Finding an average core structure: Application to the globins

Russ B. Altman

Section on Medical Informatics
SUMC MSOB X215
Stanford University
Stanford, CA 94305
altman@camis.stanford.edu

Mark Gerstein

Beckman Center for Structural Biology
Department of Cell Biology
Stanford University
Stanford, CA 94305
mbg@cb-iris.stanford.edu

Abstract

We present a procedure for automatically identifying from a set of aligned protein structures a subset of atoms with only a small amount of structural variation, *i.e.,* a core. We apply this procedure to the globin family of proteins. Based purely on the results of the procedure, we show that the globin fold can be divided into two parts. The part with greater structural variation consists of the residues near the heme (the F helix and parts of the G and H helices), and the part with lesser structural variation (the core) forms a structural framework similar to that of the repressor protein (A, B, and E helices and remainder of the G and H helices). Such a division is consistent with many other structural and biochemical findings. In addition, we find further partitions within the core that may have biological significance. Finally, using the structural core of the globin family as a reference point, we have compared structural variation to sequence variation and shown that a core definition based on sequence conservation does not necessarily agree with one based on structural similarity.

I Introduction

A Why we created this procedure?

Proteins have long been clustered into families, such as the globins or the immunoglobulins. Members of protein families tend to have similar overall folds but differences in their detailed structure. The number of protein structures from some particular families is now quite large and is rapidly increasing (e.g. in the Protein Data Bank there are over 15 distinctly different globin structures and over 20 different immunoglobulin structures). Consequently, it becomes possible (even necessary) to summarize the structural commonalties within a family, while separating the variable features from the constant ones. One important structural feature of a family is the core set of residues or atoms which occur in every member of the family and which are located in relatively invariant three-dimensional positions within all members of the structure.

The focus of this paper is the definition of these cores of low structural variance. We have developed a method for taking a family of aligned structures, finding the most structurally conserved residues, which we call the core, and defining the average locations for these core residues. We apply our procedure to a collection of globin structures and find a structural core that is quite biologically relevant.

It is important to note that we use "core" to refer to residues common to all members of a family and which have low variance in α-carbon position. Other investigators have used the term "core" differently—often based on measures of sequence conservation or hydrophobicity. Our results indicate that a core based purely on structural considerations is not the same as one based on sequence considerations.

Coupled with a consensus sequence or profile [6], an average core structure may be useful in model-building applications. For instance, given a new sequence, a comparison with a sequence profile might indicate that the sequence is a globin. In the resulting sequence alignment, residues which fall in our average structural core would be expected to be in virtually the same position as those in other globins. The position, however, of non-core atoms could be adjusted to adapt to the particular residues that occur within the invariant core. In fact, methods already exist for elaborating upon a core structure. Some approaches use a database of know segments to match to the α-carbon positions [17, 22], while others perform an intelligent search through the many possible sidechain orientations [8, 19, 24].

B Previous Relevant Work

Past work on structural superpositions of families of structures has either focused on finding the optimum superposition of a series of nearly identical structures [9, 11, 18, 28] or on finding a structural alignment between a pair of structures that are very different [14, 31, 32]. Our core finding procedure in a sense falls between these two extremes as we want to find an optimum superposition of many moderately different structures. It uses some of the methods of the series superposition procedures and can be used to refine the alignments produced by the structural alignment procedures.

All the methods aimed at superimposing a series of nearly identical structures (usually derived from an NMR structure determination or a molecular dynamics trajectory) start by assuming that there is an alignment pairing each atom in one structure with an equivalent atom in the others. After moving the centroids of all the structures to the origin, they try to find a rotation for each structure that minimizes the sum of squares of the coordinate differences between all pairs of aligned atoms. That is, they seek to minimize:

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

where the first sum is over all pairs j, k of the N structures in the ensemble Ω, the second sum is over the M aligned positions in each structure, and $\mathbf{R}_j\mathbf{x}_{ji}$ is the rotated coordinates of structure j.

II Methods

A Overview of the Core Finding Algorithm

On a high level, our core finding algorithm has 4 steps, which we will subsequently discuss in detail:

1. Start with an aligned ensemble of structures. Initially consider each aligned residue position to be a possible core position.
2. Calculate an unbiased average of all core positions and fit each member of the original ensemble to this average structure.
3. Calculate the structural variation of each aligned atom position (measured in terms of the ellipsoid volume relating to the spread of coordinate positions) and remove the position with the largest variation from the list of candidate core atoms.
4. Go back to step 2 until all atoms have been removed from the core list. Simple analysis of the statistics resulting from this procedure allows the core to be identified.

B Calculation of an Unbiased Average (step 2)

We have developed a simple method for averaging structures in an unbiased fashion. It proceeds as follows:

1. Start with an ensemble of N structures :
$$\Omega = \{a\ b\ c\ d\ ...\} \qquad (2)$$
2. Pick the first structure (in this case *a*) as a reference and fit the remaining structures to it to create a newly oriented ensemble. Then average the coordinates of this ensemble to create an average structure, denoted a^*.
3. Repeat step 2 using each structure in turn as a reference. This will create a new ensemble of average structures. (Steps 2 and 3 require a total of N(N-1) fits and N averaging operations).
$$\Omega^* = \{a^*\ b^*\ c^*\ d^* ...\} \qquad (3)$$
4. Apply formula (1) to the ensemble of average structures to compute E(Ω*), the sum of squares of the coordinate differences between all pairs of aligned positions. (N^2 distance calculations.)
5. If E(Ω*) is below some small threshold, then all the structures in Ω* are the same to within this threshold. Consequently, one of the structures can be picked and returned as the unbiased average. Otherwise go back to step 1 using Ω* as the new starting ensemble.

All pairwise fits are done using the method of Arun & Huang [2]. The unbiased nature of the method is apparent on inspection, since the order of the N structures can be randomly shuffled at the outset, with no difference in the result. Using the unbiased average, we have a way to put all the proteins into the same coordinate system without favoring any of them.

C Calculating the Structural Variation of a given position (step 3)

After the original N structures are fit to the average structure determined in the above step, it is possible to look at the structural variation of each aligned position. We have described elsewhere [1], a representation of a structure that summarizes variability about a mean position (for an atom) in terms of a variance/covariance matrix C. The eigenvalues of this matrix (σ_x^2, σ_y^2, σ_z^2) give the variance of the distribution in its three principal directions. They roughly correspond to the axes of an "ellipsoid of errors" centered at the mean. We can average these eigenvalues in either a geometrical or arithmetic fashion and calculate the volume of the ellipsoid or its average axis length:

$$V = \tfrac{4}{3}\pi\sqrt{\sigma_x^2\,\sigma_y^2\,\sigma_z^2}\ , \qquad (4)$$

$$R = \tfrac{1}{9}\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2} \qquad (5)$$

For the procedure as applied here, we have reported the volume of the ellipsoid as a measure of structural variation. (We get virtually the same results using the mean axis length). Atoms which appear in the same relative locations over all the structures in the ensemble will have small volumes, whereas atoms which are variable in their relative location will have large volumes.

D Defining a Core Cutoff (step 4)

The atom with the largest ellipsoid volume is, in some sense, the least likely atom to be a member of a structural core. In our procedure, it is therefore discarded as "non-core." The entire averaging, fitting and volume-calculation procedure is repeated until all atoms are moved from the candidate "core" list to the "non-core" list.

There are many properties of the core and non-core regions defined at each step in the iteration that can be used to define a cutoff. Conceptually, we are seeking a measure that has the property of being optimized at the point in the process at which there is an optimal separation between core atoms and non-core atoms. Our procedure starts with an empty list of non-core atoms, and initially adds atoms with largest structural variability to this list (and removes them from the core list). At some point in this process, atoms that are appropriately considered core are transferred to the non-core list. This point can be readily identified by looking at the variance of the volumes of the non-core atoms after they have been fit to one another. The success of this measure depends on the fact that core atoms (by definition) will tend to have small volumes that cluster tightly around some mean value. Non-core atoms, on the other hand, will have a broader range of volumes, and so the distribution of their volumes has high variance. Now, consider the situation when core atoms (small average volume, low variance) are added to the list of non-core atoms (high average volume, high variance). The variance of volumes will now tend to decrease, since a cohort of relatively uniform volumes are now being added. Thus, at the point when core atoms start to be transferred to the non-core list, there should be a drop in the variance of the distribution of volumes in non-core regions. This is exactly what is observed (Figure 3).

E Measuring Sequence Variation

To measure the sequence diversity of the globins we used the structurally based alignment of 577 globin sequences reported in Gerstein et al. [12]. At a given position in this alignment, we can calculate the frequencies of occurrence of the different amino acids and then measure sequence variability through the computation of an information-theoretical entropy [26, 27, 29]. However, because of the biases in the sampling of sequences within the databanks, there is an over-representation of certain globins, such as

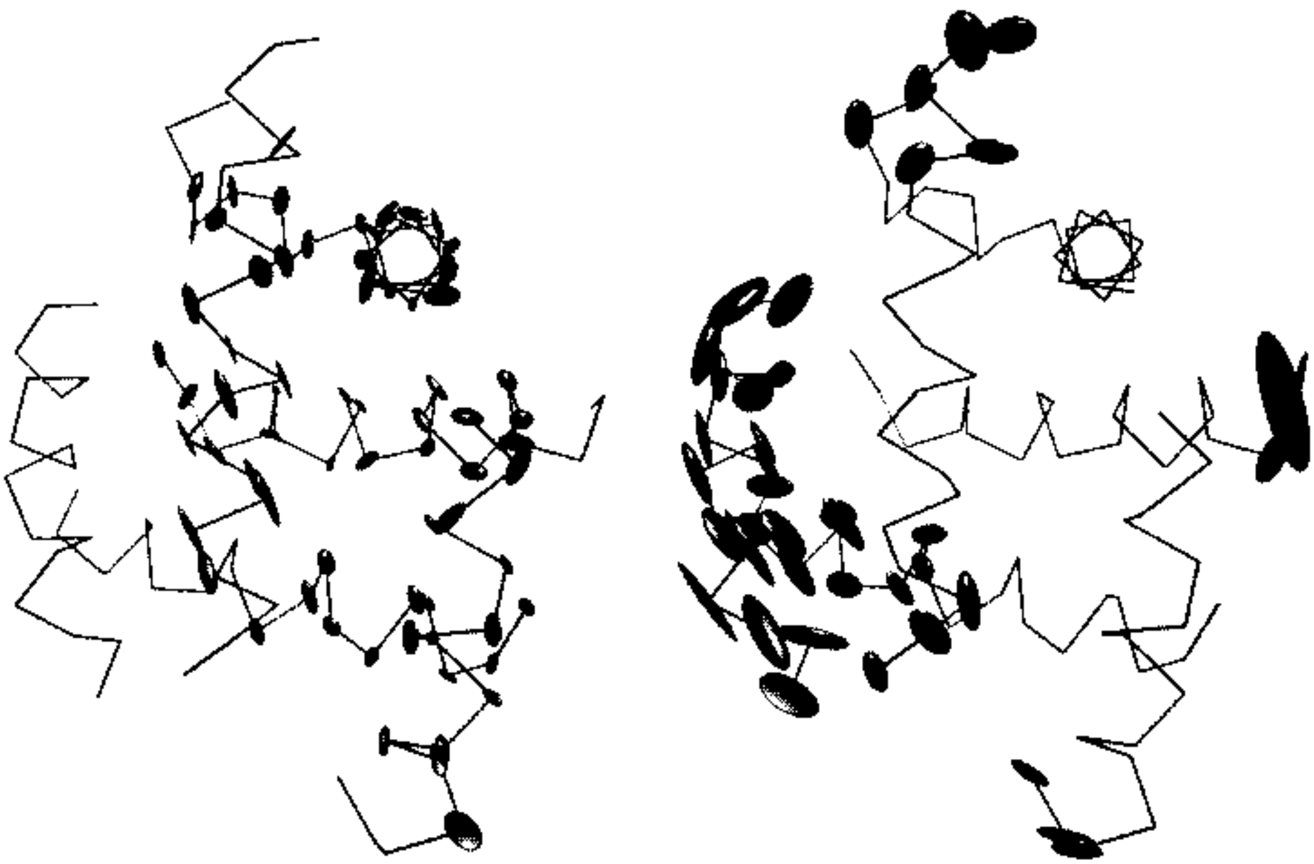


Figure 2. The ellipsoids around the 73 atoms classified as belonging to the core (LEFT) and 42 atoms classified as not belonging to the core (RIGHT). The view is roughly the same as in Figure 1. To form each ellipsoid, we created a 3x3 variance/covariance matrix from the spread of Cα coordinates at each aligned position in the 8 globin structures. Then we perform a Jacobi decomposition [25] which provides the lengths of the three ellipsoid semi-axes (from the eigenvalues) and the orientation of ellipsoid (from the eigenvectors).

mammalian hemoglobins, and under-representation of other globins, such as the plant globins, and this bias will affect the entropy calculation significantly. In constructing the alignment, a tree-based weighting scheme was developed to correct for this over- and under-representation [12]. (For a general discussion of weighting schemes see [33]). In this work in order to more accurately compare sequence variability with structure variability, we have incorporated weights into the calculation of entropy. To measure the entropy of a particular position i in a sequence alignment, we use the standard formula for Shannon entropy:

$$H(i) = -\sum_{i=1}^{20} f(i,t)\log_2 f(i,t) \ . \tag{6}$$

However, instead of the frequency of amino acid t at position i, we take $f(i,t)$ to be the normalized sum of the weights for sequences with a residue of type t in position i :

$$f(i,t) = \frac{\sum_{s(i)=t} w(s)}{\sum_{s} w(s)} \tag{7}$$

where the denominator sum is over all sequences s in the alignment and the numerator sum is over just those sequences that have a residue of type t at position i. We report sequence variability as information content relative to that if the sequences were aligned randomly:

$$R_{sequence}(i) = \sum_{t=1}^{20} \overline{f}(t)\log_2 \overline{f}(t) + H(i) \tag{8}$$

where $\overline{f}(t)$ is the average frequency of type t in the alignment (i.e. $f(i,t)$ averaged over i).

III Application to the Globins

A Choice and Alignment of Structures

From the Protein databank [5], we chose 8 structures from the globin family which have been the subject of previous investigations [4, 12, 20]: 1ECD, 1MBA, 1MBD, 2HBG, 2LH4, 2LHB and the A and B chains of 3HHB. (All structures are of the deoxy form except for 1MBA and 2LHB.) Using a canonical numbering scheme first developed by a Kendrew, Lesk & Chothia had previously aligned these structures by eye [20]. The 115 common positions in their structural alignment are shown in Figure 1.

B Progress of the Iteration

Figure 3 shows the variance of non-core volumes as a function of cycle. We defined peaks by performing a 5-atom moving average in order to smooth the curve, and then selected local maxima. The primary peak of this curve is at cycle 42 (the first 42 atoms have been removed from the original 115, and 73 putative core atoms remain). We took this peak to indicate the that the primary core segment of the globin family contains 73 atoms, as shown in Figure 1. We also noted secondary peaks at cutoffs of 64 and 84 which are discussed below.

In order to further validate the choice of our cutoff at iteration 42, we also plotted the distribution of ellipsoid volumes for the core residues and for the non-core residues in Figure 3. The two distributions are fairly well separated. The core residues clearly have a much smaller mean volume than the non-core residues. More importantly, they have a much smaller variance in volume.

C A Core Containing 73 Positions

The ellipsoids of variation for the 73 core and 42 non-core atoms are shown graphically in Figure 2. If a core is to be used as a starting point for modeling, then the quality of its structure becomes a matter of importance. We assessed the "quality" of our α-carbon structure by looking at the stereo-chemistry of the connected core residues. Three parameters characterize the geometry of an α-carbon structure [21]; for each of these parameters we compared the range of values in our structure with established norms. (1) The most important parameter is the distance between two connected Cα atoms, which should be 3.8 Å. We find the mean distance between the 67 connected core Cα atoms to be 3.8 Å with a standard deviation of 0.03 Å. (2) The next most important parameter is the angle τ between three connected Cα atoms. This angle can range between approximately 80° and 135°. For our core atoms we find 62 τ angles defined, which range between 88° and 122° with a mean of 92°. (3) The third parameter, the pseudo-torsion angle α between four connected Cα atoms, can acceptably range from -180° to +180° and so does not form a meaningful constraint on α-carbon structure.

D No correlation between sequence and structural diversity

Figure 4 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 for either all 115 aligned positions or for just the core positions.

IV Discussion

A Comparison with other methods to find an unbiased average

Central to our procedure is the finding of an unbiased average of a ensemble of structures. This is essentially the same as what is accomplished by the four previously proposed methods to perform an unbiased superposition of multiple structures (since after superposition all one need do is average the coordinates) [9, 11, 18, 28]. Our method to find an unbiased average relies solely on pairwise superpostions. Consequently, we believe it is simpler than the four methods previously proposed and easier to implement. For the globins we have extensively compared the average structure produced by our procedure against that produced by Diamond's procedure. We find that the RMS deviation between them is less than 0.001 Å/atom.

Our overall procedure would be unaffected if the step involving the calculation of an unbiased average were done with any of the above four methods. For a large number of structures it may be advisable to use Diamond's method since its calculations increase linearly with the number of structures while ours increase quadratically. However, the calculations in our procedure are easily parallelizable in a coarse-grained fashion.

B Relationship to methods that perform a structural alignment

Our procedure requires that there exist an initial structural alignment between the structures in the ensemble. Our procedure then refines this initial alignment by throwing away atoms that superpose badly. We have designed our method to work on families of relatively similar structures, such as the globins or immunoglobulins. In these cases, it may be possible to get an initial structural alignment by eye or by sequence alignment (e.g. the Kendrew sequence numbering for the globins or the Kabat sequence numbering for the immunoglobulins implies a structural alignment). Obviously, the scope of our method would increase if it could use alignments between dissimilar structures generated by automatic structure alignment procedures, such as have been previously reported [14, 31, 32]. At present there is no known method for doing multiple structural alignment. Considering the analogy with some of the known methods for multiple sequence alignment [30], we believe our method for finding and refining an unbiased average of many structures could be a useful part of an iterative multiple structural alignment algorithm.

C Practical Use of Our Procedure in Model Building

The average structures generated by our procedure exhibit acceptable stereo-chemistry. This is because we restrict ourselves to only averaging the most similar parts of the structures, and because we only average α-carbon positions and thus need never worry about averaging over a peptide flip. Because of their acceptable stereo-chemistry the structures produced by our procedure provide ideal, unbiased points to build models from. Furthermore, the structural variances that our procedure assigns to each position indicate the degree that the backbone structure in a particular region must be maintained in model-building.

D Sequence-Structure Implications

We find that the structure variation is not correlated with sequence variation. This lack of correlation is true whether we consider all 115 aligned positions, the 72 core positions, or just the 31 core positions that are buried in all the globin structures. This result has strong implications for the modeling of proteins. Many modeling studies assume that sequence conservation implies structural conservation. While this is probably true on the level of the overall fold, it is *not correct* in the globins to assume that the positions that are most conserved in sequence are those most invariant in structure.

The lack of correlation between sequence conservation and structure conservation may result from the way protein structure accommodates sequence changes. One point of views says the structural accommodation is local: a mutation in one residue is accommodated by sidechain torsion angle changes in this residue and complementary mutations in neighboring residues. The backbone stays fixed throughout. A contrasting perspective says that the structural accommodation is more global: a mutation in one residue is accommodated by backbone shifts throughout the protein (see for example the T4 lysozyme work [3, 10]). Our result provides clear evidence for this second perspective.

E Significance of the identified globin core

The most striking aspect of the core we identified in the globins is that it does not include any atoms from the F helix (or any atoms from the end of H helix that contacts the F helix). As shown in Figure 1, the heme group is bound between the E and F helices. Consequently, the pocket between these two helices is in a sense the active site of the globins. The movement of the F helix relative to the rest of globin core may be part of the mechanism by which the different globins modulate the environment of the heme and achieve different oxygen binding affinities.

Our finding that the core does not include the F helix is consistent with a number of recent NMR spectroscopy experiments [7, 16]. These experiments have shown that the F and D helices are the most mobile parts of myoglobin in solution and the last to fold. Furthermore, the folding experiments [16] indicate that the A,B,G, and H helices form a stable association before the remainder of the protein folds. This is consistent with our finding that our secondary core found at iteration 84 does not include the E helix and only includes one residue from the C helix. Thus, our core finding procedure is able to identify the most stable parts of the globin fold.

Our finding that the core does not include the F helix is also consistent with the way globin helices have been mapped onto ideal polyhedra. Muzin & Finkelstein [23] found that the helix packing geometry of most all-α proteins roughly follows the edges of ideal polyhedra. The globins, however, were an exception. They would fit ideal polyhedra only if the F helix is ignored.

F Relationship between the Globin Core and the Repressor Protein

Recently, it has been pointed out that some parts of the globin structure are similar to that of helix-turn-helix (HTH) proteins [31]. The HTH motif is one of the most common folds for DNA binding [13]. For the bacteriophage 434 repressor protein Subbiah *et al.* [31] pointed out that the two helices in the HTH motif align well with globin helices B and E and that the other three helices in the protein align well with globin helix A and parts of helices G and H (Figure 1).

The 73 positions in our globin core coincide closely with the 52 positions where the repressor aligns well to the globins. As shown in Figure 1, there is only one position (G17) where an aligned position in the repressor does not correspond to a position in the globin core. The structural similarity of the globins to other proteins beside the repressor, such as the phycocyanins and colicin A, has been pointed out [15]. These similarities, however, involve almost all of the globin fold and not distinct subsets. Consequently, they are not suitable to compare to the core.

V Conclusion

We have presented a method for finding an average structural core. We have applied this method to the globin family and found that the division between core and non-core correlates very well with other information about the globin structure, such as NMR data and the alignment with the 434 repressor. We have also shown that for the globins a core based purely on structural conservation is not the same as one based on sequence conservation.

Acknowledgments

RBA is a Culpeper Medical Scholar. 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 S Subbiah for discussions.

Availability of Results

Coordinates of the 73 residue globin core as well as further documentation in hypertext form are accessible at the following location (URL) :

ftp://cb-iris.stanford.edu/pub/mbg/ISMB-94-60/ .

References

1. Altman, R. & Jardetzky, O. 1989. The Heuristic Refinement Method for the Determination of the Solution Structure of Proteins from NMR Data. *Meth. Enzym.* 177: 177-218.

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

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

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

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

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

7. Cocco, M. J. & Lecomte, J. T. J. 1990. Characteriztion of Hydrophobic Cores in Apomyoglobin: A proton NMR Spectroscopy Study. *Biochemistry* 29: 11067-11072.

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

9. Diamond, R. D. 1992. On the multiple simultaneous superposition of molecular structures by rigid-body transformations. *Protein Science* 1: 1279-1287.

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

11. Gerber, P. R. & Müller, K. 1987. Superimposing Several Sets of Atomic Coordinates. *Acta Cryst.* A43: 426-428.

12. Gerstein, M.; Sonnhammer, E. & Chothia, C. 1994. Volume Changes on Protein Evolution. *J. Mol. Biol.* 236: 1067-1078.

13. Harrison, S. C. 1991. A structural taxonomy of DNA-binding domains. *Nature* 353: 715-720.

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

15. Holm, L. & Sander, C. 1993. Structural alignment of globins, phycocyanins and colicin A. *FEBS Lett.* 315: 301-306.

16. Jennings, P. A. & Wright, P. E. 1993. Formation of a Molten Globule Intermediate Early in the Kinetic Folding Pathway of Apomyoglobin. *Science* 262: 892-896.

17. Jones, T. A. & Thirup, S. 1986. Using known substructures in protein model building and crystallography. *EMBO J.* 5: 819-822.

18. Kearsley, S. K. 1990. An Algorithm for the Simultaneous Superposition of a Structural Series. *J. Comp. Chem.* 11: 1187-1192.

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

20. Lesk, A. M. & 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.

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

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

23. Murzin, A. G. & Finkelstein, A. V. 1988. General Architecture of the α-Helical Globule. *J. Mol. Biol.* 204: 749-769.

24. Ponder, J. W. & 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.

25. Press, W. H.; Flannery, B. P.; Teukolsky, S. A. & Vetterling, W. T. 1992. *Numerical Recipes in C*. Cambridge: Cambridge University Press.

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

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

28. Shapiro, A. & Botha, J. D. 1992. A Method for Multiple Superposition of Structures. *Acta Cryst.* A48: 11-14.

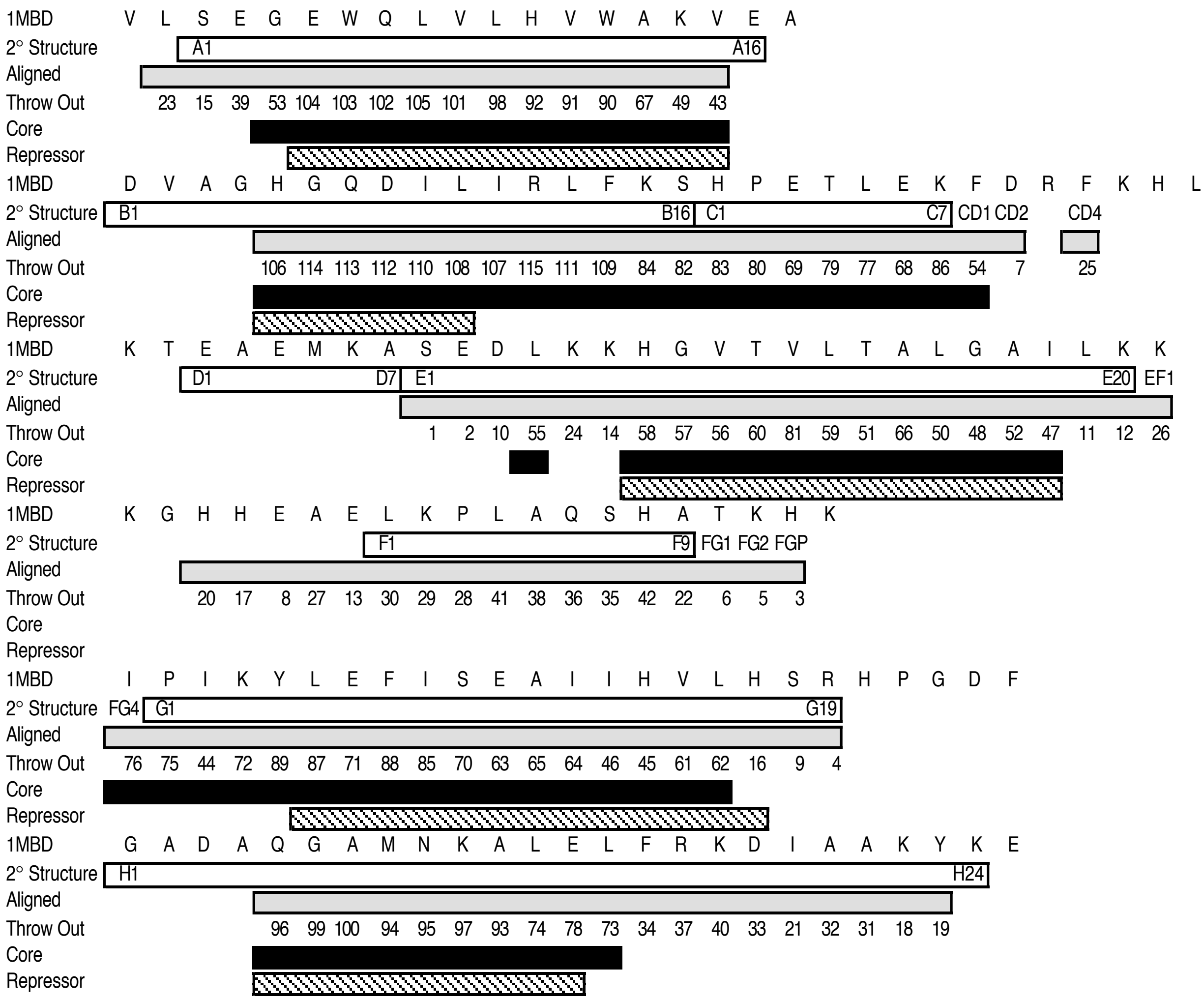
29. Shenkin, P. S.; Erman, B. & Mastrandrea, L. D. 1991. Information-Theoretical Entropy as a Measure of Sequence Variability. *Proteins: Struc. Func. Genet.* 11: 297-313.

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

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

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

33. Vingron, M. & 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.



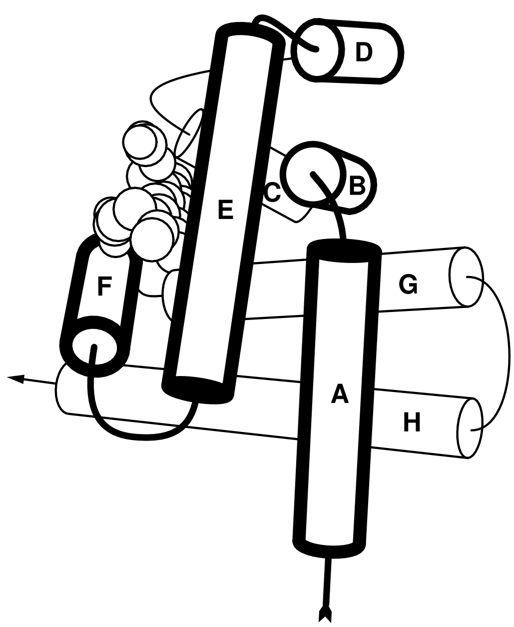


Figure 1. (LEFT) Cylinders representation of a globin (1MBD). (ABOVE) Listing of the various subsets of globin residues. From an entire globin (1MBD, which has 153 residues, is shown as an example), the set of 115 residue positions that were structurally aligned with the other globins were extracted (ALIGNED row). These aligned residues roughly correspond to the helices in the globins with the exception of the D helix (2° STRUCTURE row). The core finding procedure was applied to these 115 aligned residues to produced a core of 73 residues (CORE row). The iteration at which each of the 115 aligned residues was deleted from the putative core is shown (THROW OUT row). Finally, we compare our 73 core positions to the 52 positions in the repressor protein which are aligned to myoglobin (REPRESSOR row).

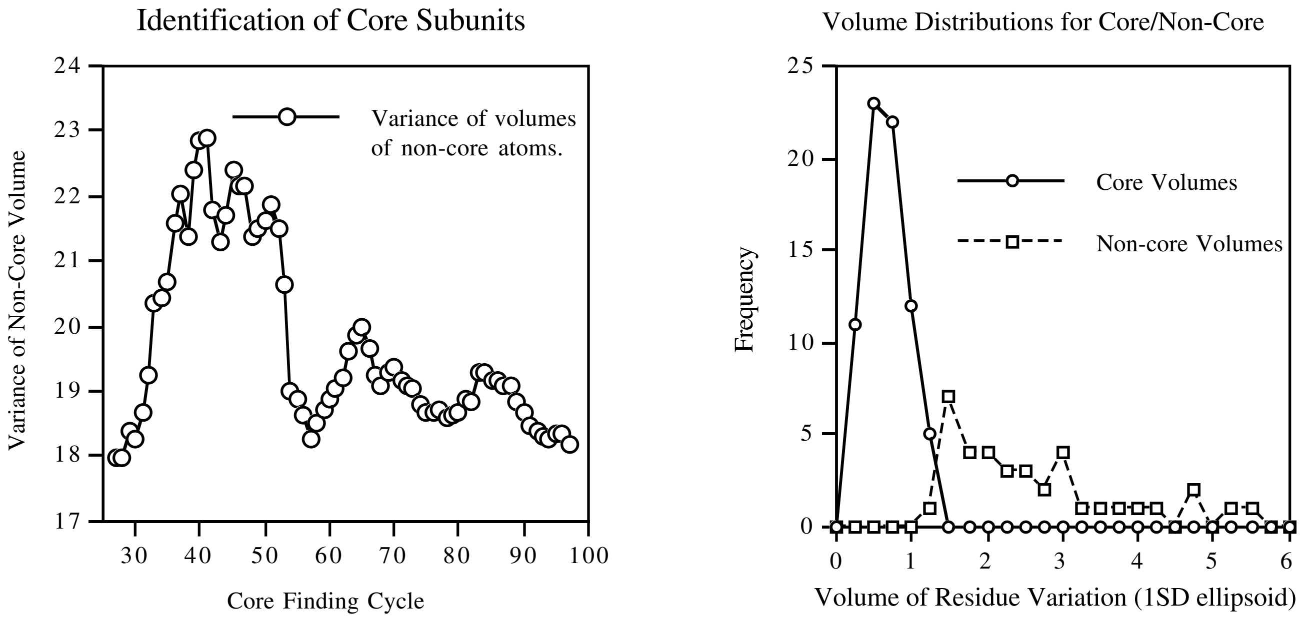


Figure 3. (LEFT) The variance in the ellipsoid volume distribution for non-core atoms (after being optimally fit to one another) is used to identify the core. This measure peaks at a "core" threshold. For the globins the primary threshold between core and non-core is at cycle 42, with secondary cutoffs at cycle 64 and 84. (RIGHT) Distribution of ellipsoid volumes (at one standard deviation) at cycle 42, the primary threshold. Each of the 73 core positions has an error ellipsoid of volume ~5 Å^3 while the 42 non-core positions have larger ellipsoids (of volume ~16 Å^3) that are more widely varied in volume.

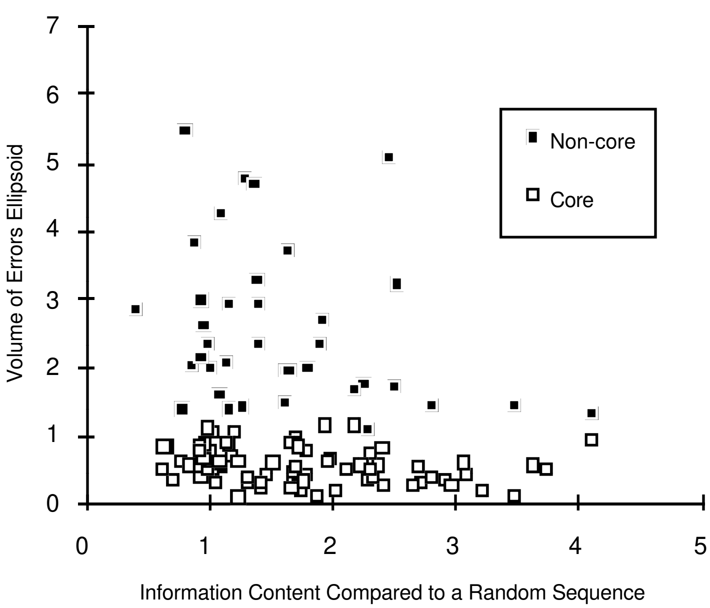


Figure 4. Graph of sequence diversity versus structural diversity for each globin position. At a particular position, sequence variation is measured by information content relative to that of unaligned sequences $R_{sequence}(i)$ (in bits per residue) and structural diversity by the volume V of the covariance matrix ellipsoid (in Å^3). There are 115 positions represented in total here and the overall Pearson correlation coefficient 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.