GIANT MOLECULES*

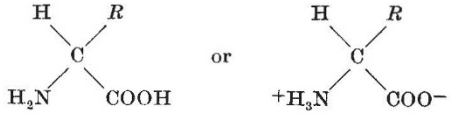
By Sir LAWRENCE BRAGG, O.B.E., F.R.S.

I PROPOSE to deal here particularly with the investigation of the structure of crystalline proteins by X-ray analysis. It is a formidable task. X-ray analysis is being extended to ever more complex molecules; but in the proteins we are confronted with a degree of complexity several orders higher than that of the most complex organic molecules which have hitherto been determined with success. The following table of molecular weights indicates the nature of the problem:

Naphthalene	128	Hæmoglobin	66,700
Penicillin	320	Tobacco seed globulin	300,000
Insulin	12,000	Hæmocyanin (whelk)	6,800,000
Myoglobin	17,000	Tomato bushy stunt virus	10,000,000
Pepsin	37,000		

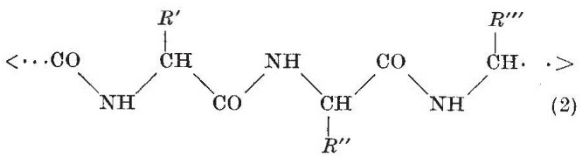
Naphthalene was the first organic molecule to be analysed, by W. H. Bragg some twenty-five years ago. Penicillin is one of the latest triumphs of X-ray analysis, described in a recent publication by Crowfoot, Bunn, Rogers-Low and Turner-Jones. Insulin, which is one of the simplest of proteins, has a molecular weight nearly forty times as great, and from it we pass to the vast molecules of such bodies as the viruses. All these substances yield highly perfect crystals, and give X-ray diffraction photographs, as was first shown by Bernal, which show a regularity of structure down to atomic dimensions. It is the task of X-ray analysis to try to interpret these photographs. One can measure the intensities of many thousands of diffracted beams from each crystal; it is like a story written in a cipher for which we are seeking the key.

The proteins are built, as was first shown by Emil Fischer in 1906, of long chains or rings of amino-acid residues. A typical amino-acid has a formula



(1)

when R represents a univalent side-chain of a kind which characterizes that particular amino-acid. The acid group of one amino-acid can be linked to the basic group of another, with the loss of water, so as to produce a chain of residues in which the different side-chains R are like differently coloured beads strung on a necklace, the 'peptide chain':



(2)

There are some twenty-three types of known amino-acid, varying in complexity from glycine, in which R is a hydrogen atom, to phenylalanine, which contains a benzene ring, or tryptophane, which has linked five- and six-membered rings. Most contain only carbon, oxygen, nitrogen and hydrogen; a few have other elements such as sulphur. In most of them the side-chain R is neutral, as in alanine, $—CH_3$, valine, $—CH(CH_3)_2$, phenylalanine, $—CH_2.C_6H_5$. It may be basic as in lysine, $—CH_2.CH_2.CH_2.CH_2(NH_2)$, or acidic as in aspartic acid, $—CH_2.COOH$. Cystine is a double-ended amino-acid, $—CH_2—S—S—CH_2—$, which can form a bridging link between one chain and another. The analysis of the amino-acid residues in a given protein is a lengthy and difficult procedure, but several have been analysed with a fair degree of completeness. Insulin (molecular weight 12,000), for example, consists, according to Chibnall and Sanger, of 106 residues, in four chains, bound together by six disulphide bridges of cystine. Myoglobin (17,000) has about 140 residues, and hæmoglobin (67,000) has about 540. If we compare the twenty-three types of residue to the letters of the alphabet, the string of residues in myoglobin is like a short statement in some twenty or thirty words, and the nature of hæmoglobin is described by a brief paragraph of about ten sentences. For some reason as yet not understood, Nature has chosen this simple way of building the structures of the molecules in the protoplasm of animals and plants. The complex and specific functions which the molecules are called upon to perform are achieved, not by building any highly complex organic structure, but by stringing these relatively simple groups in different orders and numbers, just as the same few letters of the alphabet can be used to represent Milton's "Lycidas" or a page of the telephone directory.

Certain features of protein structure deserve special mention.

(*a*) Nearly all known amino-acids occur in nearly all known proteins.

(*b*) The mean molecular weight of the residues is much the same in all proteins, lying between 110 and 120. Since the molecular weight of the atoms in the 'backbone' or 'chain' CO.CH.NH is 56, one half the molecular weight of a protein is in its chains and one half in the side-groups.

(*c*) It is nearly certain that all the naturally occurring amino-acids have the same steric configuration around the central carbon atom shown in (1). Since four different groups are attached tetrahedrally to this carbon atom, the compound can exist in dextro- or lævo-forms which are mirror images of each other, being related as a right hand is to a left hand. All amino-acids are of the lævo type, except, of course, glycine, in which R is also a hydrogen atom and there is no distinction between the right- and left-handed forms. It must have been a mere matter of chance that all forms of life started with this same left-handed thread running through them; a world passed through Alice's Looking Glass would work just as well, but some toss of the coin came up tails and decided the future form of all living matter.

(*d*) The crystalline structures of several simple amino-acids or dipeptides composed of two residues

*Evening Discourse at the Royal Institution, delivered on April 27.

linked together have been completely determined by X-rays (for example, glycine by Corey and Albrecht). They show two interesting features. They obey rather exactly the expected distances between atoms and bond-angles found in other organic compounds where strong distorting forces are absent. Huggins has given the following summary of data for an element

$$\begin{array}{c} H \quad R \quad O \\ -N-C_1-C_2- \\ H \end{array}$$

gleaned from several structures :

$$\begin{array}{c} N \text{ to } C_1,\ 1{\cdot}41 \text{ A.}; \quad C_1 \text{ to } C_2,\ 1{\cdot}52 \text{ A.}; \\ C_2 \text{ to } O,\ 1{\cdot}25 \text{ A.}; \quad C_2 \text{ to } N,\ 1{\cdot}33 \text{ A.} \\ \text{Angle } NC_1C_2,\ 112°; \text{ angle } C_1C_2N,\ 118°; \\ \text{angle } C_2NC_1,\ 118°. \end{array} \qquad (3)$$

Secondly, the structures all contain numerous hydrogen bonds between a nitrogen and oxygen atom, N—H—O. This bond appears to be of decisive importance in deciding the form of the structure, and is 2·65 A. in length. Such data are important as a guide in constructing possible models for an extended polypeptide chain.

The difficulties of applying X-ray analysis to such a complex structure as a protein may well seem insuperable at first sight. It is not possible to proceed directly from observation of diffraction to determination of structure except in very simple cases. The X-ray diffraction picture by itself is not sufficient to determine the structure; it can only be used in conjunction with other information. In the case of molecules of moderate size, this information consists in the knowledge that the unit cell contains a limited number of known atoms, and that these atoms are tied together by bonds the lengths and relative inclinations of which can be very closely anticipated from previous determinations of similar chemical combinations. It is not impossible to try a number of likely configurations, and see if one of them explains the X-ray data. But a molecule of a protein such as hæmoglobin contains 8,000 atoms, and methods of trial and error are inconceivable.

There exists, however, a method of representing the X-ray measurements which gives direct and unequivocal information about the structure, known to the crystallographer as a Patterson synthesis or vector map. The observations of intensity of diffraction are used as coefficients of a Fourier term in a series

$$\sum_h \sum_k \sum_l I_{hkl} \cos 2\pi(hx + ky + lz), \qquad (4)$$

where I_{hkl} is the square of the amplitude for a diffracted spot of order (hkl), and xyz are the co-ordinates of each point in the unit cell of the structure. The result is called a vector map. Suppose there is an atom a at $x_1y_1z_1$ and another atom b at $x_2y_2z_2$, in the crystal pattern. The vector map will then have a peak or lump of density at the point $x_1—x_2$, $y_1—y_2$, $z_1—z_2$, which is arrived at by drawing a 'vector' (that is, a line of given length and direction) from the origin equal and parallel to the vector from a to b. In other words, the vector map cannot tell us where a and b are in the crystal, but it does indicate how they lie with respect to each other. Further, the magnitude of the peak is proportional to the product of the masses at a and b. Fig. 1 illustrates the principle. It would be out of place to go into the derivation of the vector map here, though it follows very simply from the principles of optical interference ; but it is important to understand its nature, because it plays a large part in X-ray analysis. We have gained this definite information, however, by paying a heavy price, because it will easily be seen how very complicated the vector map becomes. If there are n atoms in the structure, n^2 vectors can be drawn between them, and all these are superimposed in the vector map.

A common device for analysing organic molecules of reasonable size helps to surmount this difficulty. A heavy atom such as bromine or iodine is attached to the molecule. The corresponding vectors between the heavy atoms then stand out prominently in the vector map, and can generally be recognized, as they are few in number and, being proportional to the product of the masses, they are very large. It is not hard to deduce where the iodine or bromine atoms must be in the crystal in order to give such vectors, and once they are pinned down it is considerably easier to discover where the lighter atoms are. We have, as it were, stained certain characteristic points of the molecule, much as a microscopist stains a nucleus in a cell. But the molecule of hæmoglobin contains more than 8,000 atoms, and its vector map has some seventy million superimposed peaks. No attached heavy atom could stand out in such a crowd.

There is, however, a ray of hope, provided by the polypeptide chains. Suppose the chains are arranged in a regular way, for example, like a series of parallel rods in the molecule. Many of the vectors will run from one atom to another in the same chain, and when these are drawn in the vector map, they will give a long rod of high density through the origin, and parallel to all the rods in the crystal. Similarly, the vectors from atoms in one chain to those in another will form another rod, parallel to the first, and at an appropriate distance from it.

The situation is shown diagrammatically in Fig. 1. Intra-chain vectors have been drawn as dotted lines from two atoms to their neighbours in the chain, and, of course, similar ones can be drawn from every atom in the chain. All these vectors will be crowded into the region A in the vector map. Similar vectors are shown as light lines from one chain to the next, and they all end in the regions B. If we find a dense ridge running through the origin of the vector map in a certain direction, it is reasonable to deduce that

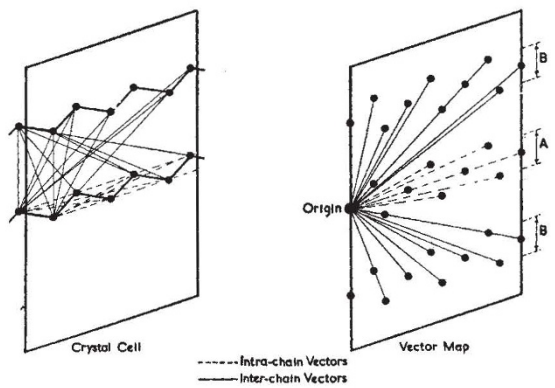


Fig. 1

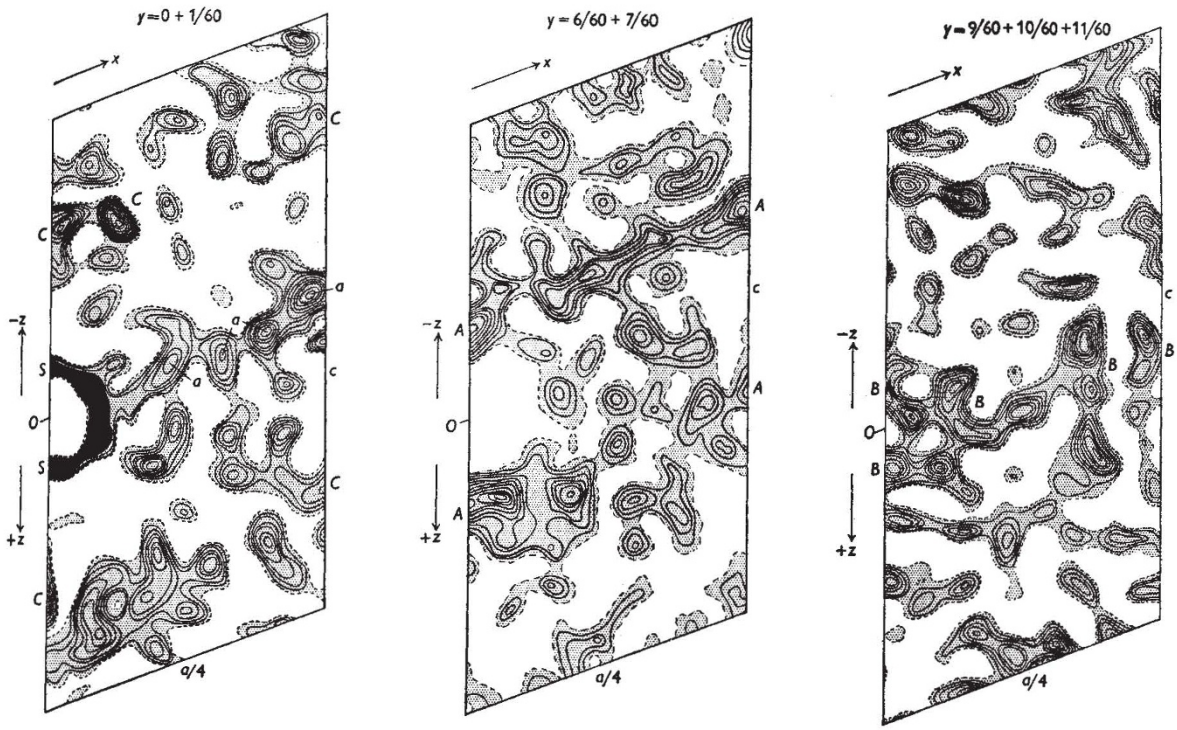


Fig. 2. Scale 1 cm. = 7 A. From *Proc. Roy Soc.*, A, 195, 474 (1949)

somewhere in the unit cell there are chains of atoms running in this direction; and if, in addition, we find parallel rods in other parts of the vector map, they give us a hint as to how the rods are packed together in the crystal.

Certain vector maps obtained by Perutz in studying crystalline hæmoglobin from horse-blood indicated the existence of such parallel rods, and it was therefore decided to make a complete three-dimensional Patterson synthesis of hæmoglobin. The principle is simple. At each point, x, y, z in the unit cell, we form expressions such as (4) for all the diffracted intensities I_{hkl} and add them, and the sum gives us the Patterson density at the point xyz. But consider the magnitude of the task. Perutz had measured some 28,000 diffracted spots given by the crystal, representing the message in cipher which has to be interpreted. In order to get a vector map with sufficient detail, it is necessary to calculate the density at intervals of $a/120$, $b/60$ and $c/60$. The total number of terms such as (4) which are required is therefore $28{,}000 \times 120 \times 60 \times 60$, or $1{\cdot}21 \times 10^{10}$ terms. Of course, the analysis does not take such a formidable form in practice, as various devices can be used which immensely shorten the labour. Even with them, the measurement of the diffracted spots and the computations for this one crystal took some four years, and the final summations were carried out with a Hollerith punched-card machine by the Scientific Computing Service, Ltd. When the project was commenced, it was by no means certain that the results would justify the effort, but fortunately this appears to be the case.

A sample of the results is shown in Fig. 2, which represents cross-sections of the vector map. The crystal cell of hæmoglobin is monoclinic, and in these pictures the a-axis slants from left to right, the b-axis is perpendicular to the plane of the diagram, and the c-axis is vertical. In the left-hand figure we have a cut through the origin O, which is half-way up the left-hand side. The density is represented by contour lines. There is an obvious ridge through the origin running parallel to the a-axis, with peaks on it at intervals of about 5 A., and there are rather less well-defined ridges above and below it. The central figure is a cut about one-tenth the way up the cell in the b direction, and two more ridges are to be seen. Higher up in the diagram (the right-hand figure) there is another ridge above the one through

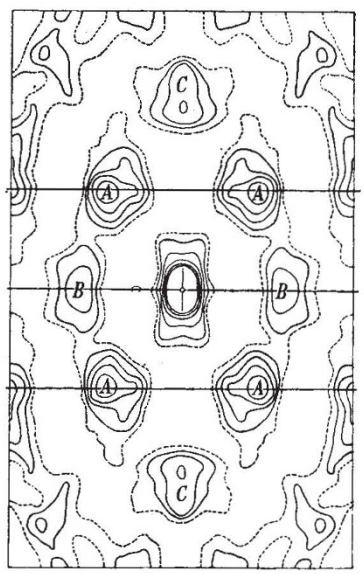


Fig. 3. Scale 1 cm. = 7 A. From *Proc. Roy Soc.*, A, 195, 474 (1949)

the origin. Fig. 3 is a projection of these ridges looking along the a-axis so that they are seen end-on, in which the ridge through the origin is now at the centre of the figure and we can see the other ridges spaced round it.

To summarize, Perutz's results indicate that in hæmoglobin the polypeptide chains run parallel to the a-axis of the crystal, and that they are stacked about 10·5 A. units apart like a series of rods in approximately close-packed array. Further, the series of peaks in each chain about 5 A. apart indicates that there is a repeat distance in the chains of 5 A., producing many vectors of multiples of this distance along the chains. From this we can deduce a further important point. If the chains have this spacing and repeat distance, the density of the crystal tells us what mass is associated with each repeat, and it turns out to be very closely the average mass of three amino-acid residues. Now three residues in a chain stretched to the utmost as in (2) would occupy 10 A., so the polypeptide chain must be folded or coiled up in some way. Again, the N—H—O bond, which is only 2·65 A. in length, cannot stretch from one chain to the next, and we have seen how important these bonds are. They must presumably be between atoms of the same chain, and it is probably these bonds which are holding the chain in its coiled or folded form.

This evidence confirms certain deductions made by Astbury from the very diffuse pictures given by wool and other forms of keratin, which is a 'denatured' fibrous protein. He has deduced that a-keratin is a folded polypeptide chain with repeats of three residues at distances of about 5 A. units.

The question might well be asked: What justification have we for thinking that the vectors in the chains will stand out prominently, when there are also all the vectors between the chains and side-groups, and from atoms in one side-group to atoms in another? Would we not expect the vector map to be a completely confused jumble of peaks? It must be remembered that one half the atoms are in the chains. Further, if these chains really are approximately rod-like and parallel, there should be a very great concentration of atoms around their central axes when one looks along them. As an example, let us take what is perhaps the simplest possible model for a chain of three repeats every five Angstrom units, and suppose it to be a spiral with three R groups attached to each turn; the bond-lengths and angles then confine the spiral to a radius of about 1·5 A. The average side-chain has about four atoms in it, and leucine (for which R is $—CH_2.CH(CH_3)_2$) may be taken as typical. Fig. 4 shows in a very idealized way the appearance of the structure when looking along the chains. For a reason explained below, the chains have been represented as being 9·5 A. apart horizontally and 14 A. apart vertically. The circles enclose the 'chain' atoms, twelve to each repeat, and three leucine groups are attached to each turn with four atoms per group (no attempt has been made to represent the actual positions of these atoms; the diagram is merely intended to show their concentration). The outline of the leucine group is drawn with a van der Waals radius suitable to the packing of groups not attached to each other by bonds. The density in the circles, where the atoms are all linked by covalent bonds, is about ten times as great as that in the spaces between them. It is reasonable to expect that these highly concentrated masses will stand out in a Patterson projection, both as regards the short vectors inside each chain and the vectors from one chain to the next.

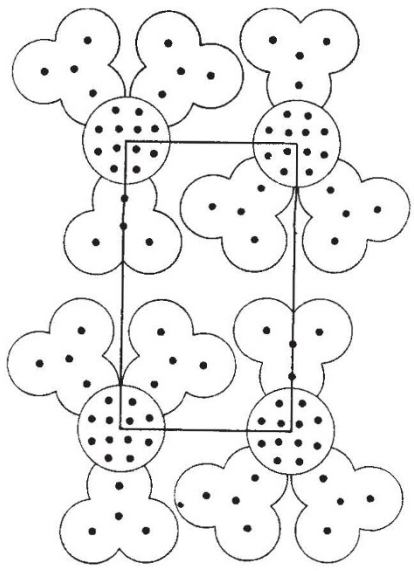


Fig. 4

The spacings of the chains in Fig. 4 are based on a recent examination of the structure of myoglobin by Kendrew, which is being published shortly. The structure is considerably simpler than that of haemoglobin, the molecule being about one quarter the size. Patterson projections on the principal planes show chains very like those in haemoglobin, and with similar repeat distances of 5 A. These chains lie in planes parallel to the b-face, and are in the direction [201] of the crystal. A Patterson projection on a plane perpendicular to the chains has turned out to be extremely simple and very suggestive. It indicates an arrangement of the chains very much like that of Fig. 4, the chains being 9·5 A. apart horizontally, and lying in sheets 14 A. apart in the vertical b direction. It further indicates that the chains are not directly above each other in the b direction as shown in Fig. 4, but slightly staggered. The fact that the side-groups clearly do not fill all the space in Fig. 4 is, of course, accounted for by the necessity of finding space for the water associated with each molecule. Reference must be made to the forthcoming paper for a complete description, but it may be said here that the myoglobin analysis has given us greater confidence in the validity of the above conclusions, and in particular that it gives us a first approach to a picture of the arrangement of the chains in this protein.

All this work on the X-ray analysis of the proteins is still of a highly speculative character. One must be prepared to abandon without hesitation any proposed model, however attractive it may appear, if further evidence indicates that it is not valid. We are like a climber, striving to conquer a most difficult pitch, and grateful for the slenderest cracks and projections which seem to offer a possible hold for hand or foot. The results obtained so far may seem meagre, but a glittering prize draws us on. There must be some very deep and fundamental reason why Nature has chosen the polypeptide chain as the building principle for all forms of living matter, a reason which should become clear if we can solve the mystery of the structure of the protein molecule.