J. Mol. Biol. (1965) **13**, 669–678

Structure and Function of Haemoglobin

II. Some Relations between Polypeptide Chain Configuration and Amino Acid Sequence

M. F. Perutz, J. C. Kendrew and H. C. Watson

Medical Research Council Laboratory of Molecular Biology
Hills Road, Cambridge, England

(*Received 18 June 1965*)

X-Ray data suggest that the globin chain has the same configuration in the myoglobins and haemoglobins of all vertebrates. Sequence data, on the other hand, show that only 9 out of more than 140 sites are occupied by the same amino acid residue in all the globin chains analysed so far. The different globins do not have any marked pattern of ionized or of polar residues in common. The most prominent common feature is the almost total exclusion of polar residues from the interior of the globin chains; this expresses itself in a pattern of 30 sites where only non-polar residues occur. Along α-helical segments these invariant non-polar sites tend to repeat on the average at regular intervals of about 3·6 residues, making the interior face of the helix non-polar.

At most sites at the surface or in surface crevices, on the other hand, replacements of many different kinds seem to occur without affecting the tertiary structure: these include replacement of non-polar by polar residues and *vice versa*.

Prolines are confined to the ends of helices or to non-helical regions; otherwise their incidence in the globins of different species is largely random. If all the proline sites observed are plotted along the sequence, they are seen to place limits on the possible lengths of helical regions. In many instances, serine, threonine, aspartic acid or asparagine occupies the first site at the amino end of the helix, followed by a proline at the second site.

In a protein of unknown structure, a regular periodicity of invariant non-polar sites might serve to recognize helical regions, and the incidence of prolines to define their lengths. A common pattern of non-polar sites might also help to identify structurally similar proteins with different enzymic function.

1. Introduction

With minor variations, the configuration of the polypeptide chain first discovered in sperm whale myoglobin (Bodo, Dintzis, Kendrew & Wyckoff, 1959; Kendrew *et al.*, 1960) is probably characteristic of the myoglobins and haemoglobins of all vertebrates (Kendrew, 1963; Perutz, 1963). Both genetic and physico-chemical evidence shows that the configuration of myoglobin and other proteins is not determined by an outside template, but is taken up spontaneously as the most stable configuration for the given amino acid sequence (Harrison & Blout, 1965; for other references, see Perutz, 1962). It seemed of interest, therefore, to ask what features the amino acid sequences of different types of globin chain have in common and which of these features might be important in determining the secondary and tertiary structures. Identical residues occupying structurally corresponding sites in all globin chains number only 9—too few to be decisive in determining the structure. The most prominent common feature is found to be a pattern of sites where only non-polar residues

occur. Most of these are located in the interior of the globin chain, away from contact with water. As a result, long α-helical segments exhibit a regular periodicity of invariant non-polar sites. Prolines are restricted to sites at the ends of helices or to non-helical regions. With one exception, the occurrence of prolines at various corners or non-helical regions of various globins is random. However, if the total incidence of prolines is plotted along the chain, it is found to define the lengths of nearly all the helical segments.

Table 1 lists the species of myoglobin and haemoglobin the amino acid sequences of which are either fully or partially known. The most recently published summary of sequences is in a review by Braunitzer, Hilse, Rudloff & Hilschmann (1964).

TABLE 1

Sources of information on amino acid sequences in myoglobin and haemoglobin

Myoglobin	Sperm whale	C	Edmundson, 1965
	Human	P	Hill, 1965 (U)
Haemoglobin	Human α and β	C	Braunitzer *et al.*, 1961 Konigsberg, Guidotti & Hill, 1961 Goldstein, Konigsberg & Hill, 1963
	γ	C	Schroeder, Shelton, Shelton & Cormick, 1963
	Lemur fulvus α and β	C	Hill, 1965 (U)
	Propithecus verreauxi	P	Hill (1965)
	Horse α	C	Braunitzer & Matsuda, 1961
	β	P	Smith, 1964
	Pig α and β	P	Braunitzer & Köhler, 1965 (U)
	Rabbit α and β	P	Naughton & Dintzis, 1962 Braunitzer & Schrank, 1965 (U)
	Lama α and β	P	Braunitzer & Hilschmann, 1965 (U)
	Carp α	P	Braunitzer & Hilse, 1965 (U)
	Lamprey	P	Braunitzer & Rudloff, 1965 (U)

C, Complete sequence; P, partial sequences of tryptic and other peptides; U, unpublished data communicated to the authors.

In cases where only the composition of tryptic peptides but not the sequence of residues along them was known, a tentative sequence was drawn up by using homologies with the most closely related chain of known sequence. This led to uncertainties in some of the very long peptides, but these hardly affect the present argument. The notation of sites adopted here follows that of Watson & Kendrew (1961) given in the preceding paper (Perutz, 1965†, Fig. 1 and Table 1). Sequential residue numbers of any site can be found by reference to Table 1 in paper I of this series, p. 650.

2. Invariant Residues (Table 2)

Residues are defined here as invariant when they occur at structurally identical sites in all the normal myoglobins and haemoglobins so far investigated. Abnormal haemoglobins have been excluded, because some of their abnormalities interfere with the oxygen-combining function, so that the protein can no longer be regarded as a haemoglobin.

† Perutz (1965) will be referred to as paper I of this series.

It might be thought that polypeptide chains which adopt a similar tertiary structure have many invariant residues in common. This holds for the family of cytochrome *c*'s, in which half the residues are invariant (Margoliash & Smith, 1964), but not for the family of globins, where the number of invariant residues has now shrunk to 9. These include the two haem-linked histidine residues E7 and F8. Two other residues which may form an essential part of the environment of the haem group are phenylalanine CD1 (see paper I, Fig. 6(b)) and leucine F4. In addition, there are four residues which

TABLE 2

The 9 invariant residues

B 6	Gly	Close contact with Gly or Ala E8
C 2	Pro	BC corner
C 4	Thr	BC corner and haem contact
CD 1	Phe	Haem contact
E 7	His	Distal haem link; absent in Hb of marine worm? (*Aplysia*†)
F 4	Leu	Haem contact
F 8	His	Haem linked
H10*	Lys	External. No visible function
H23*	Tyr	Internal H-bond with CO in FG corner

Present in primates, horse, pig, rabbit, lama, carp α and lamprey haemoglobins and sperm whale and human myoglobin, except:

A14, B6, GH5 and H10* unknown in lamprey. GH5 and H10* unknown in carp α and rabbit β.

The numbering of the residues in helix H adopted here differs from that used in earlier publications: H*n* (Watson & Kendrew, 1961) = H(n + 1)* (present paper).

† Wittenberg, Briehl & Wittenberg (1965).

may be important in determining the tertiary structure of the polypeptide chain: glycine B6 brings helix B into close contact with helix E at the site of glycine or alanine E8; proline C2 and threonine C4 form the BC corner; and tyrosine H23* forms an internal hydrogen bond with the α-carbonyl of residue FG5. Lysine H10* is external and has no obvious function. (For a discussion of these and other side-chains, see Watson & Kendrew, 1961; Kendrew, 1962.)

This list clearly shows that the invariant residues can have only a limited role in determining the secondary and tertiary structure of the globin chain.

3. Common Residues ionized at Neutral pH (Table 3)

It might be thought that the pattern of electric charges formed by the ionized side-chains helps to determine the structure of the globin chain. However, Table 3 shows only three sites which consistently carry basic side-chains ionized at neutral pH (histidines are not counted, since they are not necessarily ionized at neutral pH). Only two sites carry acidic side-chains and only one carries either acidic or basic ones. Seeing that the total number of ionized groups on the surface of sperm whale myoglobin is 43, it seems unlikely that the few sites listed in Table 3 are of decisive importance.

TABLE 3

Common residues ionized at neutral pH

Consistently basic	*Consistently acidic*
	A4 Asp, Glu
B12 Arg, Lys	B8 Asp, Glu
E 5 Arg, Lys	
H10* Lys	*Consistently either acidic or basic*
(G19 Arg, Lys, His)	
(H24* Arg, Lys, His)	G6 Arg, Lys, Glu

Present in primates, horse, pig, rabbit, lama and lamprey haemoglobin; also sperm whale and human myoglobin, except:

residues A14, B8 and H10* unknown in lamprey; H24* absent in lamprey.

4. Other Polar Residues

How many sites are consistently occupied by polar, though not necessarily by ionized, residues? There are only 17 such sites and 11 of these are connected with special functions: two haem-linked histidines, two lysines probably linked to the propionic acid side-chains of haem in haemoglobin, and seven polar residues occupying positions one and four in α-helices, as described below. Among the remaining six invariably polar sites no special function can be discerned at present.

5. Anti-helical Residues

If residues forming part of an α-helix are defined as having their α-carbonyl, or their α-imino groups or both, hydrogen-bonded within the helix, then model building shows that proline can occur in sites 1, 2, 3 or $n + 1$ of an α-helix containing n residues. Prolines are in fact observed in all these positions. A combination frequently found in haemoglobin consists of serine, threonine, aspartic acid or asparagine in position 1 followed by proline in position 2 of the helix. Kendrew & Watson have shown that in this situation the oxygen of the OH or COO^- group can be hydrogen-bonded to the α-NH group of residue 4 of the helix. An alternative combination, first found by Kendrew & Watson in the BC corner of myoglobin, consists of proline in position 2 with an internally hydrogen-bonded threonine in position 4. In one instance prolines actually occur in both positions 2 and 3 of an α-helix (helix Hβ of human haemoglobin).

Synthetic polypeptides consisting of serine, threonine, valine, isoleucine or cysteine do not form α-helices, but random coils or β-structures (Blout, de Lozé, Bloom & Fasmann, 1960; Blout, 1962). It might be expected, therefore, that these residues occur preferentially in non-helical regions. In fact, however, all the cysteines in haemoglobin lie in α-helical regions, and the other so-called "anti-helical" residues are found equally in helical and non-helical ones. The non-helical region EF in sperm whale myoglobin contains no anti-helical residue, and the neighbourhood of the AB corner in the α-chain of horse haemoglobin only one valine. On the other hand, it is true that, in sperm whale myoglobin, helix E is bent at a position where four out of five residues are of the anti-helical kind. Nevertheless, the absence of any consistent correlation implies that there must be other features of the amino acid sequence not yet considered which are powerful in determining the secondary structure.

6. Non-polar Residues (Tables 4 and 5)

These include glycine, alanine, valine, leucine, isoleucine, phenylalanine, proline, cysteine and methionine. In addition, the behaviour of tryptophan and tyrosine seems to be dominated by their non-polar aromatic rings rather than their polar groups, at least in the proteins considered here, and they will be classified as non-polar (Tanford, 1962).

TABLE 4

Replacements among the 33 internal sites

Residue	Observed	Residue	Observed
A 8	Val, Ile, Leu	E15	Val, Leu, Phe
11	Ala, Val, Leu	18	Gly, Ala, Ile
12	Trp, Phe	19	Val, Leu, Ile
15	Val, Leu, Ile		
		F 1	Leu, Ile, Phe, Tyr
B 6	Gly		
9	Ala, Ile, *Ser*, *Thr*	FG 5	Val, Ile
10	Leu, Ile		
13	Met, Leu, Phe	G 5	Phe, Leu
14	Leu, Phe	8	Val, Leu, Ile
		11	Ala, Val, Cys.H
C 4	*Thr*	12	Leu, Ile
		16	Leu, Val, Ser
CD 1	Phe		
4	Phe, Trp		
D 5	Val, Leu, Ile, Met	H 8*	Leu, Phe, Met, Trp, Tyr
		11*	Ala, Val, Phe
E 4	Val, Leu, Phe	12*	Val, Leu, Phe
8	Gly, Ala	15*	Val, Phe
11	Val, Ile	19*	Leu, Ile, Met
12	Ala, Leu, Ile	23*	Tyr

Invariably non-polar, except where indicated, in primates, horse, pig, rabbit, lama, carp α and lamprey haemoglobin; sperm whale and human myoglobin.

Not yet known:
A11, 15; B6, 9, 10; F1; G8, 11, 12, 16; GH5; H8*, 11*, 12*, 15* in lamprey (15 residues).
A8; G8, 11, 12, 16; GH5; H8*, 11*, 12*, 15*, 19* in carp α (11 residues).
G8, 11, 12, 16; GH5; H8*, 11*, 12*, 15*, 19* in rabbit β (10 residues).
H11*, 12*, 15*, 19* in lama α.
H15*, 19* in pig α.
G5, 11, 12, 16 in human myoglobin.
D5 absent in α-chains.

Examination of the models of myoglobin and haemoglobin shows that there are 33 sites which are interior in the sense that they are cut off from contact with the surrounding water. These are listed in Table 4. With three exceptions, they are invariably occupied by non-polar residues. The exceptions are threonine C4 mentioned previously, serine and threonine B9, and serine G16, which are probably also hydrogen-bonded internally. A wide variety of replacements seems to be open among the non-polar residues at many of the sites.

TABLE 5

Replacements among 44 non-polar residues at the surface or in surface crevices

Residue	Observed	Residue	Observed
NA 2	Gly, Val, Asp, His, Leu	EF 3	Gly, Asp
3	Leu, Phe	7	Gly, Ala, Asn, Lys
A 3	Gly, Ala, Asp, Glu	F 3	Ala, Ser, Thr, Asp, Gln, Lys, Pro
5	Trp, Lys, Asp	4	Leu
7	Gly, Ala, Leu, Thr, Asp, Asn, Lys, His	5	Ala, Ser
9	Gly, Leu, Thr, Asp, Asn, Lys, Arg	9	Ala, Cys.H
13	Gly, Ala, Ser, Thr, Asp, Lys	G 1	Asp, Pro
AB 1	Ala, Gly, Ser	2	Ile, Pro
		4	Tyr, Asn, Asp
B 2	Ala, Val, Glu	7	Leu, Ile, Phe
3	Gly, Ala, Asp, Glu, Lys	13	Ala, Val, Ile, Leu
4	Gly, Ala, Asp, Glu, Lys	15	Val, Thr, Glu
11	Gly, Ile, Glu	GH 2	Gly, Pro
C 2	Pro	3	Gly, Ala, Asp, Lys, His
5	Leu, Gln, Lys, Arg	5	Phe, Lys
		H 1*	Gly, Ser, Thr, Asp
CD 7	Leu, Ile, Met	2*	Ala, Pro
		4*	Val, Leu, Ala, Phe
D 3	Ala, Ser, Asp, Glu	6*	Gly, Ala
7	Gly, Ala, Lys	7*	Ala, Ser
		14*	Gly, Leu, Ser, Thr, Ala
E 9	Val, Ser, Glu, Lys	20*	Ala, Ser, Thr, Arg?
14	Gly, Ala, Ser, Glu, Asp	21*	Ala, Ser, His
16	Gly, Ser, Thr, Asp		
17	Ala, Leu, Asp, Asn, Glu, Lys		

All residues are exchanged for polar ones in one or other haemoglobin except:

NA 3 Surface crevice, holds NA segment in place.
C 2 At BC corner.
CD 7 Surface crevice, holds CD and D together.
F 4 Haem contact.
F 9 Reactive SH-group.
G 2 At FG corner.
G 7 Surface crevice.
G13 α–β contact.
H 4* Surface crevice.
H 6* Surface.

By constrast, there are only 10 sites at the surface or in surface crevices which remain consistently non-polar (Table 5). Seven of these non-polar residues have special functions. Sperm whale myoglobin contains 34 other non-polar residues at sites on or near the surface, but all these are replaced by polar residues in some of the other globin chains. The great variety of different residues which is permissible at many of the superficial sites is surprising.

Table 6 summarizes the replacements among the internal and superficial sites of sperm whale myoglobin observed in other globins.

TABLE 6

Summary of replacements of sperm whale myoglobin sites in different myoglobins and haemoglobins

		Site in sperm whale myoglobin			
		Polar		*Non-polar*	
(a)	33 internal sites	Threonine C4	1	Always non-polar	30
				Replaced by polar	2
					32
(b)	120 sites on surface or in surface crevices.	Haem-linked histidines	2	Always non-polar	
		Always basic	3	special function	7
		Always acidic	2	no visible function	3
		Always acidic or basic	1	Replaced by polar	34
		Always polar	17		
		Replaced by non-polar	45		
		Deleted	2	Deleted	4
			72		48
	Totals of (a) and (b)		73		80

Note: Replaced by polar, or non-polar, here means that such a replacement has been observed at least once.

7. Distinguishing Features of Sequences along Helical and Non-helical Regions

The only major consistent feature found here is the pattern of non-polar residues in the interior of the globin chains. In a long α-helical region, offering a regular alternation of external and internal sites, consistently non-polar residues at internal sites should recur on the average at intervals of about 3·6 residues. In Fig. 1 the secondary structure of the globin chain is summarized by representing the helical segments as sine waves and the non-helical ones as straight lines; the site of each residue is marked by a circle or cross. Black circles mark sites consistently occupied by non-polar residues, white circles those occupied by any kind of residue. Crosses mark the incidence of prolines or combinations of serines and threonines with prolines. The top of each sine wave points towards the inside of the globin chain or subunit, the bottom faces the surface.

A regular periodicity of sites occupied by non-polar residues is apparent along each of the longer α-helical regions A, B, E, G and H. There are three inconsistencies: the external sites G7, G13 and H6* are invariably non-polar; G13 probably because it lies at the αβ boundary in haemoglobin; G7 and H6* may prove to be not consistently non-polar. Segments G and H are those where least is known as yet about the sequences of many of the species included in this survey.

The structure of the globin chain is such that all non-helical regions lie at the surface, so that any of their side-chains could be exterior. Non-helical regions should therefore be

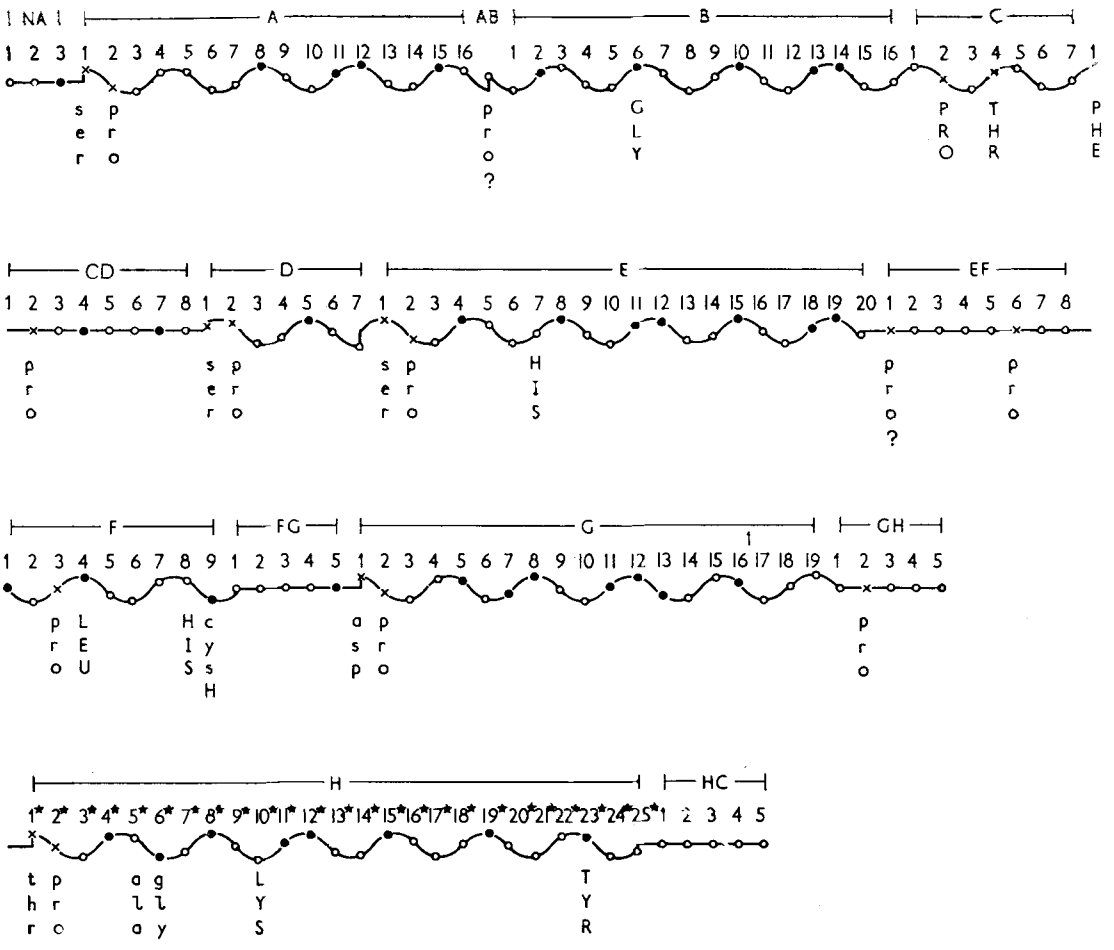

FIG. 1. Secondary structure of globin. α-Helical segments are represented as sine waves, non-helical ones as straight lines. Segments are marked as in paper I, Fig. 1 and Table 1. Black circles mark sites where only non-polar residues occur; crosses where prolines or combination of prolines with serine, threonine, aspartic acid or asparagine occur. All other sites are marked by blank circles. Residues in capital letters mark homologies, those in small letters mark other residues of special interest, such as the reactive cysteine of the β-chain which must be on the surface of the chain. Note that this diagram shows the total incidence of prolines in all species analysed so far. No single species contains prolines at all the points indicated here. Pro GH2 occurs in position $n + 1$ of helix G, since refinement of the structure of sperm whale myoglobin has shown GH1 to be the C-terminal residue of that helix. The sine waves are drawn so that the top of each wave points to the *inside* of the globin chain.

distinguished by the occasional occurrence of polar residues in all the sites. This is true of the regions EF and GH and, with one exception, of FG, but CD has two invariant non-polar sites at intervals of three; these are in fact connected by one helical turn of three residues.

With the exception of the C-termini of helices F and H, prolines mark the boundaries of all helical regions. The amino ends of six of the eight helical regions are exactly defined by the occurrence of serine, threonine, aspartic acid, or asparagine in position 1,

followed by proline in position 2. It is interesting that the boundary between the helices A and B is the site of a phase shift of 90° in the regular periodicity of non-polar residues.

8. Discussion

The most striking feature common to all globin chains is the almost complete exclusion of polar residues from interior sites. This is a remarkable vindication of the predictions by Kauzmann (1959) discussed in paper I. By contrast, only a small minority of sites at or near the surface are occupied consistently by any particular type of residue, non-polar, acidic, basic or dipolar. Only 9 out of more than 140 sites are occupied by the same residue throughout the range of globins studied, and this list may well shrink further. Prolines play a prominent part in limiting the length of the helical regions in human haemoglobin, but a less important one in myoglobin. Other so-called anti-helical residues are randomly distributed over helical and non-helical regions. Some of the corners or points of transition from helical to non-helical regions are devoid of any residues specifically designed to terminate helices.

These findings suggest that the pattern of invariantly non-polar side-chains, together with the non-polar parts of the porphyrin ring, may be decisive in determining the configuration of the globin chain. Proteins of similar structure might be recognized by the pattern of consistently non-polar side-chains which their sequences have in common. Mutations which lead to the replacement of a non-polar by a polar residue at an interior site are probably lethal. It is significant that no such replacements have been observed in any of the abnormal human haemoglobins.

The only other protein system for which a comparable amount of chemical information has been collected is cytochrome *c* (Margoliash & Smith, 1964). Here, however, the situation is quite different. Among cytochrome *c*'s of the mammalian type, half the residues are invariant. Among the variable sites a set of consistently non-polar ones is found, but a search for a regular periodicity in the incidence of such sites proved fruitless. It looks as though cytochrome *c* is devoid of long helical regions, which is consistent with the low helical content (27 to 39%) observed by optical methods (Urry & Doty, 1965).

We thank Dr G. Braunitzer and Dr R. L. Hill for letting us see the unpublished sequences of various myoglobins and haemoglobins, and for permission to use these data for Tables 2 to 6.

Note added in proof. Since this paper was written, Dr A. V. Guzzo, of the University of Chicago, has sent us a statistical analysis of the distribution of different amino-acid residues in the middle of helical regions, near the ends of helices and in non-helical regions in myoglobin and haemoglobin. His results suggest that the presence of histidine, glutamic acid or aspartic acid is a necessary, though not a sufficient, condition for the formation of corners or non-helical regions. We have had access to the amino-acid sequences of some species which were not available to Guzzo. Examination of these sequences confirms that histidine, glutamic acid or aspartic acid is indeed a constituent of every corner or of its immediate neighbourhood and of every non-helical region.

REFERENCES

Blout, E. R. (1962). In *Polyamino Acids, Polypeptides and Proteins*, ed. by M. A. Stahmann, p. 275. Madison, Wisconsin: University of Wisconsin Press.

Blout, E. R., de Lozé, C., Bloom, S. M. & Fasmann, G. D. (1960). *J. Amer. Chem. Soc.* **82**, 3787.

Bodo, G., Dintzis, H. M., Kendrew, J. C. & Wyckoff, H. W. (1959). *Proc. Roy. Soc.* A, **253**, 70.

Braunitzer, G., Gehring-Müller, R., Hilschmann, N., Hilse, K., Hobom, G., Rudloff, V. & Wittmann-Liebold, B. (1961). *Hoppe-Seyl. Z.* **325**, 283.

Braunitzer, G., Hilse, K., Rudloff, V. & Hilschmann, N. (1964). *Advanc. Protein Chem.* **19**, 1.

Braunitzer, G. & Matsuda, G. (1961). *Hoppe Seyl. Z.* **324**, 91.

Edmundson, A. B. (1965). *Nature*, **205**, 883.

Goldstein, J., Konigsberg, W. & Hill, R. J. (1963). *J. Biol. Chem.* **238**, 2016.

Harrison, S. C. & Blout, E. R. (1965). *J. Biol. Chem.* **240**, 299.

Kauzmann, W. (1959). *Advanc. Protein Chem.* **14**, 1.

Kendrew, J. C. (1962). *Brookhaven Symp. Biol.* **15**, 216.

Kendrew, J. C. (1963). *Science*, **139**, 1259.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Nature*, **185**, 422.

Konigsberg, W., Guidotti, G. & Hill, R. J. (1961). *J. Biol. Chem.* **236**, PC 55.

Margoliash, E. & Smith, E. L. (1964). Rutgers Univ. Symposium on *Evolving Genes and Proteins*, ed. by H. Vogel, in the press.

Naughton, M. A. & Dintzis, H. M. (1962). *Proc. Nat. Acad. Sci., Wash.* **48**, 1822.

Perutz, M. F. (1962). In *Proteins and Nucleic Acids*, p. 51. Amsterdam: Elsevier.

Perutz, M. F. (1963). *Science*, **140**, 863.

Perutz, M. F. (1965). *J. Mol. Biol.* **13**, 646.

Schroeder, W. A., Shelton, J. R., Shelton, J. B. & Cormick, J. (1963). *Biochemistry*, **2**, 1353.

Smith, D. B. (1964). *Canad. J. Biochem.* **42**, 755.

Tanford, C. (1962). *J. Amer. Chem. Soc.* **84**, 4240.

Urry, D. W. & Doty, P. (1965). *J. Amer. Chem. Soc.* **87**, 2756.

Watson, H. C. & Kendrew, J. C. (1961). *Nature*, **190**, 663.

Wittenberg, B. A., Briehl, R. W. & Wittenberg, J. B. (1965). *Biochem. J.* **96**, 363.