Volume Changes in Protein Evolution

Appendix:
A method to weight protein sequences
to correct for unequal representation

Mark Gerstein [1,*], Erik L L Sonnhammer [1], & Cyrus Chothia [1,2]

[1] MRC Laboratory of Molecular Biology and
[2] Cambridge Centre for Protein Engineering
Hills Road, Cambridge CB2 2QH, UK

Revised manuscript (CAM 349/93) sent to *J. Mol. Biol.* on 14 October 1993.

* Present Address: Beckman Laboratories for Structural Biology, Department of Cell Biology, Stanford Medical School, Stanford, CA 94305-8464 USA

Keywords: Residue Volumes, Random Sequences, Globins, Dihydrofolate Reductase, Plastocyanin-Azurin Family, Tree Weighting

Running Title: Volume Changes in Protein Evolution

Abstract

We have determined the variations in volume that occur during evolution in the buried core of 3 different families of proteins. The variation of the whole core is very small (~2.5 %) compared to the variation at individual sites (~13 %). However, by comparing our results to those expected from random sequence changes with no correlations between sites, we show that the small variation observed may simply be a manifestation of the statistical "law of large numbers" and not reflect any compensating changes in, or global constraints upon, protein sequences.

We have also analyzed in detail the volume variations at individual sites, both in the core and on the surface, and compared these variations with those expected from random sequences. Individual sites on the surface have nearly the same variation as random sequences (24 % vs. 28% variation). However, individual sites in the core have on average about half the variation of random sequences (13% vs. 30%). Roughly, half of these core sites strongly conserve their volume (0-10 % variation); one quarter have moderate variation (10-20%); and the remaining quarter vary randomly (20-40%).

Our results have clear implications for the relationship between protein sequence and structure. For our analysis, we have developed a new and simple method for weighting protein sequences to correct for unequal representation, which we describe in an appendix.

1. Introduction

The determination of the atomic structures of myoglobin and haemoglobin showed that very different protein sequences could produce similar three-dimensional structures. This phenomenon was explained by the hypothesis that mutations in the interior are complementary. That is, the atoms added to, or lost from, the protein core because of one mutation are compensated by a subsequent mutation in the opposite direction (Kendrew & Watson, 1966). In 1970 Lim & Ptitsyn carried out a calculation on the small number of globin sequences available at the time (52) and showed that the volume of the 31 core residues was essentially constant. At the time this result was taken to support the complementary mutation hypothesis.

Later, Lesk & Chothia (1980) analyzed the three-dimensional structures of nine different globins, which had substantial differences in sequence. (The minimum sequence identity between pairs was 16 %.) In the different globins the *mean* size of residues at homologous helix-helix interfaces varied in size by up to 50 %. This variation implied that the mutations between sequences were not locally complementary at a given interface. The proteins adapted to these changes by relative shifts in the position and orientation of the helices (up to 7 Å and 20°) in a manner that conserved the structure of the haem pocket.

The analysis of other families of proteins demonstrated that in general proteins adapt to mutations by structural changes (Chothia & Lesk, 1987), and the same view has emerged from protein engineering studies (Eriksson *et al*., 1992). However, it was not apparent how this view of protein evolution could be reconciled with the apparent "constancy" of the total volume of protein interiors found by Lim & Ptitsyn.

Ptitsyn & Volkenstein (1986) addressed this problem: by determining the volume variations at individual buried sites and for total core in the nine globin structures, they found that the core volumes of artificially generated random sequences were very close to those of observed ones. Thus, they showed that the apparent "constancy" in core volume over evolution may result from general statistical properties of sums of random numbers.

Here we greatly extend Ptitsyn & Volkenstein's work: we have determined the volume variations at the individual buried sites and for whole buried core in 568 globin sequences. We have also carried out the same calculations on two other protein families, the dihydrofolate reductases and the plastocyanin-azurin family. The three protein families have very different core sizes and structures: the plastocyanin-azurins (also known as the cupredoxins, see Adman, 1991) are all-β proteins with small cores (~4000 Å^3); the globins are all-α proteins with medium-sized cores (~6000 Å^3); and the dihydrofolate reductases are α/β proteins with large cores (~8000 Å^3).

Our calculations show that the particular results obtained by Ptitsyn & Volkenstein on a small number of globin structures are generally true for protein families with widely divergent sequences. Namely, we show that while individual sites in the protein core vary appreciably in volume (some by up to 40 % and on average by ~13 %), the volume of the whole core is nearly constant (varying ~2.5 %). Moreover, we show that uncorrelated mutations at the individual core sites can reproduce the observed variation in core volume, so it is not necessary for mutations to be locally compensating to produce the small observed variation in core volume.

We do two further sets of calculations: we determine the volume variations at surface sites and those that would be produced by random changes in sequence. Comparison of the different calculations shows the variation at individual sites on the surface is nearly what is expected for random sequences. At individual buried sites, however, the average variation is less than half of what is expected for random sequences. This disparity is due to the local size constraints imposed by the protein structure on particular buried sites. These constraints vary greatly: some sites are absolutely conserved, whilst others essentially vary like random sequences.

2. Methods

(a) Protein Sequences and their alignment *

Protein sequences were taken from SwissProt-24 (Bairoch & Boeckmann, 1992) and PIR-36 (Barker *et al*., 1992), and protein structures were taken from the Protein Data Bank (Bernstein *et al.,* 1977). In total we collected 24 plastocyanin-azurin sequences, 568 globin sequences, and 40 dihydrofolate reductase sequences.

The sequences for the three families were aligned based on "key" sequences of known structure. Bashford *et al*. (1987) described the application of this alignment procedure to 226 globin sequences. The "key" sequences corresponding to 8 known globin structures were first aligned based on structural superimposition, and then the rest of the

* The sequence alignments and computer programs used for this work will be made available electronically by email to mbg@cb-iris.stanford.edu or by anonymous ftp to cb-iris.stanford.edu or cele.mrc-lmb.cam.ac.uk (directory /pub/ProtEvol).

sequences were aligned to the keys. As shown in Table 1, for our work the globin alignment in Bashford *et al.* was expanded to include more sequences. New alignments were constructed for the dihydrofolate-reductase and plastocyanin-azurin families based on the key sequences corresponding to known structures.

Table 1 near here

Table 2 near here

The buried and surface sites in the three families are listed in Table 2. With one or two exceptions discussed in the table caption, core residues were defined as those that occupied sites whose *mean* accessible surface area (Lee & Richards, 1971) in the "key" structures was less than 15 Å^2. Various other accessibility cutoffs were tried but not found to make appreciable difference to the results. Surface sites were defined as those with more than 50 Å^2 of accessible surface.

For the structures, volumes were calculated according to the Richards' implementation of the Voronoi method (Richards, 1974). For the sequences, standard volumes for each residue type were taken from Harpaz *et al.* (1993) and are reproduced in Table 3.

Table 3 near here

(b) A weighting scheme for the comparison of sequence features. *

Table 4 near here

As described in Table 4, for a given site, we averaged the standard residue volumes over all sequences in each alignment to get a mean volume for that site and variation about this mean. Throughout the text, we express this variation as a percentage standard deviation (percent S.D.[†]). Likewise, for each sequence we summed up the standard residue volumes of the buried sites to get a core volume, and then we averaged these core volumes over all the sequences to determine a mean core volume and variations about this mean.

To compensate for the unequal representation of the sequences in our alignment, we used a weighting scheme that gave low weights to closely related sequences, such as the haemoglobin α-chains in the globin alignment. This weighting scheme is described in the appendix. It gave us greater confidence in the quantitative accuracy of our conclusions.

[†] Abbreviations Used: S.D., standard deviation; DHFR, dihydrofolate reductases; Plas.-Az., Plastocyanin-Azurin family.

However, it did not change our conclusions or results significantly as compared to those reached from doing unweighted averages.

3. The Observed Volume Variation

(a) Variation at individual sites

The variation in the volume at individual sites in the protein core is shown in Figure 1(a). The volume variation is similar in all three families. The average variation is 15 % for the globins, 13 % for the dihydrofolate reductases, 11 % for the plastocyanin-azurin family, and 13 % for all three families combined.

In all three families, roughly, half of the buried sites vary less than 10 % in volume; one quarter, between 10 and 20%; and the remaining quarter, between 20 and 40%. (As discussed below, this last quarter has a volume variation similar to that expected for random sequences.)

Figure 1 near here

Surface sites vary more in volume than core sites: 24 % on average for the globins. Figure 1(b) compares the range of volume variation at surface sites in the globins with that for core sites. None of the surface sites varies less than 10 %.

Instead of looking at the volume variation at individual sites, it is possible to look at the variation in sets of structurally neighboring sites, such as one helix-helix interface. Table 5 shows that this variation is smaller than that of individual sites (but larger than that of the whole core; see below).

Table 5 near here

(b) Volume variation of the whole core

Table 6 near here

Figure 2 near here

The variation in the total volumes of the buried cores in the three protein families is given in Table 6. The average variation over the whole range of structures is less than 2.6□%. Figure 2 shows that this apparent "constancy" in core volume covers the whole range of sequence identities.

Table 6 also shows that while the three protein families have different numbers of residues in the core, the average size of a residue in the core for each of these three families is essentially the same: ~150 Å^3 or 7 to 8 non-hydrogen atoms (*i.e.*, between Val and Leu in size).

The "constancy" of core volume discussed in the preceding paragraphs is derived from applying the standard residue volumes to protein sequences. We also carried out the calculation in the reverse direction and directly determined the core volume for 8 globin structures (Table 7). Corroborating our sequence calculations, the structure calculations have similar volume variations.

Table 7 near here

4. Volume Variations Produced by Random Sequences

In the preceding section we found that for each of the three families the large volume variations at individual core sites (11 to 14%, on average) tend to cancel to give a small variation for the total core volume (1.5 to 2.6%). To put our results in context, we calculated the volume variations that would be produced by random sequence changes that have no correlations between sites.

For these calculations, we supposed that all the sites in the core were filled with residues picked at random from identical distributions of amino acids. In this case, all the sites are uncorrelated and equivalent, so the variation in the volume of the whole core (measured as a percent S.D.) is just $\frac{1}{\sqrt{R}}$ of the variation at a single site, where R is the number of residue sites. Clearly, the crucial parameter is the variation at a single site, which, in turn, depends on the residue frequencies used in the generation of random sequences. As shown in Table 8, we tried three different types of frequency distributions:

Table 8 near here

Figure 3 near here

(i) A uniform distribution, where all residues have equal frequency.

This is the simplest scheme and introduces no *a priori* constraints. The volumes of random globin sequences constructed according to this distribution are shown in Figure 3. However, it is somewhat unrealistic because it does not exclude charged residues from the core.

(ii) General distributions that take into account the chemical character of buried residues. The simplest such distribution is another uniform distribution, which just excludes the residues that are rarely buried: charged residues and Gln and Asn. A more accurate distribution is shown in Table 3. It is derived from the frequencies of buried residues in 119 protein crystal structures. The major problem with these general buried residue distributions is that the residue frequencies found in particular proteins may be influenced by their secondary structures and topologies.

(iii) Distributions based on the specific buried residues in each of the 3 protein families. These were made by compiling frequency tables similar to the shown in Table 3 for the buried residues in each protein family. These one-family distributions obviously do not suffer from the same problems as the other two types of distributions. However, they introduce a significant amount of *a priori* bias into the generation of any random sequence.

Despite the differences between these frequency distributions, as shown in Table 8, they essentially give the same result: the calculated variations in site size (24-31 %) are roughly double the average observed variations in site size (11-14 %). The single-site variations imply that for random sequences the expected variation in core volume is 4.0 to 6.3 %, which is roughly double the observed variation (1.5 - 2.6 %).

Although the simple random distributions discussed above do not reproduce the observed variation for the core sites, they do describe the behavior at the surface sites fairly well. As shown in Table 8(b), the observed volume variation for individual sites on the surface of the globins (24 % on average) is close to that calculated from filling these sites with randomly chosen, non-hydrophobic residues (28 %).

5. The Relationship Between the Observed Single-site Variations and the Observed Core Variation

In the preceding section we assumed that all sites varied independently and randomly, according to the *same hypothetical* frequency distribution. Here we take the *observed variation* at the individual sites and calculate the expected variation for the whole core. That is, we again assume that the sites vary independently and randomly, but this time each site varies according to a frequency distribution that is derived solely from the amino acids occurring at that site in the sequence alignment. The variations in the volume of individual sites would in this case obviously be the same as the observed values shown in Figure 1(a).

Table 9 near here

Table 9 shows the result of our new set of assumptions: the calculated variation in core volume is similar to that observed. For the globins and the dihydrofolate reductases, the calculated and observed values were very close: 2.4% vs. 2.6% for the globins and 2.0% vs. 2.2% for the dihydrofolate reductases. For the plastocyanin-azurin family the calculated and observed values (2.5 vs. 1.5%) are not as close. This discrepancy may reflect the bimodal nature of the family: the different plastocyanin sequences have considerable similarity (i.e., sequence identity) to each other, as have the azurin sequences, but between the two groups the similarity are low.

Thus, provided the residues are picked according to the correct distribution, filling the sites randomly gives nearly the same variation in core volume as that observed over evolution. It is not necessary to invoke compensating changes and correlations between sites to explain the apparent "constancy" in total core volumes. It is simply a statistical effect of the "law of large numbers," i.e., in averaging random numbers, the average variation decreases as the sample size increases.

Since the average size of a buried residue in all three protein families (Table 6) is the same, our conclusion that the observed variation in core volume can be reproduced by random variations implies that the overall volume of the core is not strongly dependent on the particular sequence of the residues at the buried sites. The size of core mainly depends on the number of residues that point inward from solvent.

6. Conclusion

We have quantitatively determined the volume variations that occur in the buried cores of three protein families. Although the families have quite different structures and functions, the three families give very similar results. This suggests our results represent generally what occurs in protein families where sequence has diverged but function has been retained.

The variation in the total volume of the buried core is very small (1.5 to 2.6 %) compared the variation at individual sites (11-14% on average). However, we come to a conclusion similar to that of Ptitsyn & Volkenstein (1986) that this apparent "constancy" of core volume does *not* necessarily reflect any compensating changes in, or global constraints upon, protein sequences that have evolved from a common ancestor. Rather, it is consistent with appropriately chosen random sequence changes and as such is simply a manifestation of the statistical "law of large numbers."

We compared the volume variations observed at individual sites with those produced by the random sequences generated according to standard frequency distributions. The observed variation at individual surface sites is similar to that produced by random sequences. In contrast, the observed variation at individual core sites (~13 % on average) is roughly half that found for the random sequences (~27 %).

The smaller variation found for core sites arises from their being subject to steric constraint in differing degrees (Figure 1(b)). About half the core sites conserve their volumes (<10% variation); about a quarter have moderate volume variation (10-20%); and the remaining quarter vary as much as random sequences (20-40 %).

Proteins in the same family have a common core whose structure is shared by all family members and peripheral regions whose structures vary between different members. For the three families discussed here, roughly one third of the residues in the "common core" are buried (Chothia & Lesk, 1986). Thus, as we found (above) that half of the buried residues have strong volume constraints, approximately one sixth of the residues in the common core are greatly constrained in volume.

Our results have implications that highlight both the random and the invariant features of protein sequences:

(a) Random Features of Protein Sequences

On one hand, our conclusions imply that because of general statistical considerations, a given core volume depends mainly on the number of, and not on the type of, residues in the core. Thus, allowing for the specific size constraints on a few individual sites, residues drawn from a wide range of sequences would be acceptable for a particular core structure. These implications support the view that protein sequences are random heteropolymers, which are "edited" only slightly by evolution (Ptitsyn & Volkenstein, 1986), and that the structural features of known protein folds can accommodate a wide range of sequences (Finkelstein & Ptitsyn, 1987; Murzin & Finkelstein, 1988; Finkelstein *et al*., 1993).

(b) Sequence Determinants of Protein Folds

On the other hand, our conclusions do not imply that there are no volume constraints upon protein sequences in the core. We have found that some individual core sites are nearly invariant in volume and others have only small volume variations.

Furthermore, we have presented a procedure that gives a precise, quantitative description of this volume conservation at each site in families of sequences. The aspects of sequences that determine structure (and function) should appear as conserved features in families of related sequences. Our procedure can easily be adapted to quantify other conserved features, such as hydrophobicity and charge. Thus, it will allow one to derive a detailed picture of the conserved features in families of sequences. Examination of these features, and of their structural and functional role, should greatly extend our understanding of the relation between sequence and three-dimensional structure.

Acknowledgments

We thank R Durbin and G Mitchison for many helpful suggestions, particularly on statistics and the weighting scheme. We thank A Lesk, Y Harpaz, S Brenner, S Barrie, and J Baldwin for reading the manuscript. M G acknowledges support from a Herchel-Smith Fellowship.

References

Adman, E. T. (1992). Structure and function of copper containing proteins. *Curr. Opin. Struc. Biol.* **1**, 895-904.

Altschul, S. F., Carroll R. J. & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **207**, 647-653.

Arents, G. A. & Love, W. E. (1989). Glycera dibranchiata hemoglobin. Structure and refinement at 1.5 Å Resolution. *J. Mol. Biol.* **210**, 149-161.

Arutyunyan, E. G., Kuranova, I. P., Vainshtein B. K. & 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.

Bairoch, A. & Boeckmann, B. (1992). The Swiss-Prot protein-sequence data-bank. *Nucl. Acids Res.* **20**, 2019-2022.

Baker, E. N. (1988). Structure of azurin from *alcaligenes denitrificans*. Refinement at 1.8 A and comparison of the two crystallographically independant molecules. *J. Mol. Biol.* **203**, 1071- 1095.

Barker, W. C., George, D. G., Mewes H. W. & Tsugita, A. (1992). The PIR-international protein sequence database. *Nucl. Acids Res* **20**, 2023-2026.

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.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr, E. F., 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.

Bolin, J. T., D. J. Filman, D. A. Matthews, R. C. Hamlin & Kraut, J. (1982). Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding of methotrexate. *J. Biol. Chem.* **257**, 13650-13662.

Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi P. & Brunori, M. (1989). Aplysia limacina myoglobin. Crystallographic analysis at 1.6 Å resolution. *J. Mol. Biol.* **205**, 529-544.

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

Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **LII**, 399-405.

Davies, J. F., Delcamp, T. J., Prendergast, N. J., Ashford, A., Freisheim, H. & Kraut, J. (1993a). Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazofolate. Unpublished data bank entry.

Davies, J. F., Matthews, D. A., Oatley, S. J., Kaufman, B. T., Xuong, N. H. & Kraut, J. (1993b). Refined crystal structures of chicken liver dihydrofolate reductase. 3 Å apo-enzyme and 1.7 Å NADPH holo-enzyme complex. Unpublished data bank entry.

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.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat.* **125**: 1-15.

Fermi, G. & Perutz, M. F. (1981). *Haemoglobin and Myoglobin*. Oxford: Claredon Press.

Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at1.74 Å resolution. *J. Mol. Biol.* **175**, 159-174.

Finkelstein, A. V., Gutun, A. M. & Badretdinov, A. Y. (1993). Why are the same protein folds used to perform different functions? *FEBS Lett.* **325**, 23-28.

Finkelstein, A. V. & Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns?*Prog. Biophys. Mol. Biol.* **50**, 171-190.

Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279-284.

Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized poplar plastocyanin at 1.6 Å resolution. *J. Mol. Biol.* **169**, 521-562.

Harpaz, Y., Gerstein, M., & Chothia, C. (1993). Volume changes on protein folding. Submitted.

Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237-244.

Honzatko, R. B., Hendrickson, W. A. & Love, W.E. (1985). Refinement of a molecular model for lamprey hemoglobin from *Petromyzon marinus*. *J. Mol. Biol.* **184**, 147-164.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* **277**, 491-492.

Kendrew, J. C. & Watson, H. C. (1966). "Stabilizing interactions in globular proteins" in *Principles of Biomolecular Organization*. (ed. G. E. W. Wolstenholme & M. O' Connor) London: J & A Churchill.

Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.

Lim, V. I. & Ptitsyn, O. B. (1970). On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins. *Mol. Biol. (USSR)* **4**, 372-382.

Lim, W. A. & Sauer, R.T. (1989). Alternative packing arrangements in the hydrophobic core of lambda-repressor. *Nature* **339**, 31-36.

Morris, A. L., Macarthur, M. W. Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein-structure coordinates. *Proteins Struc. Func. Genet.* **12**, 345-364.

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

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Phillips, S. E. V. & Schoenborn, B. P. (1981). Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature* **292**, 81-82.

Ptitsyn, O. B. & Volkenstein, M. V. (1986). Protein structures and the neutral theory of evolution. *J. Biomol. Struct. Dyn.* **4**, 137-156.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.

Sibbald, P. R. & Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* **216**, 813-818.

Sneath, P. H. A. & R. R. Sokal (1973). *Numerical Taxonomy*. San Francisco: W H Freeman.

Steigemann, W. & Weber, E. (1981). Structure of erythrocruorin in different ligand states refined at 1.4 Å resolution. *J. Mol. Biol.* **127**, 309-388.

Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *CABIOS* **5**, 115-121.

Table 1 : Key Sequences and Their Associated Structures

PDB identifier	Key Sequence (SwissProt or PIR identifier)		Structure Reference	Seq. Ident. (1)	Wt. (2)
Globins					
2hhb	HAHU	Human Hemoglobin α-chain	Fermi *et al*. (1984)	100	0.10
2hhb	HBHU	Human Hemoglobin β-chain	Fermi *et al*. (1984)	45	0.16
2lhb	GGLMS	Sea Lampry Hemoglobin	Honzatko *et al*. (1985)	37	0.81
1mbd	MYWHP	Sperm Whale Myoglobin	Philips & Schoenborn (1981)	28	0.40
2hbg	GGNW1B	Bloodworm Hemoglobin	Arents & Love (1989)	24	4.35
1mba	GGGAA	Sea Hare Myoglobin	Bolognesi *et al*. (1989)	18	2.08
1ecd	GGICE3	Chironomous Hemoglobin	Steigemann & Weber (1981)	18	1.10
2lh4	GPYL2	Leghemoglobin	Arutyunyan *et al*. (1980)	18	1.05
Plastocyanin-azurin family (Plas.-Az.)					
1pcy	CUPX	Plastocyanin	Guss & Freeman (1983)	100	1.03
2aza	AZALCO	Azurin	Baker (1988)	22	1.01
Dihydrofolate reductases (DHFR)					
1dh1	DYR_HUMAN	Human DHFR	Davies *et al*. (1993a)	100	0.61
8dfr	DYR_CHICK	Chicken DHFR	Davies *et al*. (1993b)	79	0.81
4dfr	DYRA_ECOLI	E.Coli DHFR	Bolin *et al*. (1982)	35	1.08
3dfr	DYR_LACCA	L.Casei DHFR	Bolin *et al*. (1982)	22	1.33

(1) Percentage sequence identity to first key sequence in family (e.g. HAHU)

(2) The weight w(s) assigned to this particular sequence in the context of all the sequences in the family. The average weight of each sequence is 1.0 .

Table 2 : Residue Sites Constituting the Core and the Surface

globin core:

A8 A11 A12 A15 B6 B9 B10 B13 B14 C4 CD1 CD4 E4 E7 E8 E11 E12 E15 E18 E19 F1 F4 F8 FG4 G5 G8 G11 G12 G13 G15 G16 H7 H8 H11 H12 H15 H19

globin surface:

A6 A10 A13 A14 B12 C6 CD2 E2 E3 E5 E9 E13 E17 E20 F2 F3 F6 FG1 G1 G6 G10 G17 H5 H10 H13 H9

dihydrofolate reductase core:

5-11, 15-17, 24, 27, 30, 33, 34, 38, 49-53, 56-57, 60, 67, 70, 83, 86, 90, 93, 96, 100, 112-117, 120, 121, 124, 133-136, 138, 148, 156, 177,179, 181, 182.

plastocyanin-azurin core:

1, 3, 5, 14, 19, 21, 27, 29, 31, 33, 37-41, 72, 74, 80, 82, 86, 86, 92, 94, 96.

Listing of the residues forming the core in each of the three families studied and forming the surface in the globins. In the structures used for the alignment, i.e., those corresponding to the key sequences, residues in the core had, on average, less than 15 $Å^2$ of accessible surface. In addition, we included sites that had higher mean accessible surface areas, between 15 $Å^2$ and 20 $Å^2$, and whose side chains pointed into the protein interior. Surface residues had at least 50 $Å^2$ of accessible surface. For the globins, this definition closely paralleled that in Bashford *et al*. (1987).

For the globins residues are numbered according to the canonical numbering scheme introduced by Kendrew (Fermi & Perutz, 1981) and described in Bashford *et al*. (1987); for the dihydrofolate reductases, residue numbering refers to the human sequence (DYR_HUMAN); and for the plastocyanin-azurin family, numbering refers to the poplar plastocyanin sequence (CUPX).

Table 3 : Standard Residue Volumes and Frequencies

Residue Type	Our Freq. f_{tr} (1)	Janin (1979) Freq. (2)	Standard Volume V_t (3)
GLY	11.1 %	11.8 %	64
ALA	13.4 %	11.2 %	90
VAL	13.4 %	12.9 %	139
LEU	12.6 %	11.7 %	164
ILE	8.8 %	8.6 %	164
PRO	2.6 %	2.7 %	124
MET	2.5 %	1.9 %	167
PHE	6.2 %	5.1 %	193
TYR	3.1 %	2.6 %	197
TRP	1.9 %	2.2 %	231
SER	5.8 %	8.0 %	95
THR	5.5 %	4.9 %	121
ASN	1.5 %	2.9 %	126
CYS	4.0 %	1.6 %	113
HIS	1.6 %	4.1 %	159
GLU	1.4 %	2.0 %	142
ASP	3.0 %	1.8 %	118
ARG	0.9 %	2.9 %	195
LYS	0.7 %	0.5 %	170

(a) Frequency of 20 residue types in buried residues in proteins. The frequencies were defined by counting the number of buried residues in a data base of 119 protein crystal structures. Structures in the database all were solved to very high resolution (between 1.0 and 1.9 Å), had R-factors below 20 %, and had good stereochemistry as defined by Morris et al. (1992). Buried residues were defined as those with less than 15 $Å^2$ accessible surface area. This column of numbers was used for frequency distribution (ii) in Table 8(a) .

(b) Our residue frequencies shown in the first column are very similar to those in Janin (1979), an earlier determination of the frequencies of buried residues, which was based on the fewer high resolution structures known at the time.

(c) Standard Voronoi volumes for each residue type, taken from Harpaz *et al*. (1993).

Table 4 : Definitions and Averaging Scheme

Number of residue sites :	R
Number of sequences in a family :	N
Weight applied to sequence s to compensate for over-representation : where $\sum w(s) = N$.	$w(s)$,
Standard residue volume of residue at site r in sequence s :	V_{rs}
Frequency of residue type t at site r : where $\sum_t f_{tr} = 1$.	f_{tr} ,
Standard residue volume of residue type t :	V_t

For the volume of an individual site r, mean V_r and variance σ_r^2 ,

observed over all sequences in a family :

$$V_r = \frac{1}{N} \sum_{s=1}^{N} w(s) V_{rs} \qquad \sigma_r^2 = \frac{1}{N-1} \sum_{s=1}^{N} w(s) (V_{rs} - V_r)^2$$

calculated according to residue frequencies :

$$V_r = \sum_{t=1}^{20} f_{tr} V_t \qquad \sigma_r^2 = \frac{20}{19} \sum_{t=1}^{20} f_{tr} (V_t - V_r)^2$$

S.D. (%) of volume of individual site r : $\frac{\sigma_r}{V_r}$

Mean S.D. (%) of volume of all individual sites :

$$\left\langle \frac{\sigma_r}{V_r} \right\rangle = \frac{1}{R} \sum_{r=1}^{R} \frac{\sigma_r}{V_r}$$

Core volume of sequence s :

$$V_s = \sum_{r=1}^{R} V_{rs}$$

For the volume whole core, mean V and variance σ^2, observed over all sequences in a family :

$$V = \frac{1}{N} \sum_{s=1}^{N} w(s) V_s \qquad \sigma^2 = \frac{1}{N-1} \sum_{s=1}^{N} w(s) (V_s - V)^2$$

S.D. (%) of core volume : $\frac{\sigma}{V}$

Unless explicitly stated otherwise, the definitions in this table apply to all formula and mathematical expressions used throughout the text, tables, and figures.

Table 5 : Volume Variation of Interfaces

Helix	Number of residues	Volume Variation (% S.D)
A	4	6.5
B	5	9.0
E	8	8.1
F	3	4.5
G	7	4.8
H	6	5.4

The variation in volume of the core sites in the 6 main globin helices. The variation in volume is smaller than that of individual sites but larger one than that of the whole core.

Table 6 : Observed Volume Variation of the Whole Core

	Globins	Plas.-Az.	DHFR
Number of Buried Residue Sites R	37	24	56
Number of Sequences N	568	40	24
Average Core Volume V	5607	3616	8201
S.D. in Core Volume $\frac{\sigma}{V}$	2.6 %	1.5 %	2.2 %
Average Core Volume Per Residue	152	151	146

The total volume of the whole core varies less than 2.6 % in each of the three families studied. Moreover, the average core residue in each of these three families is roughly similar in volume. It contains contains 7-8 atoms and is roughly between Val and Leu in size.

Table 7 : Comparison of Sequence and Structure Calculations for the Globins

Sequence identifier (1)	Sequence Volume V_s (Å^3) (2)	Structure identifier (1)	Structure Volume V_x (Å^3) (3)	Difference (4) from mean for sequences $(V_s - V) / V$	Difference (4) from mean for structures $(V_x - \overline{V_x}) / \overline{V_x}$
HAHU	5402	2hhb	5941	-2.9 %	-1.4 %
HBHU	5544	2hhb	6001	-0.4 %	-0.4 %
GGLMS	5518	2lhb	5959	-0.9 %	-1.1 %
MYWHP	5660	1mbd	6117	1.7 %	1.5 %
GGNW1B	5600	2hbg	6169	0.6 %	2.4 %
GGGAA	5990	1mba	6419	7.6 %	6.6 %
GGICE3	5878	1ecd	6444	5.6 %	7.0 %
GPYL2	5608	2lh4	6145	0.8 %	2.0 %

(1) Sequences and structures correspond to the key sequences used in the globin alignment (Table 1).

(2) Volume (Å^3) of the core of a particular globin sequence. For each of the 37 buried sites (table 2), the standard residue volumes (table 3) were used to calculate a volume.

(3) Volume (Å^3) of the core of a particular globin X-ray crystal structure (V_x). For each of the 37 residues that formed the core, the atoms in the core were used for Voroni polyhedra calculations. Our implementation of Richards' (1974) Voronoi polyhedra program was used for the calculations. In all cases the structure volumes will be more than the sequence volumes (V_s) and are *not* directly comparable. This disparity exists because the core residue definition in table 2 is based on the average of all the structures. Any individual structure will, therefore, always expose to solvent a few of the atoms thought to be in core, and this exposure will tend to enlarge the calculated Voronoi polyhedra and increase the calculated volume.

(4) For the sequences, the percentage difference of a particular sequence volume V_s compared to the mean core volume V (defined in Table 4 and shown in Table 6) is listed. For the structures, percentage difference in a particular structure volume V_x compared to the mean value of $\overline{V_x}$ for all 8 structures is listed. As they have been normalized to compensate for the smaller average size of the structure volumes, these two percentage differences are comparable. They show that volume variations of the core calculated for the structures (-2 to +7%) is comparable to that calculated for the sequences (-3 to +8 %).

Table 8 : Comparison (1) of the Observed Variation with that of Random Sequences

Part (a): Core Sites

	Globins	Plas.-Az.	DHFR
Observed Values (from Table 6 and Figure 1a):			
Number of core sites R	37	24	56
Mean S.D. of individual sites $\langle \frac{\sigma_r}{V_r} \rangle$	14 %	11 %	13 %
S.D. in Core Volume $\frac{\sigma}{V}$	2.6 %	1.5 %	2.2 %
Assuming sites vary independently and randomly according to the same frequency distribution (same f_{tr} for all sites) (2) :			
i. Uniform distribution of all 20 amino acids :			
S.D. of individual sites	27 %	27 %	27 %
Expected S.D. in Core Volume	4.5 %	5.6 %	3.7 %
ii. Distribution of buried residues in an average protein :			
S.D. of individual sites	31 %	31 %	31 %
Expected S.D. in Core Volume	5.0 %	6.4 %	4.3 %
iii. Distribution of buried residues in a particular protein family :			
S.D. of individual sites	24 %	26 %	30 %
Expected S.D. in Core Volume	4.2 %	4.9 %	4.0 %

(1) This table involves the comparisons between the observed variance (i.e., standard deviation) of the core and some expected variances based on different random sequences. The Fischer F-test can be used to test whether the difference between the two variances is significant. Application of this test to the all comparisons in this table, using N-1 degrees of freedom for the observed variances and (R-1)(N-1) degrees of freedom for the expected variances, shows that the differences between the variances are all significant.

(2) As all sites r are equivalent for the random distributions, the expected percentage standard deviation of the whole core $\frac{\sigma}{V}$ is directly related to the percentage standard deviation of an individual site $\frac{\sigma_r}{V_r}$:

$$\frac{\sigma}{V} = \frac{1}{\sqrt{R}} \frac{\sigma_r}{V_r} \quad .$$

The expected deviation and mean volume of a site, σ_r and V_r, are, in turn, calculated according to residue frequencies f_{tr} and standard volumes V_t, as shown in table 4. Depending on one's assumptions about the distribution of amino acids in the protein core, three different sets of residue frequencies can be used.

(i) For a uniform distribution of amino acids. A residue is picked at random from twenty equi-probable amino acids, i.e., $f_{tr} = \frac{1}{20}$ for all t and r.

(ii) For the distribution of buried residues in average protein, a residue is picked from the twenty amino acids according to the frequency distribution for buried residues in 119 high-resolution crystal structures., i.e. f_{tr} is taken from the second column of table 3. This distribution gives nearly the same results as a uniform distribution of uncharged and neutral amino acids, where a residue is picked from one of the 14 equi-probable amino acids, (G A V L I P M F Y W S T C H).

(iii) For the distribution of buried residues in a particular family, for each of the three protein families, the types of residues occurring in core in all the sequences are counted and used to make a frequency distribution similar to the one shown in table 3. Then a residue is picked according to these frequencies. That is, for each protein family (i.e. for the haemoglobins), the *actual* frequencies for residues of type t to be at site r, f^A_{tr}, are averaged over sites to give f_{tr}:

$$f_{tr} = \frac{1}{R} \sum_{r=1}^{R} f^A_{tr}$$

In this expression, the actual frequencies f^A_{tr} are different at each site (as in Table 9), while the averaged frequencies f_{tr} are the same at each site. The subscript r is kept in f_{tr} to be consistent with the notation in table 4 and to emphasize that f_{tr} refers to frequencies at an individual site.

Table 8(b) : Comparison of the observed variation with that of random sequences: Globin Surface Sites

Observed Values:

Number of surface sites	26
Mean S.D. of individual sites	24 %
S.D. in total volume of the sites	4.5 %

Assuming the surface sites vary independently and randomly according to a uniform distribution of the 13 non-hydrophobic residues :

S.D. of individual sites	28 %
Expected S.D. in Total Volume	5.5 %

The calculations in this table are completely analogous to those in part (a) but they are carried out on the globin surface sites. Generating random surface sequences consists of picking a residue from the 13 non-hydrophobic amino acids (G A P Y S T N Q H E D R K), where each of these amino acids has an equal chance of being picked. Clearly there is much better agreement between the observed and calculated variation for surface sites than for buried sites.

Table 9 : Relationship of the Observed Single-Site variations to the Observed Variation of the Whole Core

	Globins	Plas.-Az.	DHFR
Observed Values (from Table 6 and Figure 1a):			
Number of Sites B	37	24	56
Mean S.D. of individual sites $\langle \frac{\sigma_r}{V_r} \rangle$	14 %	11 %	13 %
S.D. in Core Volume $\frac{\sigma}{V}$	2.6 %	1.5 %	2.2 %
Assuming sites vary independently and randomly according to the (different) observed frequency distributions for each site (different f_{tr} for each site) :			
Expected S.D. in Core Volume $\frac{\sigma}{V}$ (1)	2.4 %	2.5 %	2.0 %

(1) The variances of uncorrelated distributions add, so the expected percentage standard deviation of the whole core if the sites varied independently would be :

$$\left(\frac{\sigma}{V}\right) = \frac{1}{V}\sqrt{\sum_{r=1}^{R} \sigma_r^2} \; .$$

Figure Captions

Figure 1 : Volume Variation at Individual Sites

(a) Histogram of the variation in volume of the individual core sites ($\frac{\sigma_r}{V_r}$) in each of the three families (BLACK=dihydrofolate reductases; WHITE=globins; GRAY=plastocyanin-azurin family). Note the similarity between the families. The mean volume variation at individual sites $\langle \frac{\sigma_r}{V_r} \rangle$ is 14 % for the globins, 11 % for the plastocyanin-azurin family, and 13 % for the dihydrofolate reductases.

(b) Histogram comparing the volume variation of the individual sites ($\frac{\sigma_r}{V_r}$) in the globins in the core (BLACK) and with those on the surface (WHITE). The major difference is that there are no sites with small (<10%) volume variation on the surface.

Figure 2 : Core Volume Variation versus Sequence Identity

The volume of the core versus sequence identity to the first key sequence for all three protein families. The smallest core belongs to the plastocyanin-azurin family, followed by the globin family, and then dihydrofolate reductase family. For these three families, sequence identity was computed relative to the key sequences: AZALCO, HAHU, and DYR_HUMAN, respectively (see Table 1). For purposes of comparison, the volume of a methyl group is 26 Å^3 .

Figure 3 : Volume Variation of Random Sequences

The volume in the core versus sequence identity to one particular sequence for a random sequence. The sequence contained 37 sites, the same number as the globin core, and was generated by picking residues from a uniform distribution of amino acids, i.e., where each residue type is equi-probable. The sequence has a mean volume very similar to that of the globins (144 Å^2 per residue) but has a standard deviation roughly twice as large.

Appendix

A method to weight protein sequences
to correct for unequal representation

To avoid misleading results due to the unequal representation of sequences in a multiple sequence alignment (Felsenstein, 1985), it is desirable to reduce the weight of over-represented sequences (Altschul *et al*., 1989; Sibbald & Argos, 1990). The idea of most weighting schemes is that sequences located in densely populated regions of sequence space should get a lower weight than sequences in sparsely populated regions (assuming that the theoretical 'true' distribution is flat and has a centroid that coincides with the centroid of the observed distribution). For instance, Vingron & Argos (1989) calculate the weight of one sequence as the sum of its pairwise distances to all the other sequences, and Sibbald & Argos (1990) weight each sequence according to its Voronoi volume in sequence space.

The basis for our weighting scheme is using the distances between the sequences to cluster them in a bifurcating tree (Fitch & Margoliash, 1967). To construct the tree, we use the method of arithmetic averaging of pairwise distances (Nei, 1987; Sneath & Sokal, 1973), where the distance between sequences is measured by percentage residue identity. This method has been implemented in the program CLUSTALV (Higgins & Sharp, 1988). Since no known tree-construction method consistently makes better trees than other methods, we chose the distance averaging method because of its conceptual simplicity and modest requirements for CPU time and memory. However, our weighting scheme could in principle be applied to any rooted tree, independent of the algorithm used to construct the tree.

We call each point of bifurcation in the tree a node. The vertical length of the edges connecting two sequences to their common ancestral node represents the average sequence identity between the two clusters of sequences, i.e., the left and right subtrees of the ancestral node. In our usage, a subtree can contain many sequences or just a single sequence. An example of a bifurcating tree is drawn in Figure 4.

Figure 4 near here

To calculate weights for sequences we count distances between nodes in a tree. A shared edge, or distance, between subtrees will give rise to a shared weight increment for all the sequences in each subtree, according to the proportions already established within the subtree. At each node we calculate the weight increment for the sequences in the subtree above and continue doing so in a recursive fashion from the node closest to 100% (leaves) to the one closest to 0% (root). The relative weights of the sequences in a subtree are hence fixed when the weights for that subtree are calculated; afterwards they may change in absolute value but their relative value (to each other) will remain unchanged.

Our algorithm can be written as follows: All sequences initially have a weight of 0. We traverse the tree by visiting each node from 100% (leaves) to 0% (root) and for that node determine the weight increment for the left and right subtrees. The left subtree increment applies to all sequences in the this subtree, and likewise for right subtree increment. The weight added to a subtree is the length of the edge connecting it to the node currently being visited. The length of this edge is measured from the last, previously visited node of the subtree. It is apportioned between the sequences, so that at each node the sequence weights are updated according to the following formula:

$$w(s,b) \leftarrow w(s,b) + D(b)F(s,b),$$

where b is L or R for the left or right subtree,

s runs over all sequences in a subtree,

$D(b)$ is the edge length to be apportioned,

$w(s,b)$ is the current weight of sequence s in subtree t,

and $F(s,b)$ is the weight fraction of sequence s in the subtree t, i.e.,

$$F(s,b) = \begin{cases} 1, & w(s,b) = 0 \\ \dfrac{w(s,b)}{\sum_s w(s,b)}, & otherwise \end{cases}.$$

The first case of the formula ($w(s,b) = 0$) is used when a node is directly connected to a sequence. In this case, the connecting edge is unshared, and the formula apportions its length completely to the sequence.

Figure 4 presents a completely worked out example of the calculation of weights for a simple tree.

Table 10 near here

Table 11 near here

As shown in Table 10(a), on easy-to-classify sequences such as "AAAAAAAAAA" and "BBBBBBBBBB," our scheme gives completely intuitive weights. Furthermore, as shown in Table 10(b), in calculations on real sequences, it gives results that are similar to those of Sibbald and Argos (1990). In principle (i.e., given the same assumptions we made earlier about sequence space), Sibbald & Argos' weighting scheme should give the correct results. However, it requires a Monte-Carlo integration, which is time-consuming to perform and adds an element of randomness to the calculation (i.e., two separate calculations of the weights will not give exactly the same answers). Our method, in contrast, is fast and involves no random numbers (so it gives exactly the same results each time).

As discussed in the main body of the text and shown in Table 11, our weighting method did not qualitatively affect our conclusions about volume variation in the protein core. Rather, it gave us more confidence in the quantitative accuracy of our results.

Table 10 : Comparison of Our Weighting Scheme with Oher Methods and with Intuition

Part (a): Comparison with Intuition

Sequence	Weight
AAAAAAAAAAA	1/6
AAAAAAAAAAA	1/6
BBBBBBBBBB	1/6
BBBBBBBBBB	1/6
CCCCCCCCCC	1/3

Part (b) : Comparison with Other Methods

Sequence (PIR identifier)	Weight from Vingron & Argos (1989)	Weight from Sibbald & Argos (1990)	Weight from our method
HAGY	0.4631	0.4613	0.4751
HAHOZ	0.1321	0.1092	0.1013
HAHO	0.1338	0.1464	0.1447
HAHOD	0.1304	0.0937	0.1013
HAHOK	0.1407	0.1894	0.1776

The sequences shown above are taken from Sibbald & Argos (1990). For simple sequences, the weights our method assigns to the sequences are are in good accord with intuition.

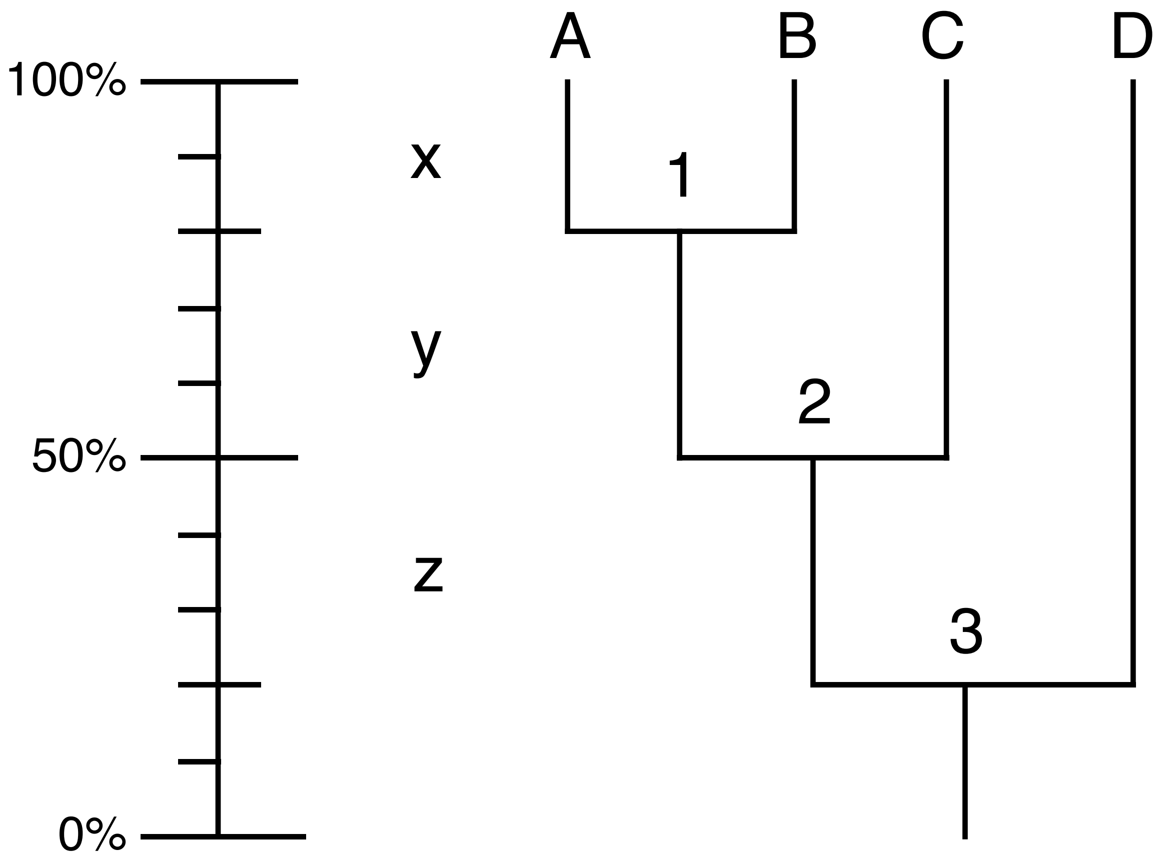
For five globins sequences, our method produces similar weights to those of Sibbald & Argos (1990). Both parts (a and b) are adapted from from Sibbald & Argos (1990). Note to make the comparison with Sibbald & Argos clearer, the weights in this table are normalized so that the sum of the weights is 1. This normalization follows a different convention from that used in the rest of the text.

Table 11 : Effect of the Weighting Scheme on our Results

	Globins	Plas.-Az.	DHFR
Calculated with weighting (from Table 6):			
Number of Sequences N	568	40	24
S.D. in Core Volume $\frac{\sigma}{V}$	2.6 %	1.5 %	2.2 %
Calculated without weighting:			
S.D. in Core Volume $\frac{\sigma}{V}$	2.2 %	1.4 %	2.0 %

This table shows the effect of the weighting scheme on our principal result, the volume variation of the whole core. In all cases, the weighting scheme tended to increase the observed volume variation. It gave greater weight to under-represented sequences in the alignment, and these sequences tended (perhaps obviously) to differ more from the mean volume than better represented sequences. Note also that the the effect of the weighting scheme was more pronounced in the larger globin alignment.

Figure 4 : A Worked Example of Our Weighting Method



The figure shows a bifurcating tree with four sequences: A, B, C, and D. A and B are 80% identical; the average identity between C and A or B is 50%; and the average identity between D and A, B, or C is 20 %. The weights for each sequence, denoted w(s), are calculated by visiting the nodes sequentially (first node 1, then 2, and finally 3), and adding increments to the total weight at each node. At the end the final weights are normalized, so that the average weight is 1. The calculation is summarized below:

	A	B	C	D
w(s) at start	0	0	0	0
Added at 1	$x = 20$	$x = 20$	$0 = 0$	$0 = 0$
Added at 2	$\frac{y}{2} = 15$	$\frac{y}{2} = 15$	$x+y = 50$	$0 = 0$
Added at 3	$z\,\frac{x+\frac{y}{2}}{3x+2y} = 8.75$	$z\,\frac{x+\frac{y}{2}}{3x+2y} = 8.75$	$z\,\frac{x+y}{3x+2y} = 13$	$x+y+z = 80$
w(s) at end	43.8	43.8	63	80
normalized	0.76	0.76	1.09	1.39