CABIOS

Vol. 11 no. 6 1995
Pages 633–644

Using a measure of structural variation to define a core for the globins

M.Gerstein and R.B.Altman[1,2]

Abstract

As the database of three-dimensional protein structures expands, it becomes possible to classify related structures into families. Some of these families, such as the globins, have enough members to allow statistical analysis of conserved features. Previously, we have shown that a probabilistic representation based on means and variances can be useful for defining structural cores for large families. These cores contain the subset of atoms that are in essentially the same relative positions in all members of the family. In addition to defining a core, our method creates an ordered list of atoms, ranked by their structural variation. In applying our core-finding procedure to the globins, we find that helices A, B, G and H form a structural core with low variance. These helices fold early in the folding pathway, and superimpose well with helices in the helix-turn-helix repressor protein family. The non-core helices (F and the parts of other helices that interact with it) are associated with the functional differences among the globins, and are encoded within a separate exon. We have also compared the variablity measure implicit in our core structures with measures of sequence variability, using a procedure for measuring sequence variability that helps correct for the biased sampling in the databanks. We find, somewhat surprisingly, that sequence variation does not appear to correlate with structural variation.

Introduction

The number of three-dimensional protein structures available in the structural database continues to increase, but the number of new folds per year is not increasing at the same rate (Orengo, 1994). As a result, the database is accumulating structures that are members of the same structural family (Levitt and Chothia, 1976; Richardson, 1981; Chothia and Finkelstein, 1990). There have been a number of efforts aimed at automatically identifying these families (Johnson *et al.*, 1990; Sander and Schneider, 1991; Pascarella and Argos, 1992; Holm *et al.*, 1993; Orengo *et al.*, 1993; Orengo *et al.*, 1994; Murzin *et al.*, 1995). One important opportunity that arises with the accumulation of large numbers of related structures is the ability to characterize them statistically. This paper is concerned with one such characterization, based on the positional variability of conserved atoms throughout members of the family. Given an alignment of a family of proteins (which establishes the correspondences between equivalent residues in each structure), we define a subset of atoms that have essentially fixed relative positions in all members of the family and call these the invariant structural core. The remaining non-core atoms have more variable relative positions, which may explain the functional differences between members of the family.

Protein cores, as we have defined them, are not precisely the same as the core definitions used in other work. Others have used measures of sequence conservation (Greer, 1990), conservation of structural and functional properties (Liebman, 1986), hydrophobic packing (Swindells, 1995), or density of contacts (Bryant and Lawrence, 1993) to define cores. Our definition is based purely on the observation that the relative positions of the core atoms are essentially fixed. Our cores have a number of potential uses. First, they can be used as a starting point in model building exercises. Once a new sequence has been aligned with any member of the family, then the core positions can be used to estimate the expected position of a subset of the residues. These positions provide an accurate scaffolding upon which the rest of the molecule can be modeled, using methods for elaborating the structure of loops (Jones and Thirup, 1986; Levitt, 1992) and for positioning sidechains given starting alpha-carbon positions (Lee and Levitt, 1991; Desmet *et al.*, 1992; Lee, 1994). Second, average core structures can be used as part of a library for inverse folding (or threading) applications, in which sequences are tested for compatibility with known folds. Many of these methods are sensitive to small variations in the backbone positions (Ponder and Richards, 1987; Sippl, 1990; Jones *et al.*, 1992; Bryant and Lawrence, 1993; Madej and Mossing, 1994). By using only those atoms whose structural variability is low, we can perhaps increase the sensitivity and specificity of the threading function. Finally, core structures may be useful in understanding the evolutionary relationships both within and between

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

[2]To whom correspondence should be addressed. Email: altman@camis.stanford.edu

Table 1. Families, structures, and ensembles
A Structures used

PDB	Protein	Species	Reference	chains	
2HHB	Hemoglobin (α & β chains)	Human	Fermi *et al.*, 1984	α	β
2LHB	Hemoglobin	Sea Lampry	Honzatko *et al.*, 1985		
1MBD	Myoglobin	Sperm Whale	Philips & Schoenborn, 1981		
2HBG	Hemoglobin	Bloodworm	Arents & Love, 1989		
1MBA	Myoglobin	Sea Hare	Bolognesi *et al.*, 1989		
1ECD	Hemoglobin	Chironomous	Steigemann & Weber, 1981		
2LH4	Leghemoglobin	Plant	Arutyunyan *et al.*, 1980		

All globin structures are of the deoxy form except for 1MBA and 2LHB. All the structures were taken from the protein databank (Bernstein *et al.*, 1977)

B Listing of structural ensembles used

Ensembles	Number of aligned atoms	Number of structures	Average, Min, and Max RMS between structures in ensemble (Å per atom)		
Globins					
α-carbons	115	8	2.19	1.22	3.16
mainchain atoms	460	8	2.18	1.22	3.13
all common atoms[a]	516	8	2.17	1.21	3.10

[a] 'All common atoms' means mainchain atoms for all 115 aligned positions plus the β and γ carbons that were conserved in all eight globin structures.

families. For example, shared core structures may be observed embedded within apparently different structural families. Others have created fragment libraries for proteins, but these are not usually at the level of entire folds (Prestrelski *et al.*, 1992). Core structures may also help distinguish regions that serve primarily structural roles from those that serve primarily functional roles.

In this paper, we extend a preliminary report on the analysis the core regions of eight globin molecules (Altman and Gerstein, 1994) and apply methods we previously used in the study of the immunoglobulins (Gerstein and Altman, 1995). We analyse five hemoglobin chains, two myoglobin chains and a plant leghemoglobin (detailed in Table 1). We demonstrate that the core defined using only alpha carbons is the same as that defined using all the backbone atoms, or all the backbone atoms plus all conserved sidechain atoms. We show that the core makes biological sense. In addition, we have used the spatial probability distributions for individual atoms to apply a distance measure between family members that is more sensitive than the traditional root-mean-square (RMS) measure. In particular, our 'calibrated' distance metric compensates for the observed variability that occurs within the globin family, and highlights differences in atomic positions that are unusual given the normal variation in position throughout the family. Finally, we show that our representations allow a comparison of the sequential diversity of an aligned set of residues (from a multiple alignment of protein family members) with their structural diversity. Using a procedure that helps correct for biases in the sequence databanks, we find that sequential diversity is not significantly correlated with structural diversity and discuss the implications of this finding.

The representation used in our method for defining core positions is based on three-dimensional mean and variance (Gaussian) in atomic positions, reminiscent of the anisotropic thermal ellipsoids that are sometimes used to summarize the position of atoms in crystallographic analyses. We have reported previously an algorithm that uses this representation for computing structure from numerous, uncertain data sources using a strategy of Bayesian combination of evidence (Altman and Jardetzky, 1989; Altman *et al.*, 1993). This algorithm has been compared with other methods for computing structure from distance information (Liu *et al.*, 1992), and has been used to compute a structure for the trp-repressor dimer (Arrowsmith *et al.*, 1991), the lac-repressor headpiece (Altman *et al.*, 1993) and cyclosporin (Pachter *et al.*, 1991) using NMR data. We have also described software for displaying structures represented using these probabilistic concepts (Altman *et al.*, 1995). The work reported here demonstrates the utility of this representation for representing and analysing aligned protein structures.

Systems and methods

The computations described in this paper were performed with Lucid Common Lisp, Perl, and C programs running in a unix environment. Much of the code was prototyped and developed in MacIntosh Common Lisp 2.0, and subsequently recompiled on a Hewlett Packard-720 (HP-720) for production runs. We are currently coding the whole method in ANSI C for general distribution.

Algorithms

The algorithms used in this paper fall into three categories: finding an average core, using the core to define a better RMS, and relating structural variation to sequence variation. Our core-finding algorithm starts with an ensemble of aligned structures, such as all the globin structures after they have been structurally aligned (Lesk and Chothia, 1980; Gerstein *et al.*, 1994) Such an alignment often contains some columns that are not completely populated (they may be the sites of deletions in some family members), and others for which every member of the family has an aligned residue. By definition, core positions should be present in all members of a family. Thus, we first remove all columns of an alignment that do not have representatives from every family member. The remaining set of positions is the set of positions which may be part of a structural core. In general, some of these aligned positions will have a high structural variation, and are thus not appropriately considered core atoms. Our technique identifies conserved atoms with high structural variation and removes them from the putative core.

Finding an average core

Our algorithm iteratively identifies the atom which is least likely to be core, and removes it from the list of candidate core atoms. The 'least likely' core atom is that atom which has the highest positional variation. We are then left with a list of the remaining atoms, from which the next noncore atom can be identified and thrown out. By repeating this procedure, we produce a rank order of atoms based on structural variability. The core of the family can then be defined by deciding the point at which noncore atoms are all thrown out, and only core atoms remain. We make this decision retrospectively after sequentially throwing out all atoms, and then examining the statistics of the core/noncore distributions that result at each iteration. The criterion for separating core from noncore atoms may vary, depending upon the uses to which the core will be put. The order of atom removal (the 'throw out order'), however, remains constant. The core-finding procedure is a generalization to multiple structures of the 'sieve-fit' procedure, previously developed for analysing protein motions (Chothia and Lesk, 1986; Gerstein and Chothia, 1991; Gerstein *et al.*, 1993a, 1993b). There are three key computations performed in core finding: (i) computing an unbiased average of a set of structures, (ii) computing the structural variability for each atom, and (iii) selecting a dividing point between core and noncore atoms.

1. Computing an unbiased average of an ensemble

A number of methods have been developed for superimposing an ensemble, Ω, of structures (Gerber and Müller, 1987; Kearsley, 1990; Diamond, 1992; Shapiro *et al.*, 1992). All these methods require an alignment which pairs each atom in one structure with an equivalent atom in the others. The methods then superimpose the centroids of the ensemble of structures, and determine a rotation for each structure such that the sum of squares of differences in coordinates between aligned atoms is minimized:

$$E(\Omega) = \sum_{j<k}^{N} \sum_{i=1}^{M} (\mathbf{R}_j \mathbf{x}_{ji} - \mathbf{R}_k \mathbf{x}_{ki})^2$$

where the outer sum is over all pairs j, k of the N structures in the ensemble Ω, the inner sum is over the M aligned positions in each structure, and $R_j x_{ji}$ are the rotated coordinates of structure j. The previously reported methods are difficult to program and may not parallelize well. We have developed a new method which is less efficient, but which uses only repeated calls to a basic RMS-fitting routine.

1. Start with an ensemble of N structures
2. For each structure in the ensemble,
   - A. perform a standard RMS fit of all other (N−1) structures to it (Arun *et al.*, 1987).
   - B. Compute the average coordinates of the selected structure, and the N−1 fitted structures.
3. Compare the minimal RMS deviation between the N average coordinates that result from fitting to each of the N structures in the ensemble. If the coordinates are all the same, to within some predefined threshold, then they constitute an unbiased average. If the coordinates are not the same, then using the average structures computed in 2B, loop to the top of step 2.

This procedure works because the average structures computed in step 2B are different from one another, but are more similar to each other than were members of the original ensemble. By repeating this procedure of fitting to each structure (and averaging), we can create a set of structures that are essentially identical, and are an unbiased average of the original starting structures. We use a predefined threshold value of 10^{-6} Å as the stopping condition. The unbiased nature of the method is evident, since there is no order dependence in the procedure or in the way in which the list of structures is ordered. The computational complexity of this approach requires $O(N^2)$ RMS fits (where N is the number of structures), since each structure is fit to all other structures. In practice, no more than three iterations are required for convergence.

2. Computing the structural variation for each atom

Given N structures that contain M conserved atoms, we summarize the structural variation for the conserved

atoms by fitting them to an unbiased average, and then calculating a three-dimensional ellipsoid volume that encloses the atoms. The volume is computed from a 3×3 variance/covariance matrix for the coordinates (x, y and z) of each atom. This matrix contains the variance of each individual coordinate in its diagonals, and the covariances between coordinates in its off-diagonals. The covariance matrix is symmetric, positive definite and can be diagonalized to give the variances along the principle axes of the constellation of atoms (Altman *et al.*, 1995). Assuming a three-dimensional normal distribution of atoms, the volume V that contains more than 96% of the atoms at two standard deviations is:

$$V = \frac{4}{3}\pi(2\sigma_x)^2(2\sigma_y)^2(2\sigma_z)^2$$

Atoms with large spatial variations after alignment of the N structures will have large volumes, and those with small spatial variation will have small volumes. The atom with the highest spatial variation is least likely to be part of core, and is removed from the list of candidate core atoms.

3. Selecting a core cutoff

The process of defining the structural variation for each atom, and removing the atom with the largest variation, results in a rank ordering of atomic variability from most variable to least variable. This order is intrinsic to the family of proteins and specific alignment used. However, for some purposes, it may be useful to define a threshold for separating atoms that should be considered core from those that should be considered noncore. The criterion used for this threshold may vary, and is somewhat arbitrary. The simplest criterion is one based only on the size of the ellipsoid enclosing the positions for an atom. Thus, we could choose an ellipsoid volume (such as $1.0\ \text{Å}^3$) as a threshold and include those atoms whose spatial variation occurs in this or smaller volumes. This criterion suffers because it does not recognize more natural divisions between core and noncore populations. Thus, we might choose a criterion based on the properties of atoms that have been discarded. We have previously suggested that the variance in noncore ellipsoid size would have a maximum when atoms that are properly considered 'core' are added to the list of noncore atoms (Altman and Gerstein, 1994). For example, if we assume that core atoms have small, homogenous ellipsoids of variation, then adding members of this homogenous population to a heterogeneous population of highly variable ellipsoids will reduce the overall variation. We showed that this criterion yields a reasonable core definition for the globins. A third criterion for a core cutoff combines elements of the first two criterion: we seek a threshold that maximizes the separation between the distribution of the volumes of the ellipsoids of variation for core and noncore atoms, and that yields a relatively homogenous population of core ellipsoids. In the case of the globins, all three of these criteria yield very similar core/noncore thresholds.

Using the Core to Calculate a 'Better RMS'

Having defined a set of core atoms for a family, we can use the core atom positions to get a high quality superposition of the family members—and thus highlight the regions which differ in detailed structure. If we superpose with all the conserved atoms (instead of only those conserved atoms with low structural variability) then our superposition distributes errors across all the atoms, and can not distinguish between structurally conserved regions, and those that are variable. Such a superposition would not be useful for understanding the detailed ways in which two members of a family differed. For example, a position by position analysis of the deviations would be relatively uninformative because the error that is primarily due to highly variable regions is distributed over the entire structure. The standard RMS deviation that would be reported from such an alignment would reflect the average deviation of all atoms, without recognizing that some atoms have very low deviations, and others have much higher deviations.

An alignment of structures using only core atoms allows us to identify and examine the structural deviations of variable regions, and provides a much more useful position by position analysis. In fact, the measured deviations between atoms can be calibrated by scaling the deviation between two atoms at a position by the statistical variation in the family at that position. For example, if the vector separating two atoms has a length of 1.0 Å in a certain direction, and if the known variance along that direction is $4\ \text{Å}^2$, then the calibrated distance between the two atoms would be 0.5 standard deviations ($1\ \text{Å} \times 1\ \text{SD}/\sqrt{4\ \text{Å}^2} = 0.5\ \text{SD}$, well within the normal variation seen in this family. If, on the other hand, the vector separating the atoms has a length of 4.0 Å, then the calibrated distance would be 2.0 SD, indicating that this difference is large, even by the standard of usual variation within the family. Thus, we can plot a position by position analysis of the distance in units of standard deviation, and determine which atoms are farther apart (or closer together) than is usual within the family. As in the case of unscaled distances, we can summarize all the standard-deviation distances between two structures as in terms of a single number, the RMS of all these SD-distances (i.e. the SD-RMS). It is interesting to note that the SD-RMS value between a structure and the average core measures the degree to which the structure is a typical member of the protein family.

Although SD units and the SD-RMS are generally useful, it may be desirable to also have some measurement in units of Ångstroms that can be related back to structural units. We have found it particularly useful to scale all the SD deviations in order to produce a 'calibrated' deviation whose RMS value is the same (in units and in value) as the standard RMS value used when summarizing the deviations between two aligned structures. In particular, the calibrated distance, d_{cal}, between two corresponding atoms in two structures is given by:

$$d_{cal} = \frac{D_{RMS}}{D_{SD}} d_{sd}$$

where D_{RMS} is the conventional RMS distance (in Ångstroms) computed over all atoms, D_{SD} is SD-RMS value computed over all atoms, and d_{SD} is the scaled distance (in SD units) between the two corresponding atoms. The calibrated distance reflects more accurately which atoms should be assigned responsibility for the overall deviations.

Relating sequence variation to structural variation

The ellipsoid volumes provide a measure of the structural variability for each aligned position in a protein family. Analogously, a variety of approaches can be used to quantify the degree of sequential variability for each aligned position. Most commonly, these are based on the concept of an information-theoretic entropy (Schneider *et al.*, 1986; Schneider and Stephens, 1991; Shenkin *et al.*, 1991). The entropy of column i in a multiple sequence alignment is derived from calculating frequencies f(i,t) of amino acids of a given type (t) in this column:

$$H(i) = -\sum_{t=1}^{20} f(i,t) \log_2 f(i,t).$$

However, the sequence databanks typically contain a biased representation of sequences, which adversely affect the computation of reasonable frequencies. That is, for a given protein, some species are over-represented and others are under-represented. There are, for instance, usually many more human sequences than dog sequences. Methods have been developed to correct for this 'biased sampling' within a multiple alignment. We have previously described one such method (Gerstein *et al.*, 1994) which is based on weighting each sequence by its position within an evolutionary tree. (See Vingron and Sibbald, 1993 for a general discussion of weighting schemes.) To incorporate our weights into calcuation of sequence variability, we simply take f(i,t) in the above formula to be the normalized sum of the weights w(j) for sequences with a residue of type t in position i:

$$f(i,t) = \frac{\sum_{j(t \text{ fixed at } i)} w(j)}{\sum_{j} w(j)}$$

where the denominator sum is over all sequences j in the alignment and the numerator sum is over just those sequences that have a residue of type t at position i.

Results

As shown in Table 1A, we chose eight structures from the globin family for our calculations. This set of structures had been previously aligned manually (Lesk and Chothia, 1980)

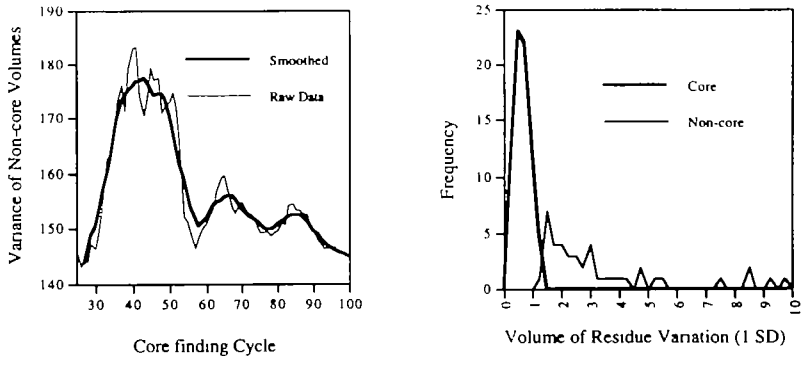


Fig. 1. Progress of the core finding procedure in the globins. (**A**, left) After fitting the noncore atoms to one another, we plot the variance in the volume of the ellipsoids in order to identify the core for the globins. We perform a 5-residue moving average in order to smooth the curve, and then selected local maximal. As discussed in the text, the variance of the noncore ellipsoids peaks at the 'core' threshold. For the globin family, we observed maxima at cycles 42, 64 and 84, corresponding to our cores containing helices A, B, E, G, and H, helices A, B, G and H, and helices A and B, respectively. (**B**, right) In order to confirm the choice of core cutoff at cycle 42, we plot the distribution of ellipsoid volumes for core atoms and noncore atoms, to evaluate the degree to which the cutoff separates two populations with distinct volume distributions. The average core ellipsoid has a volume of volume ~ 0.8 Å^3 while the non-core ellipsoids have an average volume of 3 Å^3. In addition, the noncore volumes are broadly distributed, while the core ellipsoids are tightly distributed.

Fig. 2. Average core structures of the globins The globin core. The view is roughly the same as the schematic drawing shown in Fig. 3. The mean position of all 115 atoms is shown along with their ellipsoids of variation drawn at two standard deviations (left) Relatively large ellipsoids are shown around the 42 Cα atoms classified as not belonging to the globin core (center). Smaller ellipsoids are shown around the 73 atoms that are classified as belonging to the core (right). The core structure has acceptable stereochemistry The Cα–Cα virtual bond length averages 3.76 Å, with a standard deviation of 0.034 Å.

using a canonical numbering scheme, and had been subject to a number of subsequent investigations (Bashford *et al.*, 1987; Gerstein *et al.*, 1994). We first ran the core finding algorithm on 115 α-carbons corresponding to the aligned positions. To test the sensitivity of our method to larger sets of atoms, we ran our core finding procedure on an ensemble containing more than just α-carbons. We ran the procedure on the full set of backbone atoms from the 115 aligned residues (a total of 460 atoms), as well as on a set that included all backbone atoms in addition to conserved sidechain (β and γ) carbons (a total of 516 atoms). These data sets are summarized in Table 1B. After calculating two new globin cores, we compared them to our original α-carbon core and found them to be almost identical. In particular, we performed a Spearman rank correlation (Press *et al.*, 1992) on the 'throw-out' order of the 115 α-carbons in both runs and found an almost perfect correlation (0.99). We also correlated the throw-out order of different types of atoms (i.e. mainchain C with mainchain O) in the all-atom run. We found that the correlation of α-carbon throw-out order with any of the other atoms in a residue was greater than 0.93, demonstrating that all the atoms in residue tended to be thrown out as a unit. This correlation in throw out order, in turn, suggests that α-carbons are sufficient to define the core structure.

Defining a globin core

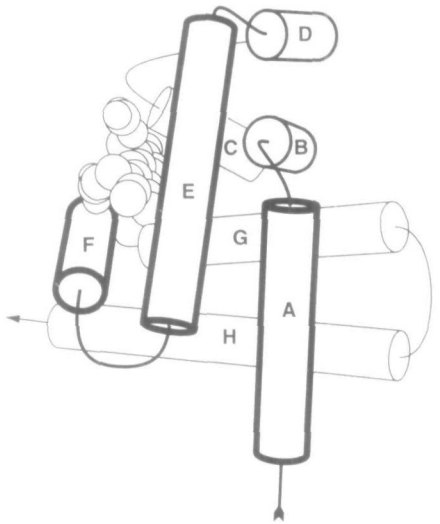
Figure 1 demonstrates two lines of evidence indicating a natural division between core and noncore atoms at cycle 42 of core finding. Figure 1A plots the variance of the atoms that have been removed (fitting them to their unbiased average), and shows a peak at cycle 42. In subsequent cycles (43 and beyond) the variance of the noncore atoms decreases, suggesting that a population of homogenous atomic volumes is being added to the list of noncore atoms. Thus, cycle 42 marks the point at which the noncore list contains the most variation, and the core list has a relatively homogenous population of 73 remaining ellipsoids. Figure 1B compares the distribution of ellipsoid volumes for the 73 core atoms and the 42 noncore atoms. The core atoms form a spike with average ellipsoid volume around 0.8 $Å^3$, while the noncore atoms have much broader distribution with an average volume around 3 $Å^3$. The overlap between these two distributions is quite small. Thus, there seems to be a reasonable core threshold at cycle 42.

In Figure 1A, we also note peaks at cycles 64 and 84, in addition to the primary core peak at cycle 42. These peaks suggest that there are two "secondary cores" within the primary core. That is, there are subpopulations of atoms which have still smaller spatial variation than the primary core, and whose variances cluster even more tightly. The smallest core contains 31 α-carbons from helices A, B and part of G; the next, intermediate core is a superset of this, containing only the A, B, G and H helices.

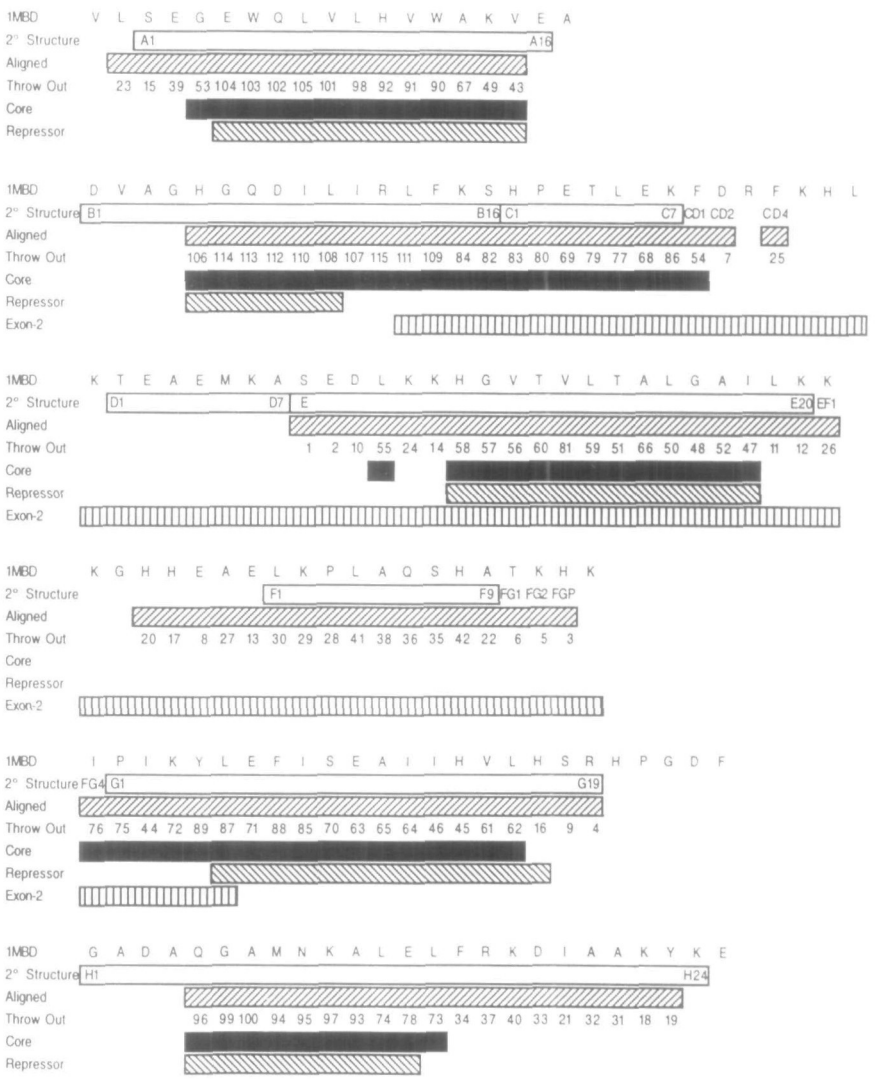
The error ellipsoids for the 73 core and 42 non-core atoms are shown graphically in Figure 2. The core

Fig. 3. Biological significance of globin core. (**A**) Cylinders representation of a globin showing standard helix labeling scheme (1MBD). (**B**) Graphical depiction of the relevant subsequences of globin family. (The residues of 1 MBD are shown for reference purposes.) The set of 115 conserved globin residue positions in the standard alignment of Lesk and Chothia (1980) are labeled (ALIGNED row). The conserved residues encompass all the globin helices, except the D helix, which is often not present. The boundaries of the helical secondary structures are labeled (2° STRUCTURE row). If the core cutoff is set at cycle 42, as discussed in the text, then there are 73 core residues for the globins, which are labeled (CORE row). The iteration at which each of the 115 aligned residues was removed during the corefinding procedure is also labeled (THROW OUT row). The 52 positions in the repressor protein which align well with the globins is shown (REPRESSOR row). Finally, the location of the second exon for myoglobin, which primarily codes for the noncore segments of the globins, is also labeled (EXON-2 row).

A


D
C
B
E
G
F
A
H

B


1MBD V L S E G E W Q L V L H V W A K V E A
2° Structure A1 A16
Aligned
Throw Out 23 15 39 53 104 103 102 105 101 98 92 91 90 67 49 43
Core
Repressor
1MBD D V A G H G Q D I L I R L F K S H P E T L E K F D R F K H L
2° Structure B1 B16 C1 C7 CD1 CD2 CD4
Aligned
Throw Out 106 114 113 112 110 108 107 115 111 109 84 82 83 80 69 79 77 68 86 54 7 25
Core
Repressor
Exon-2
1MBD K T E A E M K A S E D L K K H G V T V L T A L G A I L K K
2° Structure D1 D7 E E20 EF1
Aligned
Throw Out 1 2 10 55 24 14 58 57 56 60 81 59 51 66 50 48 52 47 11 12 26
Core
Repressor
Exon-2
1MBD K G H H E A E L K P L A Q S H A T K H K
2° Structure F1 F9 FG1 FG2 FGP
Aligned
Throw Out 20 17 8 27 13 30 29 28 41 38 36 35 42 22 6 5 3
Core
Repressor
Exon-2
1MBD I P I K Y L E F I S E A I I H V L H S R H P G D F
2° Structure FG4 G1 G19
Aligned
Throw Out 76 75 44 72 89 87 71 88 85 70 63 65 64 46 45 61 62 16 9 4
Core
Repressor
Exon-2
1MBD G A D A Q G A M N K A L E L F R K D I A A K Y K E
2° Structure H1 H24
Aligned
Throw Out 96 99 100 94 95 97 93 74 78 73 34 37 40 33 21 32 31 18 19
Core
Repressor

contains all atoms from helices the A, B and most atoms from helices C, E, G and H. The core does not contain helix F, the conserved loop regions, or the ends of helices E, G and H, which are near the heme group. If one computes the order in which helices are removed from the putative core (by computing the average throw out position of residues within the helices), then the helices can be ranked from most core-like to least. The helix that is most positionally invariant throughout the globin family is helix B, followed by helices A, H, G, C, E and F.

As mentioned in the introduction, one use of our core positions for alpha carbons is as starting points for model building. As such, they should satisfy the most basic requirements in terms of stereochemistry. Thus, we confirmed that alpha carbons of neighboring residues occur at the standard distance of 3.8 Å (with a standard deviation of 0.03 Å). We also confirmed that the bond angle between three neighboring alpha carbons ranges between 88° and 122° (with the normal range being 80° to 135°).

Figure 3 provides a summary of the key biological features of the globin family, as they relate to our core computation. The sequence of a representative myoglobin, 1MBD, is shown in the first row. The position of conserved secondary structural elements, as determined by manual structural alignments, is shown in the second row, along with the location of the common 115 residues in the standard alignment of Lesk and Chothia (1980) in the third row. Not unexpectedly, the common residues occur in the regions of secondary structure for the most part. Only the D helix of the globins is 'optional' in the sense that some globin family members do not have a D helix. The rank order of spatial variation for each residue in the common alignment of 115 residues is shown in the fourth row. Residues with high spatial variation have low ranks, and residues with low spatial variation have high ranks (since they are thrown out last). The 72 core residues in the globins, as defined by our procedure, are highlighted in the fifth row of Figure 3. The core residues are not randomly dispersed throughout the structure, but are grouped together in segments that are structurally related. No residue within helix F is part of the core. In contrast, almost the entirety of helices A, B, C, and G are contained in the core. Most of helix E and half of helix H are part of the core.

In order to evaluate the concordance of our core residues and the core elements of the helix-turn-helix repressor motif, the sixth row of Figure 3 marks the atoms in the globins that have structural equivalences in the repressor proteins. Large portions of helices A, B, E, G and H are present both in the repressor proteins, and in

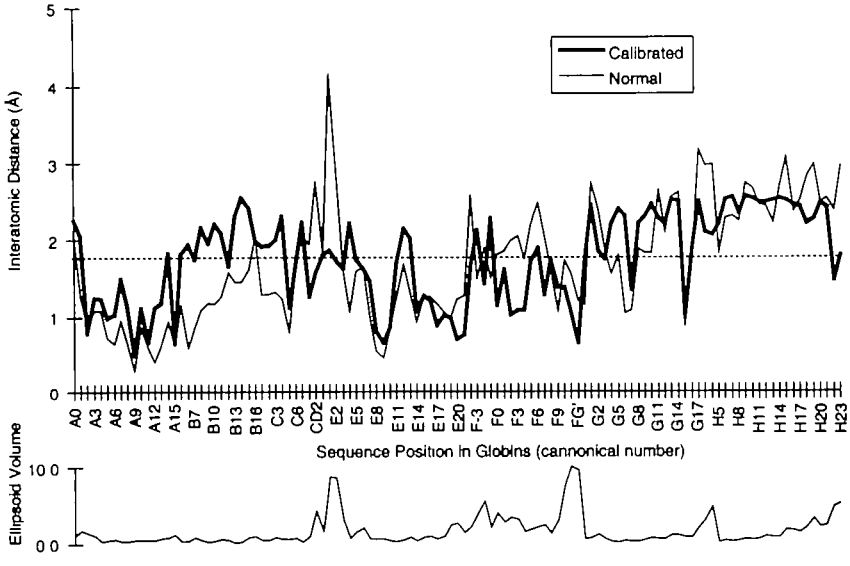


Fig. 4. Structural deviations in calibrated Ångstroms versus real Ångstroms. The relationship between normal distance in Ångstroms and a 'calibrated Ångstroms' distance, normalized using the ellipsoids of variation for each atom. The thin line shows absolute distance deviations (in Å) between the line corresponding atoms in two globins (1MBD and 1ECD) after they have been fit to the core structure The trace at the very bottom of the graph shows the volume of variation (expressed as the volume of the error ellipsoid in cubic Å) for each position in the globins, calculated after superimposing all structures using the core atoms. 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 the same (by definition, as discussed in the text) and are represented here by the horizontal dotted line. Note that in the B helix, a region of the globin structure where there is little variation within the family, the normal distances between 1ECD and 1MBD 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 F helix, where the calibrated distances are less than the normal distances. This indicates that 1ECD and 1MBD have significant relative differences in the position of the normally invariant B helix, but that the apparently large differences in position of the F helix are representative of the observed variation in F helix position.

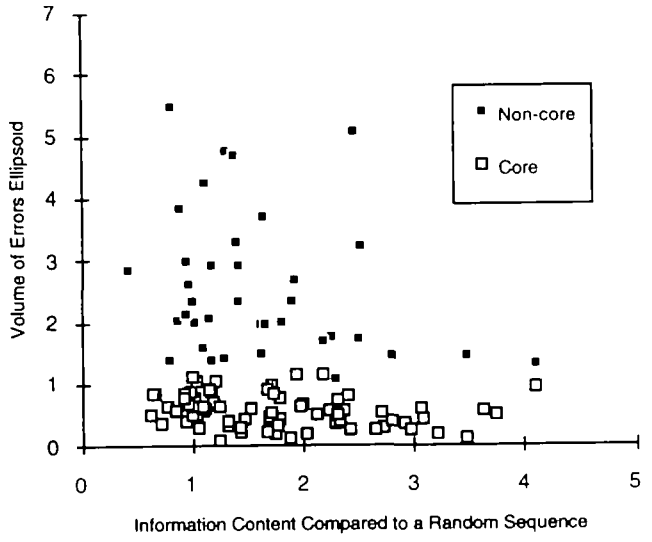


Fig. 5. Sequence variation versus structure variation in the globins. Graph of structural variability versus corrected sequential entropy for each aligned globin position. At each position, the structural variability is the volume of the ellipsoid of variation for each position (after the structures are superimposed using the core atoms) Sequence variability is computed, as described in the methods section, from a structurally based alignment of 577 globin sequences reported in Gerstein *et al* (1994). The entropy is measured in bits per residue as the information content of a given position i in the alignment relative to that if the sequences were aligned randomly;

$$R_{sequence}(i) = \sum_{t=1}^{20} \bar{f}(t)\log_2 \bar{f}(t) + H(i)$$

where $\bar{f}(t)$ is the average frequency of residue type t in the alignment, and $H(i)$ is the entropy of position i as defined above. For the globins, there are 115 positions represented in total here and the overall Pearson correlation coefficient between structural and sequential variability is 0.12. The 73 core positions are highlighted by white boxes. The correlation between information content and ellipsoid volume for just the core positions is 0.25. 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).

the core as we have defined it. In fact, the section of helix H that is part of our core is also present in the repressor proteins. Helices C and F are distinctly absent from both the repressor proteins, and from our globin core.

In order to evaluate the concordance of our core residues with the exon structure of the globins, the final row of Figure 3 marks the location of the second exon for the globins in humans. Helices F, E an C are encoded by the second exon. The other helices, A, B, G and H, are encoded by the flanking exons, and are part of the 'core of the core' mentioned above.

Using the core to compare family members

Figure 4 dramatically illustrates the value of the calibrated Ångstrom measure in comparing two structures on an atom by atom basis: we compare 1MBD and 1ECD at each of the 115 aligned positions using the standard Cα-Cα distance and the calibrated Ångstroms distance. In regions of little structural variability between the globins, such as in the B helix at position B4, 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 variable regions, such as the F helix at position F3, where the calibrated distance is less than the normal distance. The RMSD between 1MBD and 1ECD is 1.61 Å which corresponds to an SD-RMS value of 2.28.

Using the core to compare sequential and structural variation

Figure 5 shows the relationship between sequence variation, measured by our weighted entropy, and structural variation, measured by ellipsoid volume. As discussed in the figure caption, there is no significant correlation between them. This is true whether we consider all 115 aligned positions, the 73 core positions, or just the 31 core positions that are buried in all the globin structures.

Discussion

The impressive concordance of the throw-out order of residues using either alpha-carbons only, all backbone atoms, or backbone plus conserved sidechain atoms indicates that our procedure is detecting an important biological signal. Indeed, our results indicate that not only is there a very strong correlation in the throw-out position of atoms within residues (that is, the atoms in a particular residue tend to be thrown out as a group) but also residues within helices tend to be thrown out as a group (as shown in Figure 3). Thus, there is considerable evidence that the throw-out order of residues reflects fundamental biological properties of the globin family, and is quite illuminating.

The most striking characteristic of the globin throw-out order is that helix F and the ends of helices E, G and H are thrown out early and are not part of the core. The primary function of the globin family of proteins is to bind and transport oxygen at the heme group. The primary role of helix F (and the ends of E, G and H) is to coordinate the heme group and provide basic structural support. Thus, the heme binding site is essentially the functional active site for the globins—the area where the detailed functional characteristics of individual globins are manifested. Differences in oxygen affinity can, in large part, be attributed to differences in the orientation and environment of the heme group. It is, therefore, not surprising that the helices which determine this environment are not part of the conserved structural core. Although there are

strongly conserved residues in helix F which affect oxygen binding, the precise position of these residues is not always the same, and may be related to the detailed differences in function. The special nature of helix F has also been demonstrated in theoretical packing studies that use ideal polyhedra (Murzin and Finkelstein, 1988). They show that most proteins made only of alpha helices obey simple packing rules, but that the position of Helix F in the globins violates these rules (while the other helices satisfy them).

It is reasonable to wonder whether the invariant structural core of the globins is also the region that is most stable or folds earliest. Recent NMR experiments on myoglobin have indicated that helices A, B, G and H (the least variable of our core helices) form a stable "molten globule" very quickly (Jennings and Wright, 1993, Loh *et al.*, 1995), and that helices F and D fold last (Cocco and Lecomte, 1990; Jennings and Wright, 1993). The degree of similarity of the initial molten globule and the final globin positions is not precisely known, but it is not surprising that the structurally conserved portions of the globins would start to fold relatively early. In addition, the heme binding helices may prefer folding in the presence of the heme group, so their relatively late folding, and the heterogeneity of their positions over the globin family is not surprising.

Our invariant structural core not only meshes nicely with the evidence from globin function and folding, but also with the gene structure of the globins. For example, human myoglobin is encoded by three exons. The first exon encodes most of helices A and B, the second exon encodes most of helices C through F, and the final exon encodes helices G and H. Thus, the most structurally invariant elements (A, B, G and H) are encoded by the first and third exons, whereas the elements that are more variable (especially the highly variable helix F) are contained in the middle exon. Our observations lend some credence to the hypothesis that the exons may provide structural units (Gilbert, 1985)—although clearly at a level below entire domains.

The idea of a reusable structural scaffolding, that is subsequently specialized with the addition of a extra segments is also supported by recent observations on the helix-turn-helix (HTH) family (Subbiah *et al.*, 1993). This family, a common structural motif used for DNA binding, has a subset of five helices that align well with globin helices A, B, E, G and H. Specifically, the two helices in the HTH motif directly correspond to globin helices B and E, while the other three structural helices correspond to helix A and parts of helices G and H. The correspondence with our core structure is shown in Figure 3. Other proteins, such as the phycocyanins and colicin A have also shown structural similarity to the globins (Holm and Sander, 1993). Unfortunately, there are not yet sufficient numbers of structures in these families to compute a reliable core and compare it with the globin core. The similarity between the globin family and the HTH family may be based on the fact that helices tend to pack together in certain ways (Chothia *et al.*, 1981), and does not necessarily imply any evolutionary relationship between these families. In either case, however, the idea of a basic structure that can be augmented as functional requirements change is appealing.

Structural similarity clearly correlates with sequence similarity at the level of the overall fold, and this is the basis for defining families of proteins based on sequence homology. However, with regard to the globins, we find that sequence variation is not correlated with structural variation in terms of the detailed positioning of atoms measured by our error ellipsoids. This result may be to some extent influenced by the fact that globins have a particularly large 'active site', i.e. the heme binding pocket, and hence an inordinate amount of residues conserved for functional as opposed to structural reasons. However, we have also observed a lack of correlation in the immunoglobulin fold (Gerstein and Altman, 1995).

We are able to detect a set of backbone atoms whose positions are relatively invariant, despite large differences in sequence, thus indicating that proteins probably do not accommodate mutations by making only local changes in the sidechain conformation. Instead, our results suggest that there is a much more subtle, global adjustment of atom positions (within a rather small volume of variation) that allows all members of the globin family to maintain a core backbone structure (and the associated pattern of hydrogen bonding) that is essentially invariant. Similar observations have been made in the case of T4 lysozyme (Eriksson *et al.*, 1992; Baldwin *et al.*, 1993). Our observation that sequential variability is not significantly correlated with structural variability demonstrates that conserved residues can have large positional variation within a family of proteins, and variable residues can have a very small positional variation—and only a small effect on the overall backbone fold. It is possible that sequential positions may sometimes be conserved precisely to maintain a degree of flexibility that is important for functional reasons. Other times, sequential conservation may be to maintain structure. In any case, our results indicate that it is dangerous to draw detailed structural conclusions based only on the presence or absence of sequential conservation.

Our core-finding procedure relies on an estimate of three-dimensional variance to determine which atoms are least likely to be part of the core. Although the calculation of variance is free of any assumption about the form of the distribution, there are parts of our method in which a

Gaussian assumption is implied. For example, our interpretation of standard deviation (1SD boundaries contain more than 70% of distribution, 2 SD boundaries contain more than 96%) for the purposes of display as in Figure 2, relies on a Gaussian assumption. In addition, interpretation of the SD-RMS values is based partly on such an assumption. It is difficult to assess accurately the validity of a Gaussian assumption for the distribution of positions of only eight globins. However, we have reported elsewhere that the deviations of atomic positions for a larger set of aligned proteins are indeed distributed in a roughly Gaussian manner (Gerstein and Altman, 1995). Although it is clearly possible for atoms to have bimodal (in general, multimodal) distributions, our results indicate that the Gaussian assumption is reasonable. We have described a method for representing multimodal atomic distributions using an extension of the representations used here (Altman *et al.*, 1994).

Conclusion

We have applied a new method for defining the average structural cores for proteins to eight members of the globin family. Our method is based on a simple statistical analysis of the variability of atoms in a structural alignment. We have provided a rank ordering of atoms for structural variability, and chosen a threshold to separate core from noncore atoms. Our cores are biologically relevant, and are consistent with our understanding of globin function, folding pathways and gene structure. By looking at the structural variability within the globin family, we are able to divide the globin fold into two parts (the variable part that coordinates the heme group and the relatively invariant part that is the conserved structural scaffolding). We observe that the less variable part is similar in structure to the repressor proteins, is reflected in the gene structure of the globins, and is the part of the globins that folds first. Conversely, the non-core atoms of the globin family are those involved in determining the detailed functional differences between family members.

Availability

We make available the coordinates of the globin core; 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 altman@camis.stanford.edu or gerstein@camis.stanford.edu through anonymous ftp to the following URL:

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

Acknowledgements

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

References

Altman,R. and Jardetzky,O (1989) The heuristic refinement method for the determination of the solution structure of proteins from nmr data. Nuclear magnetic resonance, Part B: Structure and mechanisms. In Oppenheimer,N.J and James,T.L. (eds), *Methods of Enzymology*, Vol 177. Academic Press, New York, pp. 177–218.

Altman,R.B., Pachter,R. and Jardetzky,O. (1993).Structural uncertainty of proteins in solution by NMR: a re-evaluation of the structure of the lac repressor headpiece. *Appl. Magn. Reson.*, **4**, 441–460.

Altman,R.B., Chen,C.C , Poland,W.B. and Singh,J P. (1994) Probabilistic constraint satisfaction with non-Gaussian constraint noise. In Lopez de Mantaras,R. and Poole,D. (eds), *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, pp 15–22

Altman,R and Gerstein,M. (1994) Finding an average core structure: application to the globins. *Proc. Second Int. Conf. Intell. Sys. Mol Biol.*, AAAI Press, Menlo Park, pp 19–27.

Altman,R.B., Hughes,C and Gerstein,M.B. (1995) Methods for displaying macromolecular structural uncertainty· application to the globins. *J Mol Graphics*, **13**, 142–152

Arents,G A and Love,W.E (1989) glycera dibranchiata hemoglobin. structure and refinement at 1.5 Å resolution. *J. Mol. Biol.*, **210**, 149.

Arrowsmith,C , Pachter,R., Altman,R. and Jardetzky,O. (1991) The solution structures of *E.coli* trp repressor and trp aporepressor at an intermediate resolution. *Eur. J. Biochem.*, **202**, 53–66

Arun,K S., Huang,T.S. and Blostein,S.D. (1987) Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**, 698–700

Arutyunyan,E.G , Kuranova,I.P., Vainshtein,B.K and Steigemann,W. (1980) X-ray structural investigation of leghemoglobin. VI Structure of acetate-ferrileghemoglobin at a resolution of 2.0 Å. *Kristallografiya* (USSR), **25**, 80.

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.

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

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.

Bolognesi,M , Onesti,S., Gatti,G., Coda,A., Ascenzi,P. and Brunori,M. (1989) Aplysia limacina myoglobin. Crystallographic analysis at 1.6 Å resolution. *J. Mol Biol.*, **205**, 529.

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

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

Chothia,C., Levitt,M. and Richardson,D. (1981) Helix to helix packing in proteins. *J. Mol Biol* , **145**, 215–250.

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

Cocco,M.J. and Lecomte,J.T.J (1990) Characteriztion of hydrophobic cores in apomyoglobin: a proton NMR spectroscopy study. *Biochemistry*, **29**, 11067–11072

Desmet,J., Maeyer,M.D., Hazes,B and Lasters,I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, **356**, 539–542

Diamond,R.D. (1992). On the multiple simultaneous superposition of molecular structures by rigid-body transformations. *Protein. Sci.*, **1**, 1279–1287

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.

Fermi,G., Perutz M.F , Shaanan,B and Fourme,R (1984) The crystal structure of human deoxyhaemoglobin at1.74 Å resolution. *J. Mol. Biol.*, **175**, 159

Gerber,P.R. and Muller,K (1987) Superimposing several sets of atomic coordinates. *Acta Cryst.*, **A43**, 426–428.

Gerstein,M and Altman,R. (1995) Average core structures and variability measures for protein families, Application to the immunoglobulins. *J. Mol. Biol.*, **251**, 161–175.

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. (1993a) Domain closure in lactoferrin, two hinges produce a see-saw motion between alternative close-packed interfaces *J. Mol Biol.*, **234**, 357–372.

Gerstein,M.,Schulz, G. and Chothia,C. (1993b) Domain closure in adenylate kinase, joints on either side of two helices close like neighboring fingers. *J. Mol Biol.*, **229**, 494–501.

Gerstein,M., Sonnhammer,E. and Chothia,C (1994) Volume changes on protein evolution. *J Mol Biol.*, 236, 1067–1078.

Gilbert,W. (1985) Genes-in-pieces revised. *Science*, **229**,823–824.

Greer,J. (1990) Comparative modeling methods, application to the family of mammalian serine proteases. *Proteins*, **7**, 317–334.

Holm,L., Ouzounis,C , Sander,C., Tuparev,G. and Vriend,G (1993). A database of protein structure families with common folding motifs *Protein Sci.*, **1**, 1691–1698.

Holm,L. and Sander,C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol. Biol.*, **233**, 123–128.

Honzatko,R.B., Hendrickson,W.A and Love,W E. (1985) Refinement of a molecular model for lamprey hemoglobin from *Petromyzon marinus*. *J. Mol. Biol.*, **184**, 147.

Jennings,P.A. and Wright,P E. (1993) Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science*, **262**, 892–896.

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.

Jones,T.A. and Thirup,S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.*, **5**, 819–822.

Kearsley,S K. (1990) An algorithm for the simultaneous superposition of a structural series. *J Comp. Chem.*, **11**, 1187–1192.

Lee,C. (1994) Predicting protein mutant energetics by self-consistent ensemble optimization *J. Mol. Biol.*, **236**, 918–939

Lee,C. and Levitt,M. (1991) Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature*, **352**, 449–451.

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. (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

Liebman,M.N. (1986) Structural organization in the serine proteases. I. Macromolecular specificity in limited proteolysis *Enzyme*, **36**, 115–140.

Liu,Y., Zhao,D., Altman,R. and Jardetzky,O. (1992) a systematic comparison of three structure determination methods from NMR data, dependence upon quality and quantity of data. *J. Biomol. NMR*, **2**, 373–388.

Loh,S.N., Kay,M S. and Baldwin,R L. (1995) Structure and stability of a second molten globule intermediate in the apomyoglobin folding pathway. *Proc. Natl Acad. Sci. USA*, in press

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

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

Murzin,A.G and Finkelstein,A.V (1988) General architecture of the α-helical globule. *J. Mol. Biol.*, **204**, 749–769.

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.

Pachter,R., Altman,R.B., Czaplicki,J. and Jardetzky,O. (1991) Comparison of the NMR solution structure of cyclosporin A determined by different techniques. *J Mag. Res.*, **92**, 468–479.

Pascarella,S. and Argos,P. (1992).A databank merging related protein structures and sequences. *Prot. Eng.*, **5**, 121–137.

Philips,S.E.V. and Schoenborn,B.P. (1981) Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature*, **292**, 81.

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

Prestrelski,S.J , Williams,A.L.Jr, and Liebman,M.N. (1992) Generation of substructure library for the description and classification of protein secondary structure I Overview of the methods and results. *Proteins*, **14**, 430–439.

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

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

Schneider,T.D. and Stephens,R.M. (1991) Sequence logos, a new way to display consensus sequences. *Nucleic 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

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.

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.

Steigemann,W. and Weber,E. (1981) Structure of erythrocruorin in different ligand states refined at 1 4 Å resolution. *J. Mol. Biol.*, **127**, 309.

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.

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

Vingron,M. and Sibbald,P.R. (1993) Weighting in sequence space, A comparison of methods in terms of generalized sequences. *Proc. Natl Acad Sci USA*, **90**, 8777–8781.

Received May 5, 1995, accepted on September 19, 1995