan-binding immunoglobulin Fab J539. An X-ray diffraction study at 2.6 Å resolution. *Prot. Struct. Func. Genet.* **1**: 74.

Suzuki, M., Gerstein, M. and Yagi, N. (1994). The stereochemical basis for DNA recognition by Zn fingers. *Nuc. Acid. Res.* **22**: 3397-3405.

Suzuki, M., Yagi, N. and Gerstein, M. (1995). DNA-recognition and superstructure-formation by helix-turn-helix proteins. *Prot. Eng.* (in press).

Swindells, M. B. (1995). A procedure for the automatic determination of hydrophobic cores in protein structures. *Prot. Sci.* **4**: 93-102.

Taylor, W. R. and Orengo, C. A. (1989). Protein Structure Alignment. *J. Mol. Biol.* **208**: 1-22.

Williams, A. F. and Barclay, A. N. (1988). The immunoglobulin superfamily — domains for surface recognition. *Annu. Rev. Immunol.* **6**: 381-405.

Yee, D. P. and Dill, K. A. (1993). Families and the Structural Relatedness Among Globular Proteins. *Protein Sci.* **2**: 884-899.

Table 1 Families, Structures, and Ensembles

A Structures used

PDB	**Immunoglobulin**	**Species**	**Reference**	**chains**	
4FAB	4-4-20	mouse	Herron *et al.,* 1989	H	κ
1HIL	17/9	mouse	Schulze-Gahmen *et al*., 1988	H	κ
1NCA	NC41	mouse	Colman *et al*., 1987	H	κ
2FB4	KOL	human	Marquart *et al*., 1980	H	λ
1DBA	DB3	mouse	Arevalo *et al.,* 1993	H	κ
1IGF	B13I2	mouse	Stanfield *et al*., 1990	H	κ
2FBJ	J539	mouse	Suh *et al*., 1986	H	κ
1MCP	McPC603	human	Satow *et al.,* 1987	H	κ
6FAB	36-71	mouse	Strong *et al.,* 1991	H	κ
1FDL	D1.3	mouse	Amit *et al*., 1986	H	κ
2HFL	HyHEL-5	mouse	Sheriff *et al*., 1987	H	κ
3HFM	HyHEL-10	mouse	Padlan *et al*., 1989	H	κ
1REI	REI (VL dimer)	human	Epp *et al*., 1975		κ

B Ensembles used

Ensembles	Number of aligned atoms	Number of structures	Average, Min, and Max RMS between structures in ensemble (Å per atom)		
VL (κ)	90	12	0.79	0.44	1.33
VH	99	12	1.13	0.49	1.74
Common to both VL and VH	89	24	1.43	0.44	2.16

All immunoglobulin structures are of uncomplexed antibodies except for 2HFL, 3HFM, and 1FDL. All the structures were taken from the protein databank (Bernstein *et al*., 1977).

Table 2 The average positioning of immunoglobulin VL and VH domains

	Translation (Å)				**Rotation (°) (d)**			
	components in each direction			magnitude	components in each direction			magnitude
	T(x)	**T(y)**	**T(z)**	**|T|**	**R(x)**	**R(x)**	**R(x)**	θ
Mean (a)	24.4	0	-0.17	24.4	0	0	173	173
Deviation from Mean (b)								
HyHEL-10	0.18	-0.18	0.00	0.26	3.9	0.7	-2.2	4.5
B13I2	0.21	-0.35	0.13	0.43	3.4	-1.0	0.3	3.6
McPC603	0.18	-0.29	0.28	0.44	-3.2	-1.7	0.5	3.6
17/9	0.30	-0.14	-0.51	0.61	1.3	2.9	-1.8	3.7
36-71	-0.12	0.43	0.63	0.77	-0.2	-2.7	3.3	4.3
KOL	-0.36	-0.64	-0.46	0.86	-2.0	2.3	-3.3	4.5
HyHEL-5	-0.11	0.62	-0.63	0.89	4.2	0.8	-1.4	4.5
DB3	0.51	0.23	0.78	0.96	1.3	-4.9	5.2	7.3
D1.3	-0.38	-0.90	0.30	1.02	-2.3	1.4	-5.4	6.1
J539	0.28	0.23	-1.03	1.09	-5.8	3.5	-3.0	7.4
4-4-20	-0.06	0.18	-1.11	1.13	-2.6	4.0	3.0	5.6
NC41	0.25	0.97	1.49	1.80	2.1	-5.4	4.9	7.6
RMSD (c)	0.28	0.51	0.74	0.94	3.1	3.0	3.3	5.4

The results of applying our orientation averaging procedure to the immunoglobulins: the mean rotation and translation necessary to superimpose a VL domain onto a VH domain and deviations about this mean. (a) The mean transformation was derived from fitting the VH domain of a particular structure to the average VL domain (using the VL-VH alignment in Figure 1) and then refitting on VH. The VL domain was positioned in the coordinate system so that its centroid is at the origin; the rotation, which is applied before the translation, is around the z axis; and as much as possible of the translation is along the x axis. With these conventions, one would *roughly* expect the z axis to be parallel to the β-strands in the immunoglobulins; the x-axis to be perpendicular to the plane of the β-sandwich in each domain; and the y axis to be parallel to the plane of the sandwich but perpendicular to the strand direction.

(continued....)

Table 1 (continued)

(b) Once the average transformation is applied one can determine the incremental rotation and translation necessary to perfectly superimpose particular VH domains. These deviations from the mean transformation are shown for each Fv fragment. (c) The average of deviations from the mean transformation in (b) are zero. However, one can measure the spread of each component deviation by computing its RMS, which is shown. |T| is the magnitude of the translational component. Analogously, θ is the magnitude of the rotational component. (d) The rotation is expressed in terms of three components and an overall rotatation angle. The three components are the direction cosines scaled by the angle of rotation. They are the usual way to describe small rotations (Goldstein, 1980; Altmann, 1986). Because $\sin\theta \approx \theta$ for small rotations, the four numbers are easily related to the four quaternion parameters:

$$\mathbf{q} = (\lambda, \mu, \nu, \sigma) = \left(\frac{R(x)}{2}, \frac{R(y)}{2}, \frac{R(z)}{2}, \cos\frac{\theta}{2} \right)$$

As should be apparent the way the translations and rotation are expressed in this table is completely equivalent. This one of the strengths of the quaternion representation.

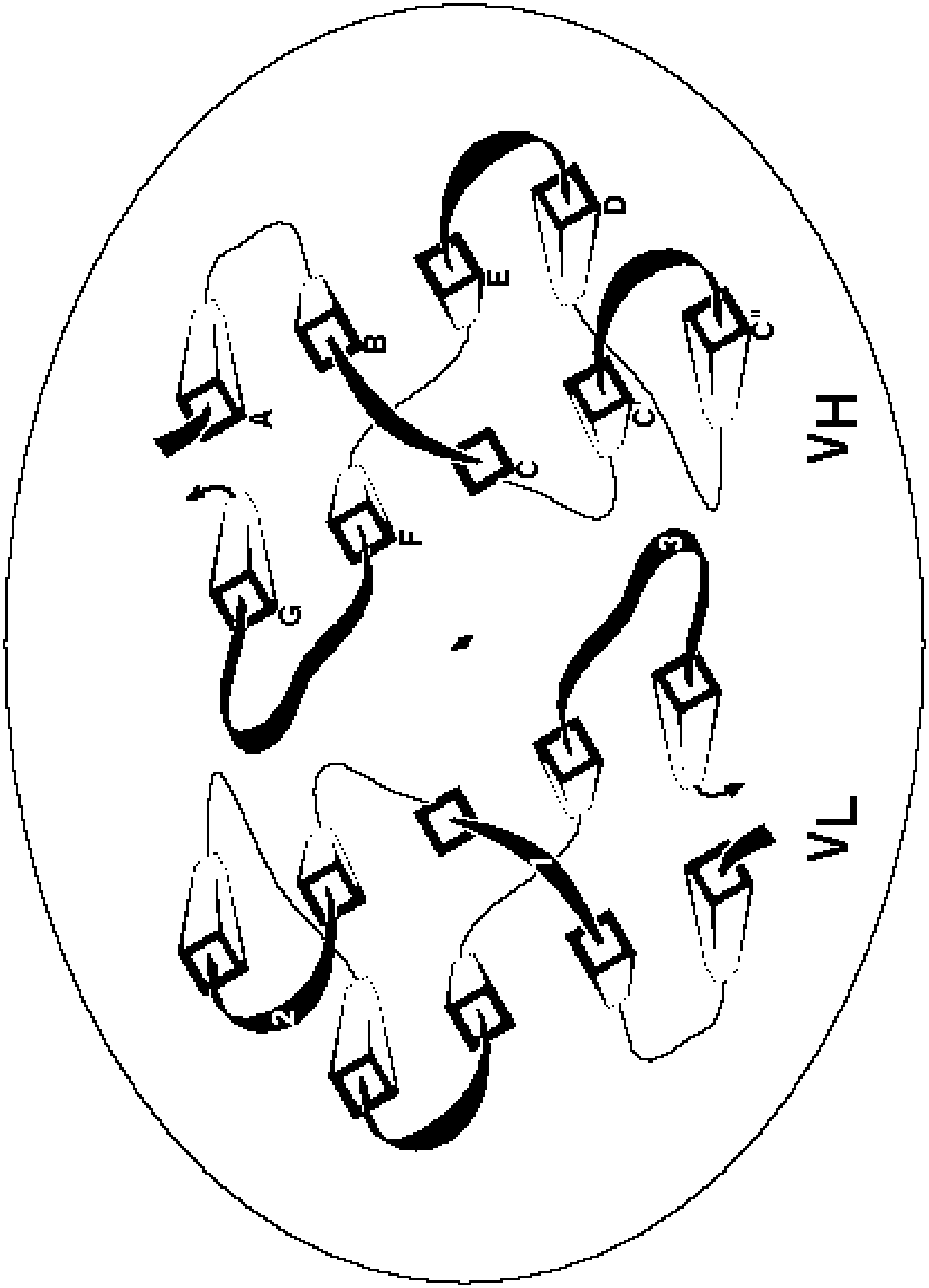
Figure 1 Outline of immunoglobulin structure

Part (A) shows a schematic outine of the structure of an immunoglobulin Fv fragment, looking down the pseudo-twofold axis at the hypervariable loops. The 9 principal strands in each domain are indicated by boxes, along with the standard strand names (A-G). (Loops 1-3 are also labeled, and strand A' is omitted for clarity.) Part (B) is a table that gives a detailed overview of the structural role of each immunoglobulin residue in conjuction with the degree of structural variability found for it by our core-finding procedure. The leftmost two columns (a) show the Kabat numbering (Kabat *et al.*, 1983) for the antibodies and the alignment we used between VL and VH. Black bars mark loops 1, 2, and 3, and the 3 hypervariable loops. The first annotation column (b) shows a canonical labeling scheme for the residues within the β-sheets, which is derived from the work of Chothia and others (Lesk & Chothia, 1982; Chothia *et al.*, 1985; Chothia & Lesk, 1987; Chothia *et al.*, 1988; Chothia *et al*., 1989; Harpaz & Chothia, 1994). The “inner” sheet that forms that VL-VH interface consists of strands A’, C, C’, C”, F, and G and the “outer” sheet consists of strands A, B, D, and E. The next annotation column (c) gives additional information about the structural role of the residues: x, •, and o are β-strand residues that, respectively, participate in the VL-VH interface (x), pack between the 2 sheets (•), or are not in the two preceding categories (o); b are residues in a β-bulge; C and W are the conserved Cys and Trp; p are the residues (with the conserved C and W) that form the conserved “pin” identified by Lesk & Chothia (1982) that holds the two sheets together; v and ^ are residues in interstrand loops on the side of the molecule near or not-near the hypervariable loops. The columns marked “throw-out order” (d) give the order that residues were thrown out in the core finding procedure when it was applied only to the VL structures, only to the VH structures, and to the VL and VH structures taken together. For the latter calculation, the 12 VL and 12 VH structures were aligned based on this table and then the core-finding routine was applied to this whole group of 24 structures to generate an average structure that is representative of both VL and VH. The throw out numbers marked with filled-in black boxes correspond to residues that were not included as part of the core and those simply boxed correspond to the residues that were the last 15 residues to be thrown out last. It is notable that the two conserved Cys residues and one conserved Trp were among this residues last to be thrown out.

Average Core Structures

Figure 1 (continued)

A



Kabat [a]		Annotation		Throw-out order		
VL	VH	(b)	(c)	VL	VH	VLVH [d]
1	1		^	**1**	**1**	**2**
2	2		^	**8**	19	**10**
3	3		^	36	70	38
4	4	**A1**	•	51	71	55
5	5	**A2**	o	38	61	47
6	6	**A3**	•	**89**	49	50
7	7			21	51	**23**
8				29		
9	8				**8**	
10	9	**A'1**	o	46	**3**	**5**
11	10	**A'2**	o	45	14	**21**
12	11	**A'3**	o	22	30	**25**
13	12	**A'4**	•	**18**	45	**33**
14	13		v	**19**	46	41
15	14		v	**20**	47	42
16	15		v	26	42	40
17	16		v	25	26	**32**
18	17		v	30	38	**34**
19	18	**B1**	•	60	37	64
20	19	**B2**	o	69	58	**77**
21	20	**B3**	•p	57	**91**	**76**
22	21	**B4**	op	**81**	**97**	**78**
23	22	**B5**	**C**	74	**94**	**79**
24	23	**B6**	o	52	72	53
25	24	**B7**	•	34	53	46
	25				21	
Loop 1						
	33				32	
33	34	**C2**	•	40	65	65
34	35	**C3**	x	**78**	74	**88**
35	36	**C4**	**W**	**86**	**89**	**85**
36	37	**C5**	x	**85**	**88**	**83**
37	38	**C6**	•	**90**	**87**	**84**
38	39	**C7**	x	**88**	**86**	**87**
39	40	**C8**	o	62	54	**27**
40	41		v	**15**	55	**15**
41	42		v	**4**	23	**6**
42	43	**C'1**	o	**12**	**10**	**9**
43	44	**C'2**	ob	**14**	27	**18**
44	45	**C'3**	xb	47	79	62
45	46	**C'4**	o	73	**85**	**80**
46	47	**C'5**	xb	55	69	73
47	48	**C'6**	•b	72	34	72
48	49	**C'7**	•	71	28	**28**
49	50	**C'8**	o	66	33	**22**
	51				66	
	52				18	
Loop 2						
53	56	**C"1**		**11**	**11**	**1**
54	57	**C"2**		**7**	31	**8**
55	58	**C"3**		**9**	56	**7**
	59				57	
	60				22	
56	61		v	**2**	**7**	**3**
57	62		v	**10**	**5**	**4**
58	63		v	42	**4**	**12**
59	64		v	41	**9**	**13**
60	65		v	43	**2**	**14**
61	66		v	63	13	**24**
62	67		v	65	29	**35**
63	68	**D1**	o	64	59	51
64	69	**D2**	•	59	64	61
65	70	**D3**	o	32	60	49
66	71			24	52	48
67	72			**16**	41	**17**
68	73			53	16	**16**
	74		^		**6**	
	75		^		15	
69	76		o	75	40	**37**
70	77	**E1**	o	70	76	56
71	78	**E2**	•p	**80**	81	70
72	79	**E3**	op	**79**	77	74
73	80	**E4**	•p	**77**	75	**75**
74	81	**E5**	o	**76**	50	57
75	82	**E6**	•	58	36	**36**
76	82a		v	31	35	**29**
77	82b		v	33	20	**31**
78	82c		v	37	39	43
79	83		v	27	43	44
80	84		v	**13**	44	**30**
81	85		v	23	83	45
82	86		v	28	**99**	52
83	87		v	35	**98**	54
84	88	**F1**	•	48	**96**	69
85	89	**F2**	o	**83**	**95**	**89**
86	90	**F3**	•	**87**	**93**	**86**
87	91	**F4**	xp	**82**	**92**	**82**
88	92	**F5**	**C**	**84**	**90**	**81**
89	93	**F6**	x	56	73	71
90	94	**F7**	o	**5**	25	**20**
	95				**12**	
Loop 3						
97	102	**G1**	o	39	24	**26**
98	103	**G2**	x	49	67	63
99	104	**G3**	•	50	62	60
100	105	**G4**	ob	61	63	59
101	106	**G5**	•b	68	80	66
102	107	**G6**	•	54	78	67
103	108	**G7**	o	67	84	68
104	109	**G8**	•	44	82	58
105	110	**G9**	o	**17**	68	39
106	111	**G10**	•	**6**	48	**19**
107	112	**G11**	o	**3**	17	**11**

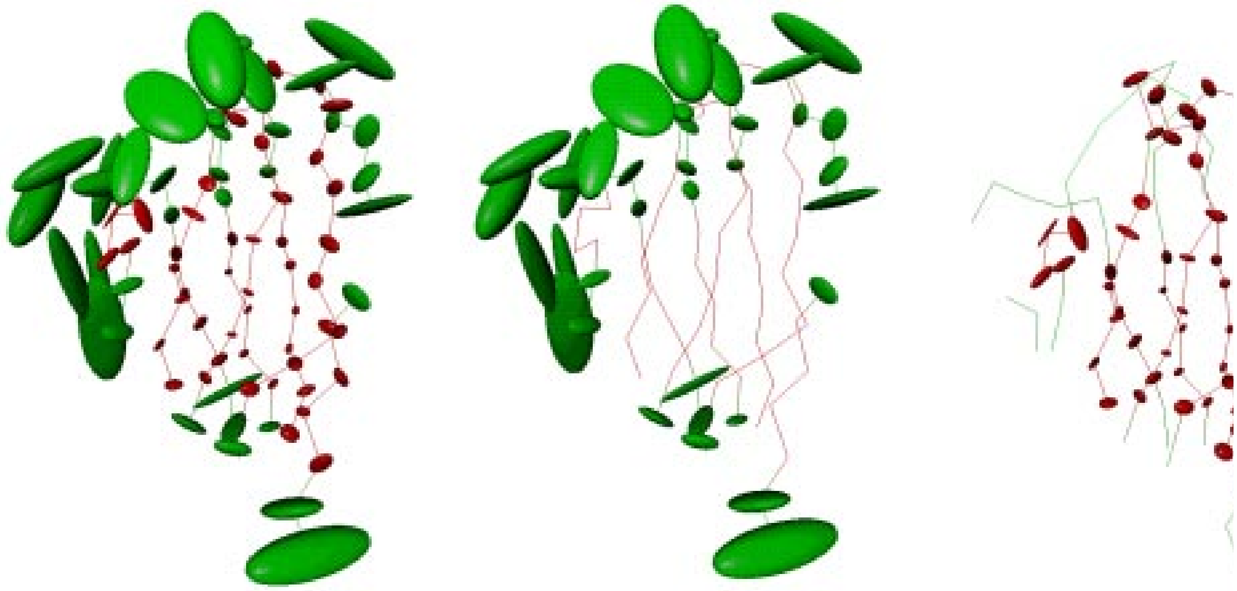
Rough **Figure 1-B**

Figure 2 Average core structures of the immunoglobulins

(A) The mean positions and ellipsoids are shown for the 89 atoms in the common alignment of the combined VL-VH domain (LEFT). The non-core Cα atoms (with ellipsoids at two standard deviations) are shown (CENTER). The core atoms are also shown (RIGHT). The central portions of the beta-strands make up the core. Part (B) shows the cores of the VL and VH domains positioned by the average transformation (discussed in Table 2) to yield an average Fv fragment. The view is the same as in the schematic in Figure 1, down the pseudo-two fold axis.

Figure 2 (continued)

A



B

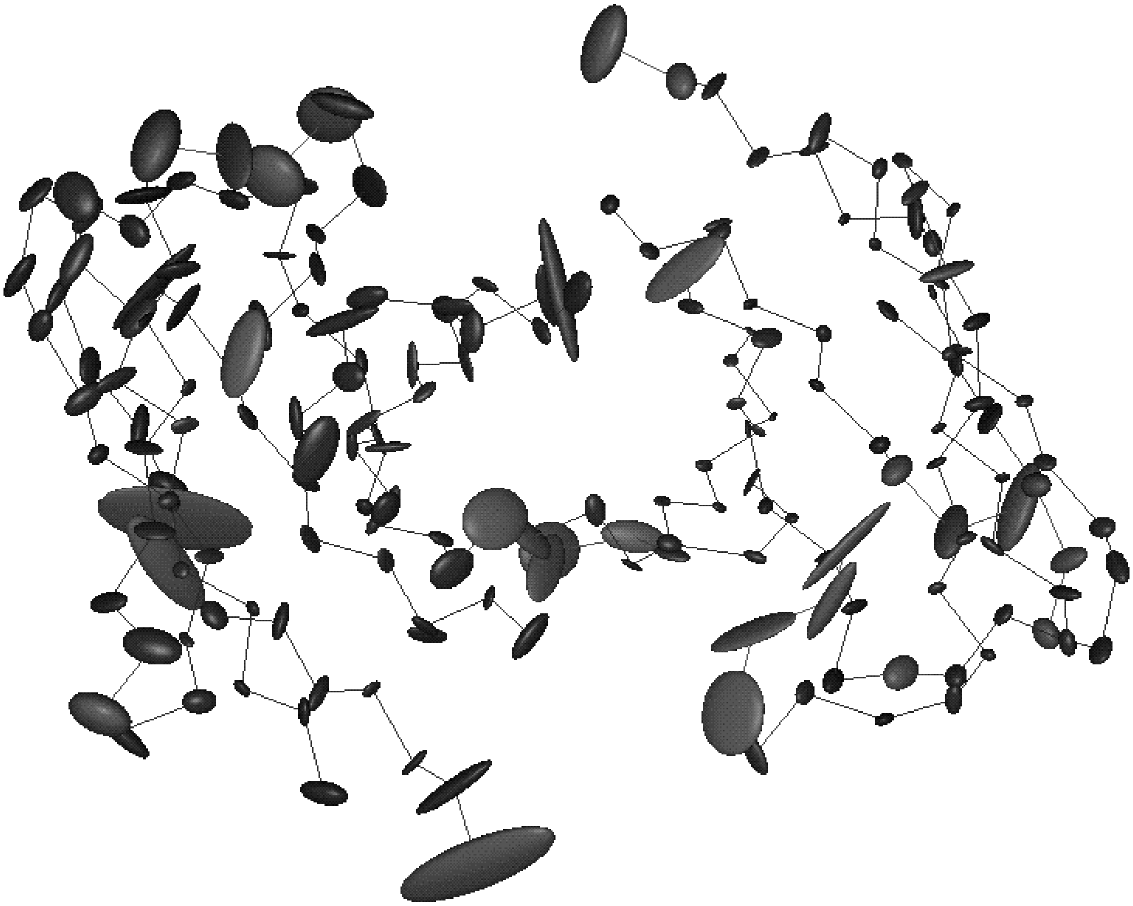
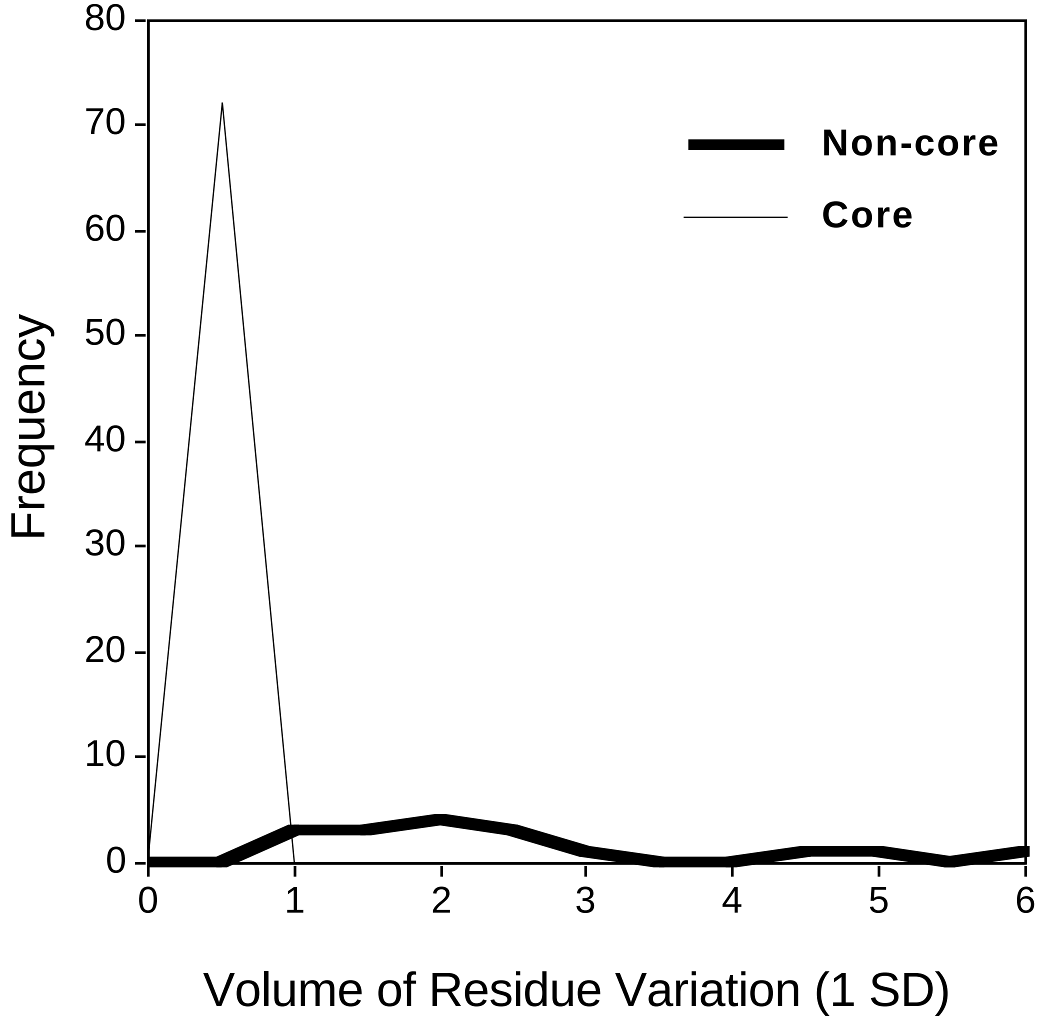


Figure 3 Determining a core vs non-core threshold



Depending on the specific application, a variety of measures can be used position the a threshold within the throw-out order to separate core from non-core. Most of these methods focus on means and variances derived from the distribution of core and non-core ellipsoid volumes. The distribution of these core and non-core volumes is shown here at cycle 17 in the core-finding procedure. This cycle was picked because it maximizes the difference between the average core and non-core volume. This often is a useful core vs. non-core threshold. However, for consistency with previous work (Altman & Gerstein, 1994), we used a slightly different threshold.

(continued...)

Figure 3 (continued)

We placed the threshold at a point that maximizes the variance in the ellipsoid volumes of the non-core atoms (i.e., the width of the thick-line distribution). This threshold implies that core atoms have a uniformly small ellipsoid size and this, in turn, makes it possible for the core structure to have particularly good Cα stereochemistry.

Three parameters characterize the stereochemistry of an α-carbon structure (Levitt, 1976): (1) the distance between two connected Cα atoms, which should be 3.8 Å; (2) the angle τ between three connected Cα atoms, which can range between approximately 80° and 135°; and (3) the pseudo-torsion angle α between four connected Cα atoms, which can acceptably range from -180° to +180° and so does not form a meaningful constraint on a-carbon structure. We tabulate statistics on the Cα—Cα virtual bond length and the τ angle below.

	Bond Length			Bond Angle τ	
Core Structure	**Average (Å)**	**Standard Deviation (Å and %)**		**Number within acceptable range (±5°)**	**Number outside of acceptable range**
Immoglobulin (Combined VL and VH)	3.77	0.051	1.4%	35	0
Immunoglobulin VL	3.78	0.061	1.6%	53	3
Immunoglobulin VH	3.76	0.039	1.0%	68	5

Figure 4 Sequence variation versus structure variation

Graph of sequence variation versus structural variation for each immunoglobin position in the combined VL-VH alignment. At a particular position, structural variation is measured by the volume of the covariance matrix ellipsoid relative to that of the smallest ellipsoid, expressed on a log-scale (so the variation in core and non-core volumes can be shown together). Sequence variation is measured in bits per residue as the information content of a given position in the alignment relative to that if the sequences were aligned randomly (as described in the methods section). There are 89 positions represented in total here and the overall Pearson correlation coefficient is 0.35. The 54 core positions are highlighted by white boxes. The correlation between information content and ellipsoid volume for just the core positions is 0.19; and for just the non-core positions, 0.10. If structural variation were correlated with sequence variation, one would expect the points to lie on a line such that small ellipsoids would be associated with a large difference in information content relative to the random sequence (and vice versa for large ellipsoids).

Figure 4 (continued)

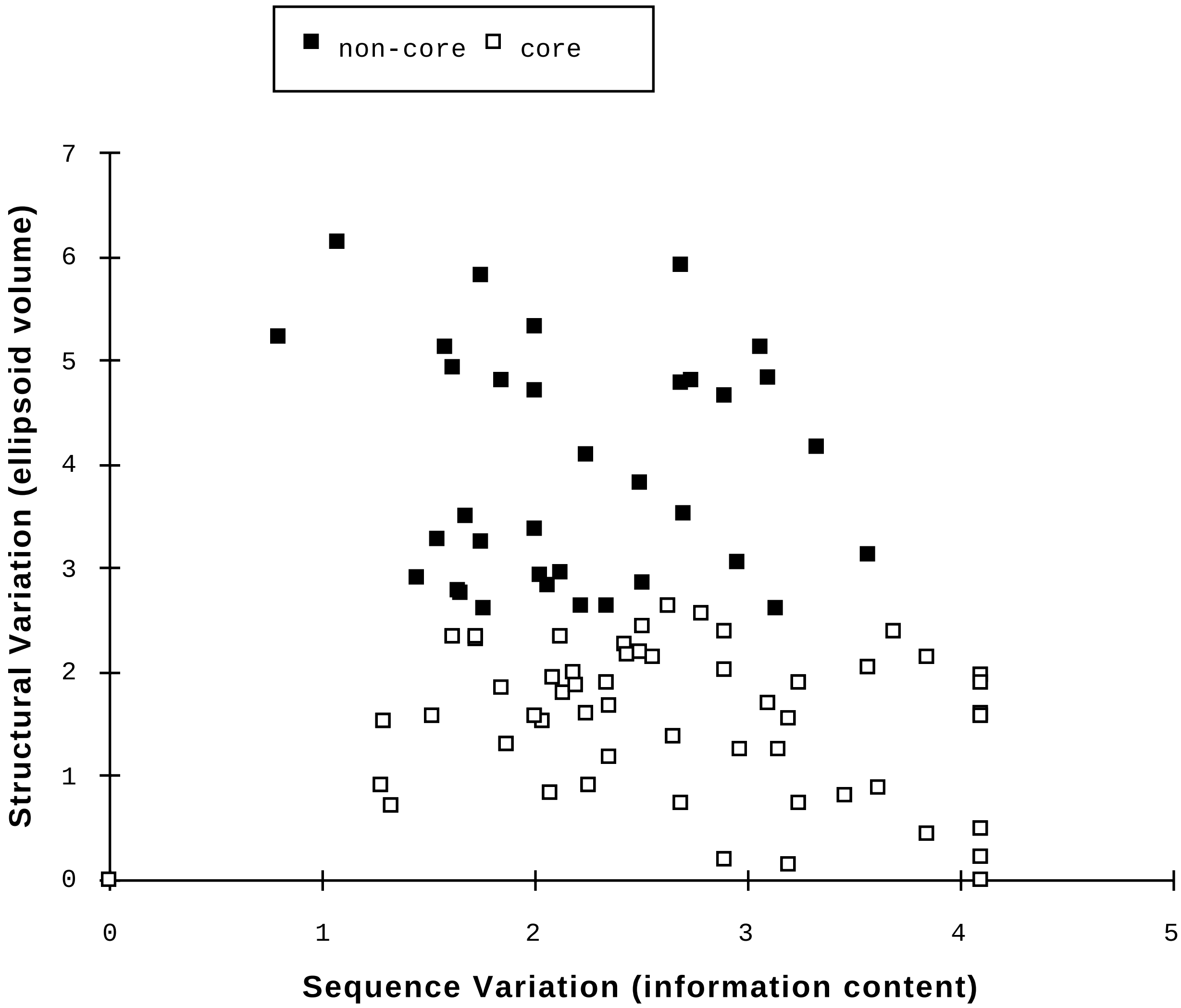


Figure 5 Structure clustering determined by normal RMS versus by SD-RMS

We fit each of the 24 immunoglobulin variable domains onto the VL-VH combined core. Then, for each of the 24 × 23 / 2 pairs of structures, we computed the normal RMS measure of similarity and the SD-RMS. Trees are a way of visually displaying these pairwise distances. We made trees clustering the immunoglobulin structures using normal RMS values (TOP) and our SD-RMS (BOTTOM). Since these trees were made only on the basis of structural similarity, they are not expected to directly comparable to phylogenetic trees made on the basis of sequences. The trees were made with the PHYLIP package (Felsenstein, 1989; Felsenstein, 1993) . The Spearman rank order correlation between the ordering of the structures based on normal RMS and on SD-RMS is 0.80 (with a probability of less than .00001 of the null hypothesis that the two orderings are the same.)

Figure 5 **(continued)**

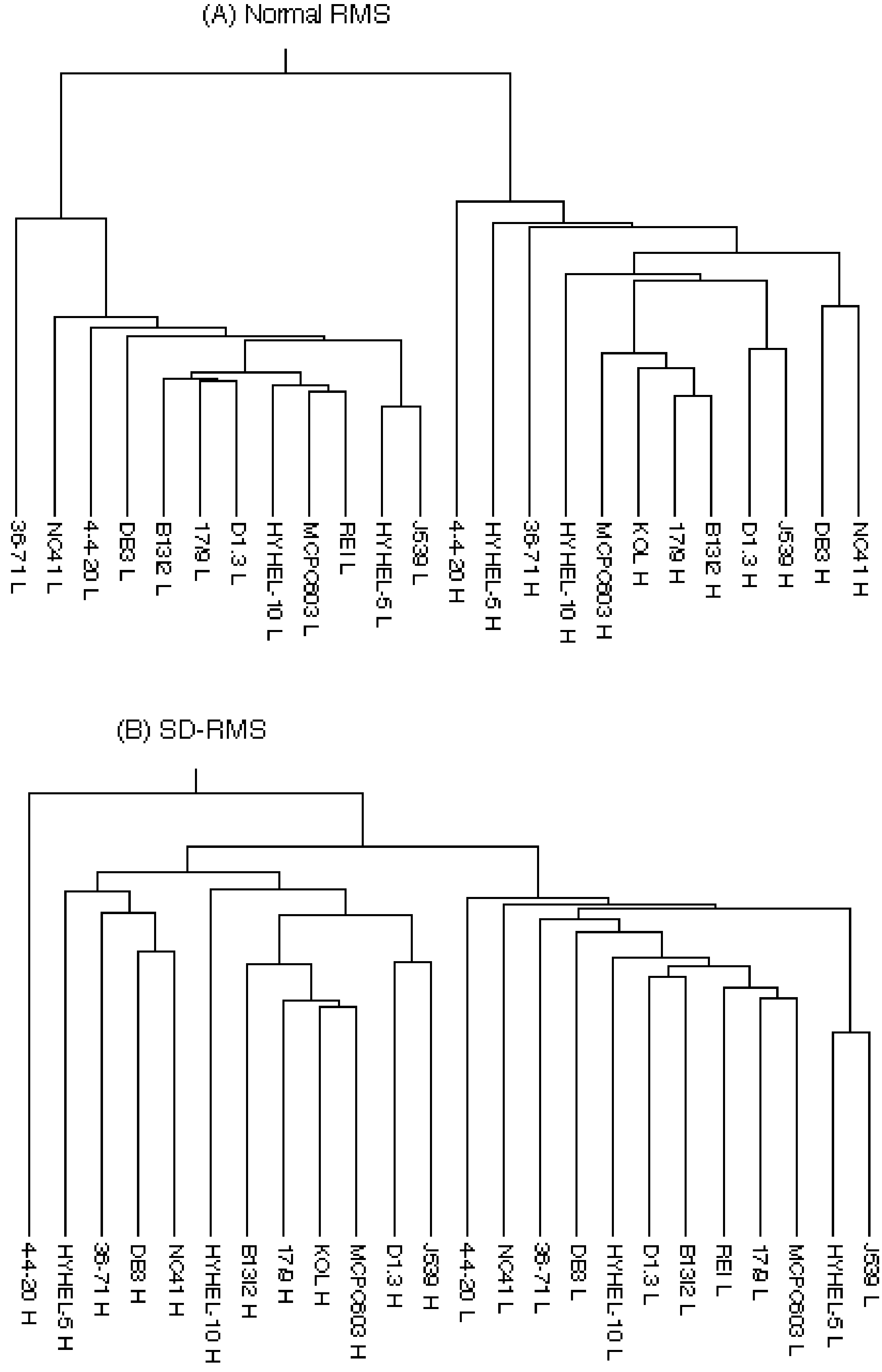


Figure 6 Measuring Structural Deviation with Calibrated Ångstroms

The relationship between normal distance in Ångstroms and “calibrated Ångstroms” distance, which is scaled according to the size of the errors ellipsoid. The thick line in the top part of the graph shows normal distance deviations (in Å) between two representative immuoglobulin structures (the VH of 4-4-20 and the VH of NC41) after they both have been fit onto the combined VL-VH core structure. The thin line at the bottom of the graph shows the average structural variation assigned to each position in computing the core structure. This variation is expressed as the volume of an error ellipsoid (in cubic Å) at each position. The thick line shows the same distance deviations as the thin line, but now calibrated according to the amount of variation at each position. The RMS value of these calibrated Ångstroms deviations and the normal distance deviations are necessarily the same and are represented here by the horizontal dotted line. Note that in the B sheet, which runs from 18 to 24 and is a region of the immunoglobulin structure where there is little variation within the family, the normal distances between the two structures are small and beneath the overall RMS value, but the calibrated distances are large and above the line. The converse is true for positions in the highly variable C”D turn (from 59 to 67), where the calibrated distances are less than the normal distances.

Figure 6 (continued)

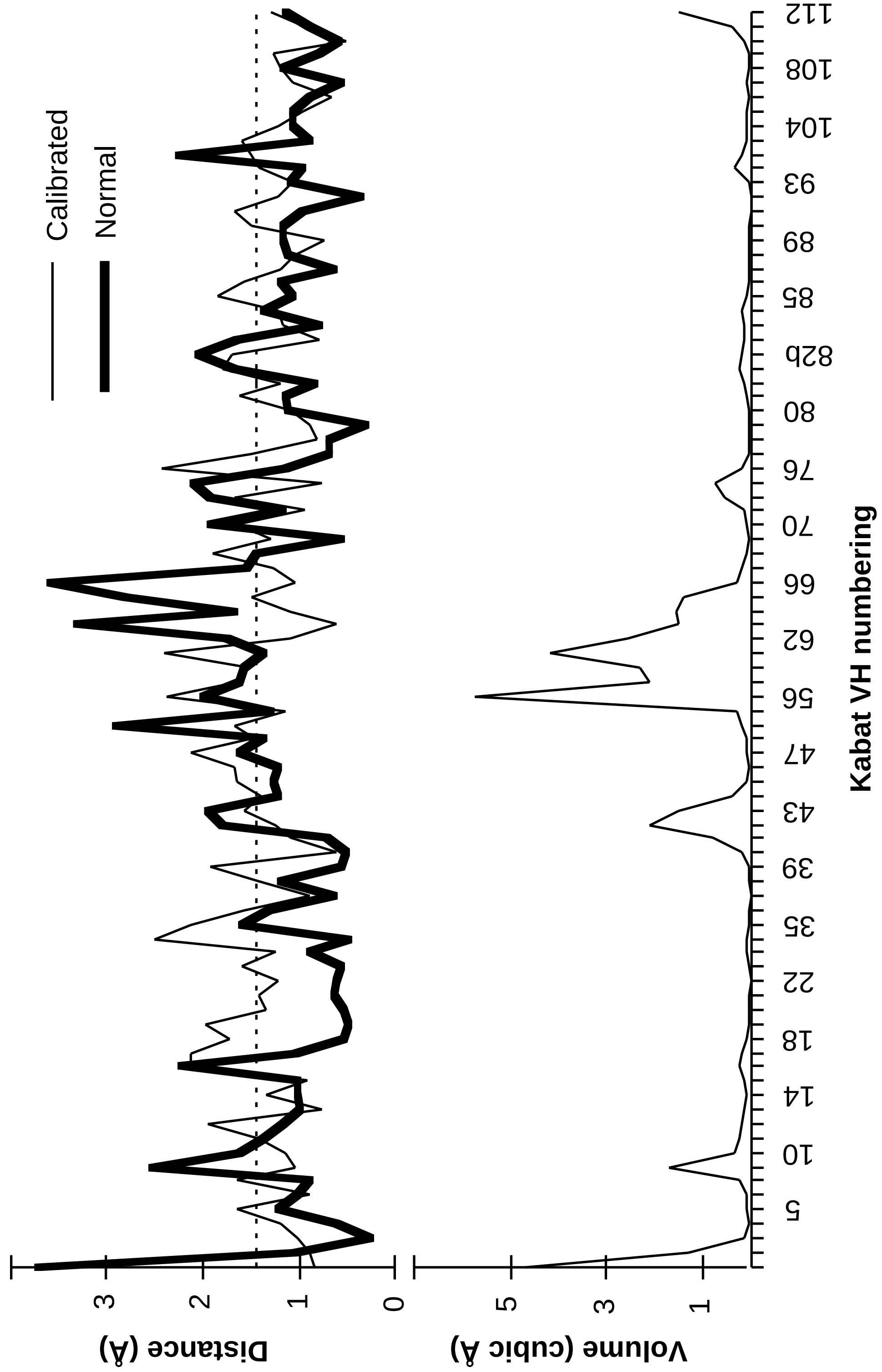


Figure 7 Distribution of atomic positions about the average structure

The distribution of displacements from the average coordinates in the 24 immunoglobulin variable domains. We fit each structure to the core and then determined the weighted coordinate differences from the mean for each atom. The thin lines show a histogram of the x component of these differences for Cα atoms at positions C2, A2, C”2, and F5 (using the strand numbering shown in Figure 1). The thick line shows a histogram of all the weighted coordinate differences aggregated together. Thus, this histogram was constructed from 6408 specific coordinate differences (= 3 coordinates × 89 atoms × 24 structures).

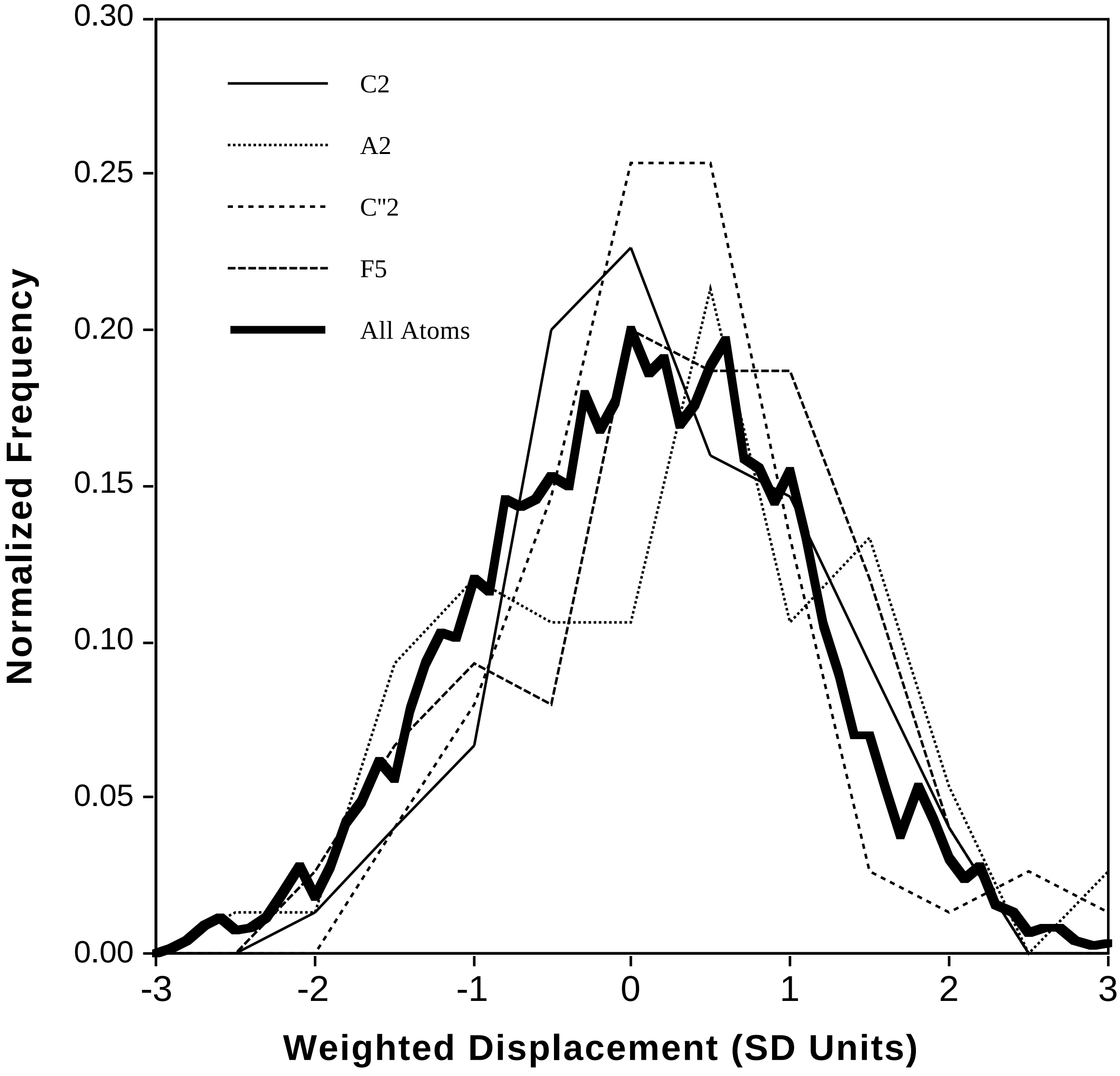
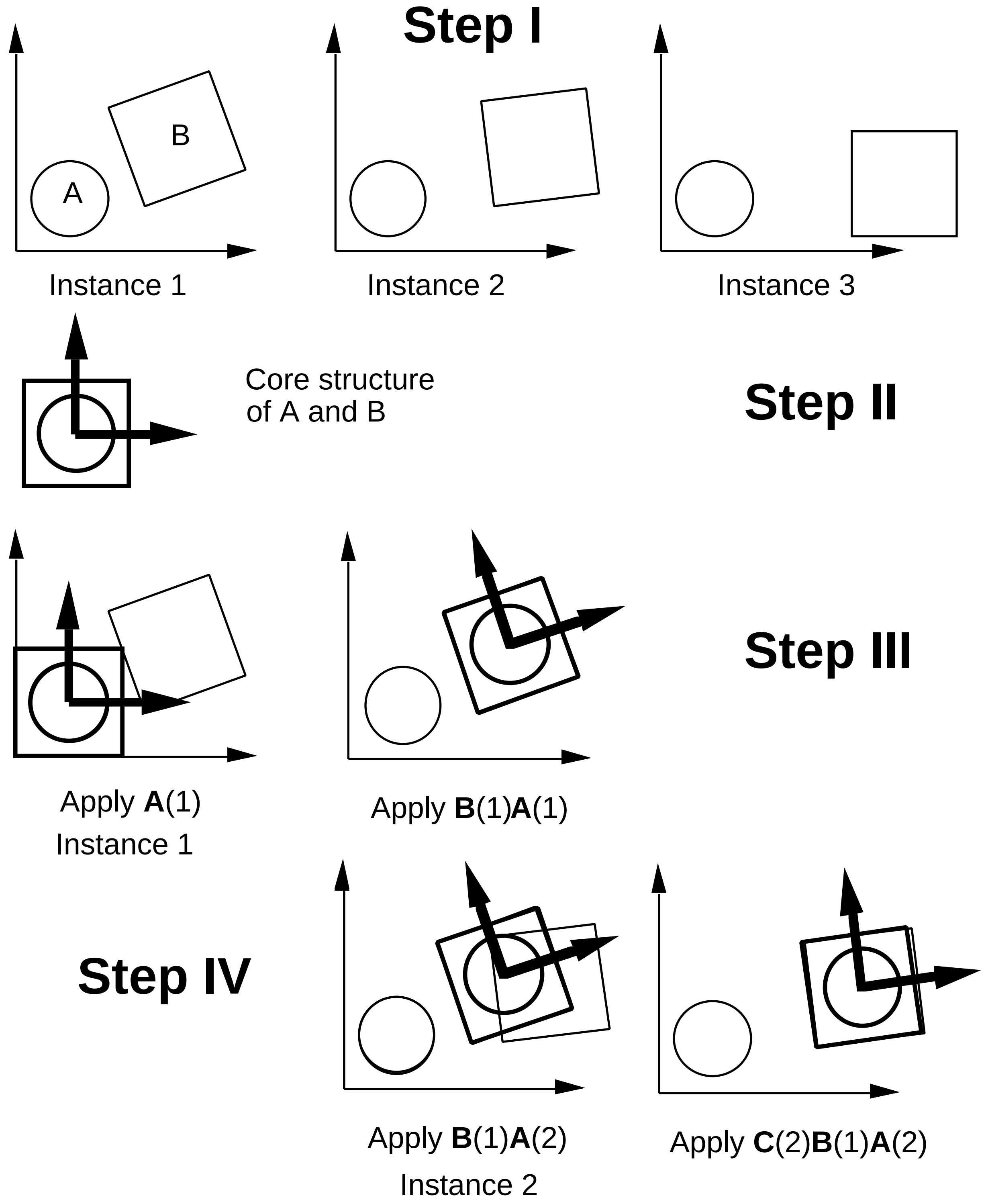


Figure 8 Schematic showing orientation averaging

The details of the orientation averaging procedure. As shown in Step I (at top), we start with a number of instances of a two-domain protein with the domains having different relative orientations (called instance 1, instance 2, and so forth). We then form the core structures of each domain (i.e. domain A and domain B), which are represented in Step II positioned on top of each other with centroids at the origin. Then in Step III, we fit the first core structure (denoted A) to first instance of the A domain. This involves a transformation $\mathbf{A}(1)$. Then we refit to superimpose the B core structure on the B domain. This involves transformation $\mathbf{B}(1)$. We would like to do the same thing for all other instances of domain A and domain B and then average all the $\mathbf{B}$ transformations to get an average transformation $\bar{\mathbf{B}}$. However, this leads to a problem since the $\mathbf{B}$ transformations often involve large rotations — for instance, in the case of the immunoglobulins, an almost 180° rotation is needed — and large rotations are not compatible with our averaging procedure. What is small and compatible with our procedure are the “differences” amongst the various $\mathbf{B}$ transformations. Consequently, as shown in Step IV, for all instances after the first instance, after applying the $\mathbf{A}$ transformation, we apply $\mathbf{B}(1)$. Using this positioning we then fit the B core structure to the B domain. This involves a third transformation $\mathbf{C}(i)$, which is the difference between the transformation required to superpose the B domain for the first instance and for instance i. After computing the $\mathbf{C}$ transformations for all instances except the first, we average them with our orientational averaging averaging procedure. Since the rotation in each $\mathbf{C}(i)$ is small we can average the corresponding quaternion just like a normal vector. After finding the average quaternion rotation and average transformation, we recombine them to produce the average transformation $\bar{\mathbf{C}}$. The average transformation relating the core structures of domains A and B is $\bar{\mathbf{B}} = \mathbf{B}(1)\bar{\mathbf{C}}$. To calculate the variation about the mean transformation $\bar{\mathbf{B}}$, we repeat step IV (above), but this time using $\bar{\mathbf{B}}$ in place of $\mathbf{B}(1)$. The “differences” we calculate, which we now denote $\mathbf{D}(i)$, give the transformation needed to superpose the B core on the B domain after applying the average transformation. (Note that unlike the procedure depicted in step IV, where $\mathbf{C}(1)$ was the identity transformation, it is necessary and possible to find a non-trivial value for $\mathbf{D}(1)$.)

Figure 8 (continued)



Instance 1

Instance 2

Instance 3

Apply **A**(1)
Instance 1

Apply **B**(1)**A**(1)

Apply **B**(1)**A**(2)
Instance 2

Apply **C**(2)**B**(1)**A**(2)