J. Mol. Biol. (1968) **33**, 491–497

Solvent Content of Protein Crystals

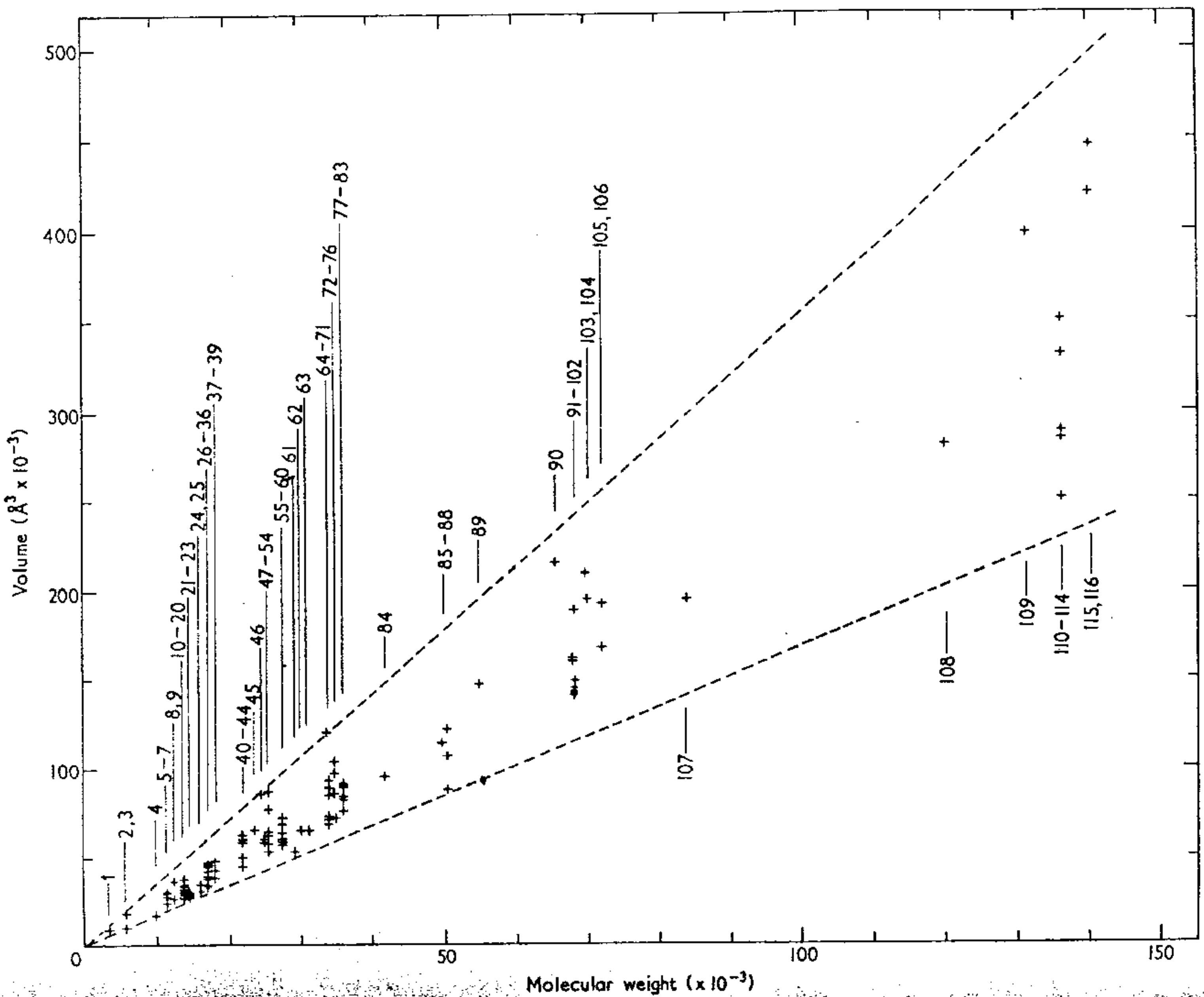
B. W. MATTHEWS

Laboratory of Molecular Biology
National Institute of Arthritis and Metabolic Diseases
National Institutes of Health, Bethesda, Md. 20014, U.S.A.

(*Received 1 January 1968*)

An analysis has been made, from the data which are currently available, of the solvent content of 116 different crystal forms of globular proteins. The fraction of the crystal volume occupied by solvent is most commonly near 43%, but has been observed to have values from about 27 to 65%. In many cases this range will be sufficiently restrictive to enable the probable number of molecules in the crystallographic asymmetric unit to be determined directly from the molecular weight of the protein and the space group and unit cell dimensions of the crystal.

For protein crystals the determination of n, the number of molecules in the crystallographic asymmetric unit, is complicated by the presence in the crystals of a variable amount of solvent. It is therefore common to supplement the measurement of the unit cell dimensions of the "wet" crystals with measurements of the cell parameters of the "air-dried" crystals, or with various density measurements in order to obtain an unequivocal value for n. However, in some cases cell dimensions cannot be obtained from air-dried crystals, and if only small crystals are available an accurate determination of the density of the wet crystals is difficult. On the other hand it is well known that for most protein crystals the number of molecules per asymmetric unit is one, although crystals are also found with $n = 1/4$, $1/2$, 2, . . . etc. (e.g. see Crick & Kendrew, 1957). In any event, to decide between possible alternatives, it is usually necessary only to know the solvent content of the crystals between rather coarse limits. Crick & Kendrew (1957) point out that for most protein crystals the solvent content is between 40 and 60%, but there are exceptions to this general rule. An analysis has been made of all the data currently available in order to find to what extent the over-all range of values observed is sufficiently restrictive to be of use in determining n in further studies. The relevant data from 116 distinct crystal forms which have been reported for a variety of globular proteins are summarized in Figure 1 by plotting the volume of the asymmetric unit, as obtained directly from the X-ray diffraction measurements, against the molecular weight of protein contained in the asymmetric unit. It is convenient to define V_M as the ratio of these two quantities, i.e. V_M is the crystal volume per unit of protein molecular weight. It will be shown that V_M bears a simple relationship to the fractional volume of solvent in the crystal. From Figure 1 it is seen that the range of values assumed by V_M is essentially independent of the volume of the asymmetric unit. The over-all distribution of values observed for V_M is shown in Figure 2, with lowest extremes being 1·68 Å^3/dalton for high potential iron protein (Kraut, Strahs & Freer, *A.C.A. Abstr.* Meeting, August 1967) and 1·72 Å^3/dalton for excelsin (Drenth & Wiebenga, 1955), and highest 3·53 Å^3/dalton and



Volume (Å^3 x 10^{-3})
Molecular weight (x 10^{-3})
0
50
100
150
200
300
400
500
1
2,3
4
5-7
8,9
10-20
21-23
24,25
26-36
37-39
40-44
45
46
47-54
55-60
61
62
63
64-71
72-76
77-83
84
85-88
89
90
91-102
103,104
105,106
107
108
109
110-114
115,116

Fig. 1

FIG. 1. Plot to illustrate the relative variation of the volume of protein crystals. The dashed lines indicate the upper and lower limits observed for V_M. Data are included for the following crystalline proteins which are listed in order corresponding to the numbering in the Figure. Except where stated otherwise, the references may be obtained from the summary of King (1963).

1. Glucagon
2. Ferredoxin (Sieker & Jensen, 1965)
3. Insulin, cubic (Harding, Crowfoot-Hodgkin, Kennedy, O'Connor & Weitzmann, 1966)
4. High potential iron protein (Kraut, Strahs & Freer, *A.C.A. Abstr.* Meeting, Aug. 1967)
5. Insulin, rhombohedral (Harding *et al.*, 1966)
6. Insulin B
7. Insulin A
8. Horse ferricytochrome C
9. Ferricytochrome C (Dickerson, Kopka, Weinzierl, Eisenberg & Margoliash, 1967)
10. Ribonuclease IX
11. Ribonuclease S, Y (Wyckoff *et al.*, 1967)
12. Ribonuclease VII
13. Ribonuclease II (Kartha, Bello & Harker, 1967)
14. Ribonuclease VIII
15. Ribonuclease V
16. Ribonuclease XI
17. Ribonuclease I
18. Ribonuclease XIV
19. Ribonuclease VI
20. Ribonuclease S, W (Wyckoff *et al.*, 1967)
21. Lysozyme, triclinic
22. Lysozyme, orthorhombic
23. Lysozyme, tetragonal (Blake *et al.*, 1965)
24. Erythrocruorin, trigonal (Huber, Formanek, Braun, Braunitzer & Hoppe, 1964)
25. Erythrocruorin, hexagonal (Huber *et al.*, 1964)
26. Myoglobin J
27. Myoglobin A (Kendrew, Dickerson, Strandberg, Hart & Davies, 1960)
28. Myoglobin G
29. Myoglobin B
30. Myoglobin D
31. Myoglobin I
32. Myoglobin F
33. Myoglobin K
34. Myoglobin C2
35. Myoglobin C1
36. Myoglobin H
37. β-Lactoglobulin Y
38. β-Lactoglobulin A
39. β-Lactoglobulin Z
40. Papain S
41. Papain D
42. Papain A
43. Papain C
44. Papain B
45. Apo ferritin B
46. Chymotrypsinogen B, B (Matthews, 1968)
47. Chymotrypsinogen A, D
48. CS_2-chymotrypsinogen A (Kraut, Sieker, High & Freer, 1962)
49. DIP-chymotrypsin, tetragonal (Corey, Battfay, Brueckner & Mark, 1965)
50. γ-Chymotrypsin (Sigler *et al.*, 1964)
51. Chymotrypsinogen A, G (Matthews, 1968)
52. Chymotrypsinogen A, F (Kraut *et al.*, 1962)
53. Chymotrypsinogen A, E (Kraut *et al.*, 1962)
54. Chymotrypsinogen A, B
55. Ribonuclease S, Z (Wyckoff *et al.*, 1967)
56. Ribonuclease X
57. Ribonuclease XIII
58. Ribonuclease IV
59. Ribonuclease III
60. Ribonuclease XII
61. Lysozyme, monoclinic
62. Carbonic anhydrase (Fridborg *et al.*, 1967)
63. Ferritin B
64. Haemoglobin 10
65. Pig haemoglobin 1
66. Carboxypeptidase (Ludwig *et al.*, 1967)
67. Reduced horse haemoglobin (Perutz, Bolton, Diamond, Muirhead & Watson, 1964)
68. Haemoglobin 6
69. Horse oxyhaemoglobin
70. Myoglobin C3
71. Ox haemoglobin, cubic
72. Insulin, monoclinic (Harding *et al.*, 1966)
73. Pepsin, monoclinic (Borisov, Melik-Adamjan, Suckever & Andreeva, 1966)
74. Pepsin, hexagonal
75. Dogfish lactic dehydrogenase 1 (Rossmann, Jeffery, Main & Warren, 1967)
76. Dogfish lactic dehydrogenase 2 (Rossmann *et al.*, 1967)
77. β-Lactoglobulin U (Aschaffenburg, Green, & Simmons, 1965)
78. β-Lactoglobulin R
79. β-Lactoglobulin X
80. β-Lactoglobulin T
81. β-Lactoglobulin S (Aschaffenburg *et al.*, 1965)
82. β-Lactoglobulin W (Aschaffenburg *et al.*, 1965)
83. β-Lactoglobulin P
84. Alcohol dehydrogenase, orthorhombic (Brändén, 1965)
85. Excelsin
86. α-Chymotrypsin (Matthews, Sigler, Henderson & Blow, 1967)
87. Chymotrypsinogen B, C (Matthews, 1968)
88. Chymotrypsinogen A, C
89. Fc γ-G-immunoglobulin (Poljak, Dintzis & Goldstein, 1967)
90. Human mercaptalbumin Hg dimer
91. Haemoglobin 8
92. Haemoglobin H (Perutz & Mazzarella, 1963)
93. Reduced human haemoglobin (Muirhead, Cox, Mazzarella & Perutz, 1967)
94. Ox haemoglobin D (Dunnill, Green & Simmons, 1966)
95. Ox haemoglobin A (Dunnill *et al.*, 1966)
96. Haemoglobin 7
97. Pig haemoglobin II
98. Rabbit haemoglobin 1
99. Ox haemoglobin C (Dunnill *et al.*, 1966)
100. Haemoglobin 2
101. Haemoglobin 4
102. Haemoglobin 5
103. Horse serum albumin
104. Glyceraldehyde phosphate dehydrogenase, PCMB (Watson & Banaszak, 1964)
105. β-Lactoglobulin N (Aschaffenburg *et al.*, 1965)
106. β-Lactoglobulin Q (Aschaffenburg *et al.*, 1965)
107. Alcohol dehydrogenase, monoclinic (Brändén, 1965)
108. Hemerythrin
109. Human serum decanol albumin
110. Lamprey oxyhaemoglobin (Greer, Perutz & Rumen, 1966)
111. Reduced lamprey haemoglobin (Greer *et al.*, 1966)
112. Sickle-cell haemoglobin
113. Haemoglobin 3
114. Haemoglobin 11
115. Lobster glyceraldehyde phosphate dehydrogenase (Watson & Banaszak, 1964)
116. Pig lactic dehydrogenase (Rossmann *et al.*, 1967)

3·43 Å^3/dalton, respectively, for cubic ox haemoglobin (North, 1959) and tetragonal chymotrypsinogen B, type B (Matthews, 1968). It will be noted that although relatively few examples are found toward the limits of the range, the frequency distribution within the range is not a symmetric one. The most commonly observed values of V_M are near 2·15 Å^3/dalton, whereas the median is at $V_M = 2{\cdot}61$ Å^3/dalton. The rather sharp cut-off at the lower end of the range can presumably be considered as corresponding to the closest packing possible for the roughly spherical protein molecules, while the tail of the distribution toward higher V_M values must correspond to progressively looser packing of the molecules, but still with sufficient intermolecular contacts to stabilize the crystal lattice.

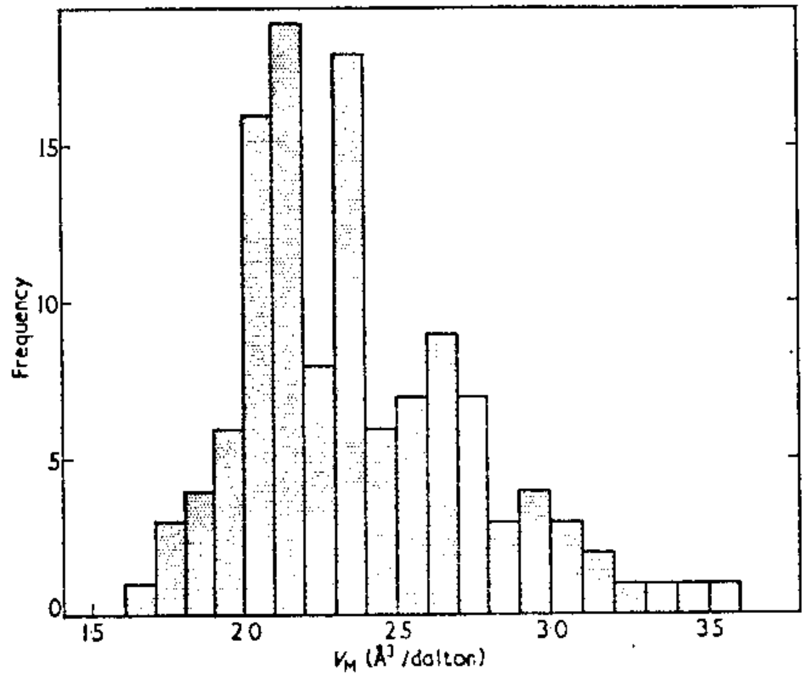


FIG. 2. Frequency distribution of values observed for V_M.

In addition to the data from Fig. 1, the following observations are included: ferritin A, V_M = 2·08 Å^3/dalton; chicken lactic dehydrogenase, V_M = 2·84 Å^3/dalton (Rossmann et al., 1967); tobacco seed globulin, V_M = 1·76 Å^3/dalton; edestin, V_M = 1·75 Å^3.

Although the extreme values of V_M vary by a factor of approximately two, the range is still narrow enough to be useful in preliminary studies of most protein crystals. From the knowledge of the unit cell of the crystal, and of the molecular weight of the protein, the values of V_M corresponding to potential values of n may be obtained. In the majority of cases only one V_M value will lie within the acceptable range, so that n is obtained directly. In the event that there are two or more alternatives for n, it will be necessary to resolve this by some other method. While it must also be anticipated that as more protein crystals are studied, examples will be found with V_M lying outside the range quoted above, there is no reason to believe that the limits quoted will need substantial modification, at least in the molecular weight region below about 70,000 for which a reasonably large number of observations have been made. Nevertheless, borderline cases should be regarded with caution.

To express the range of values of V_M in terms of the percentage of solvent contained in the crystals, it may readily be shown that $V_{prot.}$, the fraction of the crystal volume occupied by protein, is given by

$$V_{prot.} = \frac{1{\cdot}66\,\bar{v}}{V_M}$$

where $\bar{v}$ is the partial specific volume of the protein in the crystal, and V_M (as defined above) is the crystal volume in Å^3 per unit of protein molecular weight. For most proteins $\bar{v}$ is approximately 0·74 cc/g, so that unless there is reason to believe that the protein in question has an abnormally low or high partial specific volume, or that the partial specific volume has a different value in the crystal from that in dilute solution, we have the simple approximation that

$$V_{prot.} = \frac{1{\cdot}23}{V_M}.$$

By difference, the fractional volume occupied by the solvent is given by

$$V_{solv.} = 1 - \frac{1{\cdot}66\,\bar{v}}{V_M} \simeq 1 - \frac{1{\cdot}23}{V_M}.$$

On the basis of this approximation, the protein crystals already examined have a solvent content ranging from about 27 to 65%, with values near 43% occurring most frequently.

Plots similar to Figures 1 and 2 were also made using cell dimensions from air-dried crystals. In this case, as has been noted before (e.g. see Crick, 1957) there is a smaller range of values of V_M—from 1·26 Å^3/dalton and 1·28 Å^3/dalton for edestin and excelsin (Drenth & Wiebenga, 1955), to 1·77 Å^3/dalton and 2·03 Å^3/dalton for orthorhombic insulin R (Einstein & Low, 1962) and hexagonal pepsin (Perutz, 1949). It should be borne in mind that in this case the range of values is based on only 34 sets of cell dimensions, some of which are unavoidably of low accuracy, so that the limits of the range should be regarded with caution.

It must be emphasized that the results quoted here are intended only as a guide in preliminary investigations of protein crystals, and are not intended to be a substitute for measurements of crystal density or solvent content where these can be made. In particular, from the few data which are available, there appears to be a tendency for molecules of higher molecular weight to form crystals containing a relatively higher fractional volume of solvent. Further data will be needed in order to determine the range of values for V_M which might be expected for such proteins. Very recently Rossmann & Labaw (1967) and Longley (1967) have studied by electron microscopy and X-ray diffraction a trigonal modification of crystalline catalase (mol. wt about 250,000). They conclude that there are six catalase molecules in a trigonal cell of dimensions $a = 178 \pm 3$, $c = 241 \pm 4$ Å. This very open arrangement corresponds to $V_M = 4{\cdot}42$ Å^3/dalton, i.e. about 72% solvent, a value higher than has been observed for any smaller crystalline protein. Two other recent studies of different modifications of crystalline catalase (Glauser & Rossmann, 1966; Burgner & McGandy, *A.C.A. Abstr.* Meeting, January 1967) lead to V_M values of 2·52 and 2·86 Å^3/dalton, both of which are above average, although within normal limits for smaller proteins. It must be emphasized strongly, however, that the results quoted here cannot be expected to apply in all other cases. For example, the muscle protein tropomyosin can be crystallized in a three-dimensional lattice, but these crystals have the remarkably high content of about 95% solvent (C. Cohen & D. L. D. Caspar, personal communication).

Relatively few X-ray diffraction studies of crystals of spherical viruses have been reported, and these are not included in the present summary. Although a complete survey was not made of the recent literature, the following examples, which include all X-ray studies listed by King (1963), are probably fairly representative: Southern bean mosaic virus, mol. wt $= 6\cdot63 \times 10^6$, $V_M = 2\cdot68$ Å^3/dalton (Magdoff, 1960). Polio virus type 1, mol. wt $= 6\cdot7 \times 10^6$, $V_M = 3\cdot2$ Å^3/dalton (Finch & Klug, 1959). Tomato bushy stunt virus, mol. wt $= 9 \times 10^6$, $V_M = 3\cdot2$ Å^3/dalton (Caspar, 1956). Broad bean mottle virus, mol. wt $= 5 \times 10^6$, $V_M = 3\cdot4$ Å^3/dalton (Finch, Leberman & Berger, 1967). These have a mean value of $V_M = 3\cdot1$ Å^3/dalton, and all are considerably above the average for small protein molecules, but not outside the upper limits which have been observed. In addition, several reports have been made of the crystal structure of turnip yellow mosaic virus (Klug, Longley & Leberman, 1966, and references quoted therein), but in this case the number of virus particles contained in the unit cell does not seem to be well defined. Klug *et al.* (1966) propose that there are usually eight virus, particles per unit cell arranged in a diamond-type lattice, but that in some preparations this number may be increased up to a maximum of sixteen particles per cell to give a double diamond lattice. Assuming a molecular weight of $5\cdot5 \times 10^6$ for the virus, the corresponding values of V_M would range from 7·6 to 3·8 Å^3/dalton. In this case it is also possible to obtain isomorphous crystals of the "top component", i.e. of the protein shell of the virus with the core of nucleic acid removed. Clearly if one assumed a molecular weight for the top component particles equal to that of the protein shell, i.e. about 63% of the molecular weight of the intact virus, even higher values for V_M would be obtained.

The majority of the crystals described were obtained by salting out with concentrated solutions of ammonium sulfate or of phosphate. For those crystals grown from solutions containing alcohol the mean value of V_M is 2·3 Å^3/dalton, which is not significantly different from the over-all mean value of 2·37 Å^3/dalton. In other words, there is no indication that the different crystallization systems which have been used have any systematic effect on the solvent content of the crystals. Also there does not appear to be any correlation between the degree of symmetry of the crystals and the amount of solvent contained in them; neither is there any obvious relation between the solvent content and the "polarity ratio" of the protein (Fisher, 1964).

REFERENCES

Aschaffenburg, R., Green, D. W. & Simmons, R. M. (1965). *J. Mol. Biol.* **13**, 194.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature*, **206**, 757.

Borisov, V. V., Melik-Adamjan, V. R., Suckever, N. E. & Andreeva, N. S. (1966). *Acta Cryst.* **21**, A157.

Brändén, C. I. (1965). *Arch. Biochem. Biophys.* **112**, 215.

Caspar, D. L. D. (1956). *Nature*, **177**, 475.

Corey, R. B., Battfay, O., Brueckner, D. A. & Mark, F. G. (1965). *Biochim. biophys. Acta*, **94**, 535.

Crick, F. H. C. (1957). In *Methods in Enzymology*, ed. by S. P. Colowick & N. O. Kaplan, vol. 4. New York: Academic Press.

Crick, F. H. C. & Kendrew, J. C. (1957). *Advanc. Protein Chem.* **12**, 134.

Dickerson, R. E., Kopka, M. L., Weinzierl, J. V., Eisenberg, D. & Margoliash, E. (1967). *J. Biol. Chem.* **242**, 3015.

Drenth, J. & Wiebenga, E. H. (1955). *Rec. Trav. Chim. Pays-Bas.* **74**, 813.

Dunnill, P., Green, D. W. & Simmons, R. M. (1966). *J. Mol. Biol.* **22**, 135.

Einstein, J. R. & Low, B. W. (1962). *Acta Cryst.* **15**, 32.
Finch, J. T. & Klug, A. (1959). *Nature*, **183**, 1709.
Finch, J. T., Leberman, R. & Berger, J. E. (1967). *J. Mol. Biol.* **27**, 17.
Fisher, H. F. (1964). *Proc. Nat. Acad. Sci., Wash.* **51**, 1285.
Fridborg, K., Kannan, K. K., Liljas, A., Lundin, J., Strandberg, B., Strandberg, R., Tilander, B. & Wirén, G. (1967). *J. Mol. Biol.* **25**, 505.
Glauser, G. & Rossmann, M. G. (1966). *Acta Cryst.* **21**, 175.
Greer, J., Perutz, M. F. & Rumen, N. (1966). *J. Mol. Biol.* **18**, 547.
Harding, M. M., Crowfoot-Hodgkin, D., Kennedy, A. F., O'Connor, A. & Weitzmann, P. D. J. (1966). *J. Mol. Biol.* **16**, 212.
Huber, R., Formanek, H., Braun, V., Braunitzer, G. & Hoppe, W. (1964). *Ber. Bunsenges. Phys. Chemie*, **68**, 818.
Kartha, G., Bello, J. & Harker, D. (1967). *Nature*, **213**, 862.
Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G. & Davies, D. R. (1960). *Nature*, **185**, 422.
King, M. V. (1963). In *Crystal Data*, ed. by J. D. H. Donnay & G. Donnay, *A.C.A. Monograph No. 5*, p. 1263. American Crystallographic Association.
Klug, A., Longley, W. & Leberman, R. (1966). *J. Mol. Biol.* **15**, 315.
Kraut, J., Sieker, L. C., High, D. F. & Freer, S. T. (1962). *Proc. Nat. Acad. Sci., Wash.* **48**, 1417.
Longley, W. (1967). *J. Mol. Biol.* **30**, 323.
Ludwig, M. L., Hartsuck, J. A., Steitz, T. A., Muirhead, H., Coppola, J. C., Reeke, G. N. & Lipscomb, W. N. (1967). *Proc. Nat. Acad. Sci., Wash.* **57**, 511.
Magdoff, B. S. (1960). *Nature*, **185**, 673.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 499.
Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Nature*, **214**, 652.
Muirhead, H., Cox, J. M., Mazzarella, L. & Perutz, M. F. (1967). *J. Mol. Biol.* **28**, 117.
North, A. C. T. (1959). *Acta Cryst.* **12**, 512.
Perutz, M. F. (1949). *Research*, **2**, 52.
Perutz, M. F., Bolton, W., Diamond, R., Muirhead, H. & Watson, H. C. (1964). *Nature*, **203**, 687.
Perutz, M. F. & Mazzarella, L. (1963). *Nature*, **199**, 639.
Poljak, R. J., Dintzis, H. M. & Goldstein, D. J. (1967). *J. Mol. Biol.* **24**, 351.
Rossmann, M. G., Jeffery, B. A., Main, P. & Warren, S. (1967). *Proc. Nat. Acad. Sci., Wash.* **57**, 515.
Rossmann, M. G. & Labaw, L. W. (1967). *J. Mol. Biol.* **29**, 315.
Sieker, L. C. & Jensen, L. H. (1965). *Biochem. Biophys. Res. Comm.* **20**, 33.
Sigler, P. B., Skinner, H. C. W., Coulter, C. L., Kallos, J., Braxton, H. & Davies, D. R. (1964). *Proc. Nat. Acad. Sci., Wash.* **51**, 1146.
Watson, H. C. & Banaszak, L. J. (1964). *Nature*, **204**, 918.
Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Tsernoglou, D., Johnson, L. N. & Richards, F. M. (1967). *J. Biol. Chem.* **242**, 3749.