4

Chemical Characterization of Proteins, Carbohydrates, and Lipids

J. L. ONCLEY

Department of Biological Chemistry, Harvard University Medical School, Boston 15, Massachusetts

PROTEINS

POLYMER molecules of high molecular weight but with relatively simple repeating units of one or two kinds are well known today. Proteins are naturally occurring polymers in which amino-acid residues are joined through peptide bonds and certain types of cross-linking bonds. They vary widely in physical and biological properties—from the insoluble keratin of hair to small soluble hormones like insulin—and are essential components of all living matter.

It has been found that only certain amino-acid residues—those from the "natural amino acids"—are present in proteins. These amino-acid residues are shown in Table I, and can be considered in three categories. The first category represents the neutral amino-acid residues—residues which do not have acidic or basic properties over the normal *p*H range. In the second category are the acidic and the basic amino-acid residues, listed in order of decreasing acidity or increasing basicity (as indicated by the pK_a values). A third category represents amino-acid residues found only in certain proteins. Neglecting this last category of rarely found residues, 21 amino-acid residues are listed. The number of fundamental residues can be reduced to 18 if one considers the cystinyl residue as the oxidation product of two cysteinyl residues, and asparaginyl and glutaminyl residues as simple amide derivatives of aspartyl and glutamyl residues.

All of the residues except prolyl are derived from α-amino acids. The prolyl residue is derived from an α-imino acid (i.e., an N-substituted α-amino acid). An important feature of these 18 or 21 natural amino-acid residues is that in proteins they all seem to occur only in the L-configuration. The naturally occurring (protein) forms of all of the amino acids have been related to L-glyceraldehyde, and recent studies by Bijvoet *et al.*[1] have established the correctness of the Fischer convention which assumed the absolute configuration illustrated in Fig. 1. To date, no significant concentrations of D-amino-acid residues have been found in any of the proteins that have been studied.*

The peptide bond, proposed independently by Fischer and Hofmeister in 1902, is the principal linkage of the various amino-acid residues making up protein structures. It is formed by the splitting-out of water between the α-amino group of one amino acid and the carboxyl group of the next, with the formation of an amide linkage between the two residues:

$$NH_2\text{—}\underset{R_1}{CH}\text{—}COOH + NH_2\text{—}\underset{R_2}{CH}\text{—}COOH = NH_2\text{—}\underset{R_1}{CH}\text{—}CO\text{—}NH\text{—}\underset{R_2}{CH}\text{—}COOH + H_2O.$$

Peptide formulas are customarily written with the free α-amino group at the left, and are named as substitution products of the amino acid furnishing the free α-carboxyl group.

The peptide bond is fairly stable, but its hydrolysis is catalyzed by acid, base, or various proteolytic enzymes. The peptide bond of the various amino-acid residues differ in stability during acid or base catalyzed hydrolysis, and there is a high degree of residue specificity in the enzymatic hydrolysis (see Neurath, p. 185). It is possible to form amide linkages involving the ϵ-amino group of lysine or the ω-carboxyl group of aspartic or glutamic acid. These other amide bonds, with properties very much like the peptide bond, have not been observed to form an important part of any protein structure. On the other hand, considerable amounts of glutathione, a dipeptide amide with the structure γ-glutamyl-cysteinyl-glycine, are found in many organisms, and poly-glutamic acid of certain bacteria contains the γ-glutamyl amide bond. Figure 2, from a more comprehensive discussion of protein structure by Low,[2] shows the detailed configuration of a fully extended polypeptide chain. The amide group, —CO—NH—, is planar, and the bond distances observed in various peptides (αC—C′ = 1.53 A, C′—O = 1.24 A, C′—N = 1.32 A, N—αC = 1.47 A, with angles N—αC—C′ = 110°, αC—C′—N = 114°, O—C′—N = 125°, and C′—N—αC = 123°) indicate that the αC—N bond has about 40% double-bond character.[3]

A number of other bonds in addition to those making up the peptide linkage are found in most proteins. The disulfide linkage, where two cysteinyl residues are

* However, it might be argued that this statement needs a little more proof. Too often one assumes without proof that the amino acids are all in the L-configuration, particularly if one is working with small amounts of the amino acids. Since many peptide antibiotics have been found to contain the D-forms of these natural amino acids (as well as other amino-acid residues), the complete absence of the D-configuration in all proteins might be questioned.

TABLE I. Natural amino acids.

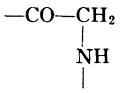
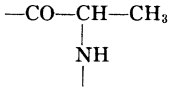
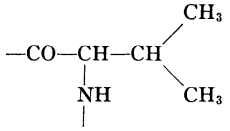
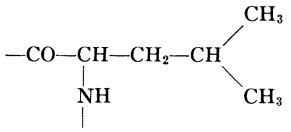
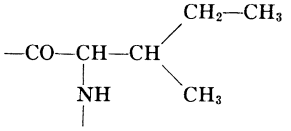
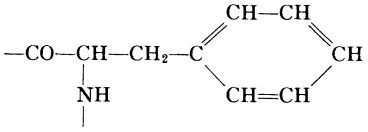
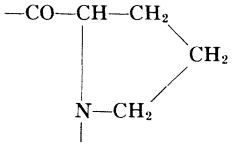
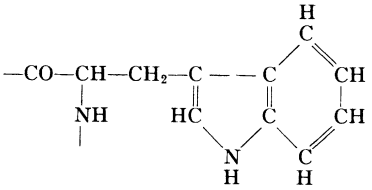
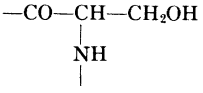
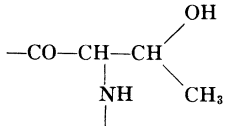
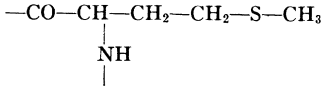
Residue symbol	(a) Residue name (b) Amino-acid name (trivial) (c) Amino-acid name (systematic)	Residue formula
	Amino-acid residues usually found in proteins	
	Neutral	
-gly-	(a) glycyl (b) glycine (c) aminoacetic acid	
-ala-	(a) alanyl (b) alanine (c) α-aminopropionic acid	
-val-	(a) valyl (b) valine (c) α-aminoisovaleric acid	
-leu-	(a) leucyl (b) leucine (c) α-aminoisocaproic acid	
-ileu-	(a) isoleucyl (b) isoleucine (c) α-amino-β-methylvaleric acid	
-phe-	(a) phenylalanyl (b) phenylalanine (c) α-amino-β-phenylpropionic acid	
-pro-	(a) prolyl (b) proline (c) pyrrolidine-2-carboxylic acid	
-try-	(a) tryptophanyl (b) tryptophan (c) α-amino-β-indolylpropionic acid	
-ser-	(a) seryl (b) serine (c) α-amino-β-hydroxypropionic acid	
-thr-	(a) threonyl (b) threonine (c) α-amino-β-hydroxybutyric acid	
-met-	(a) methionyl (b) methionine (c) α-amino-γ-methylthiobutyric acid	

TABLE I.—*Continued.*

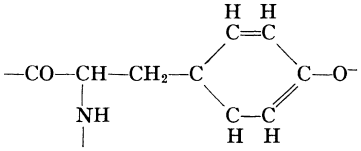
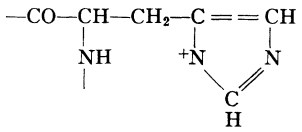
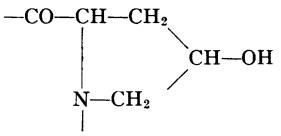
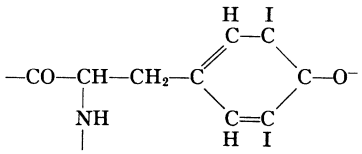
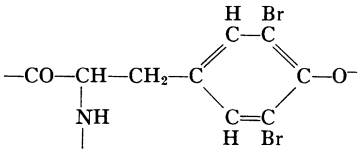
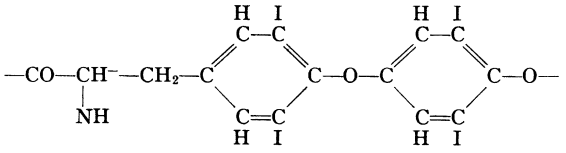
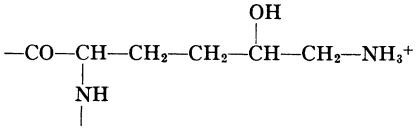
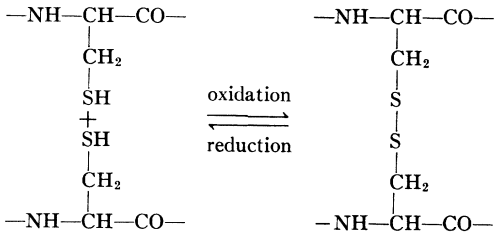
Residue symbol	(a) Residue name (b) Amino-acid name (trivial) (c) Amino-acid name (systematic)	Residue formula	
-cyS-* | -cyS-	(a) cystinyl (b) cystine (c) β-β′-dithiobis (α-aminopropionic acid)	—CO—CH—CH$_2$—S—S—CH$_2$—CH—NH— | NH | CH |	
NH$_2$ | -asp-	(a) asparaginyl (b) asparagine (c) α-aminosuccinamic acid	—CO—CH—CH$_2$—CO—NH$_2$ | NH |	
NH$_2$ | -glu-	(a) glutaminyl (b) glutamine (c) α-aminoglutaramic acid	—CO—CH—CH$_2$—CH$_2$—CO—NH$_2$ | NH |	
	Amino-acid residues usually found in proteins Acidic (ionized form)		
-OH-*	terminal carboxyl	—O$^-$ (—CO$_2^-$)	pK_a=3.6–4.0
-asp-	(a) aspartyl (b) aspartic acid (c) aminosuccinic acid	—CO—CH—CH$_2$—CO$_2^-$ | NH |	pK_a=3.9–4.7
-glu-	(a) glutamyl (b) glutamic acid (c) α-aminoglutaric acid	—CO—CH—CH$_2$—CH$_2$—CO$_2^-$ | NH |	pK_a=3.9–4.7
-tyr-	(a) tyrosyl (b) tyrosine (c) α-amino-β-(*p*-hydroxyphenyl) propionic acid		pK_a=8.5–10.9
-cySH-*	(a) cysteinyl (b) cysteine (c) α-amino-β-mercaptopropionic acid	—CO—CH—CH$_2$—S$^-$ | NH |	pK_a=*ca* 10
	Basic (ionized form)		
-his-	(a) histidyl (b) histidine (c) α-amino-β-imidazolyl-propionic acid		pK_a=6.4–7.0
-H-*	terminal amino	—H$_2^+$ (—NH$_3^+$)	pK_a=7.4–8.5
-lys-	(a) lysyl (b) lysine (c) α-ε-diaminocaproic acid	—CO—CH—CH$_2$—CH$_2$—CH$_2$—CH$_2$—NH$_3^+$ | NH |	pK_a=8.5–10.9
-arg-	(a) arginyl (b) arginine (c) α-amino-γ-guanidino-valeric acid	—CO—CH—CH$_2$—CH$_2$—NH—CH=NH$_2^+$ | NH | NH$_2$	pK_a=11.9–13.3

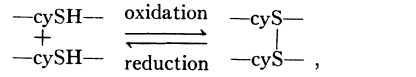
TABLE I.—*Continued.*

Residue symbol	(a) Residue name (b) Amino-acid name (trivial) (c) Amino-acid name (systematic)	Residue formula	
	Amino-acid residues occasionally found in proteins		
	Neutral		
-hypro-	(a) hydroxyprolyl (b) hydroxyproline (c) 4-hydroxypyrroli-dine-2-carboxylic acid (Found only in collagen)		
	Acidic (ionized form)		
	(a) diiodotyrosyl (b) diiodotyrosine (c) 3,5-diiodotyrosine (Found only in marine organisms and thyroglobulin)		$pK_a = 6.5$
	(a) dibromotyrosyl (b) dibromotyrosine (c) 3,5-dibromotyrosine (Found only in marine organisms)		$pK_a = ca\ 7$
	(a) thyroxyl (b) thyroxine (c) 3,3′,5,5′-tetra-iodothyronine (Found only in thyroglobulin)		$pK_a = ca\ 6.5$
	Basic (ionized form)		
-hylys-	(a) hydroxylysyl (b) hydroxylysine (c) α, -diamino-hydroxy-caproic-acid (Found only in collagen)		$pK_a = ca\ 11$

oxidized to form the disulfide linkage,

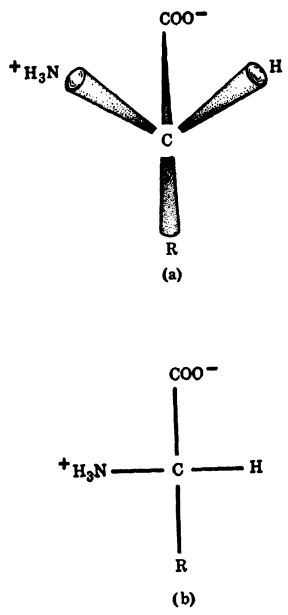


or



is found in all but a few proteins. This linkage may be between two otherwise separate peptide chains, or it may be an additional intrachain link. Both of these types of disulfide bonds are found in the insulin molecule, shown diagrammatically in Fig. 3. Calvin[4] has recently discussed the geometry of such C—S—S—C linkages. The actual spatial arrangement of a disulfide bond is shown in Fig. 4. The intrachain disulfide of insulin leads to a "link" in the polypeptide chain. This link contains six amino-acid residues (20 atoms), and it is of great interest to find that links of exactly the same number of residues are found in the peptide hormones oxytocin, arginine vasopressin, and lysine vasopressin (see Stetten, p. 563).

Recent studies by Perlmann[5] have done much to elucidate the various types of crosslinkages which involve phosphoric-acid residues. Orthophosphate —O—PO_2^-—O— and pyrophosphate —O—PO_2^-—



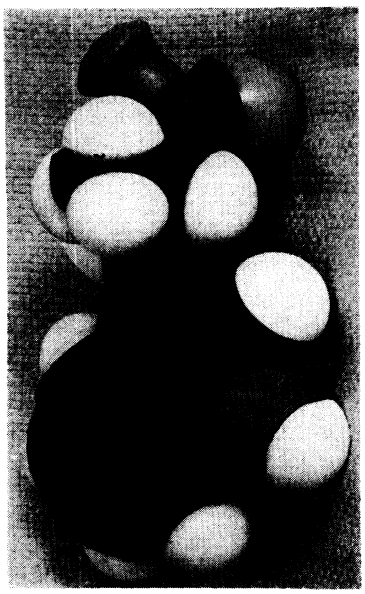
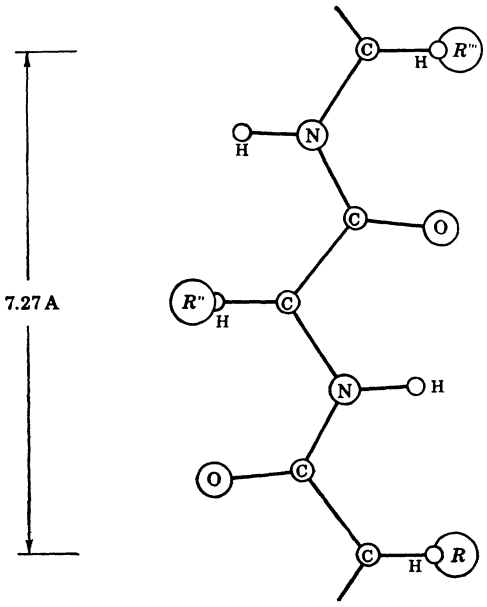

(c)

FIG. 1. Absolute configuration of the L-amino acids. (a) Geometrical representation. The COO^-, NH_3^+, and H groups are represented on a face plain of a tetrahedral carbon atom, with the R-group at the opposite apex. (b) Conventional chemical representation. (c) Photograph of a three-dimensional model of L-tryptophan.

O—PO_2^-—O— crosslinkages have been demonstrated to occur in pepsin and in α-casein, respectively. The phospho-amide crosslinkage —O—PO_2^-—NH— has also been found in α-casein, and the —O—$PO_3^=$ terminal residue is present in α-casein and in pepsin. The —NH—$PO_3^=$ terminal residue may also occur in phosphoproteins, but has not so far been demonstrated. These phosphoric-acid linkages would seem to involve seryl and threonyl residues in the case of linkage to oxygen, and possibly arginyl and lysyl residues in the case of linkage to nitrogen.

The hydrogen bond also seems to be an element in intra- and inter-peptide chain crosslinkage. Intrachain hydrogen bonding of the type —C—O—H—N— has been proposed by a number of workers, and makes up the important element of the Pauling α-helix, discussed in a later paper (see Rich, p. 50). Hydrogen bonding of the same type, but between different peptide chains, is the important crosslinking element in the pleated-sheet structures of Pauling and others, and in various double- and triple-stranded structures [for example,



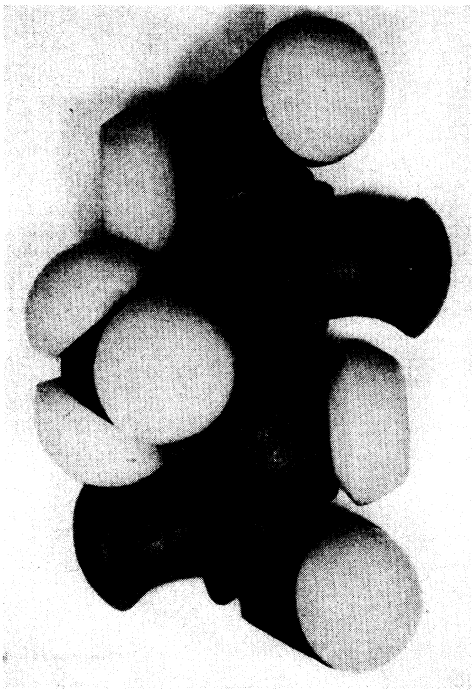


FIG. 2. Photograph of scale model of extended polypeptide chain. The R- and R'''-groups lie at the rear of this model [from B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 259].

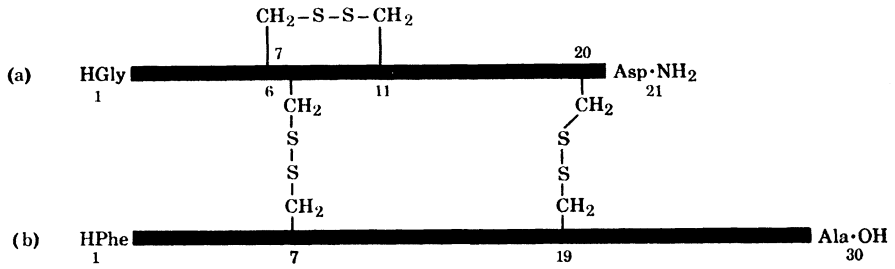


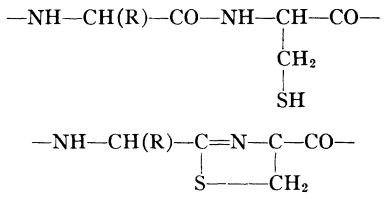
FIG. 3. Outline of the structure of insulin, showing the intrapeptide- and interpeptide-chain disulfide linkages. The numbers refer to the sequence of residues, as shown in Fig. 6 [from B. W. Low and J. T. Edsall in *Currents in Biochemical Research*, 1956, D. E. Green, editor (Interscience Publishers, Inc., New York, 1956), p. 379].

collagen (Rich, p. 58)]. Hydrogen bonds may also exist between carboxyl, amino, and tyrosyl residues. Although the energies involved in any one hydrogen bond are low (2000 to 6000 cal/mole, and even much lower if one considers the differences in energy between peptide-peptide hydrogen bonds and the same groups hydrogen bonded to water[6]), they are of great importance in structures where large numbers of hydrogen bonds can operate cooperatively. Linderstrøm-Lang[7] has reviewed an important method for the estimation of the extent of hydrogen bonding in a protein structure. (A further discussion of the hydrogen bond is found in the article by Orgel, p. 100.)

Another type of secondary bonding, leading both to intra- and inter-peptide chain crosslinkages, results from the action of van der Waals forces between the hydrocarbon-like residues. Waugh[6] has recently called attention to the large "nonpolar side-chain volume" contributed by such residues, and has stressed the point that allowance must be made for the fact that water must first be removed from contact with such groups before they can interact, and that this effect makes the interaction energies considerably larger than would be calculated solely on the basis of van der Waals attractive energies. It seems likely that the magnitude of such attractive forces within the protein molecule is quite comparable to those contributed by hydrogen bonding, and certain positions of large nonpolar residues on the peptide chains might lead to the stability of structures not compatible with the α-helix configuration. These considerations are treated in the later contribution by Waugh (p. 84).

Electrostatic forces between positively and negatively charged groups of the protein molecule can give rise both to attractive and to repulsive forces. The magnitude of such electrostatic forces depends upon the ionic strength of the solution. Most of the physicochemical studies of solutions of globular proteins (see Doty, p. 61) do not show a great dependence upon the ionic strength if the net charge of the protein is not too great, and can probably be interpreted as indicating that electrostatic forces are usually not of great importance in determining the conformation of the molecule. In the case of certain elongated protein molecules, the electrostatic forces may be of more importance in accounting for the molecular conformation, and it must be remembered that only electrostatic forces are likely to account for interactions acting over moderately large distances. Recent studies by Tanford have helped in the consideration of these electrostatic forces.[8]

Another type of intrachain linkage that may exist in proteins is the thiazoline ring, found in the peptide antibiotic bacitracin A. This linkage can be formed by a rearrangement of a peptide bond adjacent to a cysteinyl residue:



The same type of rearrangement could occur with a peptide bond adjacent to a seryl residue, to form an oxazoline ring where the sulfur atom is replaced by oxygen. Neither of these ring structures has been shown to exist in proteins.

The behavior of amino acids, peptides, and proteins is strongly dependent on the ionic or dipolar-ionic con-

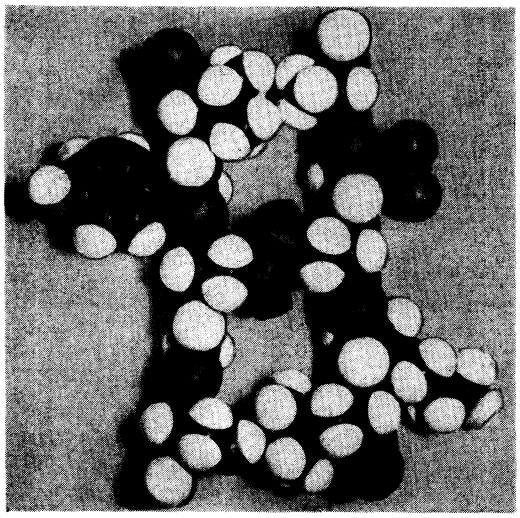


FIG. 4. Photograph of scale model of interchain disulfide linkage with antiparallel chain directions. The peptide chain on the left has the sequence -CO-leu-lys-cyS-asp-ala-NH- (bottom to top), and the chain on the right has the sequence -CO-val-tyr-cyS-ala-ser-NH- (top to bottom). The two peptide chains are about 6.3 A apart at the disulfide bond [from B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 258].

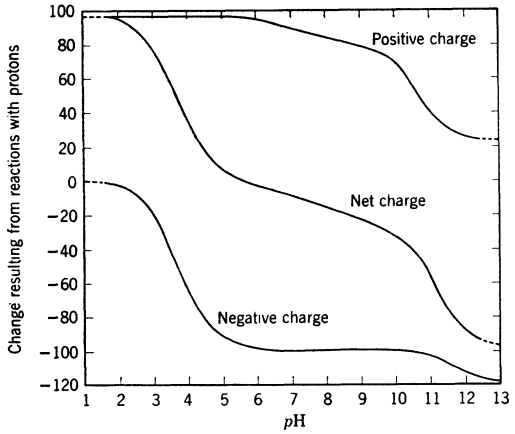
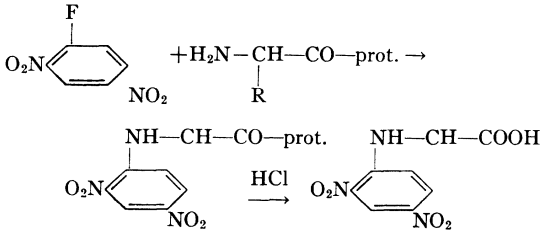


FIG. 5. Titration curve of bovine serum albumin. Also represented is the total positive charge and the total negative charge. These values are computed for 1 C-terminal carboxyl residue, 99 aspartyl and glutamyl residues, 19 tyrosyl residues, 1 cysteinyl residue, 16 histidyl residues, 1 N-terminal amino residue, 57 lysyl residues, and 22 arginyl residues.

figuration of these molecules. Each amino acid contains α-NH_3^+ and α-COO^- groups, and the acidic or basic amino acids also contain an additional charged group. When amino acids are combined in peptides or proteins, the α-NH_3^+ and α-COO^- groups are lost through the formation of the peptide bond, and only the terminal NH_3^+ and -COO^- residues can become ionized. But the acidic and basic groups of the amino-acid residues of category two (Table I) remain capable of being ionized, and it is primarily these residues which give the dipolar and ionic properties to the protein molecule. Titration of a protein solution with strong acid or base allows a calculation of the number of available acidic and basic groups and an estimation of their *p*K's. Such a titration curve for bovine serum albumin calculated from the measurements of Tanford[9] is shown in Fig. 5. It can be seen that, when the net charge of the protein is zero, there are approximately 94 positively charged groups (basic groups of histidyl, terminal amino, lysyl, and arginyl residues) and an equal number of negatively charged groups (carboxyl groups of terminal carboxyl, aspartyl, and glutamyl residues). The properties of a protein are greatly influenced by the presence of these many charged groups, and the dipole moment of the uncharged protein is a measure of the symmetry of the distribution of these charges. Studies of the titration curves of many proteins have indicated that most of the acidic and basic amino-acid residues are available for reaction. [Rice (p. 69) discusses some of the difficulties in the analysis of titration curves of proteins and other polyelectrolytes.] In certain instances, however, some of these residues appear to be unavailable for reaction unless the protein is taken to extremes of *p*H.[8] Two well-established examples of this behavior are seen in hemoglobin and in ribonuclease. In hemoglobin, about 36 groups (probably 18 carboxyl groups of aspartyl and glutamyl residues, and 18 ϵ-amino groups of lysyl residues) are found to be unreactive at neutral *p*H values,[10] whereas in ribonuclease, 3 of the 6 tyrosyl residues are found to be unreactive at their normal *p*K.[11]

One of the most exciting developments concerning the structure of proteins has been the evolution of methods which have led to the identification of the sequence of amino-acid residues in the peptide chains of a number of proteins. These developments have been made possibly largely through the work of Sanger and his colleagues who blazed the way by determining the sequence in insulin, a protein of about 6000 molecular weight.[12] Sanger's approach was to develop a method for the identification and estimation of the N-terminal residues of proteins and peptides. The reaction of the N-terminal residue with 1.2,4-fluorodinitrobenzene (FDNB) leads to the yellow dinitrophenyl (DNP) compound,



This reaction is carried out under mild (slightly alkaline) conditions where the peptide bonds are quite stable. Acid hydrolysis of the resulting DNP compound leads to a mixture of the various amino acids present in the protein and the DNP-amino acid from the N-terminal residue. The DNP-amino acids are reasonably stable under the conditions of acid hydrolysis, and can usually be obtained in good yield. There are certain DNP-amino acids, for example DNP-proline and DNP-cysteine, which are considerably less stable than others, however, and substantial corrections must be made for the destruction of such DNP-amino acids during hydrolysis.

If the DNP-protein is only partially hydrolyzed, DNP-peptides can be isolated, and subsequent complete hydrolysis of these purified DNP-peptides reveals the nature of the amino-acid residues and the N-terminal residue; whereas a partial hydrolysis of the purified DNP-peptide can lead to the arrangement of the residues in the N-terminal peptide. By this method, the four or five residues adjoining the N-terminal residue can be arranged in sequence. In the studies on insulin, the N-terminal sequences DNP-phe-val-asp-glu- and DNP-gly-ileu-val-glu-glu- were identified in this way. Also, this same method of attack has been used to determine the residue sequence in various purified peptides obtained from either acid or enzymatic hydrolysis of the protein. After a large number of such peptides were completely identified (about 65 in the case of the *B*-chain of insulin), it was found that there

TABLE II. Peptides identified in hydrolyzates of fraction B of oxidized insulin.[a]

Dipeptides from acid and alkaline hydrolyzates

H-phe-val-OH H-his-leu-OH H-his-leu-OH H-ala-leu-OH H-gly-glu-OH H-thr-pro-OH
H-val-asp-OH H-leu-cySO$_3$H-OH H-leu-val-OH H-leu-val-OH H-glu-arg-OH H-lys-ala-OH
H-asp-glu-OH H-cySO$_3$H-gly-OH H-val-glu-OH H-val-cySO$_3$H-OH H-arg-gly-OH
H-glu-his-OH H-ser-his-OH H-glu-ala-OH H-cySO$_3$H-gly-OH H-gly-phe-OH

Tripeptides from acid and alkaline hydrolyzates

H-phe-val-asp-OH H-leu-cySO$_3$H-gly-OH H-ala-leu-tyr-OH H-gly-glu-arg-OH H-pro-lys-ala-OH
H-val-asp-glu-OH H-ser-his-leu-OH H-tyr-leu-val-OH
H-glu-his-leu-OH H-leu-val-glu-OH H-leu-val-cySO$_3$H-OH
H-his-leu-cySO$_3$H-OH H-val-glu-ala-OH H-val-cySO$_3$H-gly-OH

Higher peptides from acid and alkaline hydrolyzates

H-phe-val-asp-glu-OH H-ser-his-leu-val-glu-OH H-tyr-leu-val-cySO$_3$H-OH H-thr-pro-lys-ala-OH
H-phe-val-asp-glu-his-OH H-ser-his-leu-val-glu-ala-OH H-leu-val-cySO$_3$H-gly-OH
H-glu-his-leu-cySO$_3$H-OH H-his-leu-val-glu-OH
H-his-leu-cySO$_3$H-gly-OH H-leu-val-glu-ala-OH
H-ser-his-leu-val-OH

Sequences deduced from above peptides

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly- -tyr-leu-val-cySO$_3$H-gly- -thr-pro-lys-ala
-ser-his-leu-val-glu-ala- -gly-glu-arg-gly-

Peptides identified in peptic hydrolyzate

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly-ser-his-leu-OH H-leu-val-cySO$_3$H-gly-glu-arg-gly-phe-OH
H-val-glu-ala-leu-OH H-tyr-thr-pro-lys-ala-OH
H-his-leu-cySO$_3$H-gly-ser-his-leu-OH

Peptides identified in chymotryptic hydrolyzate

H-phe-val-asp-glu-his-leu-cySO$_3$H-gly-ser-his-leu-val-glu-ala-leu-tyr-OH H-tyr-thr-pro-lys-ala-OH
H-leu-val-cySO$_3$H-gly-glu-arg-gly-phe-phe-OH

Peptides identified in tryptic hydrolyzate

H-gly-phe-phe-tyr-thr-pro-lys-ala-OH

Structure of the B-(phenylalanyl terminal) chain of insulin

H-phe-val-asp-glu-his-leu- (cyS-) -gly-ser-his-leu-val-glu-ala-leu-tyr-leu-val- (cyS-) -gly-glu-arg-gly-phe-phe-tyr-thr-pro-lys-ala-OH

[a] Taken from F. Sanger, Advances in Protein Chem. **7**, 56 (1952).

was only one sequence which would fit all of the experimental results, assuming that there was a single polypeptide chain of about 30 residues. It was further assumed that no rearrangements of the residues had occurred during the hydrolyses.

The attack outlined in the foregoing is applicable only to single peptide chains devoid of interchain linkages, such as the disulfide link described earlier. Since insulin and most other proteins contain interchain and/or intrachain disulfide linkages, these must be broken before the sequence studies can be undertaken by the foregoing methods. In the case of insulin, oxidation of all of the disulfide linkages to sulfonic-acid groups was carried out with performic acid. This reagent thus converts cystine to cysteic acid (H—cySO$_3$H—OH), a strong acid. It also converts methionine to the corresponding sulfone, and tryptophan to unidentified products. Since insulin contains no methionine or tryptophan, performate oxidation was capable of converting the insulin to two chains, called the *A*-(glycine terminal) chain and the *B*-(phenylalanine terminal) chain, with each half-cystine residue (-cyS-) converted to a cysteic-acid residue (-cySO$_3$H-). The *A*- and *B*-chains were then purified, and the sequences of amino-acid residues in each of the two peptide chains were determined by the method outlined in the following. Table II records a number of the peptides obtained from such a study of the *B*-chain, and Fig. 6 shows the entire sequence for the insulin molecule. The acid hydrolysis used in the determination of the amino-acid sequence in insulin caused the liberation of six moles of ammonia, originally present as the amide groups of the

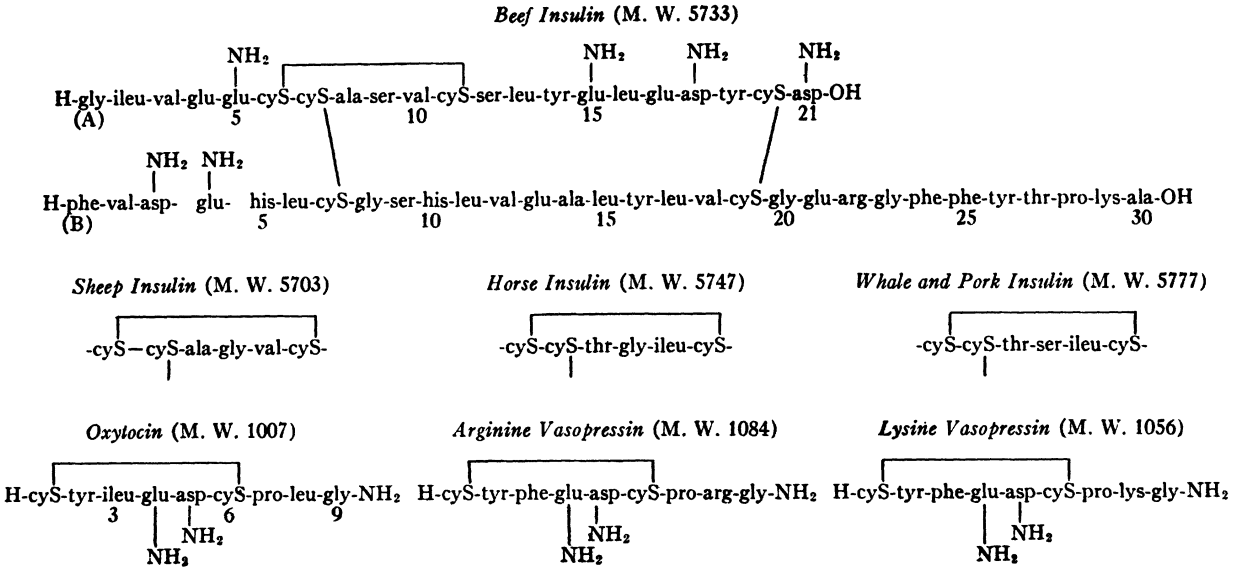


FIG. 6. Amino-acid sequences in insulin and some hormone peptides.

asparaginyl and glutaminyl residues. Reduction of the insulin by lithium borohydride ($LiBH_4$) in tetrahydrofuran converted the ω-carboxyl groups to primary alcohol groups, and acid hydrolysis of the reduced insulin then showed that three asparaginyl residues, three glutaminyl residues, and four glutamyl residues were present in insulin. In order to locate these amide groups on the individual glutamic-acid residues, two methods were used which depend on the fact that enzymatic hydrolysis does not break the ω-amide linkage. The purified peptides obtained after enzymatic hydrolysis could then be tested for amide content, either by studies of their electrophoretic behavior (since the amide derivatives had one negative charge less than the corresponding ω-acid derivatives), or by studies of the extent of ammonia liberation during their acid hydrolysis. In this way, it was found that residues 5, 15, 18, and 21 of the *A*-(glycyl terminal) chain and residues 3 and 4 of the *B*-chain were present in the amide form.

In order to find the distribution of the disulfide bridges, Sanger submitted intact insulin to partial enzymatic hydrolysis. The resulting cystine-containing peptides were then purified and sequences of these peptides were determined. This gave Sanger the final data required to completely identify the sequences and cross-linkages in beef insulin,[13] and resulted in the formula given in Fig. 6. The intrachain disulfide linkage found in the *A*-(glycyl terminal) chain is of special interest. The half-cystinyl residue at position 6 is combined with the position 11 half-cystinyl residue through a disulfide linkage. Studies on horse, sheep, whale, and pork insulin indicate sequences similar to those found in beef insulin, except for the amino acids in positions 8, 9, and 10 of the *A*-(glycyl terminal) chain. These residue differences are all within the hexapeptide disulfide link. Figure 6 also shows the sequences of the insulins of the other species studied. Whale and pork insulin are seen to be identical.

As mentioned in the discussion of the intrachain disulfide linkage, three peptide hormones isolated from the anterior pituitary and synthesized by du Vigneaud —oxytocin, arginine vasopressin, and lysine vasopressin —all contain a similar hexapeptide disulfide link. The formulas of these peptides are shown in Fig. 6 also, and the possible significance of this structure in terms of hormone structure is discussed by Stetten (p. 563).

The methods developed by Sanger and his colleagues for the study of the amino-acid sequence in insulin have subsequently been applied by a number of laboratories to the elucidation of the structure of several other proteins. Other methods of attack involving replacement of the DNP compounds by other types of derivatives have been developed, and the stepwise freeing of N-terminal and C-terminal residues by enzymatic hydrolysis catalyzed by aminopeptidases and carboxypeptidase has proved to be fruitful. Methods have also been developed to replace the performate oxidation procedure for the elimination of disulfide bonds, and these can be applied in the study of proteins containing tryptophan and other easily oxidized amino acids. A short summary of these methods for amino-acid sequence determinations can be found in a recent review by Anfinsen and Redfield.[14]

Figures 7–9 summarize some of the more complete amino-acid sequence studies on other proteins and large peptides. The most complete amino-acid sequence study of a series of related hormones is the work on the melanocyte-stimulating hormones (MSH) and adrenocorticotropic hormones (ACTH). Figure 7 records the

Beef Melanotropin (*Seryl-β-MSH*) (M.W. 2134)

H-asp-ser-gly-pro-*tyr*-lys-*met-glu-his-phe-arg-try-gly*-ser-*pro*-pro-lys-asp-OH
5 10 15 18

Pork β-Melanocyte-Stimulating Hormone (*Glutamyl-β-MSH*) (M.W. 2176)

H-asp-glu-gly-pro-*tyr*-lys-*met-glu-his-phe-arg-try-gly*-ser-*pro*-pro-lys-asp-OH
5 10 15 18

Pork α-Melanocyte-Stimulating Hormone (*α-MSH*) (M.W. 1665)

CH_3-CO-ser-*tyr*-ser-*met-glu-his-phe-arg-try-gly*-lys-*pro*-val-NH_2
5 10 13

Sheep α-Corticotropin (*α-ACTH*) (M.W. 4540)*

H-ser-*tyr*-ser-*met-glu-his-phe-arg-try-gly*-lys-*pro*-val-gly-lys-lys-arg-arg-pro-val-
5 10 15 20

NH_2
|
-lys-val-tyr-pro-ala-gly-glu-asp-asp-glu-ala-ser-glu-ala-phe-pro-leu-glu-phe-OH
21 25 30 35 39

ACTH Preparation	Residue No. 25 26 27 28 29 30 31 32 33
Sheep α-corticotropin	-ala-gly-glu-asp-asp-glu-ala-ser-glu-(NH_2)
Pork corticotropin-A (Cyanamid)	-asp-gly-ala-glu-asp-glu(NH_2)-leu-ala-glu-
Pork corticotropin-A (Armour)	-gly-ala-glu-asp-asp-glu-leu-ala-glu-

FIG. 7. Amino-acid sequences in melanocyte-stimulating and adrenocorticotropic hormones. Sheep α-corticotropin contains one more amide group, probably in position 27, 28, or 29. Beef α-corticotropin appears from preliminary structural investigations to be identical with sheep α-ACTH. All of these ACTH preparations seem to have no loss of biological potency when residues 29 to 39 (11 residues) are removed by limited enzyme hydrolysis.

results of extensive studies on these hormones in some five or six laboratories. Recent reviews of this work have been presented by Li,[15–17] Harris,[18] and by Anfinsen and Redfield.[14] The amino-acid sequences of all of the MSH and ACTH materials studied to date have shown almost identical sequences . . .-*tyr*-.X.-*met-glu-his-phe-arg-tyr-gly*-.Y.-*pro*-. . . . This sequence has been shown in italics in Fig. 7, and is seen to occur at different distances from the N-terminal residue in the β-MSH preparations (residues 5–15) and in α-MSH and the various ACTH preparations (residues 2–12). The residues X and Y in this unique sequence are seen to be lysyl or seryl, and seryl or lysyl, respectively (see Stetten, p. 563). The two pork corticotropins-A, as prepared in the Cyanamid and in the Armour Laboratories, are known to differ with respect to the total content of amide groups (none for the Armour product, and one for the Cyanamid product). The sequence differences in residues 25–28 of the two corticotropin-A preparations have not been definitely shown to indicate different sequences in the two preparations, since technical difficulties may have occurred in the sequence determination. The sheep α-corticotropin preparation of Li differs from the corticotropin-A, in that Li's preparation contains two amide groups (one is glutaminyl residue 33, and one not yet located, but probably in position 27, 28, or 29), and also one more seryl and one less leucyl residue. It may be noted that the component amino acids in positions 25–28 are the same in all three ACTH preparations, but differ in order. The different location of the amide residue (position 33 in sheep α-corticotropin and 30 in pork corticotropin-A) appears to represent a real difference in sequence. Preliminary structural studies by Li have indicated that the amino-acid sequence for beef α-corticotropin is identical with that for sheep α-corticotropin.[17]

The two β-MSH preparations are seen to differ only in the second residue, seryl for the beef hormone and glutamyl for the pork hormone. These materials are, therefore, often referred to as seryl-β-MSH and glutamyl-β-MSH. The acetylated amino group of the N-terminal seryl residue in α-MSH has only recently been identified,[18] and its presence is rather unusual. The presence of this group, as well as of the amide group of the C-terminal valyl residue in α-MSH, causes a drastic change in the hormonal activity of this compound as compared with that of the longer α-ACTH peptide.[18] This effect, as well as other aspects of the relation between the structure of these hormones, and their physiological activities, is discussed by Stetten (p. 563).

The amino-acid sequence in glucagon, a small protein involved in glucose metabolism, has recently been determined and is shown in Fig. 8. This protein, like the ACTH and MSH preparations, contains no cystinyl residues. Unlike these hormones, however, the sulfur-containing methionyl residue is present in glucagon. The relationship between glucagon and insulin in glucose metabolism is discussed by Stetten (p. 563). The

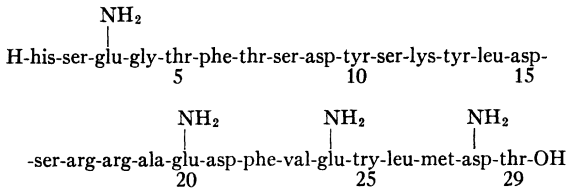


FIG. 8. Amino-acid sequence in glucagon (M. W. 3647).

sequence arrangements in all of the unrelated proteins so far evaluated show no more similarity than could be expected from simple probability considerations, and none of them indicates that repeating sequences are important elements in the structure of these small globular proteins. Whether or not repeating sequences occur in the larger globular proteins is as yet unknown. Some strong evidence of repeating elements has, however, been found in certain fibrous-protein structures, and in histone and protamine, proteins found associated with nucleic acid in the nucleoproteins.[14]

Ribonuclease is the most complicated protein in which an almost complete amino-acid sequence has been determined. This enzyme, containing 124 amino-acid residues and four disulfide bonds (molecular weight 12 000), contains more than twice the number of residues found in insulin. The presently known sequence as shown in Fig. 9 has recently been discussed by Hirs *et al.*[19] and by Anfinsen.[14,20] About 23 residues remain to be located definitely in the sequence. No unusual partial sequences seem to occur, with the possible exception of a number of repeats (-ala-ala-ala- in positions 4–7, and -met-met- in positions 29–30). These partial sequences can be compared with the unusual sequence -phe-phe-tyr- in the *B*-chain of insulin (positions 24–26).

Short, partial-residue sequences have been determined for a large number of proteins. A few of the more interesting partial sequences are reported in Figs. 10 and 11. The partial sequence shown for beef cytochrome-*c* (Fig. 10) is interesting in that this sequence contains the residues to which the porphyrin-*c* prosthetic group is attached. The exact attachment of the porphyrin-*c* residue may be either that shown in Fig. 10, or that of an otherwise identical structure where the porphyrin-*c* residue is rotated about the plane of the page by 180°. The corresponding partial sequences in a series of cytochrome-*c* preparations from sources other than beef are also shown in Fig. 10. Here, alanyl residues are sometimes replaced by seryl or glutamyl residues, lysyl residues by arginyl residues, valyl residues by lysyl residues, and glutaminyl residues by threonyl residues. The italicized sequence *-cyS-.A.-.B.-cyS-his-thr-val-glu-* has been found to occur in this order in each of the six cytochrome-*c* molecules. This study, involving work by Theorell, Tuppy, and others, has been reviewed by Tuppy[21] and by Anfinsen and Redfield.[14] The partial sequences shown for lysozyme and for human serum albumin (Fig. 11) are very incomplete. Many disulfide bonds are known to be present in these molecules, and serum albumin is somewhat unusual in that it contains a single cysteinyl residue. Many other partial sequences have been found for egg-white lysozyme,[14,22] but a study of these sequences shows that certain of them must be spurious, and it has not been possible to locate the various partial sequences in relation one to another. Anfinsen and Redfield[14] suggest that the study of lysozyme illustrates the limit of structural information that can be obtained by acid hydrolysis alone, and that more specific degradative methods must be applied before the complete amino-acid sequence can be obtained

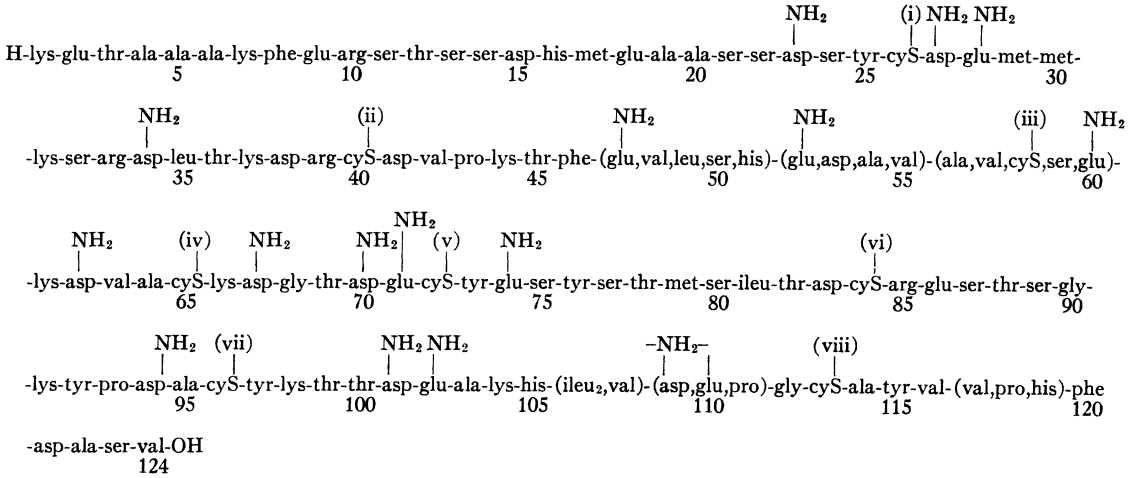
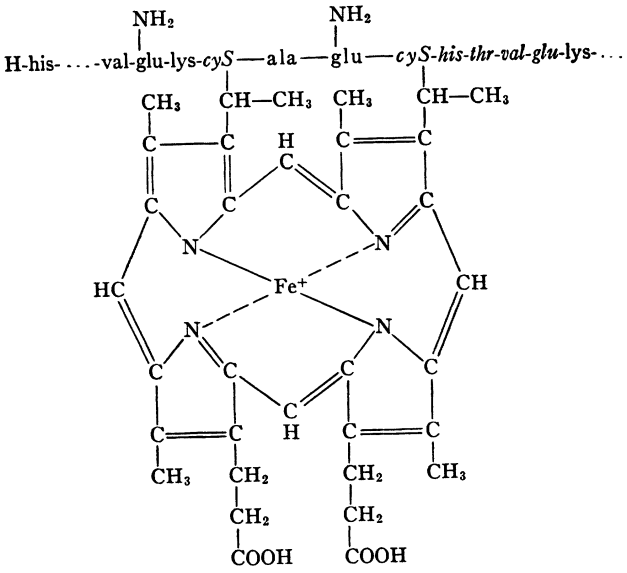


FIG. 9. Amino-acid sequence in ribonuclease (M. W. 12 000). Where the composition and position of a group of residues is known, but not the relative positions of the individual residues within the group, the listing of the group is enclosed in parentheses. The half-cystinyl residues are identified by roman numerals. Work of Anfinsen[20] has shown that the four disulfide bridges in ribonuclease are paired as follows: i–vi, iii–vii, ii–viii and iv–v (after studies by C. H. W. Hirs *et al.*[19]).

Other species of serum albumin have been studied, and it has been shown that bovine serum albumin, while containing the same N-terminal residue (H-asp-), has threonyl as the second residue and ends with a different C-terminal sequence, possibly -(ala, leu, thr, val, ser)-ala-OH.[14] Porter[23] has obtained a fragment of the bovine serum-albumin molecule after mild chymotryptic hydrolysis which seems to represent about one-fifth of the entire molecule (molecular weight about 12 000).

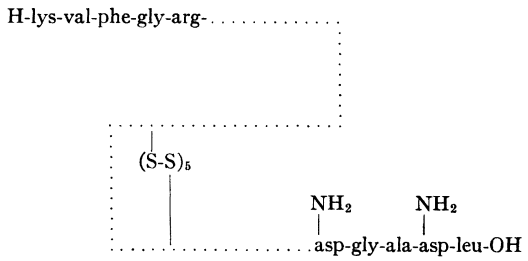
Beef Cytochrome-c (M.W. 13 100)



Preparation	*Observed residue sequence*
Beef	-val-glu-lys-*cyS*-ala-glu-*cyS*-*his-thr-val-glu*-lys-
Horse and pork	-lys-*cyS*-ala-glu-*cyS*-*his-thr-val-glu*-lys-
Salmon	-val-glu-lys-*cyS*-ala-glu-*cyS*-*his-thr-val-glu*-
Chicken	-val-glu-lys-*cyS*-ser-glu-*cyS*-*his-thr-val-glu*-
Silkworm	-val-glu-arg-*cyS*-ala-glu-*cyS*-*his-thr-val-glu*-
Yeast	-phe-lys-thr-arg-*cyS*-glu-leu-*cyS*-*his-thr-val-glu*-

FIG. 10. Partial amino-acid sequence in cytochrome-*c*. The exact attachment of the porphyrin-*c* residue may be either the residue shown or the structure where the porphyrin-*c* residue is rotated by 180° [from H. Tuppy in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 71].

Egg-white Lysozyme (M.W. 14 900)



Human Serum Albumin (M.W. 68 000)

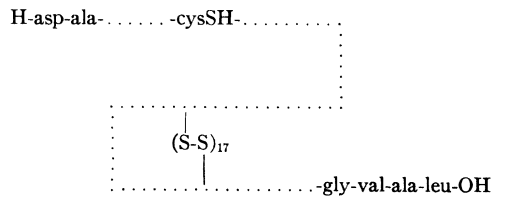


FIG. 11. Partial amino-acid sequences in lysozyme and serum albumin [from C. B. Anfinsen and R. R. Redfield, Advances in Protein Chem. 11, 1 (1956)].

This fragment appeared to have very nearly the same configuration as part of the original albumin molecule, since it effectively inhibited the reaction of a specific antiserum to bovine serum albumin with its antigen (see Kauzmann, p. 549). This fragment contained a single N-terminal amino acid (H-phe-), the unique cysteinyl residue, and a single disulfide bond.

Further structural studies of amino-acid sequences are reported in other contributions to the Paris *Symposium on Protein Structure*.[24] Considerable sequence data exist for many other proteins, notably papain, growth hormone, trypsin, chymotrypsin, pepsin, hemoglobin, and tobacco mosaic virus.

It is necessary to always bear in mind the difficulties in the isolation of purified and homogeneous components in their native state. Proteins occur in nature as complex mixtures, and are often found in the presence of highly charged polysaccharides and/or other macromolecules. These naturally occurring colloidal mixtures are often enclosed by membranes of varying stability. The extraction of a particular protein from such a mixture is always difficult, and is often impossible without the use of procedures which may cause permanent changes in the configuration or even in the covalent linkages of the resulting protein preparation. A number of useful methods for the isolation and purification of such protein systems are currently available for the separation of these components. These methods include precipitation and differential extraction by high concentrations of salts or organic solvents, adsorption and partition between immiscible solvents, as well as such physical methods as electrophoresis and ultracentrifugation. A discussion of these methods cannot be undertaken here, but it is important to remember that the purified components isolated by these methods may

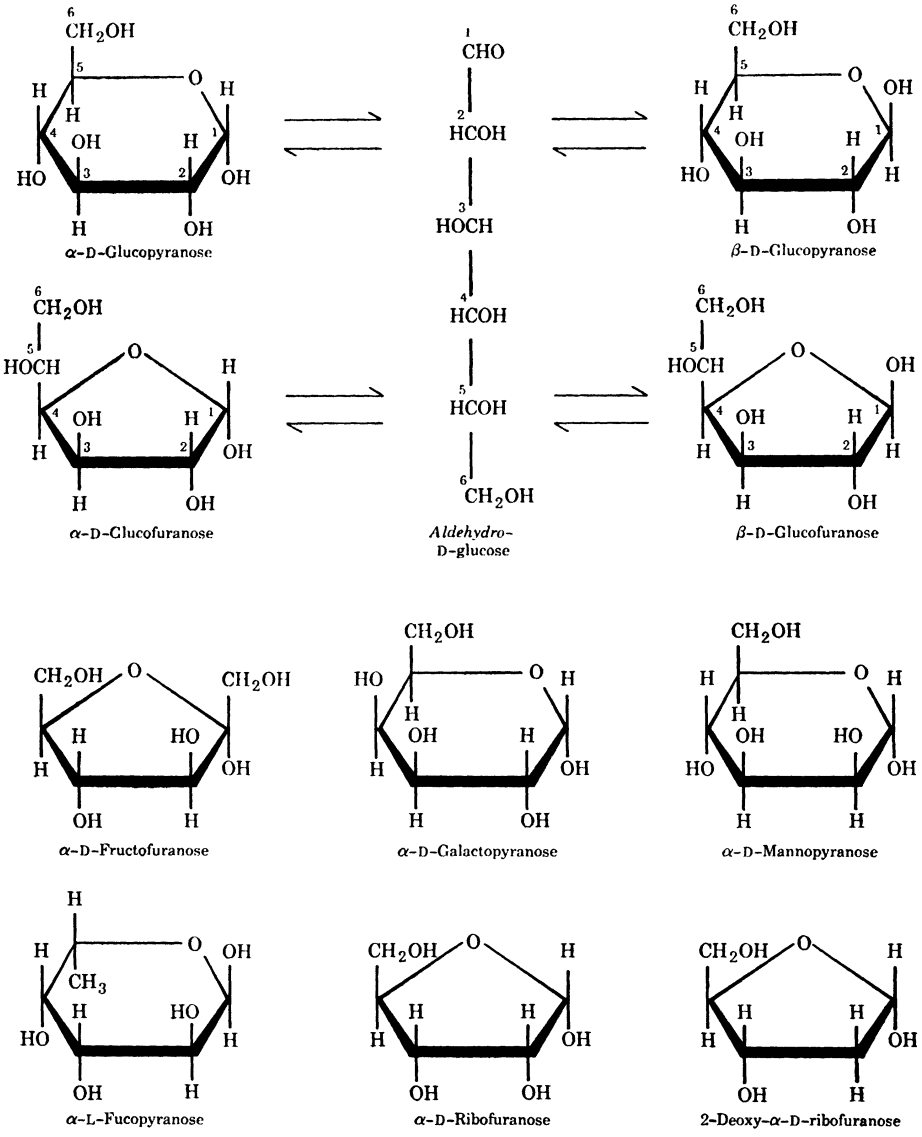


FIG. 12. Configuration of some monosaccharides.

not be truly homogeneous, and artifacts may be introduced by the isolation procedures.

Tests for the homogeneity of the purified components must be evaluated critically. Physical methods are most useful; they include studies of behavior in the ultracentrifuge and in electrophoresis, measurements of solubility (akin to the phase-rule methods of constant melting or boiling points as applied in the case of smaller molecules), and determination of the distribution coefficient between various solvents. Chemical methods showing the constancy of composition (in regard to the individual amino-acid residues, for example) are also of much value. Biological assay methods measuring activities ascribed to various components and quantitatively evaluating the activities of the purified components in terms of the activity of the native tissues are of special importance.

CARBOHYDRATES

Carbohydrates comprise another of the major groups of naturally occurring organic materials. They are often found in combination with proteins as glycoproteins and mucoproteins, especially in the higher animals. They also occur in combination with lipids to give cerebrosides, and in combination with purines and phosphoric acid to give nucleotides and nucleic acids. They form the structural elements of plants, bacteria, and certain animals in the forms of cellulose, chitin, and other high molecular-weight polysaccharides. They provide a means for the storage of chemical energy in the forms of amylose and amylopectin in plant starches, and glycogen in animals.

Although most monosaccharides are fairly stable when in the crystalline condition, they undergo many

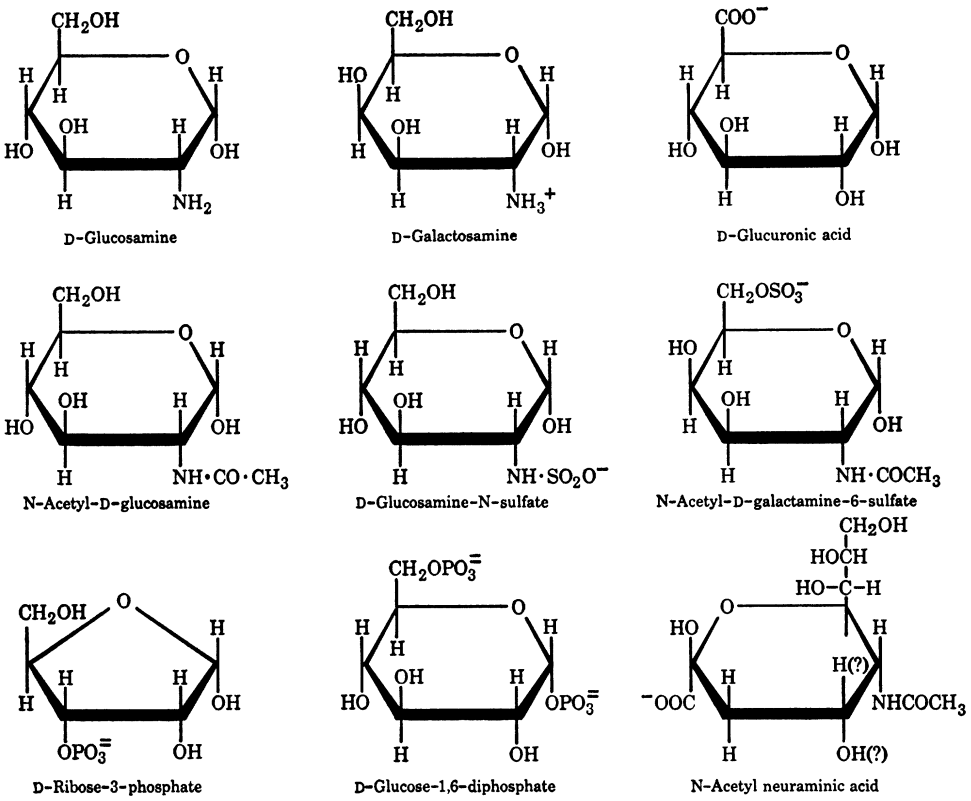


FIG. 13. Configuration of some charged sugar derivatives.

transformations when dissolved in water, particularly in the presence of acids or bases. Although the chemical formulas of the monosaccharides are often shown in a linear form, much evidence suggests that, in the main, they exist in the form of five- and six-membered rings called the furanose and pyranose forms. Each of these forms exists as α- and β-isomers, termed anomers, and interconversions between anomers and between ring isomers occur even under the mildest possible conditions of acidity and temperature. One of the simplest methods for demonstrating and studying the equilibria between the various forms is by the measurement of changes of optical rotation with time ("mutarotation"), which can easily be observed with freshly prepared sugar solutions. These various equilibria (for D-glucose) are illustrated in Fig. 12. The linear forms are shown in the Fischer formulation where the carbon atoms must be thought of as tetrahedrons, bonded together by angles projecting into the page and with the H— and —OH groups projecting out from the page. The ring forms shown in the Haworth formulation are to be viewed as a perspective representation with the heavy bonds extending out from the page. Side chains in the Haworth formulas are written according to the Fischer convention. The rings, shown here as planar, are oversimplified, since the valence angles in a single coplanar ring would be appreciably greater than those in a "strainless" structure having valence angles of 109°. The usual conformation of the pyranose ring is probably a chain form, but the free sugars in solution probably can occur to some extent in any of the possible conformations, since there is a spontaneous equilibrium with the aldehyde form. Derivatives in which the ring is fixed (because of substitution on C-1) probably are stabilized in the chair form.[25]

The aldehydo-hexoses such as D-glucose and D-galactose contain four asymmetric carbon atoms, while the keto-hexoses such as D-fructose and aldehydo-pentoses such as D-ribose contain three. Thus, many optical isomers exist (16 for the aldehydo-hexoses and 8 for the keto-hexoses and aldehydo-pentoses), and a number of these are found in various natural products. The ring structures contain an additional center of asymmetry at carbon atom 1, and the anomers of each ring structure are usually designated α- when the hydroxyl group of carbon atom 1 is *cis*- to the hydroxyl group at carbon atom 2, and β- when these two hydroxyl groups are *trans*-. In the case of 2-deoxy-D-ribose, the anomeric forms are designated like the D-ribose forms. α-D-mannose does not follow the (*cis*-) (*trans*-) rule given in the foregoing, and the commonly used nomenclature of Hudson is to be found in Pigman's review.[25] The designation L- indicates that the asymmetric carbon atom most remote from the reference group (e.g., aldehyde, keto, carboxyl, etc.) has the same configuration as L-glyceraldehyde (see Fig. 1). The L- and D-forms differ in the configuration of all asymmetric carbon atoms. In addition to the three monosaccharides

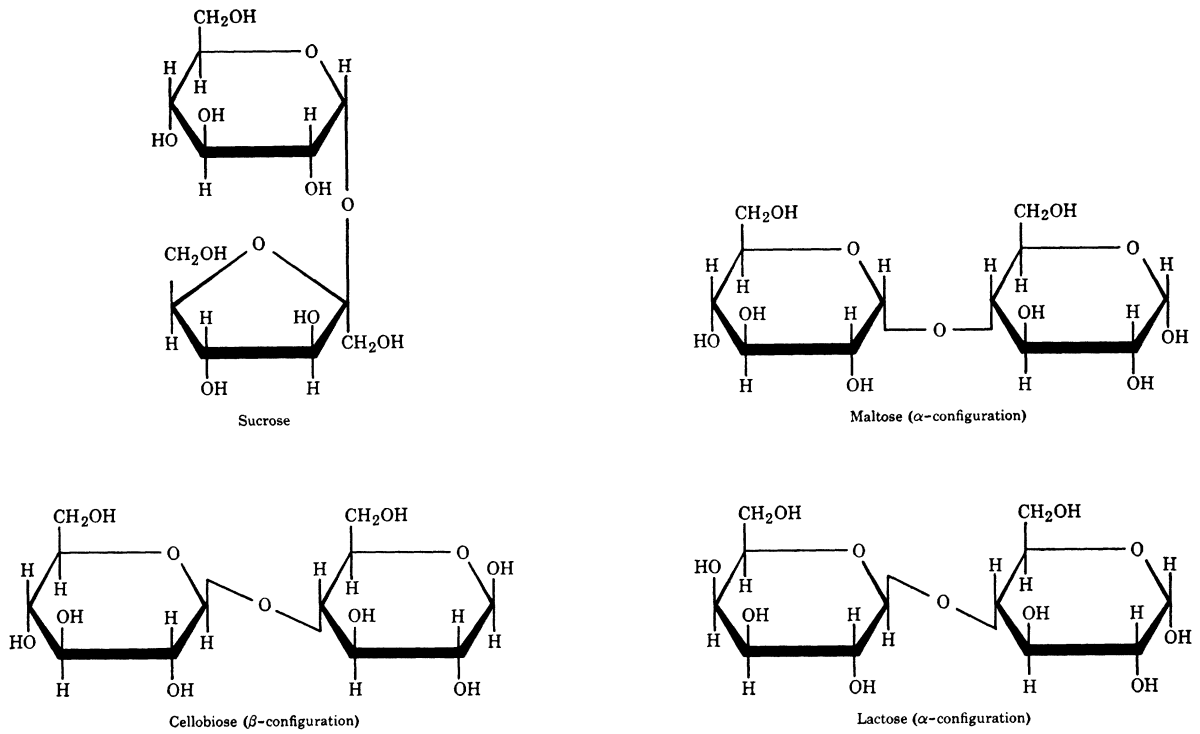


FIG. 14. Configuration of some disaccharides.

mentioned above, those most commonly found in bacterial and animal sources include D-galactose (like D-glucose with the configuration of C-4 reversed) and D-mannose (like D-glucose with configuration of C-2 reversed). Three other monosaccharides of major interest include L-fucose (6-deoxy-L-galactose), L-rhamnose (6-deoxy-L-mannose), and 2-deoxy-D-ribose, where the designation -deoxy- indicates that the appropriate —OH is replaced by —H.

Haworth formulas for most of the aforementioned monosaccharides are shown in Fig. 12. Each of these sugars would exist in the same sorts of conformations as are shown for D-glucose, and the ring structures commonly found in oligo- and poly-saccharides are portrayed in the diagrams. Only the α-anomer is depicted. The relative amounts of the four-ring conformations which will exist at equilibrium in aqueous solutions of the various monosaccharides vary considerably. In the case of neutral solutions of D-glucose, nearly two-thirds of the sugar seem to exist as β-glucopyranose, and just over one-third as α-glucopyranose. The two glucofuranose forms and the *aldehydo*-glucose form are thought to exist in minor amounts. In the case of D-mannose, over two-thirds exist as α-mannopyranose and just under one-third as β-mannopyranose. Solutions of D-ribose probably contain considerable quantities of the furanose and *aldehydo*- forms, and the mutarotation is complex, and exhibits a minimum.

There is also a number of biologically important derivatives of the monosaccharides, a few of which are listed in Fig. 13. Certain of these derivatives contain ionizable groups. Thus, glucosamine and galactosamine provide primary amine groups with a pK_a near 7.8. Glucuronic acid and galacturonic acid provide carboxyl groups with pK_a about 3.3; N-acetyl-neuraminic acid (sialic acid) has a stronger carboxyl group with pK_a 2.7; and the phosphate esters have two potential hydrogen ions, with pH_a values from 0.9 to 1.5 and from 5.9 to 6.3. The sulfate group of the amino-N-sulfate and the sulfate esters have a very strongly acidic hydrogen with $pK_a < 1$. The phosphate esters of the monosaccharides are products of intermediary metabolism, discussed in a number of other papers (see Lehninger, p. 136; Calvin, p. 147; Roberts, p. 170; Meister, p. 210).

The monosaccharide units of disaccharides may be alike as in maltose and cellobiose, or different as in sucrose and lactose. The hydrolysis of maltose and of cellobiose yields two molecules of D-glucose, whereas sucrose yields D-glucose and D-fructose, and lactose yields D-glucose and D-galactose. The monosaccharide residues are combined through an oxygen bridge of the hemiacetal hydroxyl to a second hydroxyl from another residue. The possible number of combinations of two monosaccharide units is large, since there are usually four or five free hydroxyl groups in the monosaccharide, and the oxygen bridge can come from either the α- or β-anomer of the other sugar unit. Figure 14 shows the

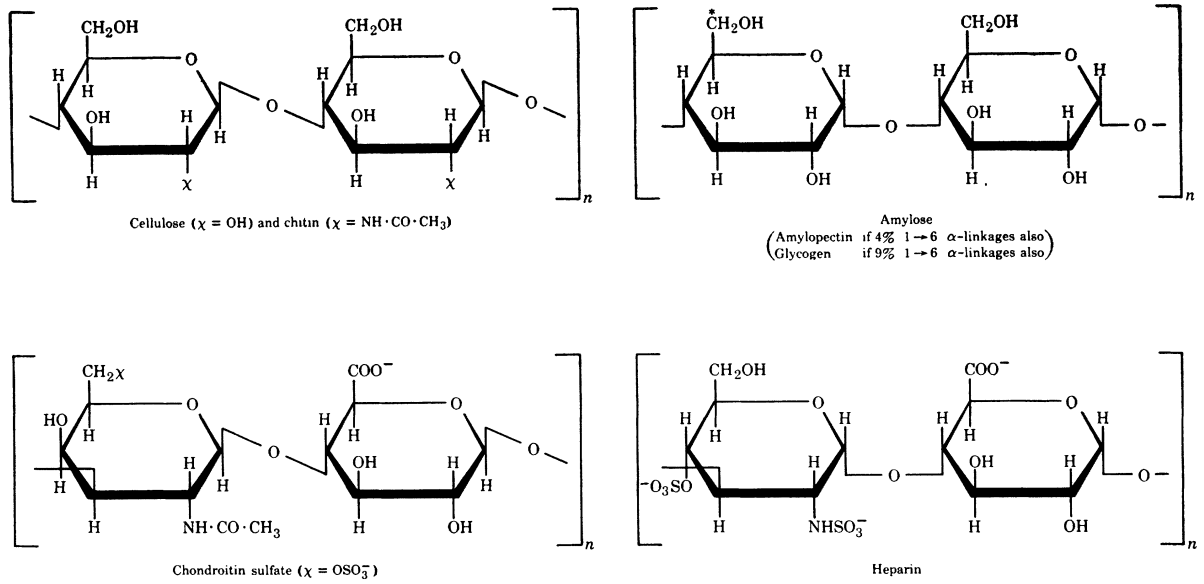


FIG. 15. Structural elements in selected polysaccharides.

structure of four common disaccharides, and illustrates some of the linkages possible. In sucrose, the two hemiacetyl hydroxyls are combined through an oxygen bridge between the two anomeric carbons, with the D-glucose being in the α-pyranose configuration, and the D-fructose in the β-furanose form. In lactose, the hemiacetyl hydroxyl (C-1) of β-D-galactopyranose bridges to the C-4 hydroxyl of D-glucopyranose. Both maltose and cellobiose contain 1 → 4 oxygen bridges between D-glucopyranose units, but in maltose there is an α-linkage while in cellobiose there is a β-linkage.

There are numerous polysaccharides of biological importance, with monosaccharide units or their derivatives as the polymer unit, often repeating either a single unit, ...A-A-A-A-..., or alternating two units, ...A-B-A-B-A-B-.... The type of glycosidic linkage between these units has the same diversity as was found in the disaccharides. There are certain of the polysaccharides that show a more complex structure owing to branching of the polymer chain by means of linkages through a third hydroxyl group of the polymer unit. Usually, the naturally occurring polysaccharides contain many hundreds of residues. The lability of the linkages makes the isolation and purification of undegraded materials difficult, and many polysaccharide preparations thus show molecular weights and polydispersity not characteristic of the native product.

Cellulose and amylose provide two of the simplest polysaccharide structures (Fig. 15). Both of these polymers are made up of D-glucose residues in the pyranose conformation, with 1 → 4 linkages between the units. In cellulose, the pyranose ring has the β-configuration and each linkage is like that in the disaccharide cellobiose; in amylose, the rings are in the α-configuration and each glycosidic linkage is like that in maltose. Chitin resembles cellulose, and is a polymer of N-acetyl-D-glucosamine units linked by a 1 → 4 β-glycosidic bond. This 1 → 4 β-glycosidic linkage leads to polymers of low solubility, whereas the 1 → 4 α-glycosidic linkage gives polymers of much higher solubility, owing to a spiraling of the macromolecule in a helix-like fashion induced by this type of bond.

Amylopectin and glycogen have a basic structure like amylose, but about five percent of the residues branch by means of 1 → 6 α-glycosidic bonds in the amylopectins, and about nine percent in the glycogens. The branched structure leads to solutions of much lower viscosity than would occur with amylose macromolecules of the same molecular weight. Detailed structures of glycogen and amylopectin have been established by the use of specific enzymes which differentiate between the 1 → 4 and 1 → 6 linkages, and Fig. 16 shows a representation of a segment of a glycogen molecule as indicated by such studies.

The aforementioned polysaccharides have been made up of uncharged polymer units. A more reactive class of polymers is made up from sugar derivatives. Less is known of their detailed structure, because the high charge density in the polymer makes the application of physicochemical methods much more difficult, and because the modifying groups are split from the polymerizing unit by many of the same reagents that are used to degrade the polysaccharide. Among the more important materials of this type, one finds hyaluronic acid, heparin, and chondroitin sulfate. These highly reactive polysaccharides are found in various animal tissues, and the structures now thought most likely are recorded in Fig. 15. None of these structures has been

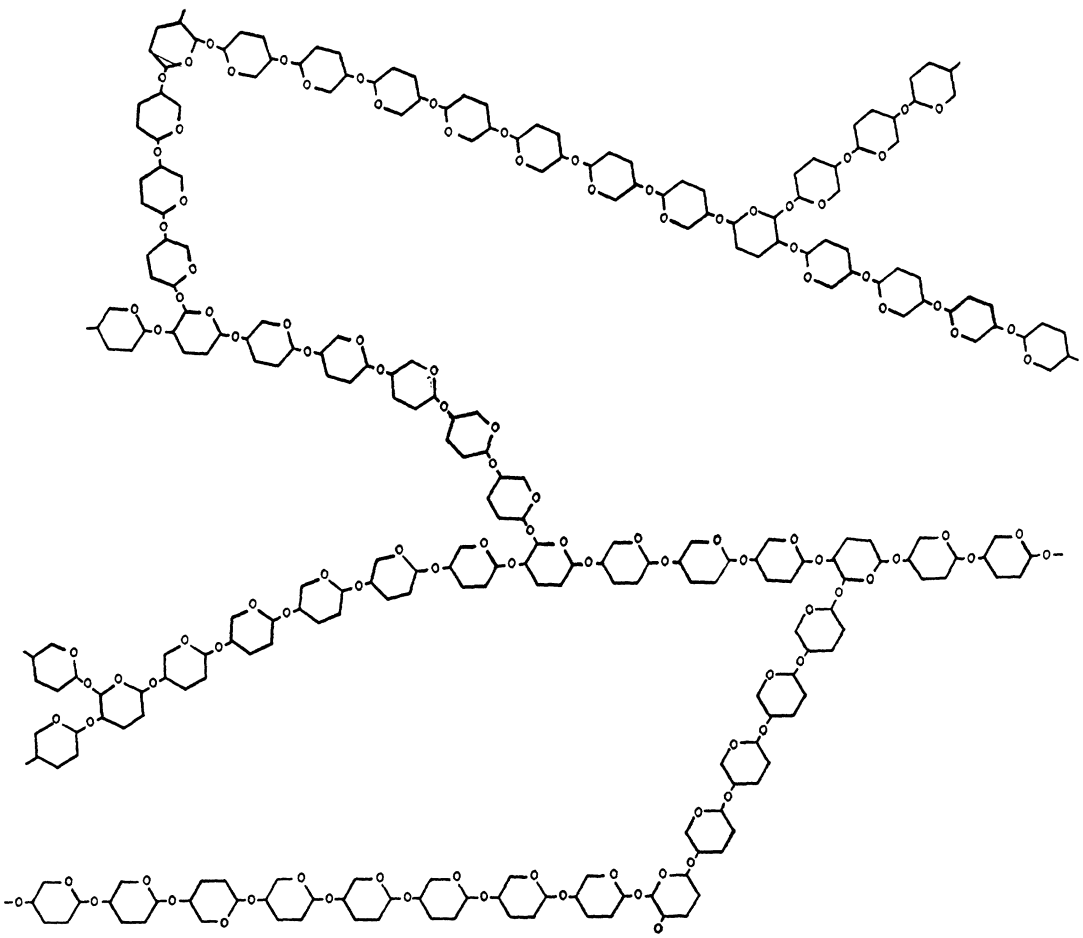


FIG. 16. Representation of a segment of a glycogen or amylopectin macromolecule.

definitely established at the present time. Other polysaccharides of this general type are to be found in bacterial systems, but none of these has structures that have been completely elucidated.

In the glycoproteins and mucoproteins, one finds large amounts of carbohydrate combined with protein.[26,27] Thus, the α_1-acid glycoprotein, a homogeneous crystallized material contains about 42% of carbohydrate with the following composition (expressed as moles per mole of glycoprotein, molecular weight 45 000): N-acetyl-D-glucosamine, 32; N-acetyl-nuraminic acid, 16; D-galactose, 18; D-mannose, 18; L-fucose, 2. Recent studies by Eylar indicate that most of this carbohydrate moiety can be removed in association with a small peptide fragment containing glutamyl, seryl, threonyl, and isoleucyl residues. Earlier studies by E. Smith showed that the smaller carbohydrate moiety of serum γ-globulin could be removed in association with a peptide fragment containing glutamyl, aspartyl, and seryl residues. These interesting results indicate that covalent linkages exist between the carbohydrate residues and certain of the aforementioned amino-acid residues.

LIPIDS

The classification "lipids" is taken to embrace the "fatty acids," all actual or potential esters of fatty acids, and often includes other materials soluble in "fat solvents," such as triterpenes, carotenoids, and fat-soluble vitamins. This large field is only touched upon here, since a recent short review by Lovern,[28] as well as the comprehensive reference book by Deuel[29] provide adequate background material. The simple fundamental lipid structures are often found to occur in complex structures of intermediate molecular weight, but to date no really high molecular-weight lipids have been discovered. On the other hand, these complex lipid molecules are often found in association with polysaccharides and/or proteins to form high molecular-weight materials.[30]

Although complex lipids often contain no charged groups (e.g., triglycerides, cholesterol, cholesterol

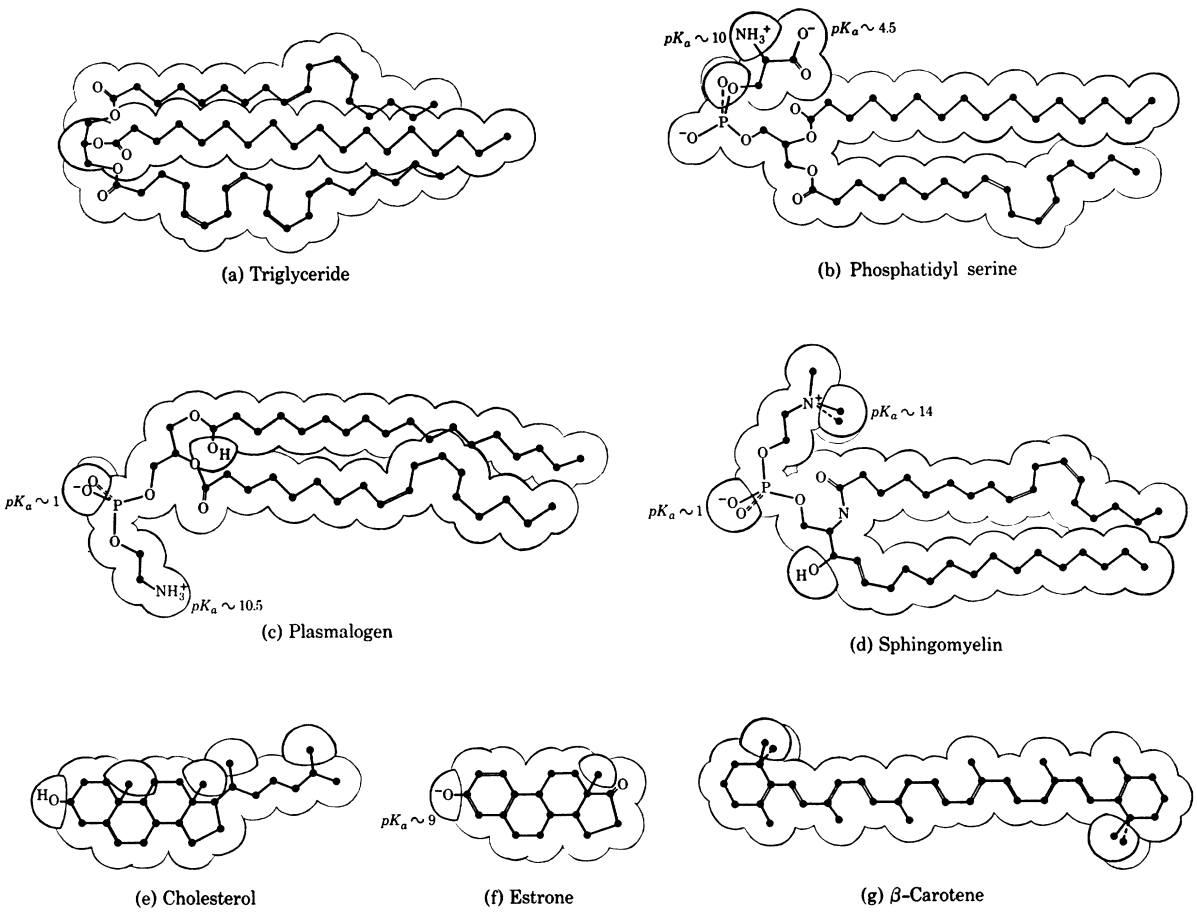


FIG. 17. Some selected lipid structures. The pK_a shown indicates roughly the acidity of the charged groups.

esters, etc.), there are lipid structures containing primary amine, tertiary amines, phosphoric-acid groups, phenolic hydroxyl groups, etc. It can easily be imagined that the lipids owe much of their importance in biological systems to the fact that they represent classes of materials which can bridge the gap from water soluble to water insoluble phases without necessitating a sharp discontinuity. They are often thought to be concentrated in biological membranes and interfaces. Lipids also provide the most compact form of storage of chemical energy.

The term "fatty acid" is not easy to define accurately, but is usually taken to include all of the straight-chain members of the acetic-acid series of carboxylic acids, and a number of naturally occurring unsaturated straight-chain acids (e.g., oleic, linoleic, arachidonic acids, etc.). The esters of these fatty acids include compounds with glycerol to form triglycerides, with α-glycerylphosphoric acid to form phosphatidic acids, with α-glyceryl phosphoryl esters to form lecithins, phosphatidyl ethanolamines, and phosphatidyl serines; with sphingosine to form sphingomyelin and cerebrosides; and with inositol N-phosphate to form phosphoinositides; etc. A few typical lipid structures involving certain of the lipid residues mentioned above are shown in Fig. 17. In (a) a typical triglyceride is shown, with arachidonic acid (4 double bonds), stearic acid (saturated), and linolenic acid (2 double bonds) joined to glycerol. The three fatty-acid residues are shown in their extended conformation, with the stearic-acid residue above the other two. The double bonds in the unsaturated fatty-acid residues are usually found in the *cis*-configuration. Phosphatidyl serine (b) represents a typical phospholipid, and is shown with one residue of linolenic acid and one of palmitic acid (saturated), joined through a glycerol residue to phosphoserine. Three ionizable groups are seen here—the strongly acidic phosphoric-acid residue (pK_a about 1), the weakly acidic carboxyl of serine (pK_a about 4.5), and the weakly basic amino residue of serine (pK_a about 10). In other phospholipids the serine residue may be replaced by ethanolamine ($HO-CH_2-CH_2-NH_3^+$) or by choline [$HO-CH_2-CH_2-N^+(CH_3)_3$]. In such phospholipids, no pK_a near 4.5 would be expected, and the basic group would be found to have a pK value of about 10 in the case of ethanolamine and above 14 with choline.

The plasmalogens [Fig. 17(c)] are glycerophosphatides containing an active group (shown here as the $-OH$ on carbon-1 of the stearic aldehyde residue) which reacts as an aldehyde group. The structure shown is one of several which have been proposed, and illustrates stearic-aldehyde and oleic-acid (1 double bond) residues joined to glycerol, which in turn is linked to phosphoethanolamine. This structure contains two ionizable residues—the phosphoric-acid group (pK_a about 1) and the amine group (pK_a about 10.5). Palmitic aldehyde sometimes replaces stearic aldehyde, and choline often replaces ethanolamine in typical plasmalogens. Other common types of lipids contain the lipid base sphingosine, $CH_3-(CH_2)_{12}-CH=CH-CH(OH)-CH(NH_2)-CH_2OH$. Illustrated here is a typical sphingomyelin [Fig. 17(d)] with a linolenic-acid residue joined through an amide linkage to the $-NH-$ of sphingosine, and phosphocholine joined through an ester linkage to the terminal hydroxyl of sphingosine. This molecule would have ionizable groups in the phosphoric-acid and the choline residues (pK_a about 1 and above 14). The cerebrosides represent another typical sphingolipid type (not illustrated). Their structure is similar to sphingomyelin, but in place of phosphocholine a sugar residue, often D-galactose (as a pyranose ring), is joined to the terminal hydroxyl of sphingosine through a glycoside linkage. Such a cerebroside would not contain ionizable groups, but the many hydroxyl groups of galactose would serve as active sites for reaction with other molecules.

Cholesterol [Fig. 17(e)] is frequently found in animal tissues, either as the free alcohol (as illustrated here) or as the ester, in which the cholesterol hydroxyl is linked to a fatty acid. Neither cholesterol nor cholesterol esters contain ionizable groups, and the solubility properties of these molecules resemble those of the triglycerides. Estrone [Fig. 17(f)], a typical female sex hormone, has a ring structure much like cholesterol except that the first (A) ring contains three double bonds, the second ring is saturated, and the C-18 methyl is lacking. Estrone contains a ketonic oxygen at C-17, and a phenolic hydroxide residue (pK_a about 9) at C-3.

β-Carotene [Fig. 17(g)] is a highly unsaturated hydrocarbon, occurring in many plants and green leaves, and also in animal systems. It contains a highly conjugated system of eleven double bonds, responsible for the intense color and the reactivity of the hydrocarbon. An alcohol (vitamin A) and an aldehyde (vitamin A aldehyde) are of great importance in visual pigments and contain a hydrocarbon chain similar to one-half β-carotene. Two closely related hydrocarbons, α- and γ-carotene usually occur with β-carotene. The double bonds in the carotenoids usually are found to be in the *trans*-configuration (in contrast to the situation in the fatty-acid residues). The carotenoid hydrocarbons, along with other terpene-type compounds, appear to be built up from repeating isoprene units $[CH_2=C(CH_3)-CH=CH_2]$. Another hydrocarbon, squalene, $[CH_3-C(CH_3)=CH-CH_2-CH_2-C(CH_3)=CH-CH_2-CH_2-C(CH_3)=CH-CH_2-]_2$, is found in large amounts in shark-liver oil, and also is an intermediate for the synthesis of cholesterol in mammalian systems.

The mode of attachment of these lipid structures to the peptide moiety of a lipoprotein is poorly understood.[30] Denaturation of the protein and extraction with lipid solvents is usually sufficient to completely break the lipoprotein complex, so that it would seem that, if any covalent linkages are present, they must be very labile. A number of lipo-polysaccharides are known, and lipids are found to be firmly linked with some of the glycoproteins. It may be that the carbohydrate moiety of such complex macromolecules provides a covalent attachment for the lipid molecules. Much of the lipid material now thought to occur in cells as "free lipid" will, in the near future, be found to have intimate molecular binding to protein, glycoprotein, or carbohydrate cellular components.

ACKNOWLEDGMENTS

The author is greatly indebted to Dr. Margaret J. Hunter, Dr. Martha L. Ludwig, Dr. Donald F. H. Wallach, and Dr. Colin Green for their constructive criticism regarding this manuscript.

BIBLIOGRAPHY

[1] J. M. Bijvoet, A. F. Peerdeman, and A. J. van Bommel, Nature **168**, 271 (1950).

[2] B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I., p. 235.

[3] L. Pauling in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 17.

[4] M. Calvin, Federation Proc. **13**, 697 (1954).

[5] G. E. Perlmann, Advances in Protein Chem. **10**, 1 (1955).

[6] D. F. Waugh, Advances in Protein Chem. **9**, 325 (1954).

[7] K. Linderstrøm-Lang in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 23.

[8] C. Tanford in *Symposium on Protein Structure*, A. Neuberger, editors (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 35.

[9] C. Tanford, S. A. Swanson, and W. S. Shore, J. Am. Chem. Soc. **77**, 6414 (1955).

[10] J. Steinhardt and E. M. Zaiser, Advances in Protein Chem. **10**, 152 (1955).

[11] C. Tanford, J. D. Hauerstein, and D. G. Rands, J. Am. Chem. Soc. **77**, 6409 (1955).

[12] F. Sanger, Advances in Protein Chem. **7**, 1 (1952).

[13] F. Sanger, *Currents in Biochemical Research*, 1956, D. E. Green, editor (Interscience Publishers, Inc., New York, 1956).

[14] C. B. Anfinsen and R. R. Redfield, Advances in Protein Chem. **11**, 1 (1956).

[15] C. H. Li, Advances in Protein Chem. **11**, 101 (1956).

[16] *Ibid.*, **12**, 270 (1957).

[17] C. H. Li in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 302.

[18] J. I. Harris in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 333.

[19] C. H. W. Hirs, W. H. Stein, and S. Moore in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 211.

[20] C. B. Anfinsen in *Symposium on Protein Structure*, A Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 223.

[21] H. Tuppy in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 66.

[22] P. Jollés, J. Jollés-Thaureaux, and C. Fromageot in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 277.

[23] R. R. Porter in *Symposium on Protein Structure*, A. Neuberger, editor (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958), p. 290.

[24] A. Neuberger, editor, *Symposium on Protein Structure* (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1958).

[25] W. Pigman, *The Carbohydrates* (Academic Press, Inc., New York, 1957).

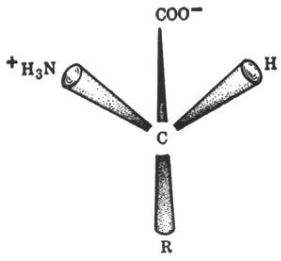
[26] K. Meyer, Advances in Protein Chem. **2**, 249 (1945).

[27] G. E. W. Wolstenholme and M. O'Connor, editors, *Chemistry and Biology of Mucopolysaccharides, Ciba Symposium* (J. and A. Churchill, London, 1958).

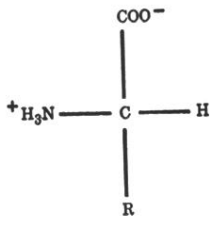
[28] J. A. Lovern, *The Chemistry of Lipids of Biochemical Significance* (Methuen and Company, Ltd., London; John Wiley and Sons, Inc., New York, 1955).

[29] H. J. Deuel, Jr., *The Lipids* (Interscience Publishers, Inc., New York, 1951), Vols. I–III.

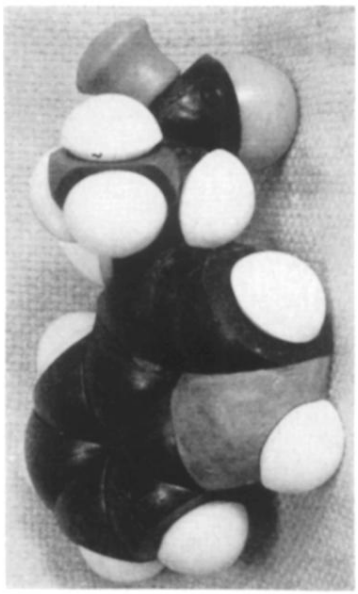
[30] E. Chargaff, Advances in Protein Chem. **1**, 1 (1944).



(a)



(b)



(c)

FIG. 1. Absolute configuration of the L-amino acids. (a) Geometrical representation. The COO^-, NH_3^+, and H groups are represented on a face plain of a tetrahedral carbon atom, with the R-group at the opposite apex. (b) Conventional chemical representation. (c) Photograph of a three-dimensional model of L-tryptophan.

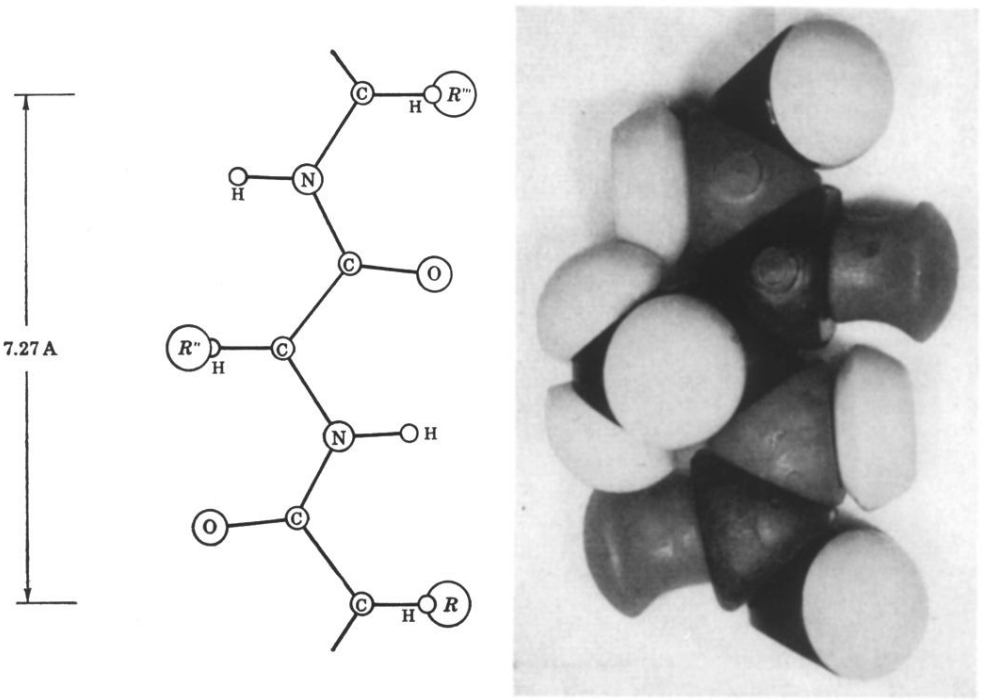


FIG. 2. Photograph of scale model of extended polypeptide chain. The R- and R'''-groups lie at the rear of this model [from B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 259].

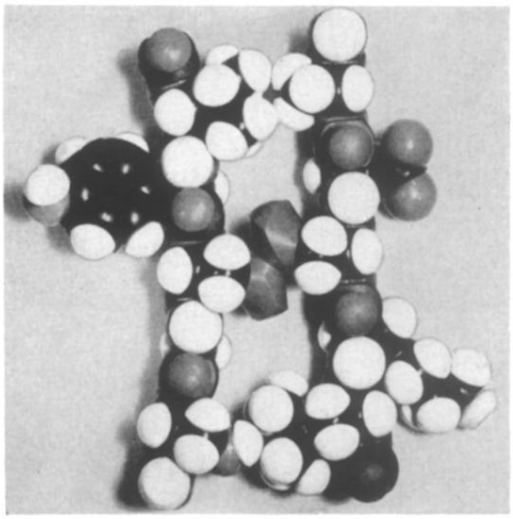


FIG. 4. Photograph of scale model of interchain disulfide linkage with antiparallel chain directions. The peptide chain on the left has the sequence -CO-leu-lys-cyS-asp-ala-NH- (bottom to top), and the chain on the right has the sequence -CO-val-tyr-cyS-ala-ser-NH- (top to bottom). The two peptide chains are about 6.3 A apart at the disulfide bond [from B. W. Low in *The Proteins*, H. Neurath and K. Bailey, editors (Academic Press, Inc., New York, 1953), Vol. I, p. 258].