SOVIET PHYSICS USPEKHI VOLUME 9, NUMBER 2 SEPTEMBER-OCTOBER 1966

621.386.8+547.96

Physics of Our Days

X-RAY STRUCTURE ANALYSIS OF GLOBULAR PROTEINS

B. K. VAĬNSHTEĬN

Usp. Fiz. Nauk 88, 527-565 (March, 1966)

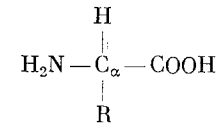
1. PRINCIPLES OF PROTEIN STRUCTURE. CRYSTALLINE PROTEINS

MANY years of effort by prominent researchers have recently culminated in noteworthy advances in solving the protein-structure problem. The deciphering of the structure of certain proteins and the determination of the structure of DNA are the most important factors in the intensive growth of molecular biology. One of the causes of this growth has been the widespread application of physical methods for studying biological macromolecules.

The most important of these methods is x-ray structure analysis. In the period of over a half century of its existence (Laue discovered x-ray diffraction in 1912), it has provided solid state physics, chemistry, mineralogy, and metallurgy with a concrete world of atomic crystal structures, and has now produced just as remarkable a result in biology.

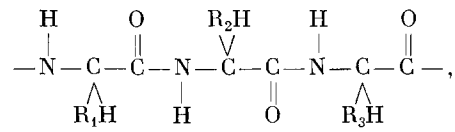
The various globular proteins bring about the innumerable set of metabolic reactions of living matter.[1,2] These reactions, e.g., the splitting of certain bonds, or conversely, the combination of certain molecules with each other, or electron transfer, etc., are highly selective, and proceed with enormous rates. Thus, enzymes accelerate certain reactions by factors of 10^6—10^{11}. Chemical and biological methods of studying enzyme activity of proteins when they interact with a substrate (the substance that a given enzyme "works on") or an inhibitor (a substance that blocks this activity) permit one to get information on the chemical structure of the active center of the protein molecule. Full understanding of the mechanism of a biological activity is impossible without a knowledge of the spatial structure. This is what defines the significance and role of x-ray structure analysis of proteins. Of course, the number of structure determinations of certain proteins done up to now is a trifle in comparison with their entire innumerable variety. For example, the number of proteins in man is estimated to be 10^5; the number of enzymes already isolated and studied by biochemists amounts to about 1000. Nevertheless, we can say that the existing results are of vast interest, especially when combined with biological results.

The proteins are chain molecules of high molecular weight, composed of amino-acid residues. The amino acids



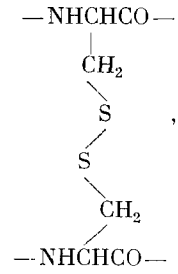
(1)

are distinguished by their radicals R. Removal of one of the H atoms of the amino group and the OH of the carboxyl (with loss of water) makes it possible to link the residues into a polypeptide chain



(2)

which is the basis of protein structure. The great majority of proteins are composed of 22 "principal" amino acids, which are often likened to an "alphabet of the protein language" (Table I). However, more than fifty rarer amino acids are also known.[3] The individuality and properties of a given amino acid or unit of the polypeptide chain are determined by the radical R. The radicals of one group of amino acids, e.g., glycine, phenylalanine, etc., are neutral; they repel water molecules, i.e., are hydrophobic. The radicals of the other residues possess active polar or charged groups OH, COO^-, NH_3^+, which can form hydrogen or ionic bonds. Water molecules easily bind to them. The cystine residue

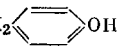
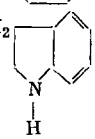
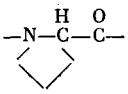
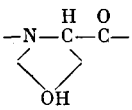


(3)

formed by combination of two cysteines plays the special role of a disulfide bridge linking polypeptide chains with one another.

The C_α atom of the amino acids is asymmetric. Hence, the amino acids are optically active molecules. Proteins are composed of levo (L) amino acids. Thus, the proteins, just like all biological molecules and structures in general, exist in only one of the two possible mirror-equivalent forms.

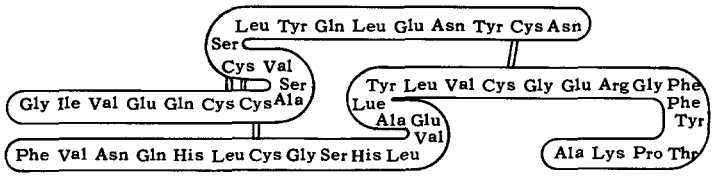
Table I

Amino acid	R	Amino acid	R
Glycine (Gly)	—H	Methionine (Met)	$-CH_2CH_2SCH_3$
Alanine (Ala)	$-CH_3$	Phenylalanine (Phe)	$-CH_2$ 
Valine (Val)	$-CH(CH_3)_2$	Tyrosine (Tyr)	$-CH_2$ 
Leucine (Leu)	$-CH_2CH(CH_3)_2$	Tryptophan (Trp or Try)	$-CH_2$ 
Isoleucine (Ileu or Ile)	$-CHCH_3CH_2CH_3$	Histidine (His)	$-CH_2$ 
Serine (Ser)	$-CH_2OH$	Proline* (Pro)	
Threonine (Thr)	$-CH(OH)CH_3$	Hydoxyproline* (Hypro)	
Aspartic acid (Asp)	$-CH_2COOH$		
Glutamic acid (Glu)	$-CH_2CH_2COOH$		
Asparagine (Asn or AspN or $AspNH_2$)	$-CH_2CONH_2$		
Glutamine (Gln or GluN or $GluNH_2$)	$-CH_2CH_2CONH_2$		
Lysine (Lys)	$-CH_2CH_2CH_2CH_2NH_2$		
Arginine (Arg)	$-CH_2CH_2CH_2NHC(=NH)NH_2$		
Cysteine (CySH)	$-CH_2SH$		
Cystine (CyS_2)	$-CH_2SSCH_2-$		

*Instead of the R group, the amino acid residues within the chain are shown as a whole. For these two so-called imino acids, the radical is attached not only to the αC atom but also to the N atom of the chain to form a five-membered ring with it.

The simplest information on the structure of any protein molecule involves its overall amino-acid composition. Data on this composition are, so to speak, a zero-order approximation to describing the structure of the protein.

Further, one distinguishes between primary, secondary, tertiary, and quaternary structure of a protein. The primary structure is the sequence of amino acids in the chains and the way that the chains are joined (if there are several of them) or that a given chain reconnects with itself by disulfide bridges. The written-out chemical formula describing the primary structure, e.g., the formula of insulin (4), contains the topology of the protein molecule. The letters are now assembled into a sentence. However, this is still far from enough to give a complete description, since the chain, each of its links, and the side groups are arranged in a definite way in three-dimensional space.



(4)

One establishes the primary structure, as F. Sanger et al.[4] first did with insulin (4) in 1955, by splitting the protein enzymatically into short peptides containing 3—5—7 residues. The sequence in these short peptides can now be determined by chemical methods. Since the chain is broken at different places under different methods of splitting, one can unambiguously establish the entire "sentence" by combining the "word fragments" obtained. The primary structures of about twenty proteins are known at present.

Now, the concepts of secondary and tertiary structure begin to involve the geometry. The studies of Pauling's school[5-7] have played a very important role in understanding the general geometric laws governing the conformations of polypeptide chains. Pauling and Corey[5,6] based their work on the determination by x-ray diffraction of the structures of certain amino acids (almost all of them have been studied by now; see Fig. 1) and peptides. Thus they established "standard" bond lengths and angles characteristic of the unit of the polypeptide chain (Fig. 2). Variety of conformations of the chains is brought about by free rotation about the C_α—N and C—C_α single bonds, whereas the amide group always remains planar. Hydrogen bonds are characteristic of proteins (as they are of biological structures in general). These NH...O bonds can arise between different chains or between the links of a given chain. Having an energy of 5—10

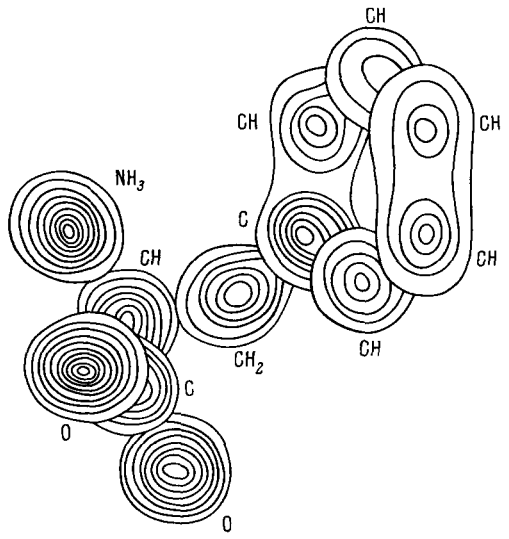


FIG. 1. Fourier map of the phenylalanine molecule.[8] Electron-density contours are drawn every electron/Å^3.

kcal/mole (the energy of covalent bonds is 50–80 kcal/mole), they are strong enough to stabilize some particular conformation of the polypeptide chains, and weak enough to permit conformational transitions under certain conditions.

X-ray diffraction study of the packing of polypeptide chains in fibrous proteins (of which hair, skin, etc., are made) was begun as early as the thirties by Astbury.[9,10] These studies, together with those of model polymers, or synthetic polypeptides containing only one type of R group,[11,12] has given evidence of two fundamental types of conformations of polypeptide chains. In one of them, the β form, the chains (2) are extended in a parallel arrangement, and are joined to one another by hydrogen bonds. In 1953, Pauling and Corey established the structure of the other, or α form (Fig. 3) from data on the stereochemistry of amino acids and the laws governing hydrogen bonds. Notions of the helical structure of polymers in general, and polypeptide chains in particular, had been running through the minds of researchers for some time, but they were not able to reconcile them with the laws of crystal symmetry. According to the latter, crystals can contain only screw axes of symmetry of integral

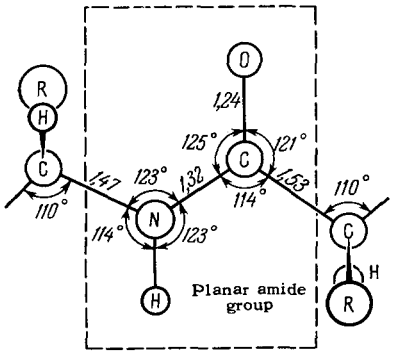


FIG. 2. Structure of the unit of the polypeptide chain.[5]

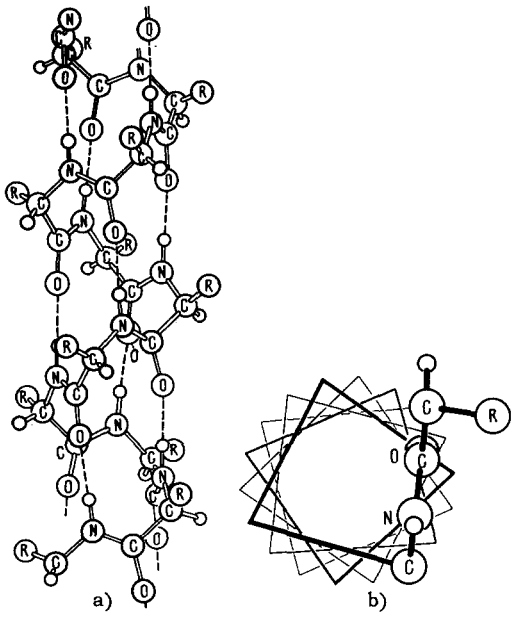


FIG. 3. α–Helical conformation of the polypeptide chain (right-hand α helix) and its projection.

order, namely two-, three-, four-, or sixfold. Pauling's revolutionary step was to reject integral-order symmetry so as to satisfy conformational requirements. Thus, the α helix (Figs. 3, 4) has 18 residues in five turns (18/5 helix). Its period is 27 Å, the pitch of the helix is 5.2 Å, and the rise per residue is 1.5 Å. The α structures studied thus far are all right-handed. The characteristics of the symmetry and structure of biological chain molecules have now been elucidated, and a theory of the diffraction patterns of these structures has been created.[13-15]

The way in which the polypeptide chain is wound up in a definite conformation stabilized by hydrogen bonds is called the secondary structure. Continuing our analogy with the alphabet, we can liken the secondary structure to the manner of writing: in parallel or antiparal-

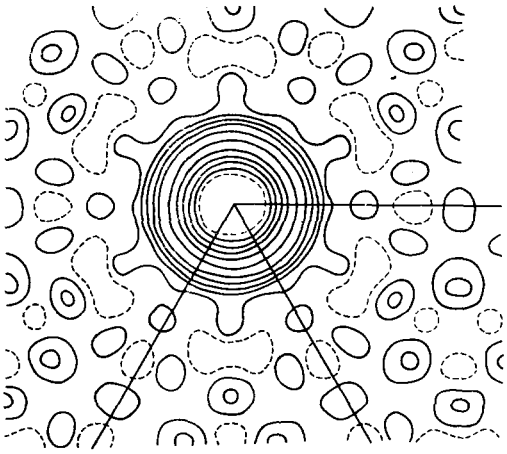


FIG. 4. Potential Fourier projection of an α helix (the synthetic polypeptide poly-γ-methyl-*L*-glutamate, electron-diffraction data[11]).

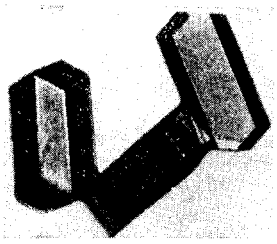


FIG. 5. Sperm-whale myoglobin crystals (type A(× 15)).

lel lines, or in a line passing helically about a cylinder.

In the globular proteins, the chain is coiled in a complex fashion and compactly folded into a globule, and has different conformations in different regions. This is no longer a question of uniform lines, but a three-dimensional crossword (chainword) puzzle. The three-dimensional spatial structure of a protein molecule with specification of the chain sequence and the positions of all the radicals and atoms is the tertiary structure of the protein.

We must note that a number of globular proteins contain so-called prosthetic groups of non-peptide nature, often containing metal atoms.

A so-called quaternary structure is also distinguished for some proteins (and other biomolecules). They are composed of several subunits or protomers which are either all identical (of one type), or different (of two or more types). The manner in which the protomers are packed in the molecule is called the quaternary structure.

Certain other methods besides x-ray diffraction analysis also furnish valuable information on the structure of globular protein molecules. These are sedimentation (molecular weight), small-angle x-ray scattering (shape and molecular weight), optical measurements, and in particular optical rotatory dispersion (percent helix in the chains), electron microscopy, etc.

If a globular protein has been well purified, then as a rule one can crystallize it (Fig. 5).[16] Depending on the solvent, the pH, etc., certain proteins can form several polymorphic modifications. The crystals contain mother liquor in the interstices between the molecules ("wet crystals"), and are stable in equilibrium with this liquid or its vapor. It has been established by x-ray diffraction that part of the water molecules are firmly bound to the surface of the protein molecule, forming a sort of "envelope" (Fig. 6), while the other,

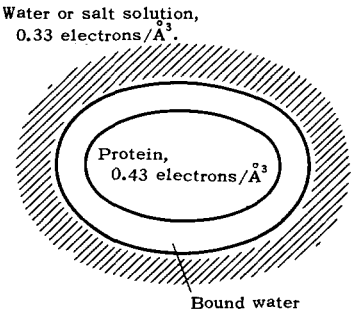


FIG. 6. Hydration of the hemoglobin molecule.[40]

free fraction of the mother liquor is disordered.[17] One can dry the crystals, and then a considerable part of the free water of crystallization escapes, the volume of the unit cell decreases,[18] and the crystals become disordered.

The first x-ray diffraction patterns of protein single crystals (pepsin and insulin) were obtained in 1934 by J. D. Bernal and his now-famous student D. Crowfoot-Hodgkin (Nobel laureate in 1964, the prize being given for her x-ray analysis of vitamin B_{12} [19,20]). The very fact of the formation of protein crystals showing thousands of reflections in their x-ray diffraction patterns (Fig. 7) indicates that all the giant molecules of a given protein are identical and that they have a fixed internal structure.

Sometimes doubt is expressed as to whether protein molecules retain the same structure in the crystal that they have in solution, where they are biologically active. Without discussing all the proofs (the simplest of them is the fact that in the wet crystal these molecules are practically surrounded by solvent), we shall state that these doubts are ungrounded.

The quantitative difficulties of the x-ray analysis of proteins (and we should immediately stipulate that they are an insignificant fraction of the entire set of difficulties in the solution of this problem) can be illustrated by the rough figures given in Table II. The information contained in an x-ray diffraction experiment is the set of intensities of the beams diffracted by the crystals. The number of them is proportional to the volume of the unit cell, but becomes less when the structure is imperfect. With "ordinary" crystals, one can perform a precise structure analysis, and determine the interatomic distances to an accuracy of

Table II

	Unit-cell periods, Å	Unit-cell volume, Å³	Number of atoms per cell	Number of reflections	d_{min}, Å
Ordinary crystals	5—15	10^3	up to 10^2	1000	0.5—1 Å
Complex organic compounds — vitamins, hormones, etc.	10—25	10^4	up to 10^3	up to 3000	1.0
Globular proteins	30—100	10^5	up to 10^4	1000—30 000	1,2—10 Å
Viruses	200—2000	10^6—10^9	up to 10^7	up to thousands	10 Å and greater

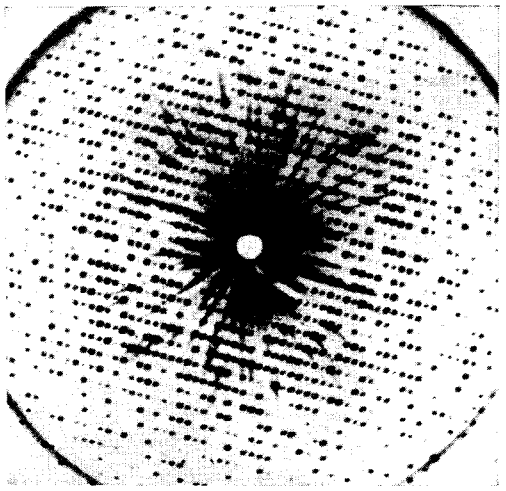


FIG. 7. Precession pattern of a myoglobin crystal (one of the sections of the reciprocal lattice).

0.005 Å, find the anisotropic thermal vibration parameters, etc. If a protein crystal gives an extensive diffraction pattern with tens of thousands of reflections, then for it also one can in principle carry the investigation to the point of determining the atomic positions.

In order to carry out an x-ray diffraction study, one encloses a single crystal of the protein of length up to 1 mm and cross-section several tenths of a millimeter in a thin-walled capillary with a drop of the mother liquor. Special x-ray cameras (precession or Weissenberg) allow one to record photographically the plane nets of the reciprocal lattice (see Fig. 7). The intensities of the spots are then photometered. Automatic x-ray diffractometers connected with computers are now coming into ever wider use.

Protein crystals are sensitive to radiation damage by x-rays. Therefore they are replaced after receiving a certain exposure.

As we shall see below, in the course of an x-ray analysis one has to study the crystals not of one given protein, but of tens of modified crystals. Thus, the scope alone of the x-ray experiment proper becomes enormous. For example, one-fourth of a million reflections were measured in the study of myoglobin from the set of x-ray diffraction patterns of this protein and its derivatives.

However, the fundamental difficulty of the problem in principle consists in something else: the problem of deriving the structure from the experimental data. In order to discuss this problem in its complexity and essence, we must take a short excursion into the theory of x-ray structure analysis in general, and crystalline proteins in particular.

2. PRINCIPLES OF X-RAY STRUCTURE ANALYSIS. METHODS OF DETERMINING THE STRUCTURE OF CRYSTALLINE PROTEINS

The coherent scattering of x-rays involves the electrons, i.e., it is determined by the electron-density function $\rho(\mathbf{r})$, which is the time-average electron distribution of the given object. The function $\rho(r)$ for a crystal is periodic in three dimensions with the periods a, b, and c of the unit cell. The amplitude of scattering is

$$F_{hkl}=\frac{1}{\Omega}\int_{\Omega}\varrho(x,\,y,\,z)\exp\left[2\pi i\left(\frac{hx}{a}+\frac{ky}{b}+\frac{lz}{c}\right)\right]dx\,dy\,dz; \quad (5)$$

Here Ω is the volume of the unit cell, and h, k, and l are the Miller indices of the reflecting planes. The latter determine the corresponding reciprocal-lattice vector $\mathbf{H}_{hkl} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, where $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ are the periods of the reciprocal lattice. The Bragg reflection angles 2θ are determined by the relation

$$\frac{\sin\vartheta}{\lambda}=\frac{|\mathbf{H}_{hkl}|}{2}=\frac{1}{2d_{hkl}}, \quad (6)$$

where d_{hkl} is the interplanar spacing. The intensities of the reflections are proportional to the squares of the moduli of the amplitudes: $I_{hkl} \sim |F_{hkl}|^2$.

Just as in optics, an x-ray diffraction experiment provides a Fourier expansion of the function $\rho(x, y, z)$ of the scattering power of the object, and the amplitudes F_{hkl} of Eq. (5) are nothing other than the coefficients of this expansion. According to Abbe's theory of the optical microscope, the focusing of the diffracted waves with lenses to give the objective image is simply equivalent to a Fourier synthesis. Since lenses for x-rays do not exist, this second stage of image formation is carried out by calculation. Thus, the three-dimensional series

$$\varrho(x,\,y,\,z)=\frac{1}{\Omega}\sum_{h}\sum_{k}\sum_{l}F_{hkl}\exp\left[-2\pi i\left(\frac{hx}{a}+\frac{ky}{b}+\frac{lz}{c}\right)\right] \quad (7)$$

gives the function $\rho(x, y, z)$ of interest to us.[21-24] The coefficients F_{hkl} in the general case of noncentrosymmetric structures are complex quantities:

$$F_{hkl}=|F_{hkl}|\,e^{i\alpha}. \quad (8)$$

If the structure is centrosymmetric ($\rho(\mathbf{r}) = \rho(-\mathbf{r})$), then the phase $\alpha = 0$ or π, and the values of $F = \pm|F|$ are real. The electron-density function can be represented as the sum of the electron densities ρ_j of the atoms contained in the unit cell:

$$\varrho(\mathbf{r})=\sum_{j}\varrho_j(\mathbf{r}-\mathbf{r}_j). \quad (9)$$

Let us discuss the nature of the function $\rho(\mathbf{r})$ reconstructed by Eq. (7) as a function of the completeness of the set of F_{hkl}. We shall assume that the series includes all reflections contained within a sphere of radius $|\mathbf{H}_{max}| = d_{min}^{-1}$. That is, according to Eq. (6), the reflections occur at Bragg angles not exceeding some angle $2\theta_{max}$. In optics this would correspond to limiting the aperture angle of the microscope, which leads to a loss in resolution in the image. An analogous effect of loss of resolution, which can be conveniently characterized by the quantity d_{min}, also

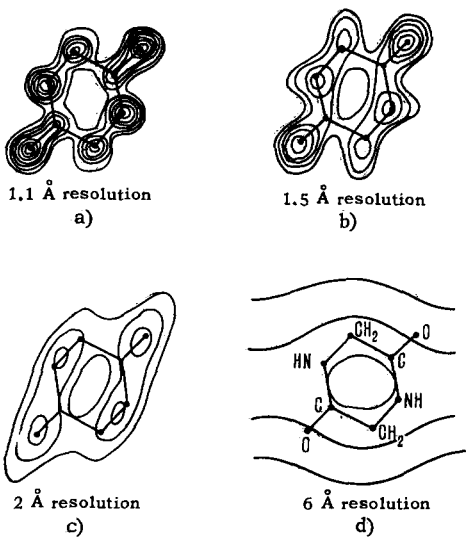


FIG. 8. The electron-density map of the diketopiperazine molecule at different resolutions (in Å).

occurs in the Fourier series.[21-23] Figure 8 shows how a Fourier synthesis changes in nature, depending on the number of F_{hkl} terms included in the series within a certain limiting $|H_{max}| = d_{min}^{-1}$. We see that we must have $d_{min} \cong 1.5$ Å in order to resolve atoms, while $d_{min} \cong 6$ Å gives a crude indication of concentrations of electron density without resolving individual atoms.[25]

The problem of resolution is important from two standpoints in analyzing proteins. First, the crystals of some proteins (including all proteins in the dry form) generally give reflections only for large d (e.g., for $d_{min} \cong 10$ Å), and the study of such proteins is of no interest. The reason for the rapid decline in the intensities is disorder in the structure (rotations and displacements of the molecules and other packing defects, as well as possible defects in the structures of the molecules themselves). Thermal motion also contributes to the decline in intensities with increasing $\sin\theta/\lambda$. One can describe the joint effect of all these factors on the intensity to a first degree of approximation by the Debye temperature factor

$$I_{hkl} = I_{hkl(0)} \exp\left[-16\pi^2\overline{u^2}\left(\frac{\sin\vartheta}{\lambda}\right)^2\right], \qquad (10)$$

where $\overline{u^2}$ is the mean-square displacement of the atoms from their equilibrium positions. For ordinary organic crystals, $\sqrt{\overline{u^2}} \cong 0.2$ Å, for "good" proteins $\sqrt{\overline{u^2}} \cong 0.5$ Å, and for disordered proteins and viruses $\sqrt{\overline{u^2}} \gtrsim 1$ Å, which is comparable with the interatomic distances. Such a quantity in the exponential causes a rapid decline in the intensities of the diffracted beams, d_{min} increases, and the resolving power of the Fourier series falls. In fact, the scattering intensities I_{hkl} of the crystal are a reflection of the electron density averaged over time and over all the unit cells. Hence, if this electron density is "blurred" by disorder, then the Fourier series can reconstitute only an equally blurred picture. However, there is a whole set of crystalline proteins having d_{min} small enough (1.2—2 Å) to permit in principle a complete structure analysis.

The other aspect of the resolution problem is methodological. The number of reflections is proportional to $|H_{max}|^3 = (d_{min}^{-1})^3$. It is hard to measure at once all of the existing tens of thousands of reflections from a crystalline protein and put them all to work at once. It is more expedient to proceed stepwise, first studying the protein at low resolution (say, 5 Å, for which one needs only about a thousand reflections), and after one has revealed the general outlines of its structure, to go on to higher resolution.

When one includes a limited number of reflections in the series, one corrects their intensities by an artificial temperature factor of the type in Eq. (10), which smoothly reduces them to a minimum at d_{min}. This prevents the appearance in the Fourier map of false details arising from sharp cutoff of the series.

The fundamental difficulty of x-ray structural analysis in general, and that of proteins in particular, is the lack of direct evidence on the phases α (Eq. (8)) of the Fourier coefficients since experiment gives only $|F_{hkl}|$.[21-24]

The phases can be easily calculated if the structure is known. The scattering amplitudes f_j of the individual atoms (the atomic form factors) have been tabulated. By using Eqs. (5) and (9), we obtain

$$F_{hkl} = \sum_{j=0}^{n} f_j \exp\left[2\pi i\left(\frac{hx_j}{a} + \frac{ky_j}{b} + \frac{lz_j}{c}\right)\right], \qquad (11)$$

where x_j, y_j, and z_j are the coordinates of the atoms in the cell. If the calculated values $|F_{calc}|$ for some test model agree well with the observed values $|F_{obs}|$, this means that also the calculated phases (or signs, for centrosymmetric structures) are close to the true values. If we assign these phases to the observed values of $|F_{obs}|$ and calculate the series (9), we get a picture of the structure. This type of approach based on fitting $|F_{obs}|$ and $|F_{calc}|$ is possible in searching for, and especially in refining structures,[21-23,26-29] but cannot be applied to proteins.

Also the so-called "direct methods" of determining the phases[30-33] from data on the moduli $|F|$ prove to be inapplicable in practice. They work only for numbers n of atoms per unit cell up to 100—200, whereas $n \cong 10^3$—10^4 for proteins.

Another group of methods is based on the joint x-ray structural analysis of crystals that are similar in structure (the isomorphous-structure method), or on the introduction into the structure of heavy, strongly scattering atoms. In the latter case, the quantities f_j of one or several of such atoms make the major contribution to the value of F_{hkl} (Eq. (11)), and a knowledge of their positions permits one to determine the phases to a first approximation. These approaches are espe-

cially effective when combined with calculating the so-called Patterson function.[34]

Let us form the self-convolution (quadratic convolution) of the electron density:

$$Q(\mathbf{r}) = \overset{2}{\varrho}(\mathbf{r}) = \int \varrho(\mathbf{r}')\,\varrho(\mathbf{r}'+\mathbf{r})\,dV_{\mathbf{r}'}. \qquad (12)$$

Its Fourier coefficients are the products $F_{hkl}F^{*}_{hkl} = |F_{hkl}|^2$, which are experimentally observable positive quantities. The Fourier series based on them is the Patterson function or self-convolution $\overset{2}{\rho}$:

$$Q(\mathbf{r}) = \frac{2}{\Omega}\sum_{h=0}^{+\infty}\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}|F_{hkl}|^2\cos\left[2\pi\left(\frac{hx}{a}+\frac{ky}{b}+\frac{lz}{c}\right)\right]. \qquad (13)$$

$Q(\mathbf{r})$ takes on large values whenever the vector $\mathbf{r}$ in (12) corresponds to distances between points $\mathbf{r}'$ and $\mathbf{r}'+\mathbf{r}$ such that the original function $\rho(\mathbf{r}')$ has large values at each of the points. However, these $\mathbf{r}$ values are the distances between the centers of the atoms. By using the representation (9), we can express the convolution (12) as a sum of the convolutions of the electron densities of pairs of atoms $\overset{\frown}{\rho_j\rho_k} = q_{jk}(\mathbf{r}-\mathbf{r}_{jk})$, where $\mathbf{r}_{jk} = \mathbf{r}_j - \mathbf{r}_k$ is the interatomic distance:

$$Q(\mathbf{r}) = \sum_{j=k}^{n} q_{jj}(\mathbf{r}-0) + \sum_{j\neq k}^{n(n-1)} q_{jk}(\mathbf{r}-\mathbf{r}_{jk}). \qquad (14)$$

Thus, $Q(\mathbf{r})$ contains the peaks q_{jk}; the vector $\mathbf{r}_{jk}$ joining each pair of atoms in the structure is also represented in $Q(\mathbf{r})$ by an oriented vector drawn from the coordinate origin to the peak q_{jk} (Fig. 9).

When there are n atoms in the structure, $Q(r)$ contains n^2 peaks, n of them coinciding at the origin.[14]

While the function $Q(\mathbf{r})$ does not give a picture of the atomic arrangement, the information that we can get from it is very valuable: the observed interatomic distances r_{jk} determine the distances existing in the structure. In principle, one can unambiguously derive the structure from $Q(\mathbf{r})$.[35-39] However, this approach also is inapplicable to proteins. The problem is that the unit-cell volume is proportional to n, while the number of peaks in the Patterson function is proportional to $n(n-1)$, and they are subject to multiple overlap. In proteins, $n > 1000$, and the $n(n-1) > 10^6-10^7$ overlapping maxima of the Patterson function can reflect only the overall character of the distribution of interatomic distances.

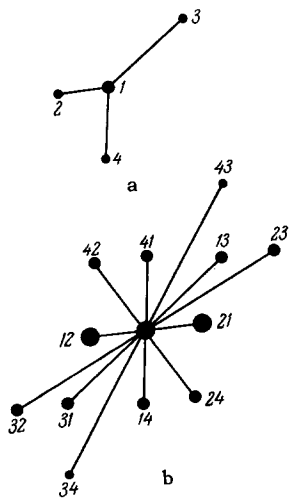


FIG. 9. A four-atom point structure (a) and its interatomic-distance function (b). Atoms 1 and 2 are heavy.

If a light-atom structure contains some heavy atoms (1 and 2 in Fig. 9), then since $q_{jk} \sim Z_jZ_k$, their corresponding peaks will be the strongest. However, in proteins even these peaks cannot be directly discerned against the background of even stronger peaks arising from overlap. Still, this difficulty can be obviated if one has two isomorphous structures. That is, the structures have identical cells, all their atoms are in the same arrangement, and all of them are identical except for one or several (heavy) ones. Then, by working with the differences of the Patterson function (13), one can eliminate the effect of the "common portion" of the structure and find the position of the heavy atom.

But how can we obtain isomorphous crystals of such complex molecules as proteins, which are produced only in the process of biosynthesis? At first, they tried to use dry and wet crystals of a given protein as the isomorphs. Since the structure of the molecules does not change under moderate drying, the difference in scattering power arises from the decrease in the amount of mother liquor. The x-ray diffraction patterns show appreciable changes in the intensities of the low-order reflections.[40,41] In the Fourier expansion (7), they correspond to the long-wave harmonics determining the coarsest details of the structure. Hence, an analysis of the intensity changes upon drying cannot give information on the fine structure of the molecules. However, it proves useful in determining their external shape and position in the unit cell.

M. Perutz was the first to call attention to the fact that while one cannot replace any of the atoms of a protein by heavier atoms, perhaps one could "attach" such atoms to a protein molecule as part of relatively small organic molecules or inorganic ions.[42] L. Bragg has called this idea of Perutz a gold mine, but as Perutz later remarked, the gold mine proved to lie so deep that the first results of its development arrived after many years. Hundreds of experiments with various substances containing heavy atoms revealed the following. The surfaces of protein molecules, being rich in active groups, sometimes adsorb such substances. However, isomorphism is often lost thereby: the unit cell and molecular packing are changed, and such substituted crystals are not suitable for x-ray diffraction analysis. In working on the structure of ribonuclease during the fifties, the group of the prominent researcher D. Harker (USA) obtained tens of derivatives containing heavy atoms, but they all proved not to be isomorphous.[43] The English have had better success: J. Kendrew, working on myoglobin, and M. Perutz himself, who has studied hemoglobin (here they tested the crystals of these proteins as obtained from animals of different species).

The status of the problem at present is as follows. There is a list of about a hundred substances that show promise as possible isomorphous additives, e.g., $HgCl_2$, $AuCl_4^-$, PtI_4^-, $PtCl_6^=$, $UO_2(OH)_n$, $IrCl_3$, $Pt(NH_2CH_2COOH)_2$, $Pb(CH_3COO)_2$, $Ta_6Cl_{12}Cl_2$, ClHg⟨⟩SO_3H, $UO_2(P_2O_7)_n$, $AgNO_3$, ClHgCH(HgCl)COOH, etc.

We see that they comprise the most varied compounds without any special features permitting us to choose them systematically (although attempts at a conscious selection are continually being made). Except for special cases (e.g., the covalent interaction of certain mercurials with SH groups), we cannot predict whether a given additive will react with a given protein, with what groups on the protein molecule it will combine, and if it combines, whether the structure will remain isomorphous with the original. The nature of the binding forces of the introduced groups with the protein surface is not clear; apparently electrostatic interaction plays a large role.

There are two ways to introduce heavy-atom additives (M) into protein crystals (P). The first is to grow the protein crystals from a solution containing M, and the second is to put a prepared crystal of the pure protein into such a solution. In the latter case, the M molecules diffuse into the crystal, and occupy positions specific for them on the protein molecules. It sometimes proves useful to introduce two additives at the same time.

Now we shall estimate the effect of a heavy atom on the intensities of the diffracted beams.[44] It would seem at first glance that this effect should be insignificant, for a protein molecule contains about 10^5 electrons scattering x-rays, whereas the introduced M atom contains only 80—90 electrons. However, this is not the case.

According to a theorem on the completeness of Fourier series, the intensities $I_{hkl} = |F_{hkl}|^2$ are related to the squared values of the electron density in the cell.[45] Hence, taking (11) into account, we can write

$$\sum_{h,k,l=-\infty}^{+\infty} I_{hkl} = \Omega \int_{\Omega} \varrho^2(\mathbf{r})\, dv_r = \Omega \sum_j \int \varrho_j^2\, dv_r. \qquad (15)$$

We see that the atoms contribute to the intensity in proportion to the squares of their electron densities, i.e., on the average,

$$\bar{I} \sim \sum_j Z_j^2, \qquad (16)$$

where the Z_j are the atomic numbers. Then, for the protein $\bar{I}_P \sim nZ_P^2$, where n is the number of atoms in the molecule, whereas for the protein with an added heavy atom $\bar{I}_{P,M} \sim nZ_P^2 + Z_M^2$.

If we write $\Delta I = \bar{I}_{P,M} - \bar{I}_P$, we find that

$$\frac{\sqrt{\overline{(\Delta I)^2}}}{\bar{I}_P} \cong \sqrt{\frac{2}{n}}\,\frac{Z_M}{Z_P}. \qquad (17)$$

Proteins consist of C, N, and O atoms, i.e., $Z_P \cong 7$ (we can neglect the H atoms), while $Z_M \cong 80$. Then, for example, for myoglobin which has $n \cong 1200$, Eq. (17) gives a relative mean-square deviation of about 45%. The whole point is that the light atoms, which are distributed throughout the unit cell, do not scatter in phase, but the electrons of the heavy atom are concentrated together, and make a constant (in phase) contribution to the scattering. Of course, not all the intensities are altered, but there is a statistical relation. The contribution of the heavy atoms to the intensity declines when they show incomplete occupancy of their positions, or when the conditions of isomorphism are not strictly met.[44]

The first stage in determining the structure of a protein is to determine the position of the heavy atom M in the substituted protein P,M.[46-49,93] As we already know, calculation of the Patterson function from $|F_{P,M}|^2$ does not reveal the position of M. However, if we measure the intensities $|F_P|^2$ from the pure protein, we can combine these quantities. Let us examine what a Patterson synthesis with coefficients $(|F_{P,M}|^2 - |F_P|^2)$ will give. According to (13), this will be the difference $Q_{P,M} - Q_P$, and according to (14), these functions are:

$$Q_P = \sum_{j,j'} q_{jj'}(\mathbf{r} - \mathbf{r}_{jj'}), \qquad (18)$$

$$Q_{P,M} = \sum_{j,j'} q_{jj'}(\mathbf{r} - \mathbf{r}_{jj'}) + \sum_{M,M'} q_{M,M'}(\mathbf{r} - \mathbf{r}_{M,M'}) + \sum_{j,M} q_{j,M}(\mathbf{r} - \mathbf{r}_{j,M}). \qquad (19)$$

The difference map of interest to us contains not only the interaction peaks $q_{M,M'}$ of the heavy atoms but also the peaks $q_{j,M}$ arising from interaction of the heavy atoms with the n atoms of the protein (n being large). While the latter give an undesirable background, nevertheless such syntheses permit one to determine the positions of the heavy atoms.

Another possibility is to calculate the Patterson function with coefficients $|\Delta F|^2 = ||F_{P,M}| - |F_P||^2$. First let us consider the case of a centrosymmetric

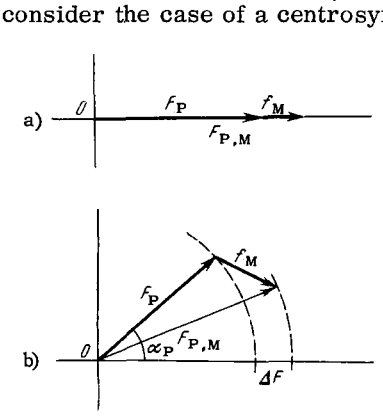


FIG. 10. The structure factors of the protein F_p, the heavy atom f_M, and the overall structure factor $F_{P,M}$. a) Centrosymmetric case (the phases α are 0 or π). b) General case of the phase diagram for a protein containing a heavy atom.

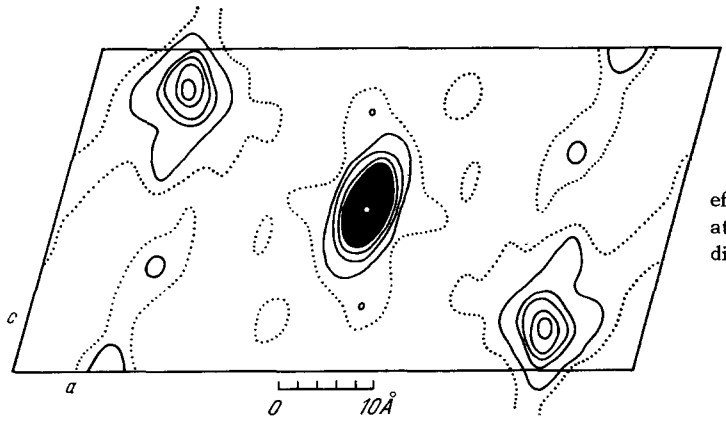


FIG. 11. Difference Patterson function calculated from the coefficients $||F_{P,M}| - |F_P||^2$ for myoglobin. The coordinate origin is at the center. The vector to the peak MM' determines the M–M' distance ($M = HgI_4^{=}$).

projection of a protein crystal (e.g., along a twofold axis). In this case, $F_{P,M}$ and F_P have the phases $\alpha = 0$ or π. Then, since f_M is small (Fig. 10a), their phases are the same, as a rule. That is, $|\Delta F| = |F_{P,M} - F_P| = |f_M|$, where f_M is the scattering amplitude of the M atoms. Consequently, the Patterson synthesis in $|\Delta F|^2$ in this case is nothing other than

$$Q_M = \overset{2}{\varrho}_M = \sum_{M,M'} q_{M,M'} \qquad (20)$$

This is the function of interest to us, from which we can easily find the coordinates of the heavy atoms.

However, in the general non-centrosymmetric case, the quantity $|\Delta F|^2$ cannot be interpreted so simply (Fig. 10b). It can be shown that a Patterson synthesis derived from such $|\Delta F|^2$ values will contain the systematic $q_{M,M'}$ peaks of (20), but it will also contain a certain background arising from $q_{j,j'}$ and $q_{j,M}$ (Fig. 11).

It is essential that the positions of the heavy atoms M_1 or M_2 found in different derivatives PM_1, PM_2, etc., should be referred to the same origin. To do this, one uses the so-called interatomic-distance correlation functions.[50-52]

When the positions of the heavy atoms have been found, their contribution f_M to the structure factor $F_{P,M}$ of the heavy-atom-labeled protein is then known:

$$F_{P,M} = F_P + f_M, \qquad (21)$$

where $F_{P,M}$ and F_P are defined by Eq. (11). Equation (21) is illustrated in the complex plane by the diagram in Fig. 10b. We know from experiment the moduli $|F_P|$ and $|F_{P,M}|$. The problem is to find the phases α_{hkl} of the reflections $|F_{P,hkl}|$ from the protein. As has been done in Fig. 12a, the intersections of the circles of radii $|F_P|$ and $|F_{P,M}|$ drawn about the points 0 and $-f_M$ give two possible values, α and α'. Hence the problem isn't solved yet. However, if we have a second isomorphous derivative, the phase α can be determined unambiguously from the quantities

$$f_{M1},\quad |F_{P,M1}|,\quad f_{M2},\quad |F_{P,M2}|,\quad |F_P| \qquad (22)$$

according to the same procedure (Fig. 12b).[53]

The quantities $f_{M,i}$ are small. Errors occur in determining them and all the other quantities entering into (22). Hence, it is better to have several derivatives to determine the phases, rather than just two (Fig. 13). However, such a procedure of graphic determination of the phase angle "by eye" is arbitrary.

One can select the phase angle in a more rigorous way.[54,55] One can calculate the probability $p(\alpha, |F_P|)$ that a given phase α and given modulus $|F_P|$ are cor-

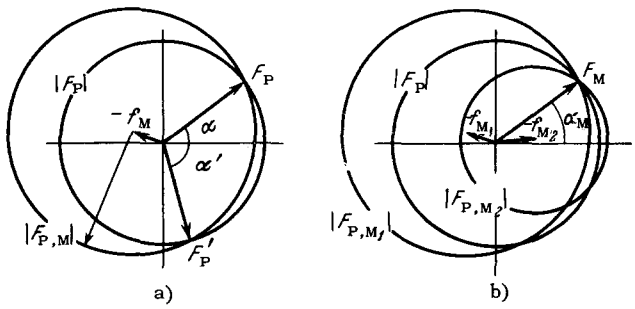


FIG. 12. a) Diagram of the determination of the two possible values of the phases α_P from the known values f_{MP} $|F_{P,M}|$, and $|F_P|$. b) Unambiguous phase determination when two heavy-atom derivatives are used. In addition to the quantities indicated in diagram (a), f_{M_2} and $|F_{P,M_2}|$ are also known.

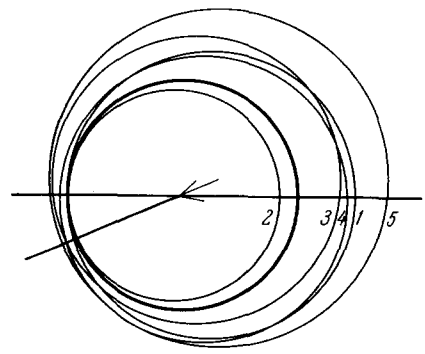


FIG. 13. An example of determination of the phase α_P for a myoglobin reflection. Five derivatives were used; the heavy circle corresponds to $|F_P|$ (the protein).

rect from the errors of determining the quantities in (22) for all the existing derivatives. Then one draws the corresponding plot on the complex plane (Fig. 14) (this distribution may have two maxima).

The errors in determining $|F_P|$ and α in calculating the Fourier synthesis (7) give the errors $\Delta\rho$ in the electron-density distribution. We can find the mean-square error $\overline{\Delta\rho^2}$ from Eq. (15); it is

$$\overline{\Delta\rho^2}=\frac{1}{\Omega}\sum_{h,k,l}(\Delta F_{hkl})^2, \qquad (23)$$

where ΔF_{hkl} is the error of determination of the experimental values F_P. The Fourier synthesis will be "optimal" when $\overline{\Delta\rho^2}$ is minimal. This requirement is satisfied under the following condition.[54] One must use in the Fourier synthesis not the most probable values of $|F|$ and α found from the highest maximum in the distribution $p(\alpha, |F|)$, as it would seem at first glance, but the quantities $r|F|$ and α_r. The latter are the polar coordinates of the centroid of the probability distribution (see Fig. 14). If the distribution has a single maximum, then it is practically unambiguous. However, if there are two separate maxima, the given amplitude provides little information of much value, and its contribution to the Fourier series is correspondingly decreased in proportion to the quantity r. The procedure for finding α_r and r is assigned to the computer (using this approach, it has been shown possible to seek protein structures based on a single isomorphous substitution[56]).

Thus, the method of introducing heavy atoms and determining the phases from isomorphous series is now the fundamental, but unfortunately very narrow doorway into the protein-structure world. There are

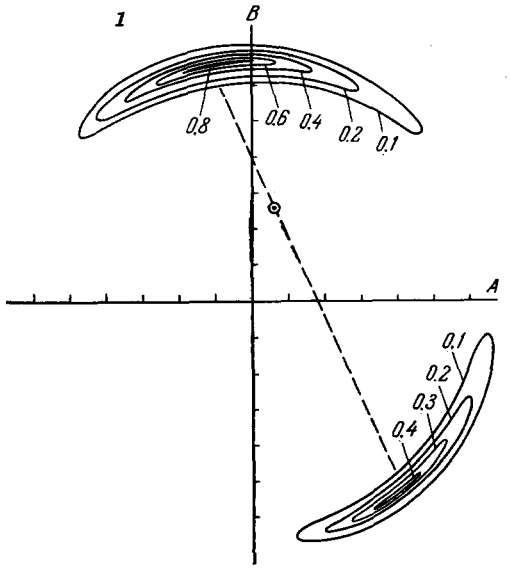


FIG. 14. The probability distribution $p(\alpha, |F_P|)$ in the complex plane. To construct the best Fourier map, one uses the value corresponding to the center of gravity of the distribution, having coordinates α_r, $r|F|$, (indicated by ⊙), instead of the most probable value α, $|F|$ (indicated by 1).

two other "transoms," which we should also discuss.

When x-rays of a wavelength near the absorption edge of a given atom are scattered, the phenomenon of anomalous dispersion takes place. In this case, we have to add to the ordinary real value of the atomic scattering factor f_0 a real component $\Delta f'$ and an imaginary one $i\Delta f''$ as well:

$$f=f_0+\Delta f'+i\Delta f''=f'+if''. \qquad (24)$$

With a suitable choice of wavelength, the quantity f'' for the heavy atoms making the major contribution to the anomalous scattering can amount to 15% of f_0.

First of all we note that for ordinary scattering, according to (5) and (11),

$$F_{\mathbf{H}}=F^*_{-\mathbf{H}}, \quad |F_{\mathbf{H}}|=|F_{-\mathbf{H}}|, \quad \alpha_{\mathbf{H}}=-\alpha_{-\mathbf{H}},$$

That is, the amplitudes for the centrally-related points hkl and $\overline{hkl}$ in reciprocal space are complex conjugates. Their moduli are identical and they are experimentally indistinguishable (Friedel's Law). However, Friedel's Law does not hold in case of anomalous scattering.[57-58] For simplicity, we shall examine the case in which the only atom scattering anomalously is the heavy atom M. Then,

$$F(\mathbf{H})=\sum_j f_j\exp(2\pi i\mathbf{r}_j\mathbf{H})+f'_M\exp(2\pi i\mathbf{r}_M\mathbf{H})$$
$$+if''_M\exp(2\pi i\mathbf{r}_M\mathbf{H})=F_{P,M}(\mathbf{H})$$
$$+f''_M\exp\left[i\left(2\pi\mathbf{r}_M\mathbf{H}+\frac{\pi}{2}\right)\right]. \qquad (25a)$$

The centrosymmetric reflection has the amplitude

$$F(\overline{\mathbf{H}})=\sum_j f_j\exp[2\pi i\mathbf{r}_j(-\mathbf{H})]+f'_M\exp[2\pi i\mathbf{r}_M(-\mathbf{H})]$$
$$+if''_M\exp[2\pi i\mathbf{r}_M(-\mathbf{H})]=F_{P,M}(\overline{\mathbf{H}})$$
$$+f''_M\exp\left[i\left(-2\pi\mathbf{r}_M\mathbf{H}+\frac{\pi}{2}\right)\right]. \qquad (25b)$$

The corresponding diagram is shown in Fig. 15. We see that if anomalous scattering is neglected, $F_{P,M}$ obeys Friedel's Law: $F_{P,M}(\mathbf{H}) = F^*_{P,M}(\overline{\mathbf{H}})$. However, the imaginary component f''_M in the two cases has a phase shift of $+\pi/2$. Thereupon,

$$|F(\mathbf{H})|\neq|F(\overline{\mathbf{H}})|. \qquad (25c)$$

When we take into account anomalous scattering, it becomes possible, first, to determine the absolute configurations of molecules in enantiomorphic crystals, and second, to determine the positions of the anomalously scattering atoms M.[59-62] A single isomorphous derivative is sufficient to do this in principle.[63] The anomalous-scattering effect is used in protein-crystal studies most often in combination with the ordinary heavy-atom method.[61]

There is still another idea on deciphering protein structures.[64,65] I have already mentioned that some protein molecules are built of identical subunits S.

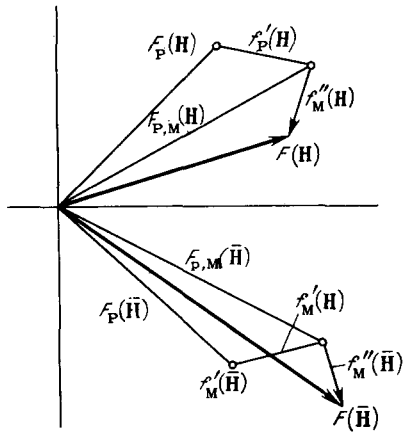


FIG. 15. Phase diagram of the centrally-related reflections F(H) and F($\bar{H}$) for the anomalous-scattering case.

Each subunit is characterized by its vector set of interatomic peaks $\sum_{S_1} q_{jk}$. The set $\sum_{S_2} q_{jk}$ of peaks of another subunit is identical, but differently oriented, and both are contained in the Patterson function (14). Let us calculate the "self-overlap" of these functions under rotation:

$$R(\vartheta, \varphi) = \int Q(\mathbf{r})\, Q(\mathbf{r}')\, dv_r, \qquad (26)$$

where $\mathbf{r}' = [C]\mathbf{r}$, and [C] is the coordination-transformation matrix for rotations. The function R has trivial maxima for rotations corresponding to the crystallographic symmetry, but will also have non-trivial maxima θ_1, φ_1 where $\sum_{S_1} q_{jk}$ and $\sum_{S_2} q_{jk}$ coincide. This determines the relative orientation of the subunits. This method has been used in studying insulin, but has thus far given no substantial results.

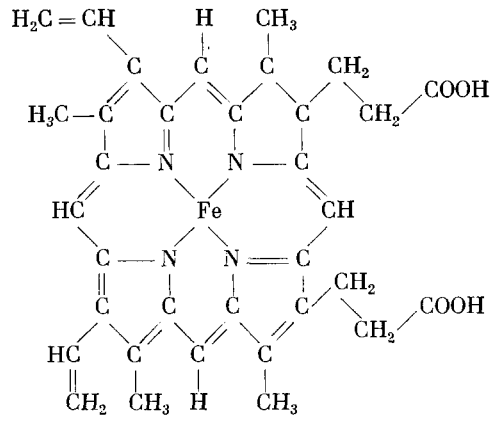
At present, the fundamental method is that of heavy isomorphous substitution.

3. THE STRUCTURES OF MYOGLOBIN AND HEMOGLOBIN

Myoglobin and hemoglobin are globular proteins that reversibly bind the oxygen molecule O_2. The hemoglobin of the red blood cells transports oxygen in the bloodstream. The combination or release of oxygen is regulated by its partial pressure. Thus, the hemoglobin is saturated with oxygen in the lungs where it is abundant, and releases it to the tissues, where the oxygen is consumed. Hemoglobin also indirectly participates in another aspect of the respiratory process: the removal of CO_2 by the venous blood.

Myoglobin is the protein that stores oxygen in the muscles and releases it as they do work. This function is especially important for animals whose breathing is interrupted. Hence, the muscles of whales, seals, dolphins, and penguins are very rich in myoglobin.

In addition to the protein component globin, myoglobin and hemoglobin contain the prosthetic group heme:



(27)

The Fe^{++} atom lying within the porphyrin ring is the agent that directly plays the role of weakly and reversibly binding the O_2 molecule. However, the Fe^{++} atom possesses this function neither by itself nor when contained in heme, but only when heme interacts with globin. However, if the protein component differs, e.g., in catalase, which also contains heme, the function of the heme also is altered.

J. Kendrew and his associates in Cambridge began the x-ray diffraction study of myoglobin shortly after the end of the Second World War.[16,55,66-70] After a series of tests, they chose for study the myoglobin of the sperm whale, which gives good crystals (see Fig. 5). The molecular weight of this protein is about 18,000, and it contains 153 amino-acid residues, i.e., about 1200 atoms not counting hydrogens. Myoglobin has one terminal amino group. This indicates that it consists of a single polypeptide chain. It has no cystine S—S bridges. The primary structure of myoglobin was not known at the time that the x-ray structural analysis was made, but it has now been determined, both from the x-ray investigation itself and by chemical means.[71]

The unit cell of wet crystals of type A sperm-whale myoglobin crystals is monoclinic: a = 64.6 Å, b = 31.1 Å, c = 34.8 Å, β = 105.5°, the space group is $P2_1$, and the cell contains two molecules of protein. Numerous tests made it possible to find five additives giving isomorphous crystals: K_2HgI_4, $AgNO_3$, p-chloromercuribenzenesulfonate, $Hg(NH_3)^{++}$, p-iodophenylhydroxylamine, and some others. The determination of the heavy-atom positions (see Fig. 11) made it possible to proceed to determine the phases (see Fig. 13).

It gave no substantial results to calculate the projections. The first view of the tertiary structure of the myoglobin molecule and protein molecules in general was given by the three-dimensional Fourier synthesis at 6 Å resolution (1957) (Figs. 16 and 17). The molecule has the shape of a somewhat flattened triangular prism of dimensions 25 × 35 × 45 Å. The map revealed rodlike condensations of electron density of diameter 5 Å, separated by distances of about 10 Å, and arranged

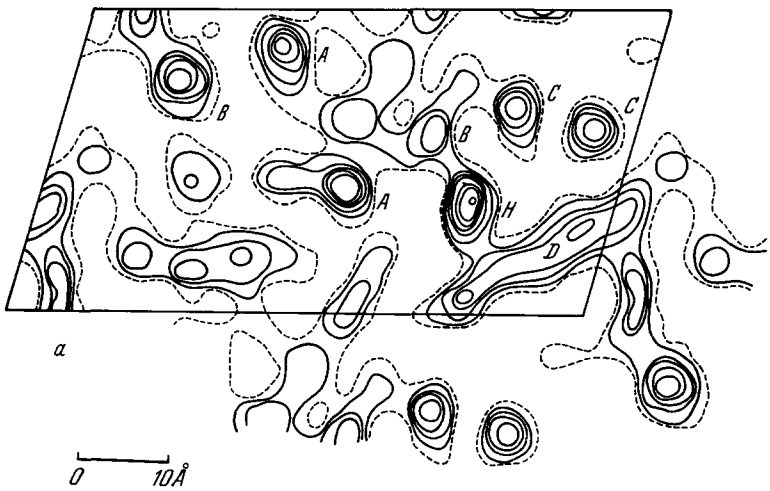


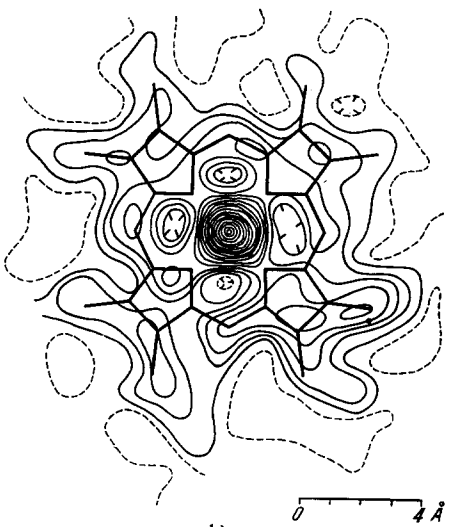
FIG. 16. One of the sections of the Fourier map of myoglobin at 6 Å resolution. A–D; cross-sections of α helices. H–heme.

in a very complicated fashion (see Fig. 17). There was no doubt even at this stage that this was the polypeptide chain in the form of an α-helix. Owing to the inadequate resolution, at first they could not determine quite exactly the course of the chain, since they observed bridges of elevated electron density between the rods at several places. However, they did this quite unambiguously at the next stage. A flat disk having a considerable electron density was identified as the heme.

The next stage was to calculate a Fourier map from the 9600 reflections at 2 Å resolution[55,68] (1960). The synthesis was carried out in the form of 96 sections perpendicular to the a axis at a spacing of 2/3 Å (Fig. 18). This map distinctly revealed the course of the polypeptide chain. In the previously-observed rod-like regions, it was now directly resolved as a right-handed α-helix of pitch 5.4 Å (Fig. 19). The side-groups R repeat after a 100° rotation of the helix and with an axial spacing of 1.5 Å. A model of the molecule is shown in Fig. 20 (see also Fig. 24a). A total of 118 of the 153 residues proved to be in the α configuration (i.e., about 70% of the residues). An α-helix is observed in eight almost exactly linear regions (the axis of the helix being linear), each consisting of from 7 to 24 residues. About 48 residues are found at the sites of the "breaks" and bends. Most of the breaks contain three residues, and one of the bends contains eight residues bent in the form of an S-shaped loop (upper right-hand corner of Fig. 17). The planar heme group was well revealed. With respect to the surface of the molecule, it lies as though in a pocket formed by the folds of the polypeptide chain (Fig. 18b). Contrary to expectation, the iron atom proved to lie 0.3 Å out of its plane. The coordination of the Fe atom is octahedral; it is surrounded by four N atoms of the porphyrin ring

FIG. 17. Three-dimensional model of the distribution of high electron density regions in the myoglobin molecule obtained at 6 Å resolution, indicating the course of the polypeptide chain (the dark disk is the heme).

a)



b)

FIG. 18. a) Three-dimensional Fourier map of myoglobin at 2 Å resolution: set of plane sections. b) The region of the map corresponding to the heme.

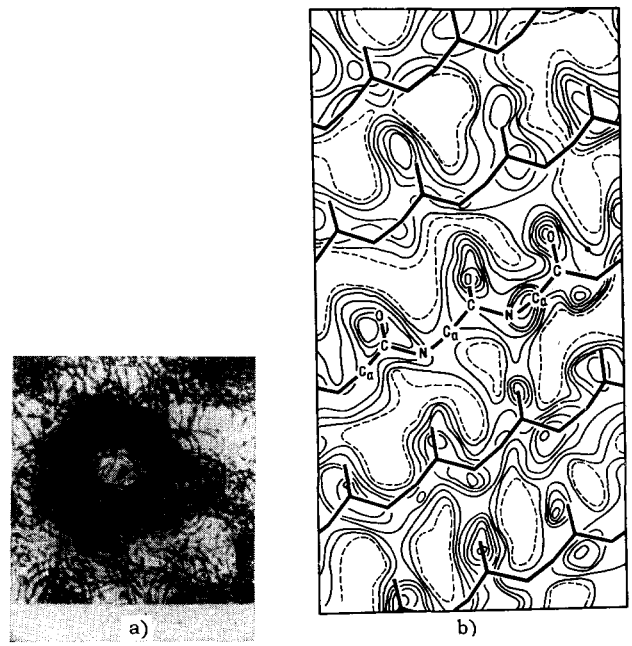


FIG. 19. a) A view of the map of Fig. 18 along the axis of one of the α helices. b) An unwound cylindrical section of the map of Fig. 18 along one of the α helices.

and is supported on the side facing the molecule by one of the N atoms of a histidine residue of the polypeptide chain. The oxygen molecule combined with the myoglobin is attached to the sixth, outer position at a distance of 2.1 Å. Study of the electron-density distribution in the side groups, even at the insufficient resolution of 2 Å, still permitted the identification of about 100 of them. These data proved to agree well with the chemical data.[71] The positions of three-fourths of the atoms in the molecule were established.

In recent years, the structure has been refined by using all 25,000 reflections at a resolution of 1.4 Å. The phases were found by cycles of successive approximations[70] (Fig. 21). The positions of 120 residues have been determined with full reliability, and the remaining 30 to a high degree of probability; 1100 of the 1200 atoms have been resolved. Most of the charged and polar side groups R (lysine, arginine, glutamic and aspartic acids, histidine, tryptophan, serine, tyrosine, and threonine) lie at the surface of the molecule, whereas the non-polar residues are inside.[72]

Scouloudi[73] has studied seal myoglobin by the isomorphous-replacement method. By calculating the Fourier projection at 4 Å resolution and comparing it with the analogous projection of sperm-whale myoglobin, she showed that the tertiary structure, or na-

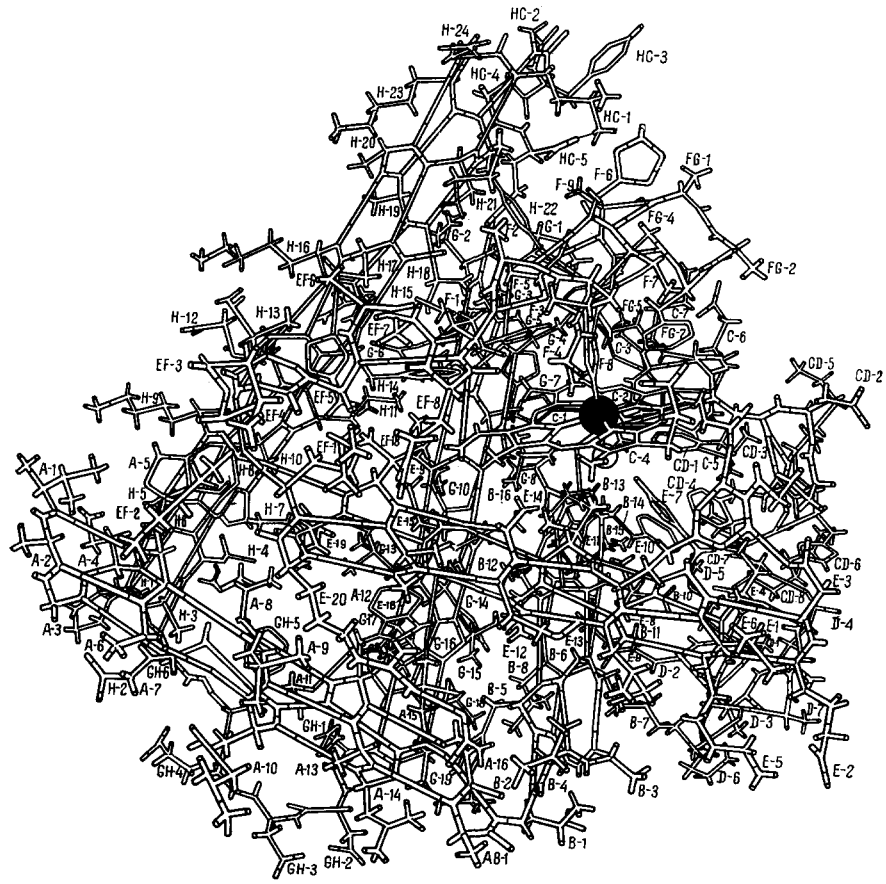


FIG. 20. Three-dimensional model of the myoglobin molecule. The rods are the covalent and hydrogen bonds, and the sphere is the Fe atom.

Table III

Hemoglobin	Symmetry, space group	a, Å	b, Å	c, Å	β	Number of molecules per cell
Horse oxy-	Monoclinic, C2	109.0	63.5	54.9	110°53′	2
Horse reduced	Orthorhombic, $C222_1$	77,0	81.8	92.7		4
Human reduced	Monoclinic, $P2_1$	63.4	83.6	53.9	99°15′	2

ture of the pathway of the chain, was the same in these proteins.

M. Perutz and his associates have conducted an extensive set of studies of different hemoglobins.[17,18,42,74-80,120] The molecular weight of hemoglobin is 64,500, and the molecule contains about 5000 C, N, and O atoms plus the same number of hydrogens and four Fe atoms. It consists of four subunits in identical pairs. Two of them (the α-chains) contain 141 residues, and two (the β-chains) contain 146, or 574 residues in all. Each subunit contains one heme. However, the hemoglobin molecule is not a mechanical quadruple of protomers, for the hemes in it interact: if three of them have taken up O_2, then the ability of the fourth one to do so is sharply increased.

Four forms of hemoglobin are known: oxyhemoglobin (containing O_2), reduced hemoglobin (without O_2), carboxyhemoglobin ("poisoned" by stable combination with CO), and methemoglobin (with trivalent Fe^{+++}).

As early as 1938, Haurowitz[81] discovered the remarkable fact that the external structure of hemoglobin crystals changes upon transition from the oxy- to the reduced state and back. This indicated a change in the structure of the molecules upon reacting with oxygen.

The most complete x-ray diffraction data have been obtained for the crystals whose characteristics are given in Table III.

The initial studies established the shape of the molecules, the nature of their hydration (see Fig. 6), the presence of α helices in the molecule, and the mutual orientation of the heme planes (the latter was determined by the paramagnetic-resonance method), and a number of other features of the structure.

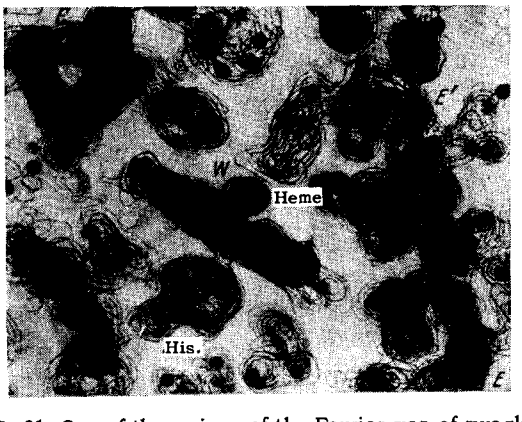


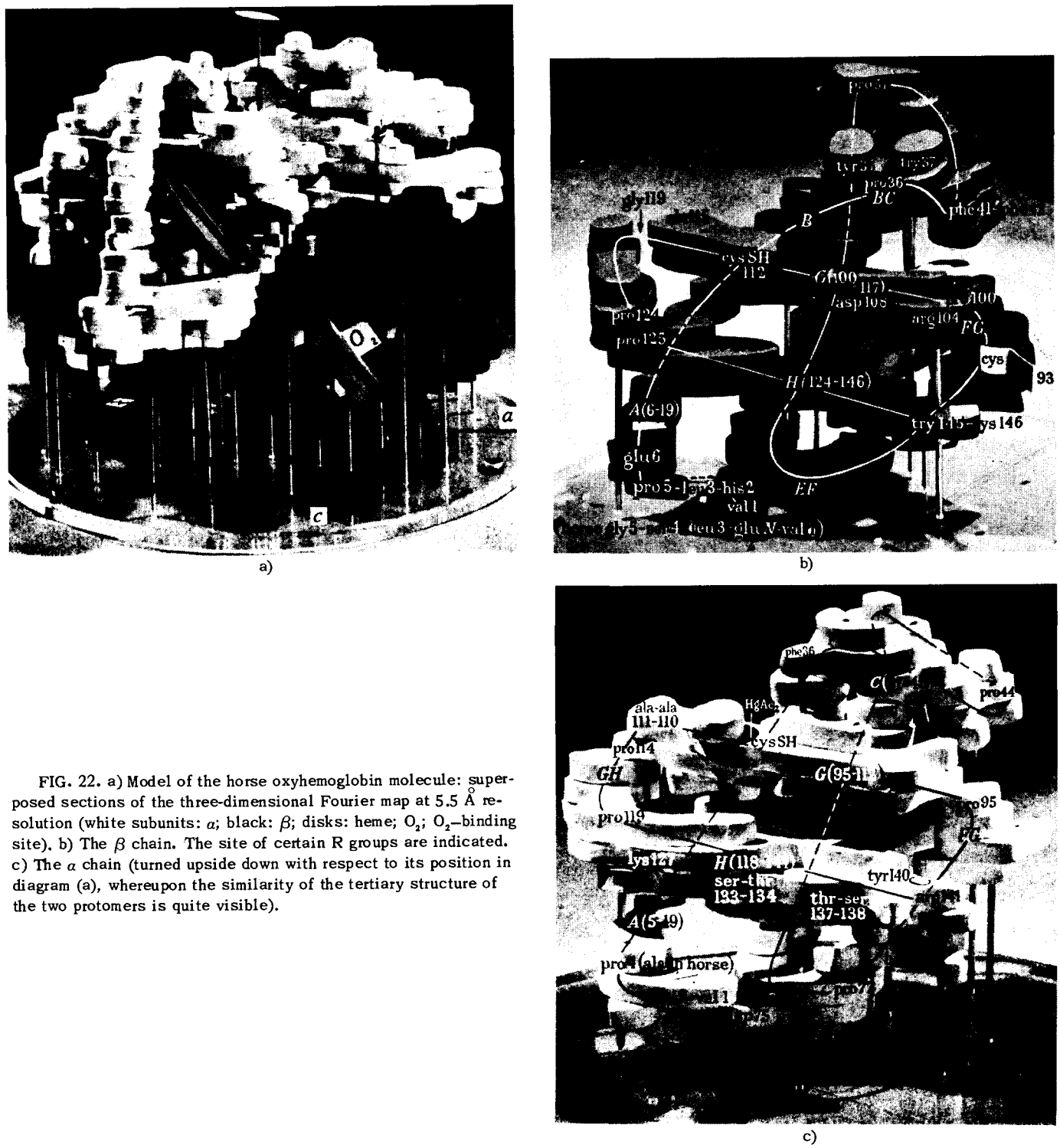
FIG. 21. One of the regions of the Fourier map of myoglobin at 1.4 Å resolution. In the center: the heme projected parallel to its plane; a histidine residue adjoins it on one side, and on the other side is a water molecule occupying the hydrogen-binding site.

The isomorphous-substitution method combined with anomalous-scattering data made it possible in 1960 to obtain a three-dimensional Fourier synthesis of horse oxyhemoglobin crystals at a resolution of 5.5 Å (Fig. 22). The series included 1200 reflections. The isomorphous substituents used were p-chloromercuribenzoate, which reacts with the SH groups of the β-chains, di(chloromercuri)acetate, mercuric acetate, and some others. The phases were determined from the probability centroid (see Fig. 14). The electron density of the polypeptide chains has values of 0.54 electron/$Å^3$ and greater. If we cut out figures on this contour line from all the calculated two-dimensional sections of the Fourier, and superimpose them on one another, we will get the model of the molecule shown in Fig. 22a. The model excellently reveals the α (white in Fig. 22c) and β (black in Fig. 22b) subunits of the molecule, the course of the polypeptide chain in them, and the arrangement of the hemes.

The molecules are situated on the twofold axes of the unit cell, and thus have twofold axes themselves. The four subunits are compactly packed in 50×55×64 Å ellipsoid. They are packed in such a way that the molecule approximately has a tetrahedral point-group symmetry. The contact between the subunits of the same type is relatively small. Consequently there is a channel along the 50 Å axis (coinciding with the true twofold axis), with recesses on both sides. Just as in myoglobin, the hemes lie in special "pockets." The iron atoms occur at the distances: Fe_α—Fe_α 36.0 Å, Fe_β—Fe_β 34.3 Å, Fe_α—Fe_β 25.2 Å.

Combination of the known data on the primary structure of the hemoglobin chains with a knowledge of the chain configurations obtained by x-ray analysis, together with the detailed known data on the structure of myoglobin, has greatly simplified the analysis of the hemoglobin structure and has made it possible to give a model of the tertiary structure with a number of details[80,120] (Fig. 24).

Comparison of the primary structure, i.e., the sequence of the amino-acid residues R in the α and β chains shows that substitutions of R have occurred in 78 of the 150 sites. Nevertheless, it turned out that the tertiary structure of the two protomers is almost identical (Figs. 22b and c). The fact was even more strik-



a) b) c)

FIG. 22. a) Model of the horse oxyhemoglobin molecule: superposed sections of the three-dimensional Fourier map at 5.5 Å resolution (white subunits: α; black: β; disks: heme; O_2; O_2–binding site). b) The β chain. The site of certain R groups are indicated. c) The α chain (turned upside down with respect to its position in diagram (a), whereupon the similarity of the tertiary structure of the two protomers is quite visible).

ing that the tertiary structure of both of these protomers resembled myoglobin just as much (Fig. 24), though the latter essentially differs from them in primary structure. In fact, if we compare these three sequences together, only 20 R groups occupy identical positions.

In their most recent studies[120,121] trying to elucidate the factors determining the tertiary structure and its relation to the primary structure, Perutz, Kendrew, and Watson have compared the 18 sequences known for respiratory proteins: two myoglobins and 16 α and β chains of hemoglobins of different animal species. It turned out that only nine R groups occupy identical positions throughout the set. Evidently, these groups are of vital importance to the tertiary structure. Indeed, five of them are in contact with the heme, and the other four form interchain stabilizing bonds or define breaks in the α-helices.

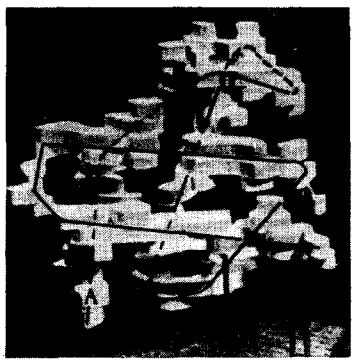


FIG. 23. The α chain of reduced human hemoglobin (5.5 Å resolution) (cf. Fig. 22c).

However, another fact apparently plays no smaller a role: only non-polar (hydrophobic) residues lie inside the protomer in all cases. The non-polar (N) residues are the following: Gly, Ala, Val, Leu, Ile, Phe, Pro, Cys, Met, Trp, and Tyr; the others are polar (P). The numbers of P and N residues in the examined proteins, as in most proteins in general, are about equal. That is, they amount to 70—80. Thirty-three residues are completely inside the molecule, i.e., out of contact with the H_2O molecules surrounding the protein. In all 18 examined cases, these 33 residues are of N type, and while among them there are many substitutions of one R for another, they are all N⟷N replacements, but not N→P. Thus, after hundreds of millions of years of evolution of the respiratory proteins, mutations, i.e., N⟷N substitutions within the protomer, have not destroyed (and perhaps have perfected) their structure and function. On the other hand, the 120 residues on the surface or in its recesses (including all the P residues) show replacements of every type: N⟷P, N⟷N, P⟷P. Below we shall consider the essence of these facts.

It was of great interest to compare oxy- and reduced hemoglobin. For detailed study they chose reduced human hemoglobin,[79] whose primary structure had been established at about that time,[82] and whose crystals are more suitable for analysis than those of reduced horse hemoglobin (see Table III). (The results obtained have also been confirmed for reduced horse hemoglobin.[80]) They obtained three isomorphous derivatives, and calculated a Fourier electron-density map at a resolution of 5.5 Å. The structure of the β and β chains (Fig. 23), i.e., the tertiary structure, proved to be the same as for horse oxyhemoglobin within the limits of error of measurement. However, the quaternary structure of reduced hemoglobin (both of man and of horse) had undergone an important change. It consists in the following: the α subunits do not change their mutual arrangement, but the β subunits move apart by 7 Å, as if sliding on the surface of the α subunits, so that the α—β distance does not vary. One can best ascertain this from the change in the Fe_β—Fe_β interatomic distance (Fig. 25), which increases from 33.4 to 40.3 Å. Perutz descriptively called this remarkable effect in which the quaternary structure of hemoglobin changes upon reacting with O_2, the "breathing" of the molecule. In fact, however, the latter does not expand while "inhaling," or combining with oxygen, but while "exhaling" or releasing it. Conversely, it contracts while "inhaling." This has concretized the cause of the change in the crystal structure upon transition from oxyhemoglobin to reduced hemoglobin.[81] Reduced horse hemoglobin crystals

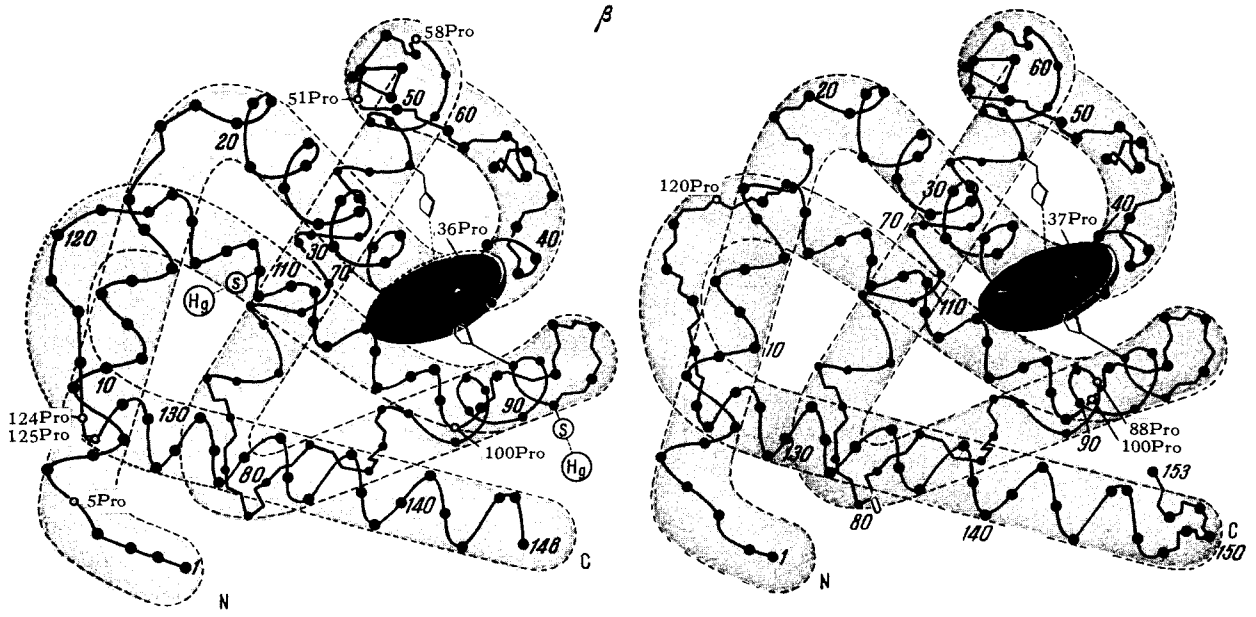


FIG. 24. Similarity of the tertiary structures of the β chain of hemoglobin (left) and the myoglobin molecule (right). The residues are numbered; the proline residues are often found at the breaks in the chain.

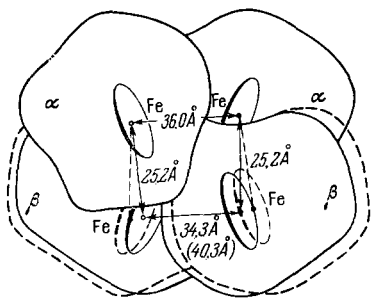


FIG. 25. Diagram of the change in the quaternary structure of hemoglobin during the oxidation-reduction process: "breathing" of the molecule (the dotted lines correspond to the reduced form).

were monitored by x-ray diffraction while transforming into crystals of the oxy- form. The appearance of new reflections made it possible to establish that this transition takes place in two stages. At first the process takes place at random, and the type of structure obtained is a disordered solid solution of oxy- molecules among the reduced ones. Ordering takes place at a certain stage.[80] Who could have supposed that these "metallurgical" notions could be applicable even to protein crystals? However, of course, the very fact of structural change in the molecule is of greatest interest to us. Apparently this effect is generally characteristic of enzymatic reactions performed by globular proteins.

The 5.5 Å resolution used in the described hemoglobin studies does not allow us to decide whether the change in the quaternary structure is due to any minor changes in the structures of the α and β subunits. Most likely, these changes are small and are a complex result of the shift in the equilibrium of electrostatic and other forces determining the arrangement of the subunits. This shift, perhaps, is induced by the binding (or release) of only one molecule of O_2 by the entire enormous protein molecule.

The studies of myoglobin and hemoglobin have permitted the drawing of another whole series of interesting conclusions.[78-80] "Molecular diseases" of the blood are known, involving the presence of certain abnormal hemoglobins in place of the ordinary form, certain residues being replaced by others. Thus, in the poorly oxidizable hemoglobin M, three residues are replaced, and these three residues are situated near the heme. We can conclude from this that they hinder contact of one of the Fe atoms with oxygen. The so-called sickle-cell anemia is due to the single replacement of residue No. 6 in the β chain (glutamic acid) by the neutral valine. Here it is impossible to explain the change in the functions of the molecule by purely geometric considerations, and it can apparently be understood only from the standpoint of the electronic structure. The decipherments of hemoglobin and myoglobin have been brilliant advances that have turned a new page in x-ray structure analysis and molecular biology. M. Perutz and J. Kendrew received the Nobel prize for these studies in 1962. Below we shall take up again the general conclusions drawn from these studies that concern the structure of proteins in general.

4. STUDIES OF LYSOZYME AND OTHER PROTEINS

Another protein has been added very recently to the set of two proteins that have been studied in detail and were described above. It is lysozyme, for which a Fourier map has been calculated at a resolution of 2 Å. Three-dimensional Fourier maps at resolutions of 3—6 Å have been made for chymotrypsinogen, ribonuclease, and carboxypeptidase, and projections for certain other proteins.

In distinction from the heme proteins, all the cited enzymes have S—S bridges and contain few α-helical regions. This makes for great difficulties in interpreting the maps at 4—6 Å resolution.

Figure 26 shows a three-dimensional Fourier map of chymotrypsinogen at 5 Å resolution according to Kraut.[83] Six types of crystals have been obtained under various conditions from this protein (of molecular weight 25,000, with 243 residues). Two of them proved suitable for introducing heavy atoms. They tested 200 different additives, four of which gave isomorphous crystals, permitting them to determine the phases and calculate the map. The ellipsoidal molecule (40×40×50 Å) proved to be a very complex interweaving of curvilinear rodlike regions of elevated electron density. Evidently, these regions correspond to polypeptide chains. However, one cannot assign their existing branchings unambiguously to S—S bridges or to the chain itself. Thus, they found it impossible to determine the path of the chain at this resolution. They could attribute only one relatively short linear rod to an α-helical region. There is a depression in the surface of the molecule, possibly corresponding to the active site.

The lack of information in this map, which had been constructed after enormous labor, was a somewhat dis-

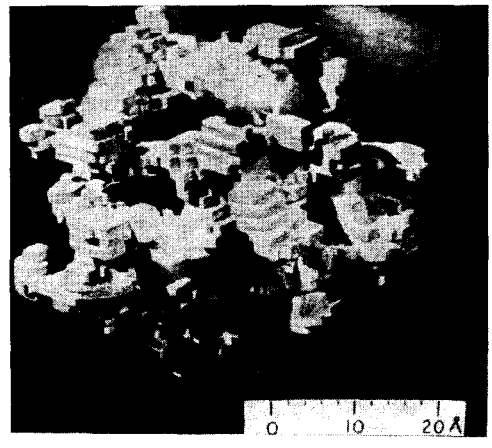
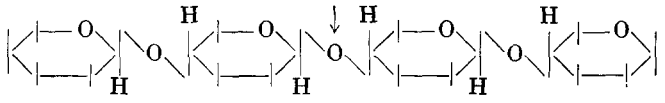


FIG. 26. Three-dimensional Fourier map of chymotrypsinogen at 5 Å resolution (superposition of sections).

FIG. 27. Three-dimensional model of the lysozyme molecule at 6 Å resolution.

appointing fact, especially after the models of myoglobin and hemoglobin derived at 6 Å resolution had furnished much more interpretable data. The fortunate circumstance proved to be the high α-helix content in the heme proteins (~ 70%, which was not previously known). This was manifested as linear rods, and the absence of S—S bridges permitted them unambiguously to determine the course of the chain and to correlate it with the primary structure.

Almost as little could be said from the Fourier maps of lysozyme[84,85] at 5 and 6 Å resolution as for chymotrypsinogen. These maps were obtained in 1962 by two groups of researchers, in England and the United States (Fig. 27). However, a recently completed study of this protein at 2 Å resolution by Blake, North, Phillips, and their associates[86] (1965) is of very great interest. This is because they have deciphered a molecule having completely different functions from the heme proteins. As it turned out, it differed considerably from the latter in structure. Lysozyme occurs in the tissues of many animals and plants, and has a protective function: it can dissolve certain bacteria. The mechanism of action of this enzyme consists in breaking the so-called β-glycosidic bonds of polysaccharides:



(28)

(the bond is indicated by the arrow), which occur in bacterial cell walls. The molecular weight of lysozyme is 14,600, and the sequence of the 129 residues in it is known. There are four S—S bridges (Fig. 28).

Egg-white lysozyme crystals are tetragonal, with space group $P4_32_12$, a = 79.1 Å, c = 37.9 Å, with eight molecules in the unit cell. Various heavy-atom additives were used at different stages in the study. Six of them proved effective at 2 Å resolution: $HgI_4^=$, $PdCl_2$, $UO_2F_5^=$, etc. The 9040 reflections were measured on the automatic linear diffractometer, and the phases were determined by the centroid method. Owing to the high symmetry, the amplitudes of the 1640 hk0, h0l, and hh$l$ reflections are real. That is, $\alpha = 0$ or π for them, thus facilitating the study.

The Fourier map was constructed in 60 sections perpendicular to the c axis. One block of ten sections is shown in Fig. 29. The map confirmed the correctness of the earlier 6 Å model (see Fig. 27), but permitted them to follow the course of the polypeptide chain exactly, to determine the α-helical regions and the S—S bridges, and to identify almost all the R groups, both in the helical and non-helical regions (although the atoms were not resolved). Figure 30 shows a model of the molecule (cf. Fig. 28).

The molecule can be described very approximately as a 45×30×30 Å ellipsoid. Fifty-five of the 129 residues occur in the α-helical conformation, there being six of such regions, some of them very short (see Figs. 28 and 30). The percent helix (~ 40%) is considerably

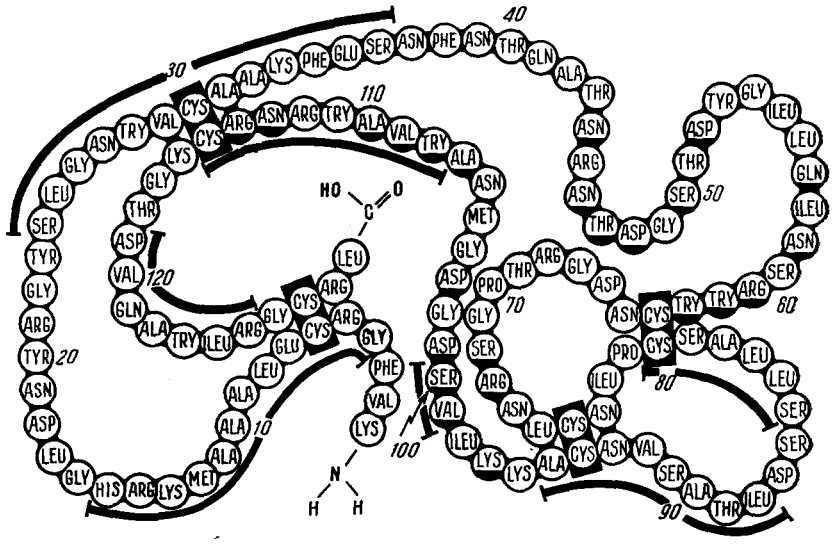


FIG. 28. Primary structure of lysozyme. The solid heavy line indicates the regions of the chain in the α conformation. The residues marked ○ are near the active site. The residues are numbered.

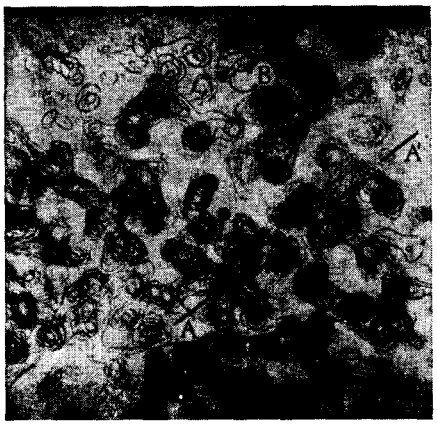


FIG. 29. A region of the three-dimensional Fourier map of lysozyme at 2 Å resolution. A – A': the axis of one of the α helix approximately perpendicular to the plane of the drawing.

less than in myoglobin. The nature of the tertiary structure of the rest of the molecule is very complex. In the region of residues 35–80 the chain is coiled very fancifully, so that three of its segments are approximately antiparallel and form something like a loop. Each disulfide bridge adjoins at least one α-helical region.

The complex nature of the molecule prevents one from distinguishing its "surface" and "interior" as clearly as for the heme proteins. If, nevertheless, one makes such a distinction, then the tendency to arrange the hydrophilic residues on the surface and the hydrophobic ones inside is not manifested so distinctly as in myoglobin.

An elongated depression of considerable size in the molecule is of especial interest, and is shown in Fig. 30 (it was visible even at 6 Å resolution). The residues bordering it belong to quite different parts of the chain,

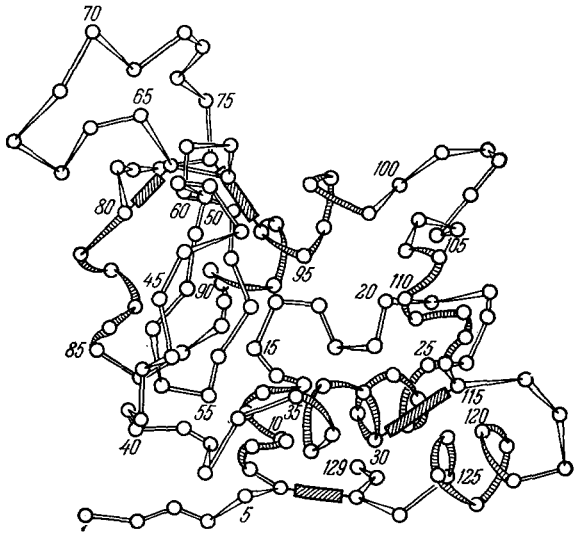


FIG. 30. A three-dimensional diagram of the arrangement of amino-acid residues (circles) in the lysozyme molecule. The residues are numbered to correspond to Fig. 28. The shaded cylinders are the S – S bridges. The α-helical regions are shaded.

and are marked in Fig. 28. This depression adjoins a very well constructed hydrophobic "pocket" made by residues 28, 108, and 111 (tryptophan), 105 (methionine), and 23 (tyrosine). To judge from all evidence including the chemical data, this depression plus the pocket is the active center of the molecule. A substrate molecule (28) entering it is subjected to the specific action.

After they have performed the enzymatic reaction, the protein molecules release the substrate. This hinders study of the labile enzyme-substrate complex. Still, there are substances similar in structure to the substrate, or inhibitors, which (like a misfitting key in a lock) "jam" the enzyme molecule and block its activity.

An elegant x-ray biochemical experiment[87] has been performed on lysozyme. It would seem to be a prototype for many such experiments in the future. Inhibitors (I) were selected of structure similar to (28), including the so-called N-acetylglucosamine, its dimer chitobiose, etc.

It had previously been established in studying certain myoglobin derivatives that if one makes a Fourier difference map from the amplitudes $||F_{P,M}| - |F_P||$ and the phases α_P of the pure protein, this map will give the electron density of the ligand molecule.[88] This method was used to determine the position of the inhibitor.

Crystals of inhibited lysozyme proved to be isomorphous with those of pure lysozyme, and x-ray patterns of them were taken. The map made from $||F_{P,M}| - |F_P||$ and α_P at 6 Å resolution gave an electron-density maximum at the position of the I molecule. It turned out (Fig. 31) that the inhibitor is actually situated in the above-described depression in the lysozyme molecule. However, of course, its exact position can be established only at higher resolution.

The decipherment of lysozyme was a new major advance in the x-ray crystallography of proteins. The overall status of the studies in this field is summarized in Table IV (see also the reviews [41,49,89,91]). We shall only make some commentaries on it.

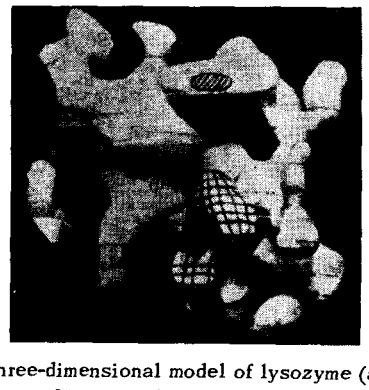


FIG. 31. Three-dimensional model of lysozyme (at 6 Å resolution). The regions of increased electron density revealed by difference synthesis are shaded. These regions correspond to the inhibitor molecule.

Table IV. X-ray studies of crystalline proteins*

Protein (+ source)	Molecular weight	Symmetry, space group	No. of molecules per cell	Status of investigation**	Researchers, references (country)***
Myoglobin (sperm whale)	17 800	Monocl., $P2_1$	2	3D, 1.4 Å (1963)	Kendrew, Watson[66-70] (England)
Myoglobin (seal)		Monocl., A2	2	2D, 2 Å (1963)	Scouloudi[73] (England)
Hemoglobin (oxy-, horse)	67 000	Monocl., C2	2	3D, 5.5 Å (1959)	Perutz[74-77] (England)
Hemoglobin (reduced, human)		Monocl., $P2_1$	2	3D, 5.5 Å (1963)	Muirhead, Perutz[79] (England)
Hemoglobin (reduced, horse)		Orthorh., $C222_1$	4	2D, 5.5 Å (1964)	Perutz[80,120] (England)
Hemoglobin (oxy-, bovine)		Orthorh., $P2_12_12_1$	4	2D, 6 Å (1960)	North[5], Green[91] (England)
Hemoglobin (lamprey)	18 400	Monocl., C2	8	Preliminary	Love (USA)
Ribonuclease (bovine)	13 683	Monocl., $P2_1$	2	3M, 3 Å (1965)	Harker, Kartha[43,92,93] (USA)
Ribonuclease (bovine)		Monocl.,	2	3D, 6 Å (1962)	Carlisle[14] (England)
Lysozyme (egg white)	14 400	Tetrag., $P4_32_12$	8	3D, 5 Å (1962)	Corey[84] (USA)
Lysozyme (egg white)		Tetrag.,	8	3D, 2 Å (1965)	Phillips, North[86] (England)
Lysozyme (inhibited)		Tetrag.,	8	3D, 6 Å (1965)	Phillips[87] (England)
Lysozyme (inhibited)		Tricl., P1	1	3D, 6 Å (1962)	Dickerson[95] (USA)
Chymotrypsinogen A (bovine)	25 000	Orthorh., $P2_12_12_1$	4	3D, 4 Å (1963)	Kraut[83] (USA)
α-Chymotrypsin (bovine)	25 000	Monocl., $P2_1$	4	Preliminary	Blow[96] (England)
γ-Chymotrypsin (bovine)	25 000	Tetrag., $P4_22_12_1$	8	Preliminary	Davis[97] (USA)
Insulin (porcine)	5 733	Rhombohed., R3	12	3D Calculations	Hodgkin[98] (England)
Insulin (porcine)		Monocl., $P2_1$	12	3D Calculations	Hodgkin[99] (England)
Insulin (bovine)		Orthorh., $P2_12_12_1$	16	3D Calculations	Low[100,101] (USA)
Carboxypeptidase (bovine)	34 300	Monocl., $P2_1$	2	3D, 5 Å (1965)	Lipscomb[102] (USA)
Lactoglobulin (cow's milk)	35 000	Orthorh., $B22_12$	4	2D, 4 Å (1963)	Green[103] (England)
Papain C (papaya juice	22 000	Orthorh., $P2_12_12_1$	4	2D, 5 Å (1962)	Drenth[104] (Holland)
Carbonic anhydrase C (human)	30 000	Monocl., $P2_1$	2	2D, 5.5 Å (1965)	Strandberg[105] (Sweden)
Alcohol dehydrogenase (horse)	83 000	Monocl., $P2_1$	2	Preliminary	Branden (Sweden)
Lactic dehydrogenase	140 000	Tetrag., I42	4	Preliminary	Rossmann (USA)
Glucagon (ovine)	3 483	Cubic $P2_13$	12	Preliminary	King[106] (USA)
Pepsin (bovine)	34 500	Monocl., $P2_1$	4	Preliminary	Andreeva [91] (USSR)
Trypsin (bovine)	24 000	Orthorh., $P2_12_12_1$	4	Preliminary	Vainshtein (USSR)
Cytochrome C (horse)	12 000	–	–	Preliminary	Dickerson (USA)
Glyceraldehyde phosphate dehydrogenase (crab)	140 000	Orthorh., $P2_12_12_1$	4	Preliminary	Watson[122] (England)
Tobacco mosaic virus protein	17 000	Helical	–	Calculations	Holmes[108] (England)
Ferritin (horse)	747 000	Cubic	432	Study of quaternary structure	Harrison[107]

*Mainly from the data of the review[90]. The table omits a number of data involving study of certain modifications of various proteins, unit cells, etc. (see the review[41]).

**3D and 2D respectively indicate three- and two-dimensional Fourier syntheses; the resolution follows.

***The directors of the research groups are named.

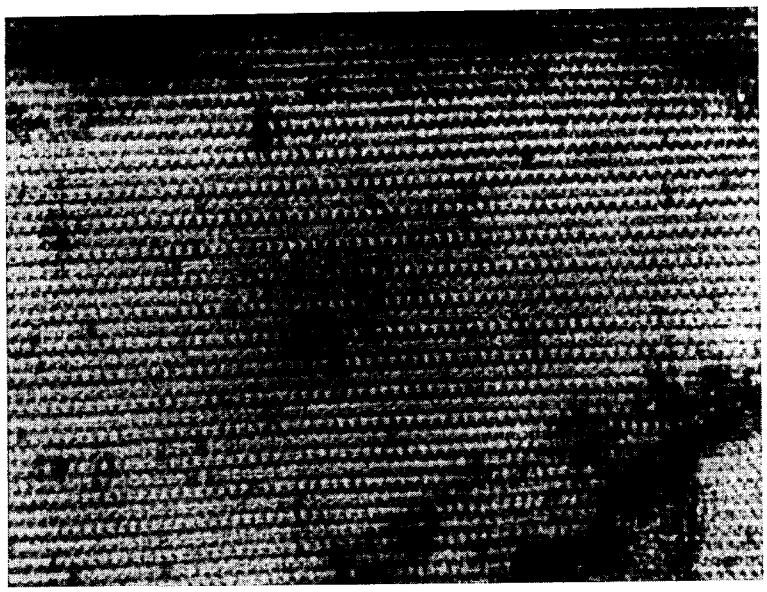


FIG. 32. Electron micrograph of catalase. Magnification 250,000 ×.

Fifteen years of studies of ribonuclease will apparently soon be completed by Harker and his associates (United States).[92,93] This is an extremely interesting protein whose enzymatic function is to break the chain of RNA (ribonucleic acid) into individual nucleotides. The molecular weight of this protein is 13,683, with 124 residues and four S—S bridges. Harker and his associates have obtained 15 types of crystals. In one of these, they could introduce five substituents, mostly organic dyes containing heavy atoms. A map at 3 Å resolution was obtained in 1964. Carlisle's group[94] working in parallel in England has been able to obtain only one derivative, and the map obtained was not reliable.

Studies of insulin (4) have been conducted for more than 20 years. This is a protein controlling sugar exchange[98,99] (molecular weight 5733). Two groups are working on it, in England (on hexagonal zinc insulin) and in the United States[100,101] (on orthorhombic insulin citrate and sulfate). The proposed models have not been confirmed by direct construction of Fourier maps.

Besides chymotrypsinogen[83] and carboxypeptidase,[102] which were cited above, encouraging results have been obtained for papain,[104] carbonic anhydrase,[105] chymotrypsin,[96] and lactoglobulin.[103] In the Soviet Union, studies of crystalline proteins have been begun in only two laboratories, a clearly inadequate number.

In closing this section, it is expedient to say a few words on the possibilities of studying protein structures by the electron-microscope method.[109-111] The 4—5 Å resolution of the best contemporary instruments formally corresponds already to the low resolutions that have permitted drawing conclusions on molecular structures. The fundamental difficulties here are methodological and concern sample preparation. The protein must crystallize in thin layers, and further, contrast must be created in the specimen by introducing substances that strongly scatter electrons (e.g., PTA = phosphotungstic acid). One of the few proteins yielding electron micrographs by the stated methods is catalase[110,111] (molecular weight 250,000, Fig. 32). The picture reveals a perfectly regular structure. However, since the picture is two-dimensional, it is hard to distinguish the images of individual molecules and draw justified conclusions on their three-dimensional structure (even their quaternary structure). Nevertheless, this approach is not without promise, and we can expect that electron microscopy of protein crystals will give interesting results in the near future, especially if we bear in mind the relative simplicity of its methodology as compared with the colossal laboriousness of x-ray diffraction analysis.

5. FACTORS DETERMINING THE STRUCTURE OF PROTEIN MOLECULES. STRUCTURE AND BIOLOGICAL ACTIVITY

Protein synthesis occurs in the ribosomal particles of the cell (Fig. 33).[112] The genetic information contained in the nucleotide sequence of DNA (deoxyribonucleic acid) is "transcribed" onto the chain of iRNA (informational RNA). The iRNA passes through the ribosome, to which amino acids are continuously supplied (by means of another RNA, the so-called transport RNA). In the genetic code, each triplet of nucleotides of the iRNA corresponds to a definite amino acid. Thus, according to the directions on the iRNA chain, the ribosome "jots down" the polypeptide chain, attaching a link at a time to it, and extruding it along with the already finished portion of the chain. Consequently, the first stage in the biosynthesis of a protein fixes its primary structure and sequence of R groups.

How, then, is the spatial structure of the globular particle formed? As we have seen, the polypeptide chains in it are folded in a very complex fashion, which

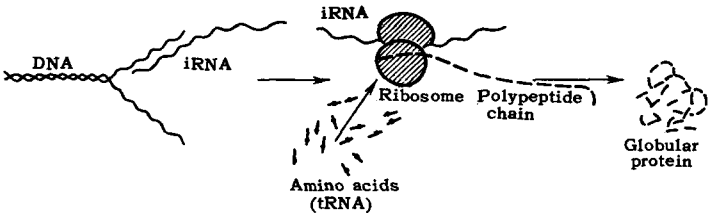


FIG. 33. Simplified diagram of protein biosynthesis.

is unique for a given protein, such that part of them are "inside" the molecule, and the rest on the "surface," forming a complex relief of radicals and functional groups. No special devices have been found in cells resembling a three-dimensional "mold" that could form the secondary and tertiary structure. A number of data indicate that this structure is formed by the principle of self-organization. That is, a given chain under given conditions can fold itself only into a definite structure, and no other. Filament—helix—disordered globule transitions are well known in the physical chemistry of polymers.[113] However, in the protein case we have to explain specifically the uniqueness and definiteness of the resultant structure of the globule.

This definiteness must be dictated by the individuality of the R groups and their sequence. For example, we know that proline residues cannot adopt the α conformation, and thus their position in the chain limits the possible length and arrangement of the α-helical regions. Indeed, all the proline residues in myoglobin and hemoglobin are arranged either at the breaks in the α helices or in the non-helical regions (although there are breaks and non-helical regions not containing these residues as well[72,78,121]). In most cases, the cysteine residues $-CH_2SH$ reconnect to form disulfide bridges S—S of cystine (biosynthesis provides only cysteine). Then (see (4) and Fig. 28), this reconnection directly limits the possible three-dimensional conformations.

Studies of synthetic polypeptides consisting of only one type of R group have shown that polypeptides made of R = Ser, Thr, Val, Ile, or Cys do not form α helices, and the stated residues have been termed antihelical. However, a statistical analysis of the distribution of R groups in myoglobin and hemoglobin[120,121] has shown that, although some breaks almost completely consist of these residues, they are also encountered no less often in the α helices. They also found that His, Glu, and Asp always occur in the breaks.

Hypotheses[114] were advanced long ago that the organizing factor in the formation of globular molecules could be the tendency to collect the non-polar (hydrophobic) groups inside the protein globule, and the polar (hydrophilic) groups on the surface.[115-117] This rule was actually found to hold in myoglobin and hemoglobin. By binding to the polar R groups, the water molecules form an electrostatic double layer around the globule. This shields the field of these R groups and decreases the free energy. When hydrophobic R groups contact H_2O molecules, the latter become ordered, so as to decrease the entropy. Hence these R groups find it "advantageous" to collect inside the molecule and avoid contact with water molecules. All this leads to a minimum free energy and an increase in the entropy of the globule-water system, i.e., it stabilizes it. At the same time, the van der Waals interaction of the non-polar R groups is enough to explain the attraction and compact packing of the chains within the globule.[72,78]

We note that if we adopt the hypothesis that the hydrophobic groups are located within the molecules, we can understand the tendency toward formation of subunits in proteins of high molecular weight.[118] The ratio of the numbers of residues of the two types in proteins varies within relatively narrow limits. This means also that the "surface-to-volume" ratio can be maintained only for a certain volume of the globule. If we exceed it, the hydrophilic R groups won't have enough room on the surface. Then proteins of high molecular weight "have no further recourse" than to form subunits and thereby increase the proportion of surface.

The hypothesis that the primary structure determines the secondary and tertiary structures is confirmed in some interesting experiments on the so-called reversible denaturation.[2,116,117] All of the S—S bonds in the protein can be broken by certain agents (this has been done with ribonuclease, trypsin, and lysozyme). Then the protein loses its enzymatic properties, and according to all data its three-dimensional structure collapses (is denatured), and the molecule swims in the solution as a random coil. However, its primary structure or sequence of R groups is preserved. Now one can reconstitute the S—S bonds by other reagents. It turns out that this leads to restoration of the enzymatic activity and all the properties of the globular protein, even its crystal structure. Hence, the spatial structure of the molecules has been completely restored.

We have briefly listed the facts favoring the hypothesis of "self-organization" based on assignment of the primary structure. However, there is another group of facts that are not very understandable from this viewpoint, or that contradict it. For example, reversible-denaturation experiments have not succeeded on insulin (4). One cannot understand what determines the correctness of pairing of the S—S bridges between a given pair of SH groups, rather than another pair, whereas there are many more possibilities of pairing

when there are, e.g., eight SH groups (as in lysozyme and ribonuclease). The tendency to show a hydrophilic surface is poorly marked in lysozyme (Sec. 5). Finally, it has remained a puzzle until recently why the tertiary structures of the heme proteins are fundamentally identical in spite of the very great differences in their primary structures.

However, all these facts are gradually being explained, and it is becoming obvious that the principle of self-organization under definite conditions of the medium must be made the basis of understanding the ways in which the three-dimensional structures of globular protein molecules and of biologic globular structures in general are formed from the chain molecules of biopolymers. We might suppose that the three-dimensional globular structure arising directly after the entire polypeptide chain emerging from the ribosome has been completed is that which is thermodynamically most favorable and stablest under the given conditions. Also we may consider that its stability and the possibility of formation of the given unique three-dimensional conformation are the result of the entire process of evolution of the given protein. Analysis of myoglobin and hemoglobin shows [120,121] that what determines the tertiary structure might not be the primary structure as such, but certain of its invariants. That is, a small fraction of the R groups remain fixed at definite sites along the chain, while only the nonpolarity (N) of the other R groups is fixed, without requiring that they be strictly individualized. On the other hand, for most enzymes (e.g., lysozyme and ribonuclease), the three-dimensional construction is less standardized (they contain little α helix), but is maintained by a more complex equilibrium of forces, the hydrophobic interactions apparently playing a relatively minor role. Hence, their possibilities for variation in primary structure must be considerably less.

In individual stages that proceed with difficulty, the self-organization principle can be supplemented by the intervention of special agents. Thus, for example, there are data indicating that rapid rearrangement of the S—S bonds is carried out by a special enzyme.

The fundamentals of structure of globular proteins that have been elucidated by biochemical and x-ray diffraction studies have raised a set of new problems. One of them is the very large fraction of the bulk of protein molecules (i.e., of the total number of residues) that takes no visible part in the function of the active site, which is formed by a relatively small number of residues. It is not clear whether this bulk is only a necessary massive foundation for the delicate enzymatic mechanism built on it, or facilitates the action of this mechanism, or is left in part as a rudiment of the evolutionary process. Various facts support the one or the other explanation. Thus, one can remove a considerable part of the chain from the molecules of certain proteins practically without changing their enzymatic activity. [1,2] However, at the same time we have seen from the example of hemoglobin the sensitivity of the entire structure of the vast molecule to combination with only one molecule of O_2. The importance of the molecular structure as a whole, in any case for certain proteins, is especially sharply evident from the example of a recently discovered phenomenon, the so-called allosteric inhibition. [119] Here the enzymatic activity is blocked by combination with a special inhibitor, not at the active site, but at a completely different site on the molecule. [2] Only when one removes this "protector" can one start the enzymatic reaction. This serves to regulate these processes. Apparently, the allosteric reaction involves reorientation of subunits or some other type of change in the quaternary or perhaps tertiary structure.

In some way or another, each molecule of a given protein is a very complex spatial mechanism, uniquely constructed especially to carry out a given specific function. The individual principles of its atomic architecture are understood, and many features of its reactions are explained by the chemical nature of the individual radicals and their combinations, and electronic-structure concepts are adduced to explain individual facts. However, a unified approach to describing the structure and properties of these molecules has not been worked out yet. A protein molecule is an ordered and condensed atomic microsystem that is stable under very strict conditions, and that can assume a number of states depending on contact with other molecules. The order of the day calls for examining it from the standpoint of solid state physics, not with periodicity as in a crystal, but with a certain internal one-dimensional pseudoperiodicity of the polypeptide chain.

[1] M. Dixon and E. C. Webb, Enzymes, Academic Press, New York, 1958; Russ. Transl., IL, 1961.

[2] Fermenty (Enzymes), collected volume, Ed. A. E. Braunshteĭn, Nauka, 1964.

[3] J. P. Greenstein and M. Winitz, Chemistry of the Amino Acids, Wiley, New York, 1961; Russ. Transl., Mir, 1965.

[4] Ryle, Sanger, Smith, and Kitai, Biochem. J. 60, 541 (1955).

[5] R. B. Corey and L. Pauling, Proc. Roy. Soc. B141, 10 (1953).

[6] R. B. Corey, in collected volume Khimiya belka (Protein Chemistry), M., IL, 1949, p. 299. [Probably: Advances in Protein Chemistry 4, 385 (1948).]

[7] L. Pauling, R. B. Corey, and H. R. Branson, Proc. Natl. Acad. Sci. USA 37, 205 (1951); L. Pauling and R. B. Corey, ibid, 236.

[8] G. V. Gurskaya, Kristallografiya 9, 839 (1964); Soviet Phys. Crystallography 9, 709 (1965).

[9] W. T. Astbury, Proc. Roy. Soc. B141, 1 (1953).

[10] W. T. Astbury, Trans. Faraday Soc. 34, 378 (1938).

[11] C. H. Bamford, A. Elliot, and W. E. Hanby, Synthetic Polypeptides, Academic Press, New York, 1956.

[12] L. I. Tatarinova and B. K. Vaĭnshteĭn, Vysokomolekulyarnye soedineniya **4**, 270 (1962).[sic!]
[13] Cochran, Crick, and Vand, Acta Cryst. **5**, 581 (1952).
[14] Klug, Crick, and Wyckoff, Acta Cryst. **11**, 199 (1958).
[15] B. K. Vaĭnshteĭn, Difraktsiya rentgenovskikh lucheĭ na tsepnykh molekulakh (Diffraction of X-rays by Chain Molecules), AN SSSR, 1963; Engl. Transl., American Elsevier, New York, 1966.
[16] J. C. Kendrew and R. G. Parrish, Proc. Roy. Soc. **A238**, 305 (1956).
[17] J. Boyes-Watson, E. Davidson, and M. F. Perutz, Proc. Roy. Soc. **A191**, 83 (1947).
[18] W. L. Bragg and M. F. Perutz, Acta Cryst. **5**, 277, 323 (1952).
[19] J. D. Bernal and D. Crowfoot, Nature **133**, 794 (1934).
[20] D. Crowfoot, Nature **135**, 591 (1935).
[21] R. W. James, The Optical Principles of the Diffraction of X-rays, G. Bell, London (1948); Russ. Transl. IL, 1950.
[22] M. T. F. von Laue, Röntgenstrahl-Interferenzen (X-ray Interference), Akademische Verlagsgesellschaft, Frankfurt am Main, 1960.
[23] H. Lipson and W. Cochran, The Determination of Crystal Structures, G. Bell, London (1953); Russ. Transl., IL, 1956.
[24] M. A. Poraĭ-Koshits, Prakticheskiĭ kurs rentgenostrukturnogo analiza (Practical Course in X-ray Structure Analysis), MGU, 1960.
[25] D. C. Hodgkin, Nature **188**, 441 (1960).
[26] V. Luzzati, Acta Cryst. **5**, 802 (1952).
[27] F. A. Brusentsev and V. F. Dvoryankin, Zhur. Strukt. Khim. **4**, 465 (1963).
[28] A. K. Bhuiya and E. Stanley, Acta Cryst. **17**, 746 (1964).
[29] B. K. Vaĭnshteĭn, I. M. Gel'fand, R. L. Kayushina, and Yu. G. Fedorov, DAN SSSR **153**, 11 (1963); Soviet Phys. Doklady **8**, 1040 (1964).
[30] J. Karle, in Advances in Structure Research by Diffraction Methods, Ed. R. Brill, Vol. I, p. 55, Interscience, New York, 1964.
[31] A. I. Kitaĭgorodskiĭ, Teoriya strukturnogo analiza (The Theory of Crystal Structure Analysis), AN SSSR, 1957; Engl. Transl., Consultants Bureau, New York, 1961.
[32] I. M. Rumanova, DAN SSSR **98**, 399 (1954); **118**, 84 (1958), Soviet Phys. Doklady **3**, 19 (1958).
[33] C. L. Coulter, J. Mol. Biol. **12**, 292 (1965).
[34] A. L. Patterson, Z. Krist. **90**, 517 (1935).
[35] M. J. Buerger, Acta Cryst. **3**, 87 (1950).
[36] M. J. Buerger, Vector Space and its Application in Crystal-Structure Investigation, Wiley, New York, 1959; Russ. Transl., IL, 1961.
[37] B. K. Vaĭnshteĭn, DAN SSSR **78**, 1137 (1951).
[38] A. I. Kitaĭgorodskiĭ, UFN **46**, 23 (1952).
[39] V. I. Simonov, DAN SSSR **136**, 813 (1961), Soviet Phys. Doklady **6**, 98 (1961).
[40] L. Bragg and M. F. Perutz, Proc. Roy. Soc. **A213**, 425 (1952); M. F. Perutz, ibid. **A225**, 264 (1954).
[41] B. W. Low, in The Proteins, Ed. H. Neurath and K. Bailey, Academic Press, New York, 1953, Vol. I, Part A, p. 235; Russ. Transl., M., IL, 1956, Vol. II, Chapter I.
[42] D. W. Green, V. M. Ingram, and M. F. Perutz, Proc. Roy. Soc. **A225**, 287 (1954).
[43] M. V. King, B. S. Magdoff, Adelman, and D. Harker, Acta Cryst. **9**, 460 (1956).
[44] F. H. C. Crick and B. S. Magdoff, ibid. **9**, 901 (1956).
[45] B. K. Vaĭnshteĭn, JETP **27**, 44 (1954).
[46] W. L. Bragg, Acta Cryst. **11**, 70 (1958).
[47] D. M. Blow, Proc. Roy. Soc. **A247**, 302 (1958).
[48] D. C. Phillips, in Advances in Protein Crystallography (in press).
[49] G. N. Ramachandran and S. Raman, Acta Cryst. **12**, 957 (1959).
[50] M. F. Perutz, Acta Cryst. **9**, 867 (1956).
[51] M. G. Rossmann, Acta Cryst. **13**, 221 (1960).
[52] L. K. Steinrauf, Acta Cryst. **16**, 317 (1963).
[53] D. Harker, Acta Cryst. **9**, 1 (1953).
[54] D. M. Blow and F. H. C. Crick, Acta Cryst. **12**, 794 (1959).
[55] R. E. Dickerson, J. C. Kendrew, and R. E. Strandberg, Acta Cryst. **14**, 1188 (1961).
[56] D. M. Blow and M. G. Rossmann, Acta Cryst. **14**, 1195 (1961).
[57] C. Bokhoven, J. C. Schoone, and G. M. Bijvoet, Acta Cryst. **4**, 275 (1951).
[58] G. Kartha and G. N. Ramachandran, Acta Cryst. **8**, 195 (1955).
[59] G. N. Ramachandran and R. R. Ayyar, in Crystallography and Crystal Perfection, Ed. G. N. Ramachandran, Academic Press, New York, 1964, p. 25.
[60] M. G. Rossmann, Acta Cryst. **14**, 383 (1961).
[61] A. C. T. North, Acta Cryst. **18**, 212 (1965).
[62] Y. Okaya, Y. Saito, and R. Pepinsky, Phys. Rev. **98**, 1857 (1955).
[63] S. Raman, Proc. Ind. Acad. Sci. V. L. N 2, Sec. A, 95 (1959).
[64] M. G. Rossmann and D. M. Blow, Acta Cryst. **15**, 24 (1962); **16**, 39 (1963); **17**, 1474 (1964).
[65] M. Rossmann, D. Blow, M. M. Harding, and E. Coller, Acta Cryst. **17**, 338 (1964).
[66] M. M. Blum, G. Bodo, H. M. Dintzis, and J. C. Kendrew, Proc. Roy. Soc. **A246**, 369 (1958).
[67] G. Bodo, H. M. Dintzis, J. C. Kendrew, and H. W. Wyckoff, Proc. Roy. Soc. **A253**, 70 (1959).
[68] J. C. Kendrew, Dickerson, Strandberg, Hart, Davies, D. C. Phillips, and Shore, Nature **185**, 422 (1960).
[69] J. C. Kendrew, M. C. Watson, Strandberg, R. E. Dickerson, D. C. Phillips, and Shore, Nature **190**, 666 (1961).
[70] J. C. Kendrew, Biofizika **8**, 273 (1963).
[71] A. B. Edmundson and C. H. W. Hirs, Nature **190**, 663 (1961).

[72] J. C. Kendrew, Brookhaven Symp. in Biology, No. 15 (1962).
[73] H. Scouloudi, Proc. Roy. Soc. **A258**, 181 (1960).
[74] L. Bragg and M. F. Perutz, Proc. Roy. Soc. **A225**, 315 (1954).
[75] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, Will, and A. C. T. North, Nature **185**, 416 (1960).
[76] A. F. Cullis, H. Muirhead, M. F. Perutz, and M. G. Rossmann, Proc. Roy. Soc. **A265**, 15, 161 (1962).
[77] M. F. Perutz, Sci. Amer. **211**, No. 5, 64 (1964).
[78] M. F. Perutz, Proteins and Nucleic Acids, Elsevier, New York, 1962.
[79] H. Muirhead and M. F. Perutz, Nature **199**, 633 (1963).
[80] M. F. Perutz, Bolton, Diamond, H. Muirhead, and M. C. Watson, Nature **203**, 687 (1964).
[81] F. Haurowitz, Z. physiol. Chem. **254**, 266 (1938).
[82] G. Braunitzer, Gehring-Müller, Hilschmann, Hilse, Hobom, Rudloff, and Wittmann-Liebold, ibid. **325**, 283 (1961).
[83] J. Kraut, Sieker, High, and Freer, Proc. Natl. Acad. Sci. USA **48**, 1417 (1962).
[84] R. H. Stanford, R. E. Marsh, and R. B. Corey, Nature **196**, 1176 (1962).
[85] C. C. Blake, R. Fenn, A. C. T. North, D. C. Phillips, and R. J. Poljak, Nature **196**, 1173 (1962).
[86] C. C. Blake, Koenig, Mair, A. C. T. North, D. C. Phillips, and Sarma, Nature **206**, 757 (1965).
[87] L. N. Johnson and D. C. Phillips, Nature **206**, 761 (1965).
[88] L. Stryer, J. C. Kendrew, and M. C. Watson, J. Mol. Biol. **8**, 96 (1964); Watson, Kendrew, and Stryer, ibid. 166.
[89] A. Rich and D. W. Green, Ann. Rev. Biochem. **30**, 93 (1961).
[90] R. E. Dickerson, The Proteins, Ed. H. Neurath, Academic Press, New York, 1964, Vol. 2, p. 603.
[91] N. S. Andreeva, Usp. sovr. biol. **58**, 3 (1964).
[92] M. V. King, J. Bello, Pignataro, and D. Harker, Acta Cryst. **15**, 144 (1962).
[93] G. Kartha, J. Bello, D. Harker, and DeJarnette, in Aspects of Protein Structure, Ed. G. N. Ramachandran, Academic Press, New York, 1964, p. 13.
[94] C. H. Carlisle and R. A. Palmer, Acta Cryst. **15**, 129 (1962).
[95] R. E. Dickerson, Reddy, Pinkerton, and Steinrauf, Nature **196**, 1178 (1962).
[96] D. M. Blow, M. G. Rossmann, and B. A. Jeffrey, J. Mol. Biol. **8**, 65 (1964).
[97] P. B. Sigler, Skinner, Coulter, Kallos, Braxton, and D. R. Davies, Proc. Natl. Acad. Sci. USA **51**, 1146 (1964).
[98] D. Crowfoot, Proc. Roy. Soc. **A164**, 580 (1938).
[99] M. Adam, D. C. Hodgkin, et al. Abstracts of Papers, 7th Congress of the International Union of Crystallographers, Moscow, 1966.
[100] B. Low and C. B. Shoemaker, Acta Cryst. **12**, 893 (1959); Shoemaker, Einstein, and Low, ibid. **14**, 459 (1961).
[101] A. S. McGavin, J. R. Einstein, and B. W. Low, Proc. Natl. Acad. Sci. U.S.A. **48**, 2150 (1962).
[102] Hartsuck, Ludwig, Muirhead, Steitz, and W. L. Lipscomb, ibid. **53**, 396 (1965).
[103] D. W. Green, Aschaffenburg, Coppola, Simmons, and Dunnil, Acta Cryst. **16**, A7.4 (1963).
[104] J. Drenth, Jansonius, Koekoek, Marrink, Munnik, and Wolthers, J. Mol. Biol. **5**, 398 (1962).
[105] B. Tilander, B. Strandberg, and K. Fridborg, J. Mol. Biol. **12**, 740 (1965).
[106] M. V. King, J. Mol. Biol. **1**, 375 (1959).
[107] P. M. Harrison, J. Mol. Biol. **1**, 69 (1959); **6**, 404 (1963).
[108] K. C. Holmes and R. Leberman, J. Mol. Biol. **6**, 439 (1963).
[109] H. Fernandez-Moran, Reed, Koike, and Willms, Science **145**, 930 (1964).
[110] R. C. Valentine, Nature **204**, 1262 (1964).
[111] B. K. Vaĭnshteĭn, N. A. Kiselev, and V. L. Shpitsberg, DAN SSSR **167**, No. 1 (1966).
[112] V. M. Ingram, The Biosynthesis of Macromolecules, Benjamin, New York, 1965.
[113] T. M. Birshteĭn and O. B. Ptitsyn, Konformatsii makromolekul (Conformations of Macromolecules) Nauka, 1964, Engl. transl., Wiley, New York, 1966.
[114] S. E. Bresler and D. L. Talmud, DAN SSSR **43**, 310 (1944).
[115] V. A. Belitser, Usp. sovr. biol. **50**, 3 (1960).
[116] W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[117] S. J. Singer, ibid. **17**, 1 (1962).
[118] H. F. Fisher, Proc. Natl. Acad. Sci. U.S.A. **51**, 1285 (1964).
[119] J. Monod, J. P. Changeux, and F. Jacob, J. Mol. Biol. **6**, 306 (1963).
[120] M. F. Perutz, J. Mol. Biol. **13**, 646 (1965).
[121] M. F. Perutz, J. C. Kendrew, and H. C. Watson, J. Mol. Biol. **13**, 669 (1965).
[122] H. C. Watson and L. I. Banaszak, Nature **204**, 918 (1964).

Translated by M. V. King