708

[Reprinted from the Journal of the American Chemical Society, 61, 1860 (1939).]

The Structure of Proteins

By Linus Pauling and Carl Niemann

[Reprinted from the Journal of the American Chemical Society, **61**, 1860 (1939).]

[CONTRIBUTION FROM THE GATES AND CRELLIN LABORATORIES OF CHEMISTRY, CALIFORNIA INSTITUTE OF TECHNOLOGY, No. 708]

The Structure of Proteins

BY LINUS PAULING AND CARL NIEMANN

1. Introduction

It is our opinion that the polypeptide chain structure of proteins,[1] with hydrogen bonds and other interatomic forces (weaker than those corresponding to covalent bond formation) acting between polypeptide chains, parts of chains, and side-chains, is compatible not only with the chemical and physical properties of proteins but also with the detailed information about molecular structure in general which has been provided by the experimental and theoretical researches of the last decade. Some of the evidence substantiating this opinion is mentioned in Section 6 of this paper.

Some time ago the alternative suggestion was made by Frank[2] that hexagonal rings occur in proteins, resulting from the transfer of hydrogen atoms from secondary amino to carbonyl groups with the formation of carbon–nitrogen single bonds. This *cyclol hypothesis* has been developed extensively by Wrinch,[3] who has considered the geometry of cyclol molecules and has given discussions of the qualitative correlations of the hypothesis and the known properties of proteins.

It has been recognized by workers in the field of modern structural chemistry that the lack of conformity of the cyclol structures with the rules found to hold for simple molecules makes it very improbable that any protein molecules contain structural elements of the cyclol type. Until recently no evidence worthy of consideration had been adduced in favor of the cyclol hypothesis. Now, however, there has been published[4] an interpretation of Crowfoot's valuable X-ray data on crystalline insulin[5] which is considered by the authors to provide proof[6] that the insulin molecule actually has the structure of the space-enclosing cyclol C_2. Because of the great and widespread interest in the question of the structure of proteins, it is important that this claim that insulin has been proved to have the cyclol structure be investigated thoroughly. We have carefully examined the X-ray arguments and other arguments which have been advanced in support of the cyclol hypothesis, and have reached the conclusions that there exists no evidence whatever in support of this hypothesis and that instead strong evidence can be advanced in support of the contention that bonds of the cyclol type do not occur at all in any protein. A detailed discussion of the more important pro-cyclol and anti-cyclol arguments is given in the following paragraphs.

2. X-Ray Evidence Regarding Protein Structure

It has not yet been possible to make a complete determination with X-rays of the positions of the atoms in any protein crystal; and the great complexity of proteins makes it unlikely that a complete structure determination for a protein will ever be made by X-ray methods alone.[7] Nevertheless the X-ray studies of silk fibroin by Herzog and Jancke,[8] Brill,[9] and Meyer and Mark[10] and of β-keratin and certain other proteins by Astbury and his collaborators[11] have provided strong (but

(1) E. Fischer, "Untersuchungen über Aminosäuren, Polypeptide und Protein," J. Springer, Berlin, 1906 and 1923.

(2) F. C. Frank, *Nature*, **138**, 242 (1936); this idea was first proposed by Frank in 1933: see W. T. Astbury, *J. Textile Inst.*, **27**, 282 (1936).

(3) D. M. Wrinch, (a) *Nature*, **137**, 411 (1936); (b) **138**, 241 (1936); (c) **139**, 651, 972 (1937); (d) *Proc. Roy. Soc.* (London), **A160**, 59 (1937); (e) **A161**, 505 (1937); (f) *Trans. Faraday Soc.*, **33**, 1368 (1937); (g) *Phil. Mag.*, **26**, 313 (1938); (h) *Nature*, **143**, 482 (1939); etc.

(4) (a) D. M. Wrinch, *Science*, **88**, 148 (1938); (b) THIS JOURNAL, **60**, 2005 (1938); (c) D. M. Wrinch and I. Langmuir, *ibid.*, **60**, 2247 (1938); (d) I. Langmuir and D. M. Wrinch, *Nature*, **142**, 581 (1938).

(5) D. Crowfoot, *Proc. Roy. Soc.* (London), **A164**, 580 (1938).

(6) In ref. 4d, for example, the authors write "The superposability of these two sets of points represented the first stage in the proof of the correctness of the C_2 structure proposed for insulin. ... These investigations, showing that it is possible to deduce that the insulin molecule is a polyhedral cage structure of the shape and size predicted, give some indication of the powerful weapon which the geometrical method puts at our disposal."

(7) A protein molecule, containing hundreds of amino acid residues, is immensely more complicated than a molecule of an amino acid or of diketopiperazine. Yet despite attacks by numerous investigators no complete structure determination for any amino acid had been made until within the last year, when Albrecht and Corey succeeded, by use of the Patterson method, in accurately locating the atoms in crystalline glycine [G. A. Albrecht and R. B. Corey, THIS JOURNAL, **61**, 1087 (1939)]. The only other crystal with a close structural relation to proteins for which a complete structure determination has been made is diketopiperazine [R. B. Corey, *ibid.*, **60**, 1598 (1938)]. The investigation of the structure of crystals of relatively simple substances related to proteins is being continued in these Laboratories.

(8) R. O. Herzog and W. Jancke, *Ber.*, **53**, 2162 (1920).

(9) R. Brill, *Ann.*, **434**, 204 (1923).

(10) K. H. Meyer and H. Mark, *Ber.*, **61**, 1932 (1928).

(11) W. T. Astbury, *J Soc. Chem. Ind.*, **49**, 441 (1930); W. T. Astbury and A. Street, *Phil. Trans. Roy. Soc.*, **A230**, 75 (1931); W. T. Astbury and H. J. Woods, *ibid.*, **A232**, 333 (1933); etc.

not rigorous) evidence that these fibrous proteins contain polypeptide chains in the extended configuration. This evidence has been strengthened by the fact that the observed identity distances correspond closely to those calculated with the covalent bond lengths, bond angles, and N-H $\cdots$ O hydrogen bond lengths found by Corey in diketopiperazine.

The X-ray work of Astbury also provides evidence that α-keratin and certain other fibrous proteins contain polypeptide chains with a folded rather than an extended configuration. The X-ray data have not led to the determination of the atomic arrangement, however, and there exists no reliable evidence regarding the detailed nature of the folding.

X-Ray studies of crystalline globular proteins have provided values of the dimensions of the units of structure, from which some qualitative conclusions might be drawn regarding the shapes of the protein molecules. An interesting attempt to go farther was made by Crowfoot,[5] who used her X-ray data for crystalline insulin to calculate Patterson and Patterson–Harker diagrams.[12] Crowfoot discussed these diagrams in a sensible way, and pointed out that since the X-ray data correspond to effective interplanar distances not less than 7 Å. they do not permit the determination of the positions of individual atoms;[13] the diagrams instead give some information about large-scale fluctuations in scattering power within the crystal. Crowfoot also stated that the diagrams provide no reliable evidence regarding either a polypeptide chain or a cyclol structure for insulin.

Wrinch and Langmuir[4] have, however, contended that Crowfoot's X-ray data correspond in great detail to the structure predicted for the insulin molecule on the basis of the cyclol theory, and thus provide the experimental proof of the theory. We wish to point out that the evidence adduced by Wrinch and Langmuir has very little value, because their comparison of the X-ray data and the cyclol structure involves so many arbitrary assumptions as to remove all significance from the agreement obtained. In order to attempt to account for the maxima and minima appearing on Crowfoot's diagrams, Wrinch and Langmuir made the assumption that certain regions of the crystal (center of molecule, center of lacunae) have an electron density less than the average, and others (slits, zinc atoms) have an electron density greater than the average. The positions of these regions are predicted by the cyclol theory, but the magnitudes of the electron density are not predicted quantitatively by the theory. Accordingly the authors had at their disposal seven parameters, to which arbitrary values could be assigned in order to give agreement with the data. Despite the numbers of these parameters, however, it was necessary to introduce additional arbitrary parameters, bearing no predicted relation whatever to the cyclol structure, before rough agreement with the Crowfoot diagrams could be obtained. Thus the peak B″, which is the most pronounced peak in the $P(xy0)$ section (Fig. 2 of Wrinch and Langmuir's paper) and is one of the four well-defined isolated maxima reported, is accounted for by use of a region (V) of very large negative deviation located at a completely arbitrary position in the crystal; and this region is not used by the authors in interpreting any other features of the diagrams. This introduction of four arbitrary parameters (the three coördinates and the intensity of the region V) to account for one feature of the experimental diagrams would in itself make the argument advanced by Wrinch and Langmuir unconvincing; the fact that many other parameters were also assigned arbitrary values removes all significance from their argument.

It has been pointed out by Bernal,[14] moreover, that the authors did not make the comparison of their suggested structure and the experimental diagrams correctly. They compared only a fraction of the vectors defined by their regions with the Crowfoot diagrams, and neglected the rest of the vectors. Bernal reports that he has made the complete calculation on the basis of their structure, and has found that the resultant diagrams show no relation whatever to the experimental diagrams. He states also that with seven density values at closest-packed positions as arbitrary parameters he has found that a large number of structures which give rough agreement with the experimental diagrams can be formulated.

We accordingly conclude that there exists no satisfactory X-ray evidence for the cyclol structure for insulin.

(12) A. L. Patterson, *Z. Krist.*, **90**, 517 (1935); D. Harker, *J. Chem. Phys.*, **4**, 825 (1936).

(13) It has also been pointed out by J. M. Robertson, *Nature*, **143**, 75 (1939), that the intensities of 60 planes could not provide sufficient information to locate the several thousand atoms in the insulin molecule.

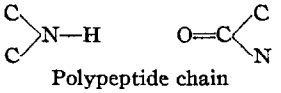
(14) J. D. Bernal, *Nature*, **143**, 74 (1939); see also D. P. Riley and I. Fankuchen, *ibid.*, **143**, 648 (1939).

3. Thermochemical Evidence Regarding Protein Structure

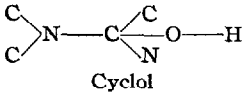
It is, moreover, possible to advance a strong argument in support of the contention that the cyclol structure does not occur to any extent in any protein.

X-Ray photographs of denatured globular proteins are similar to those of β-keratin, and thus indicate strongly that these denatured proteins contain extended polypeptide chains.[15] Astbury[16] has also obtained evidence that in protein films on surfaces the protein molecules have the extended-chain configuration, and this view is shared by Langmuir, who has obtained independent evidence in support of it.[17] Now the heat of denaturation of a protein is small—less than one hundred kilogram calories per mole of protein molecules for denaturation in solution,[18] that is, only a fraction of a kilogram calorie per mole of amino acid residues. Consequently the structure of native proteins must be such that only a very small energy change is involved in conversion to the polypeptide chain configuration.

It is unfortunate that there exist no substances known to have the cyclol structure; otherwise their heats of formation could be found experimentally for comparison with those of substances such as diketopiperazine which are known to contain polypeptide chains or rings. It is possible, however, to make this comparison indirectly in various ways. A system of values of bond energies and resonance energies has been formulated[19] which permits the total energy of a molecule of known structure to be predicted with an average uncertainty of only about 1 kcal./mole for a molecule the size of the average amino acid residue. The polypeptide chain (amide form) and cyclol can be represented by the following diagrams

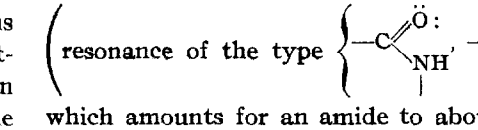


Polypeptide chain



Cyclol

The change in bonds from polypeptide chain to cyclol is N–H + C=O ⟶ N–C + C–O + O–H. With N–H = 83.3, C=O = 152.0, N–C = 48.6, C–O = 70.0, and O–H = 110.2 kcal./mole, the bonds of an amino acid residue are found to be 6.5 kcal./mole less stable for the cyclol configuration than for the chain configuration. This must further be corrected for resonance of the double bond



which amounts for an amide to about 21 kcal./mole[19]; there is no corresponding resonance for the cyclol, which involves only single bonds. We conclude that the cyclol structure is less stable than the polypeptide chain structure by 27.5 kcal./mole per amino acid residue.

This value relates to gaseous molecules, containing no hydrogen bonds, and with the ordinary van der Waals forces also neglected. It is probable that the ordinary van der Waals forces would have nearly the same value for a cyclol as for a polypeptide chain; and the available evidence[19,20] indicates that the polypeptide hydrogen bonds would be slightly stronger than the hydrogen bonds for the cyclol structure. Moreover, the observed small values (about 2 kcal./mole) for the heat of solution of amides and alcohols show that the stability relations in solution are little different from those of the crystalline substances. We accordingly conclude that the polypeptide chain structure for a protein is more stable than the cyclol structure by about 28 kcal./mole per amino acid residue, either for a solid protein or a protein in solution (with the active groups hydrated[21]).

The comparison of the polypeptide chain and cyclol can also be made without the use of bond energy values. The heat of combustion of crystalline diketopiperazine, which contains two glycine residues forming a polypeptide chain,[22] is known;[23] from its value, 474.6 kcal./mole, the heat of formation of crystalline diketopiperazine (from elements in their standard states) is calculated to be 128.4 kcal./mole, or 64.2 kcal./mole per glycine residue. A similar calculation cannot be made directly for the cyclol structure, because no sub-

(15) W. T. Astbury, S. Dickinson and K. Bailey, *Biochem. J.*, **29**, 2351 (1935).

(16) W. T. Astbury, *Nature*, **143**, 280 (1939).

(17) I. Langmuir, *ibid.*, **143**, 280 (1939).

(18) M. L. Anson and A. E. Mirsky, *J. Gen. Physiol.*, **17**, 393, 399 (1934).

(19) (a) L. Pauling and J. Sherman, *J. Chem. Phys.*, **1**, 606 (1933); (b) L. Pauling, "The Nature of the Chemical Bond," Cornell University Press, Ithaca, N. Y., 1939. The values quoted above are from the latter source; they involve no significant change from the earlier set.

(20) M. L. Huggins, *J. Org. Chem.*, **1**, 407 (1936).

(21) The suggestion has been made [F. C. Frank, *Nature*, **138**, 242 (1936)] that the energy of hydration of hydroxyl groups might be very much greater than that of the carbonyl and secondary amino groups of a polypeptide chain; there exists, however, no evidence indicating that this is so.

(22) R. B. Corey, ref. 7.

(23) M. S. Kharasch, *Bur. Standards J. Research*, **2**, 359 (1929).

stance is known to have the cyclol structure; but an indirect calculation can be made in many ways, such as the following. One hexamethylenetetramine molecule and one pentaerythritol molecule contain the same bonds as four cyclized glycine residues and three methane molecules; hence the heat of formation of a glycine cyclol per residue is predicted to have the value 32.2 kcal./mole found experimentally[24] for $\frac{1}{4}C_6H_{12}N_4(c) + \frac{1}{4}C(CH_2OH)_4(c) - \frac{3}{4}CH_4(c)$. Similarly the value for $N(C_2H_5)_3(c) + C_2H_5OH(c) - 3C_2H_6(c)$ is 40.2 kcal./mole. The average of several calculations of this type, 36 kcal./mole, differs from the experimental value of the heat of formation of diketopiperazine per residue, 64 kcal./mole, by 28 kcal./mole. This agrees closely with the value 27.5 kcal./mole found by the use of bond energies, and we can be sure that the suggested cyclol structure for proteins is less stable than the polypeptide chain structure by about this amount per amino acid residue. Since denatured proteins are known to consist of polypeptide chains, and native proteins differ in energy from denatured proteins by only a very small amount (less than 1 kcal./mole per residue), we draw the rigorous conclusion that *the cyclol structure cannot be of primary importance for proteins; if it occurs at all (which is unlikely because of its great energetic disadvantage relative to polypeptide chains) not more than about three per cent. of the amino acid residues could possess this configuration.*

The above conclusion is not changed if the assumption be made that polypeptide chains are in the imide rather than the amide form,[3b] since this would occur only if the imide form were the more stable. In this case the experimental values of heats of formation (such as that of diketopiperazine) would still be used as the basis for comparison with the predicted value for the cyclol structure, and the same energy difference would result from the calculation.

It has been recognized[25-27] that energy relations present some difficulty for the cyclol theory (although the seriousness of the difficulty seems not to have been appreciated), and various suggestions have been made in the attempt to avoid the difficulty. In her latest communication[3h] Wrinch writes, "The stability of the globular proteins, under special conditions, in solution and in the crystal, we attribute to definite stabilizing factors;[26,27] namely, (1) hydrogen bonds between the oxygens of certain of the triazine rings, (2) the multiple paths of linkage between atoms in the fabric, (3) the closing of the fabric into a polyhedral surface which eliminates boundaries of the fabric and greatly increases the symmetry, and (4) the coalescence of the hydrophobic groups in the interior of the cage." These factors are, however, far from sufficient to stabilize the cyclol structure relative to the polypeptide chain structure. (1) The hydrogen bonds between hydroxyl groups in the cyclol structure would have nearly the same energy (about 5 kcal./mole) as those involving the secondary amino and carbonyl groups of the polypeptide chain. The suggestion[26] that resonance of the protons between oxygen atoms would provide further stabilization is not acceptable, since the frequencies of nuclear motion are so small compared with electronic frequencies that no appreciable resonance energy can be obtained by resonance involving the motion of nuclei. (2) We are unable to find any aspects of the bond distribution in cyclols which are not taken into consideration in our energy calculation given above. (3) There is no type of interatomic interaction known to us which would lead to additional stability of a cage cyclol as the result of eliminating boundaries and increasing the symmetry. (4) The stabilizing effect of the coalescence of the hydrophobic groups has been estimated[26] to be about 2 kcal./mole per CH_2 group, and to amount to a total for the insulin molecule of about 600 kcal./mole. It seems improbable to us that the van der Waals interactions of these groups are much less than this for polypeptides. The maximum of 600 kcal./mole from this source is still negligibly small compared with the total energy difference to be overcome, amounting to about 8000 kcal./mole for a protein containing about 288 residues.[28]

We accordingly conclude that the cyclol structure is so unstable relative to the polypeptide structure that it cannot be of significance for proteins.

It may be pointed out that a number of experiments[29-31] have added the weight of their evi-

(24) The values of heats of combustion used are $C_6H_{12}N_4(c)$, 1006.7; $C(CH_2OH)_4(c)$, 661.2; $CH_4(c)$, 210.6; $N(C_2H_5)_3(c)$, 1035.5; $C_2H_5OH(c)$, 325.7; $C_2H_6(c)$, 370.0 kcal./mole.

(25) F. C. Frank, *Nature*, **138**, 242 (1936).

(26) I. Langmuir and D. Wrinch, *ibid.*, **143**, 49 (1939).

(27) D. Wrinch, *Symposia on Quant. Biol.*, **6**, 122 (1938).

(28) Other suggestions regarding the source of stabilizing energy which have been made hardly merit discussion. "Foreign molecules" (Wrinch, ref. 27), for example, cannot be discussed until we have some information as to their nature.

(29) G. I. Jenkins and T. W. J. Taylor, *J. Chem. Soc.*, 495 (1937).

(30) L. Kellner, *Nature*, **140**, 193 (1937).

(31) H. Meyer and W. Hohenemser, *ibid.*, **141**, 1138 (1938).

dence to the general conclusion reached in this communication that the cyclol bond and the cyclol fabric are energetically impossible.

4. Further Arguments Indicating the Non-existence of the Cyclol Structure

There are many additional arguments which indicate more or less strongly that the cyclol structure does not exist. Of these we shall mention only a few.

It has been found experimentally that two atoms in adjacent molecules or in the same molecule but not bonded directly to one another reach equilibrium at a distance which can be represented approximately as the sum of certain van der Waals radii for the atoms.[19,32] Two carbon atoms of methyl or methylene groups not bonded to the same atom never approach one another more closely than about 4.0 Å., and two hydrogen atoms not bonded to the same atom are always at least 2.0 Å. apart. It has been pointed out by Huggins[33] that the cyclol structure places the carbon atoms of side chains only 2.45 Å. apart, and that in the C_2 structure for insulin there are hydrogen atoms only 0.67 Å. apart. We agree with Huggins that this difficulty alone makes the cyclol hypothesis unacceptable.

A closely related argument, dealing with the small area available for the side chains of a cyclol fabric, has been advanced by Neurath and Bull.[34] The area provided per side chain by the cyclol fabric, about 10 sq. Å., is far smaller than that required; and, as Neurath and Bull point out, the suggestion[3e,26] that some of the side chains pass through the lacunae of the fabric to the other side cannot be accepted, because this would require non-bonded interatomic distances much less than the minimum values found in crystals.

One of the most striking features of the cyclol fabric is the presence of great numbers of hydroxyl groups: in the case of cyclol C_2 there are 288 hydroxyl groups exclusive of those present in the side chains. Recently Haurowitz[35,36] has subjected the cyclol hypothesis to experimental tests on the basis of the existence or non-existence of cage hydroxyl groups. In the first communication[35] Haurowitz concludes on the basis of his and previous experiments[37-39] on the acylation and alkylation of proteins that the experimental evidence is in decided opposition to the conception that proteins possess great numbers of hydroxyl groups and therefore to the cyclol hypothesis. It seems to us that the objection raised by Haurowitz[36] is worthy of consideration and it certainly cannot be disposed of on the grounds that the original structure has been destroyed unless some concrete evidence can be submitted to indicate that this is the case. In a second communication Haurowitz and Astrup[36] write that "According to the classical theory of protein structure the carboxyl and amino groups found after hydrolytic splitting of a protein come from —CO—NH— bonds. According to the cyclol hypothesis, however, the free carboxyl and amino groups must be formed, during the splitting, from bonds of the structure =C(OH)—N=. The classical theory would predict on hydrolysis no great change in the absorption spectrum below 2400 Å. because the CO groups of the amino acids and of the peptide bonds both are strongly absorbing in this region.[40] On the other hand, the cyclol hypothesis would predict a greatly increased absorption because of the formation of new CO groups. . . . The absorption for genuine and for hydrolyzed protein is about equal. This seems to be in greater accordance with the classical theory of the structure of proteins than with the cyclol theory."

Mention may also be made of the facts that no simple substances with the cyclol structure have ever been synthesized[29] and that in general chemical reactions involving the breaking of covalent bonds are slow, whereas rapid interconversion of polypeptide and cyclol structure must be assumed to occur in, for example, surface denaturation. These chemical arguments indicate strongly that the cyclol theory is not acceptable.[41]

5. A Discussion of Arguments Advanced in Support of the Cyclol Theory

Although a great number of papers dealing with the cyclol theory have been published, we have

(32) N. V. Sidgwick, "The Covalent Link in Chemistry," Cornell University Press, Ithaca, N. Y., 1933; E. Mack, Jr., THIS JOURNAL, **54**, 2141 (1932); S. B. Hendricks, *Chem. Rev.*, **7**, 431 (1930); M. L. Huggins, *ibid.*, **10**, 427 (1932).

(33) M. L. Huggins, THIS JOURNAL, **61**, 755 (1939).

(34) H. Neurath and H. D. Bull, *Chem. Rev.*, **23**, 427 (1938).

(35) T. Haurowitz, *Z. physiol. Chem.*, **256**, 28 (1938).

(36) T. Haurowitz and T. Astrup, *Nature*, **143**, 118 (1939).

(37) J. Herzig and K. Landsteiner, *Biochem. Z.*, **61**, 458 (1914).

(38) B. M. Hendrix and F. Paquin, Jr., *J. Biol. Chem.*, **124**, 135 (1938).

(39) K. G. Stern and A. White, *ibid.*, **122**, 371 (1938).

(40) M. A. Magill, R. E. Steiger and A. J. Allen, *Biochem. J.*, **31**, 188 (1937).

(41) Another argument against cyclols of the C_2 type can be based on the results reported by J. L. Oncley, J. D. Ferry and J. Shack, *Symposia on Quant. Biol.*, **6**, 21 (1938), H. Neurath, *ibid.*, **6**, 196 (1938), and J. W. Williams and C. C. Watson, *ibid.*, **6**, 208 (1938), who have shown that dielectric constant measurements and diffusion measurements indicate that the molecules of many proteins are far from spherical in shape.

had difficulty in finding in them many points of comparison with experiment (aside from the X-ray work mentioned above) which were put forth as definite arguments in support of the structure.

One argument which has been advanced is that the cyclol theory "readily interprets the total number of amino acid residues per molecule, without the introduction of any *ad hoc* hypothesis"[3e] and that "The group of proteins with molecular weights ranging rom 33,600 to 40,500 are closed cyclols of the type C_2 containing 288 amino acid residues."[3e] Now the presence of imino acids (proline, oxyproline) in a protein prevents its formation of a complete cyclol such as C_2, and many proteins in this molecular weight range are known to contain significant amounts of proline: for insulin 10% is reported,[3f] for egg albumin 4%,[42] for zein 9%,[43] for Bence–Jones protein 3%,[44] and for pepsin 5%.[45] Wrinch has stated that "a future modification" (in regard to the number of residues) "is also introduced if imino acids are present"[3c]; "these numbers perhaps being modified if imino acids are present";[3e] and "if certain numbers of imino acid residues are present, these numbers" (of residues) "may be correspondingly modified."[3g] This uncertainty regarding the effect of the presence of imino acids in cyclols on the expected number of residues leaves the argument little force. In fact, even the qualitative claim that the cyclol hypothesis implies the existence of polyhedral structures containing certain numbers of amino acid residues and so predicts that globular proteins have molecular weights which fall into a sequence of separated classes can be doubted for the same reason.

It has been claimed[36] that the cyclol hypothesis explains the facts that proteins contain certain numbers of various particular amino acid residues and that these numbers are frequently powers of 2 and 3,[46] and it is proper that we inquire into the nature of the argument. Wrinch states[27] "An individual R group" (side chain) "is presumably attached, not to just any α-carbon atom, but only to those whose environment makes them appropriate in view of its specific nature. As an example of different environments, we may refer to the cyclol cages; here the pairs of residues at a slit have 'different environments' and the residues not at a slit fall into sets which again have 'different environments.' We therefore expect characteristic proportions to be associated with aromatic, basic, acidic, and hydrocarbon R groups, respectively, even perhaps with individual R groups. In any case a non-random distribution of the proportions of each residue in proteins in general is to be expected on any fabric hypothesis. On the cyclol hypothesis, for example, α-carbons having equivalent environments occur in powers of 2 and 3. . . . It is difficult to avoid interpreting the many cases which have recently been summarized in which the proportions of many types of residue are powers of 2 and 3 as further direct evidence in favor of the cyclol fabric. This fabric consists of an alternation of diazine and triazine hexagons, with symmetries respectively 2 and 3." Also it has been said by Langmuir[47] that "The occurrence of these factors, 2 and 3, furnishes a powerful argument for a geometrical interpretation such as that given by the cyclol theory. In fact, the hexagonal arrangement of atoms in the cyclol fabric gives directly and automatically a reason for the existence of the factors 2 and 3 and the non-occurrence of such factors as 5 and 7."

On examining the cyclol C_2, however, we find that these statements are not justified. The only factors of 288 are of the form $2^n\ 3^m$; moreover, the framework of the cyclol C_2 has the tetrahedral symmetry T, so that if the distribution of side chains conforms to the symmetry of the framework the amino acid residues would occur in equivalent groups of twelve. But in view of the rapid decrease in magnitude of interatomic forces with distance there would seem to be little reason for the distribution of side chains over a large protein molecule to conform to the symmetry T; it is accordingly evident that any residue numbers might occur for the cyclol C_2. We conclude that the cyclol hypothesis does not provide an explanation of the occurrence of amino acid residues in numbers equal to products of powers of 2 and 3.

Although there is little reason to expect that the distribution of side chains would correspond to the symmetry of the framework, it is interesting to note that the logical application of the methods of argument used by Wrinch suggests strongly that sixty residues of each of two amino acids should be present in a C_2 cyclol. This cyclol contains twenty lacunae of a particular type—each surrounded by a nearly coplanar border of twelve diazine and triazine

(42) H. O. Calvery, *J. Biol. Chem.*, **94**, 613 (1931).

(43) T. B. Osborne and L. M. Liddle, *Am. J. Physiol.*, **26**, 304 (1910).

(44) C. L. A. Schmidt, "Chemistry of Amino Acids and Proteins," C. C. Thomas, Springfield, Ill., 1938

(45) Unpublished determination by one of the authors.

(46) M. Bergmann and C. Niemann, *J. Biol. Chem.*, **115**, 77 (1936); **118** 307 (1937).

(47) I. Langmuir, *Symposia on Quant. Biol.*, **6**, 135 (1938).

rings. Each of these has trigonal symmetry so far as this near environment is concerned. Hence it might well be expected that a particular amino acid would be represented by three residues about each of these twenty lacunae, giving a total of sixty residues. But the number 60 cannot be expressed in the form $2^n 3^m$, it is not a factor of 288, and the integer nearest the quotient 288/60, 5, also cannot be expressed in the form $2^n 3^m$.

One of the most straightforward arguments advanced by Wrinch[3c,d] is that a protein surface film must have all its side chains on the same side, which would be the case for a cyclol fabric but not for an extended polypeptide chain. This argument now has lost its significance through the recently obtained strong evidence that proteins in films have the polypeptide structure,[16,17] and not the cyclol structure.

There can be found in the papers by Wrinch many additional statements which might be construed as arguments in support of the cyclol structure. None of these seems to us to have enough significance to justify discussion.

6. The Present State of the Protein Problem

The amount of experimental information about proteins is very great, but in general the processes of deducing conclusions regarding the structure of proteins from the experimental results are so involved, the arguments are so lacking in rigor, and the conclusions are so indefinite that it would not be possible to present the experimental evidence at the basis of our ideas of protein structure[48] in a brief discussion. In the following paragraphs we outline our present opinions regarding the structure of protein molecules, without attempting to do more than indicate the general nature of the evidence supporting them. These opinions were formed by the consideration not only of the experimental evidence obtained from proteins themselves but also of the information regarding interatomic interactions and molecular structure in general which has been gathered by the study of simpler molecules.

We are interested here only in the role of amino acids in proteins—that is, in the simple proteins (consisting only of α-amino and α-imino acids) and the corresponding parts of conjugated proteins; the structure and linkages of prosthetic groups will be ignored.

The great body of evidence indicating strongly that the amino acids in proteins are linked together by peptide bonds need not be reviewed here.

The question now arises as to whether the polypeptide chains or rings contain many or few amino acid residues. We believe that the chains or rings contain many residues—usually several hundred. The fact that in general proteins in solution retain molecular weights of the order of 17,000 or more until they are subjected to conditions under which peptide hydrolysis occurs gives strong support to this view. It seems to us highly unlikely that any protein consist of peptide rings containing a small number of residues (two to six) held together by hydrogen bonds or similar relatively weak forces, since, contrary to fact, in acid or basic solution a protein molecule of this type would be decomposed at once into its constituent small molecules.

There exists little evidence as to whether a long peptide chain in a protein has free ends or forms one more peptide bond to become a ring. This is, in fact, a relatively unimportant question with respect to the structure, as it involves only one peptide bond in hundreds, but it may be of considerable importance with respect to enzymatic attack and biological behavior in general.

A native protein molecule with specific properties must possess a definite configuration, involving the coiling of the polypeptide chain or chains in a rather well-defined way.[49] The forces holding the molecule in this configuration may arise in part from peptide bonds between side-chain amino and carboxyl groups or from side-chain ester bonds or S–S bonds; in the main, however, they are probably due to hydrogen bonds and similar interatomic interactions. Interactions of this type, while individually weak, can by combining their forces stabilize a particular structure for a molecule as large as that of a protein. In some cases (trypsin, hemoglobin) the structure of the native protein is the most stable of those accessible to the polypeptide chain; the structure can then be reassumed by the molecule after denaturation. In other cases (antibodies) the native configuration is not the most stable of those accessible, but is an unstable configuration impressed on the molecule by its environment (the influence of the antigen) during its synthesis; denaturation is not reversible for such a protein.

Crystal structure investigations have shown

(48) We believe that our views regarding the structure of protein molecules are essentially the same as those of many other investigators interested in this problem.

(49) H. Wu, *Chinese J. Physiol.*, **5**, 321 (1931); A. E. Mirsky and L. Pauling, *Proc. Nat. Acad. Sci.*, **22**, 439 (1936).

that in general the distribution of matter in a molecule is rather uniform. A protein layer in which the peptide backbones are essentially coplanar (as in the β-keratin structure) has a thickness of about 10 Å. If these layers were arranged as surfaces of a polyhedron, forming a cage molecule, there would occur great steric interactions of the side chains at the edges and corners. (This has been used above as one of the arguments against the C_2 cyclol structure.) We accordingly believe that *proteins do not have such cage structures.*[50] A compact structure for a globular protein might involve the superposition of several parallel layers, as suggested by Astbury, or the folding of the polypeptide chain in a more complex way.

One feature of the cyclol hypothesis—the restriction of the molecule to one of a few configurations, such as C_2—seems to us unsatisfactory rather than desirable. The great versatility of antibodies in complementing antigens of the most varied nature must be the reflection of a correspondingly wide choice of configuration by the antibody precursor. We feel that the biological significance of proteins is the result in large part of their versatility, of the ability of the polypeptide chain to accept and retain that configuration which is suited to a special purpose from among the very great number of possible configurations accessible to it.

Proteins are known to contain the residues of some twenty-five amino acids and it is not unlikely that this number will be increased in the future. A great problem in protein chemistry is that of the order of the constituent amino acid residues in the peptide chains. Considerable evidence has been accumulated[46] suggesting strongly that the stoichiometry of the polypeptide framework of protein molecules can be interpreted in terms of a simple basic principle. This principle states that the number of each individual amino acid residue and the total number of all amino acid residues contained in a protein molecule can be expressed as the product of powers of the integers two and three. Although there is no direct and unambiguous experimental evidence confirming the idea that the constituent amino acid residues are arranged in a periodic manner along the peptide chain, there is also no experimental evidence which would deny such a possibility, and it seems probable that steric factors might well cause every second or third residue in a chain to be a glycine residue, for example.

The evidence regarding frequencies of residues involving powers of two and three leads to the conclusion that there are 288 residues in the molecules of some simple proteins. It is not to be expected that this number will be adhered to rigorously. Some variation in structure at the ends of a peptide chain might be anticipated; moreover, amino acids might enter into the structure of proteins in some other way than the cyclic sequence along the main chain.[51] The structural significance of the number 288 is not clear at present. It seems to us, however, very unlikely that the existence of favored molecular weights (or residue numbers) of proteins is the result of greater thermodynamic stability of these molecules than of similar molecules which are somewhat smaller or larger, since there are no interatomic forces known which could effect this additional stabilization of molecules of certain sizes. It seems probable that the phenomenon is to be given a biological rather than a chemical explanation—we believe that the existence of molecular-weight classes of proteins is due to the retention of this protein property through the long process of the evolution of species.

We wish to express our thanks to Dr. R. B. Corey for his continued assistance and advice in the preparation of this paper, and also to other colleagues who have discussed these questions with us.

Summary

It is concluded from a critical examination of the X-ray evidence and other arguments which have been proposed in support of the cyclol hypothesis of the structure of proteins that these arguments have little force. Bond energy values and heats of combustion of substances are shown to lead to the prediction that a protein with the cyclol structure would be less stable than with the polypeptide chain structure by a very large amount, about 28 kcal./mole of amino acid residues; and the conclusion is drawn that proteins do not have the cyclol structure. Other arguments leading to the same conclusion are also presented. A brief discussion is given summarizing the present state of the protein problem, with especial reference to polypeptide chain structures.

PASADENA, CALIF. RECEIVED APRIL 22, 1939

(50) Wrinch recently has suggested[3b] that even if proteins are not cyclols the cage structure might be significant.

(51) H. Jensen and E. A. Evans, Jr., *J. Biol. Chem.*, **108**, 1 (1935), have shown that insulin probably contains several phenylalanine groups attached only by side-chain bonds to the main peptide chain.