Protein Science (1994), *3*:1706–1711. Cambridge University Press. Printed in the USA.

INVITED PAPER, SPECIAL SECTION IN HONOR OF MAX PERUTZ

Quantification of tertiary structural conservation despite primary sequence drift in the globin fold

HANS-ERIK G. ARONSON,[1] WILLIAM E. ROYER, JR.,[2] AND WAYNE A. HENDRICKSON[1]

[1] Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032
[2] Program in Molecular Medicine and Department of Biochemistry and Molecular Biology, University of Massachusetts Medical Center, Worcester, Massachusetts 01605

(RECEIVED July 20, 1994; ACCEPTED July 25, 1994)

Abstract

The globin family of protein structures was the first for which it was recognized that tertiary structure can be highly conserved even when primary sequences have diverged to a virtually undetectable level of similarity. This principle of structural inertia in molecular evolution is now evident for many other protein families. We have performed a systematic comparison of the sequences and structures of 6 representative hemoglobin subunits as diverse in origin as plants, clams, and humans. Our analysis is based on a 97-residue helical core in common to all 6 structures. Amino acid sequence identities range from 12.4% to 42.3% in pairwise comparisons, and, despite these variations, the maximal RMS deviation in α-carbon positions is 3.02 Å. Overall, sequence similarity and structural deviation are significantly anticorrelated, with a correlation coefficient of −0.71, but for a set of structures having under 20% pairwise identity, this anticorrelation falls to −0.38, which emphasizes the weak connection between a specific sequence and the tertiary fold. There is substantial variability in structure outside the helical core, and functional characteristics of these globins also differ appreciably. Nevertheless, despite variations in detail that the sequence dissimilarities and functional differences imply, the core structures of these globins remain remarkably preserved.

Keywords: hemoglobin; molecular evolution; protein fold; sequence divergence; structural superposition

With the first 2 crystal structures of protein molecules, namely those of sperm whale myoglobin (Kendrew et al., 1958, 1960) and of horse hemoglobin (Perutz et al., 1960, 1968), it became clearly evident that 3-dimensional structure is remarkably conserved in the course of a molecular evolution that results in substantial divergence of amino acid sequence. As structures were determined for other, more distantly related members of the globin family, this preservation of tertiary structure was seen to extend to instances where the primary structures had diverged nearly to the point of unrecognizable similarity (Huber et al., 1971; Hendrickson & Love, 1971; Padlan & Love, 1974). Qualitative comparisons of these globin structures provided a compelling illustration of this conservation (Hendrickson & Love, 1971; Love et al., 1971), and the basis for this structural integrity has been analyzed (Lesk & Chothia, 1980; Bashford et al., 1987). The principle of structural inertia at the tertiary level is now known in other protein families such as in the immunoglobulin superfamily (Williams & Barclay, 1988), in the helical cytokines (Rozwarski et al., 1994), in type III domains from fibronectin (Bork & Doolittle, 1992), and in the cystine-knot growth factors (Wu et al., 1994) where the sequence similarity can fall well below the "twilight" zone of homology comparison. In light of apparent generality of this feature of molecular evolution, and given more recent additions to the globin family of structures, we feel that an updated and better quantified comparison of globin structures is in order.

Although a globin-like fold is found in the phycocyanins and phycoerythrins (Schirmer et al., 1985; Ficner et al., 1992), as well as in those heme proteins involved in the transport and storage of oxygen, we concentrate here on these hemoglobin and myoglobin molecules. To date, many of these proteins have been characterized by high-resolution X-ray crystallography (Takano, 1977; Steigemann & Weber, 1979; Arutyunyan et al., 1980; Phillips, 1980; Shaanan, 1983; Fermi et al., 1984; Honzatko et al., 1985; Kuriyan et al., 1986; Arents & Love, 1989; Bolognesi et al., 1989; Hubbard et al., 1990; Braden et al., 1994; Kolatkar et al., 1994; Royer, 1994), yielding complete 3-dimensional structural information. A number of other structures have been studied

Reprint requests to: Wayne A. Hendrickson, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, New York 10032; e-mail: aronson@cuhhca.hhmi.columbia.edu.

at lower resolution. The basic globin fold consists of 7 helices, labeled A, B, C, E, F, G, and H. These helices, which are present in all known globin structures, are sometimes supplemented with a D helix, as in the β-chain of human hemoglobin. The helices form a cavity that contains the heme group, protoporphyrin IX complexed with a hexacoordinated iron ion. Four of the iron coordination sites are occupied by pyrrole nitrogen atoms from the protoporphyrin group, the fifth coordination site is held by a histidine (the proximal histidine) from the F helix of the protein, and the sixth site takes on exogenous ligands. In the normal ferrous state, this site can reversibly bind oxygen (its primary role), as well as carbon monoxide and nitric oxide. In the ferric (met) state, it can bind cyanide, azide, nitric oxide, or water. In addition, in both oxidation states the iron can, in some cases, bind more exotic ligands, such as nicotinic acid (Appleby et al., 1973). Close to the sixth coordination site, there is usually a histidine (the distal histidine) that is part of the E helix. This residue can play a crucial role in the regulation of oxygen binding (Bashford et al., 1987; Olson et al., 1988).

In this study, we have compared 6 diverse representatives of the globin family: the α- and β-chains of human hemoglobin, lamprey (*Petromyzon marinus*) hemoglobin, clam (*Scapharca inaequivalvis*) hemoglobin, insect (*Chironomus thummi thummi*) larval hemoglobin, and yellow lupine (*Lupinus luteus* L.) root nodule leg hemoglobin. Many globins function as subunits of larger complexes. For those analyzed here, human hemoglobin is an $\alpha_2\beta_2$ tetramer; the clam hemoglobin is a homodimer; and the insect, plant, and lamprey hemoglobins are all monomers. Based on the comparative superposition of these high-resolution crystal structures of phylogenetically divergent globins, we show that the globin structure has remained remarkably unaffected by evolutionary mutations and functional variation.

Results and discussion

We have made a structural alignment of the sequences associated with the 6 globin structures selected for analysis and this result is shown in Figure 1. For each molecule, the best refined structure in a ligated state was chosen for comparison. Following an initial pairwise superposition of Cα-backbone structures, a 97-residue core was defined from helical elements that are in common to all 6 structures (Fig. 1), and this was then used to compute the degree of sequence similarity (Table 1) and the RMS structural deviation (Table 2) in pairwise comparisons. The results show that there is a high degree of structural conservation at the 3-dimensional level despite sometimes very little sequence similarity. Among these 6 globins from 5 phyla, the sequence identity can be as low as 12.4% (Table 1), but the RMS deviation of the Cα backbones is never greater than 3.02 Å (Table 2). Structural conservation is, of course, not independent of sequence similarity. The correlation coefficient between the results in Tables 1 and 2 is $C = -0.71$; that is, as sequence similarity goes down, structural deviation goes up. The correlation tends toward insignificant, however, when sequence similarity is very low. Thus, whereas $C = -0.75$ among the human, lamprey, and clam structures, where sequence comparisons range from 18.6% to 42.3% identity, the anticorrelation falls to $C = -0.38$ among the lamprey, clam, insect, and plant set, where identities range from 13.4% to 19.6%.

Stereo drawings of the α-carbon backbones for the 6 superimposed structures are shown in Figure 2 after translational displacement. Although the most striking aspect of this comparison is the structural similarity quantified in Table 2, clearly the structures do differ substantially in detail even at the Cα-backbone level. First, in the cases of lamprey and clam hemoglobins, there are substantial N-terminal extensions in comparison to the others. In the clam hemoglobin (Fig. 1D), this extension is helical and is called the pre-A helix (Royer, 1994), whereas in lamprey hemoglobin (Fig. 1C), this segment has an extended conformation looping back around the H helix. Second, 3 of the globins, human-β (Fig. 1B), lamprey (Fig. 1C), and insect (Fig. 1E), have a D helix; the others (Fig. 1A,D,F) do not. As is also evident from the sequence alignment, there are many other insertions and deletions at interhelical junctions. Only the B–C junction is immune from such changes. The E–F and G–H junctions are particularly variable. Finally, some distinction in heme tilts is noticeable. Despite these variations, the overall impression that one gains from Figure 2 is that of a remarkable structural integrity in the helical core of globins given the sequence diversity documented in Table 2.

In addition to variability in sequence and in structure outside the helical core, there is also substantial variation in function among these globins. On a detailed level, in insect hemoglobin the role of the distal histidine is played by isoleucine 62, rather than by the apparently conserved histidine 58 (Huber et al., 1971; Steigemann & Weber, 1979). Although the Cα of His 58 superimposes onto the Cα position of the distal histidines of the other globins much better than to any other Cα position, the side chain of Ile 62 has taken the position and role normally occupied by a histidine. In *Glycera* hemoglobin (Padlan & Love, 1974), the distal histidine position is occupied by a leucine residue.

Perhaps the most important differences among these globins is that 3 of them are part of oligomeric complexes that exhibit

Table 1. *Amino acid sequence identity*[a]

	Human-β	Lamprey	Clam	Insect	Plant
Human-α	42.3	36.1	20.6	12.4	17.5
Human-β		27.8	23.7	15.5	18.6
Lamprey			18.6	16.5	19.6
Clam				13.4	14.4
Insect					15.5

[a] Values cited are percentages for 97 residues in core segments identified in Figure 1.

Table 2. *Structural comparison*[a]

	Human-β	Lamprey	Clam	Insect	Plant
Human-α	1.18	1.45	1.62	2.08	2.73
Human-β		1.22	1.96	2.00	2.44
Lamprey			1.88	1.84	2.39
Clam				2.18	3.02
Insect					2.88

[a] Values cited are RMS deviations (Å) in the Cα positions for the 97 core residues identified in Figure 1.

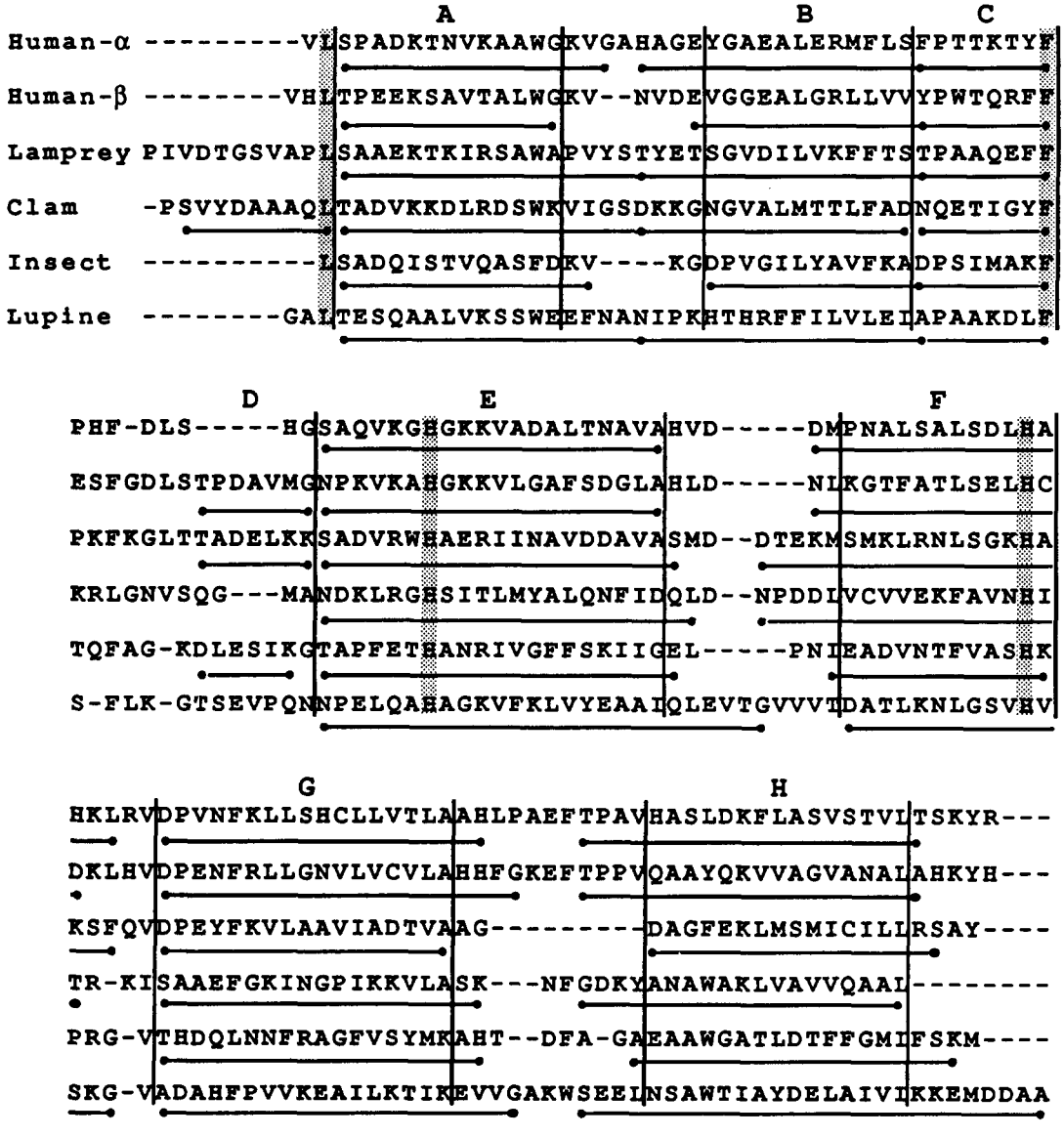

Fig. 1. Structural alignment of globin-chain amino acid sequences. Sequences of the 6 globin chains, as defined in the corresponding PDB entries specified in the Materials and methods, have been aligned in correspondence with superpositions of the 3-dimensional structures. The 4 invariant residues are shaded. The spans of helices identified by DSSP (Kabsch & Sander, 1983) are underlined and identified by the standard letter designations (Kendrew et al., 1960). Core helical segments in common to all 6 structures, and used in comparisons for Tables 1 and 2, are delimited by the vertical lines.

cooperativity. The α and β chains of human hemoglobin are part of an $\alpha_2\beta_2$ tetramer, whereas clam hemoglobin is a homodimer. Moreover, lamprey hemoglobin functions through ligand-induced dissociation of oligomers. The assembly of subunits in human and clam hemoglobins differs significantly. In clam, the E and F helices are on the inside of the assembly and the G and H helices are on the outside, whereas the reverse is true for mammalian hemoglobins (Royer, 1994). The alternate assembly of subunits in clam and mammalian hemoglobins reflects a very different mechanism for cooperativity. Thus, cooperativity in vertebrates and invertebrates must have developed independently, as has been noted earlier (Royer & Love, 1986).

Another interesting observation is the role of leghemoglobin in the root nodules of the leguminous plant, such as yellow lupine. This leghemoglobin (Fig. 1F) has been thought to take part in the process of nitrogen fixation, which takes place in the root nodules. It is believed that leghemoglobin modulates the availability of free oxygen to prevent oxidation of the sensitive metal clusters of bacterial nitrogenase. At the same time, leghemoglobin facilitates the oxygen flux necessary for bacteroid respiration (Appleby, 1984). This would then be the only globin involved in sequestering of oxygen as well as in transport and supply.

The comparison of the 6 globins in this study illustrates the great structural similarity of the proteins within this family. It is evident that low sequence homology does not preclude structural conservation. With the widespread sequence variability and the extensive functional variation within the globin family,

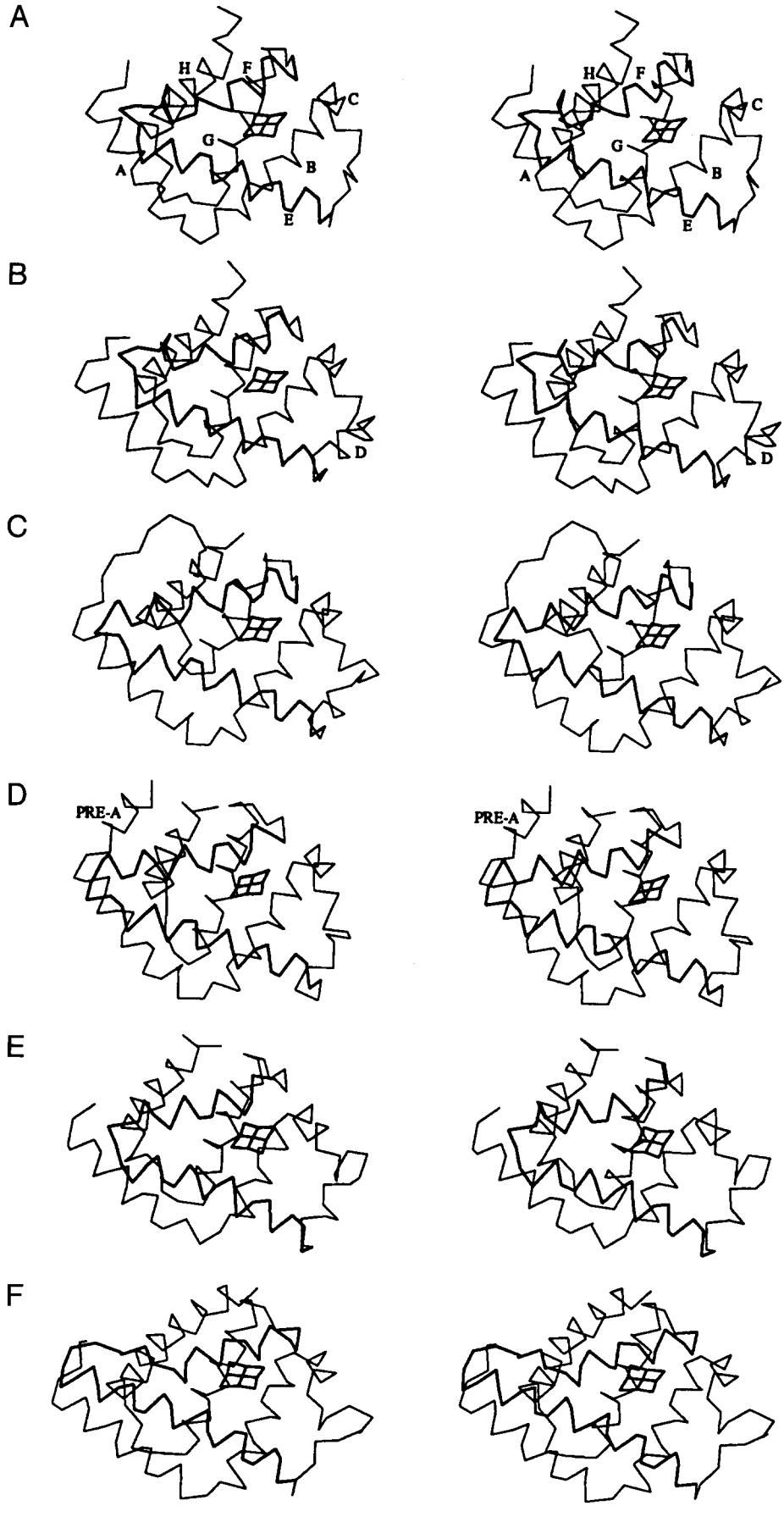

Fig. 2. Stereo drawings of α-carbon backbones for 6 globin structures: **(A)** α-subunit of human hemoglobin, **(B)** β-subunit of human hemoglobin, **(C)** lamprey hemoglobin, **(D)** *Scapharca* clam hemoglobin, **(E)** *Chironomus* insect hemoglobin, and **(F)** yellow lupine leghemoglobin. Atomic coordinates are from the corresponding PDB entries as identified in the Materials and methods. The Cα positions from core segments (Fig. 1) have been superimposed by TOSS (Hendrickson, 1979) onto lamprey hemoglobin in a standard view looking along the crystallographic *c*-axis (Hendrickson & Love, 1971). The superimposed models were then separated vertically for comparative display. Heme groups are represented by the iron atom, pyrrole nitrogen atoms, and methene-bridge carbon atoms of the group.

it is remarkable how well conserved the basic structure is. The great ability of the globin fold to endure, while accommodating significant alteration in the mode of its use, predicts that little variation will be found in the 3-dimensional structures of the more than 250 wild-type globins that have been sequenced.

Materials and methods

Sequence data and atomic coordinates

Refined coordinates for 5 crystallographically determined globin structures were obtained from the Protein Data Bank (PDB) (Bernstein et al., 1977): human (oxy) hemoglobin (Shaanan, 1983) (PDB file 1hho), lamprey (cyanomet) hemoglobin (Honzatko et al., 1985) (PDB file 2lhb), insect (carbonmonoxy) hemoglobin (Steigemann & Weber, 1979) (PDB file 1eco), yellow lupine root nodule (cyanomet) leghemoglobin (Arutyunyan et al., 1980) (PDB file 2lh3), and dimeric clam (carbonmonoxy) hemoglobin (Royer, 1994) (PDB file 3sdh). The sequence data for these structures were derived from these same PDB files.

Sequence alignment and comparison

The sequences were aligned (Fig. 1) by first finding the conserved structural features (i.e., the helices) and then aligning these according to conserved residues (a leucine at the start of the A helix, the proximal and distal histidines, and a phenylalanine at the end of the C helix). The 7 helices found in each of the structures were defined using DSSP (Kabsch & Sander, 1983), a program that relies strongly on hydrogen bonding patterns to define the secondary structural motif for each residue in a protein. In some cases, DSSP did not find clear breaks in the hydrogen bonding at the ends of a helix. Backbone torsion angles were then used to determine the helix boundaries. We allowed for no deletions within the helices, arguing that evolutionary pressures will favor deletions within loop regions rather than within highly conserved secondary structures (Brändén & Jones, 1990), such as the 7 helices found in all globins. Similarly, with one exception, we allowed only a single gap (insertion or deletion) within each interhelical region.

The aligned sequences were used to determine 7 segments that are helical in all 6 globins. These common segments, which correspond to core portions of the 7 conserved helices, are identified in Figure 1. They comprise a total of 97 residues, representing at least 63% of all residues in any of the 6 globins. The percent sequence identity (Table 1) was determined for the core helical segments only.

Structural comparison

The Cα backbones of the proteins were compared quantitatively using a modified version of TOSS (Hendrickson, 1979). This program performs a least-squares superposition of structurally equivalent points. In our case, these points were the α-carbon positions within the 7 core helical segments. In this manner, we obtained the RMS deviation (Table 2) between the superimposed Cα-backbone segments from the cores of pairwise compared structures. The correlation coefficient between sequence identity (q) and structural deviation (r) was computed as $C = \sum(q-\bar{q})(r-\bar{r})/[\sum(q-\bar{q})^2 \times \sum(r-\bar{r})^2]^{1/2}$. The structural display (Fig. 2) was generated using a locally modified version of the plotting routine PLUTO (Motherwell, 1976) that supports PostScript.

Acknowledgments

We thank Steven Sheriff and Arno Pähler for help with computer programs.

References

Appleby CA. 1984. Leghemoglobin and *Rhizobium* respiration. *Annu Rev Plant Physiol 35*:443–478.

Appleby CA, Wittenberg BA, Wittenberg JB. 1973. Nicotinic acid as a ligand affecting leghemoglobin structure and oxygen reactivity. *Proc Natl Acad Sci USA 70*:564–568.

Arents G, Love WE. 1989. *Glycera dibranchiata* hemoglobin. Structure and refinement at 1.5 Å resolution. *J Mol Biol 210*:149–161.

Arutyunyan ÉG, Kuranova IP, Vainshtein BK, Steigemann W. 1980. X-ray structural investigation of leghemoglobin. VI. Structure of acetateferrileghemoglobin at a resolution of 2.0 Ångstroms. *Sov Phys Crystallogr 25*(1):43–58.

Bashford D, Chothia C, Lesk AM. 1987. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol 196*:199–216.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, 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.

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.

Bork P, Doolittle RF. 1992. Proposed acquisition of an animal protein domain by bacteria. *Proc Natl Acad Sci USA 89*:8990–8994.

Braden BC, Arents G, Padlan EA, Love WE. 1994. *Glycera dibranchiata* hemoglobin. X-ray structure of the carbonmonoxide hemoglobin at 1.5 Å resolution. *J Mol Biol 238*:42–53.

Brändén CI, Jones TA. 1990. Between objectivity and subjectivity. *Nature 343*:687–689.

Fermi G, Perutz MF, Shaanan B, Fourme R. 1984. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J Mol Biol 175*:159–174.

Ficner R, Lobeck K, Schmidt G, Huber R. 1992. Isolation, crystallization, crystal structure analysis and refinement of B-phycoerythrin from the red alga *Porphyridium sordidum* at 2.2 Å resolution. *J Mol Biol 228*:935–950.

Hendrickson WA. 1979. Transformations to optimize the superposition of similar structures. *Acta Crystallogr A 35*:158–163.

Hendrickson WA, Love WE. 1971. Structure of lamprey haemoglobin. *Nature New Biol 232*(32):197–203.

Honzatko RB, Hendrickson WA, Love WE. 1985. Refinement of a molecular model for lamprey hemoglobin from *petromyzon marinus*. *J Mol Biol 184*:147–164.

Hubbard SR, Hendrickson WA, Lambright DG, Boxer SG. 1990. X-ray crystal structure of a recombinant human myoglobin mutant at 2.8 Å resolution. *J Mol Biol 213*:215–218.

Huber R, Epp O, Steigemann W, Formanek H. 1971. The atomic structure of erythrocruorin in the light of the chemical sequence and its comparison with myoglobin. *Eur J Biochem 19*:42–50.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*:662–666.

Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, Shore VC. 1960. Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. *Nature 185*:422–427.

Kolatkar PR, Hackert ML, Riggs AF. 1994. Structural analysis of *Urechis caupo* hemoglobin. *J Mol Biol 237*:87–97.

Kuriyan J, Wilz S, Karplus M, Petsko GA. 1986. X-ray structure and refinement of carbon-monoxy (FeII)-myoglobin at 1.5 Å resolution. *J Mol Biol 192*:133–154.

Lesk AM, 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.

Love WE, Klock PA, Lattman EE, Padlan EA, Ward KB Jr, Hendrickson WA. 1971. The structures of lamprey and bloodworm hemoglobins in

relation to their evolution and function. *Cold Spring Harbor Symp Quant Biol 36*:349–357.
Motherwell WDS. 1976. *PLUTO. Program for plotting molecular and crystal structures*. University of Cambridge, England.
Olson JS, Mathews AJ, Rohlfs RJ, Springer BA, Egeberg KD, Sligar SG, Tame J, Renaud J-P, Nagai K. 1988. The role of the distal histidine in myoglobin and haemoglobin. *Nature 336*:265–266.
Padlan EA, Love WE. 1974. Three-dimensional structure of hemoglobin from the polychaete annelid, *Glycera dibranchiata*, at 2.5 Å resolution. *J Biol Chem 249*(13):4067–4078.
Perutz MF, Muirhead H, Cox JM, Goaman LCG. 1968. Three-dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 Å resolution: The atomic model. *Nature 219*:131–139.
Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. 1960. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature 185*:416–422.
Phillips SEV. 1980. Structure and refinement of oxymyoglobin at 1.6 Å resolution. *J Mol Biol 142*:531–554.
Royer WE Jr. 1994. High-resolution crystallographic analysis of a cooperative dimeric hemoglobin. *J Mol Biol 235*:657–681.
Royer WE Jr, Love WE. 1986. The low resolution structures of the cooperative hemoglobins from the blood clam *Scapharca inaequivalvis*. In: Linzen B, ed., *Invertebrate oxygen carriers*. Berlin: Springer-Verlag. pp 111–115.
Rozwarski DA, Gronenborn AM, Clore GM, Bazan JF, Bohm A, Wlodawer A, Hatada M, Karplus PA. 1994. Structural comparisons among the short-chain helical cytokines. *Structure 2*:159–173.
Schirmer T, Bode W, Huber R, Sidler W, Zuber H. 1985. X-ray crystallographic structure of the light-harvesting biliprotein C-phycocyanin from the thermophilic cyanobacterium *Mastigocladus laminosus* and its resemblance to globin structures. *J Mol Biol 184*:257–277.
Shaanan B. 1983. Structure of human oxyhaemoglobin at 2.1 Å resolution. *J Mol Biol 171*:31–59.
Steigemann W, Weber E. 1979. Structure of erythrocruorin in different ligand states refined at 1.4 Ångstroms resolution. *J Mol Biol 127*:309–338.
Takano T. 1977. Structure of myoglobin refined at 2.0 Å resolution. II. Structure of deoxymyoglobin from sperm whale. *J Mol Biol 110*:569–584.
Williams AF, Barclay AN. 1988. The immunoglobulin superfamily—Domains for cell surface recognition. *Annu Rev Immunol 6*:381–405.
Wu H, Lustbader JW, Liu Y, Canfield RE, Hendrickson WA. 1994. Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure 2*:545–558.