Figure 52. A model of the myoglobin molecule based on X-ray crystallographic data. We can deduce from the dimensions of this model, and from the known number of amino acid residues in myoglobin, that certain portions of the polypeptide chain of myoglobin must be coiled to form an internal secondary structure. The gray disk-like structure shown in the photograph represents the porphyrin prosthetic group, and the light- and dark-colored spheres represent heavy-metal derivatives which were attached to the protein to facilitate analysis of the diffraction data. Dr. J. C. Kendrew of the University of Cambridge kindly furnished this photograph.
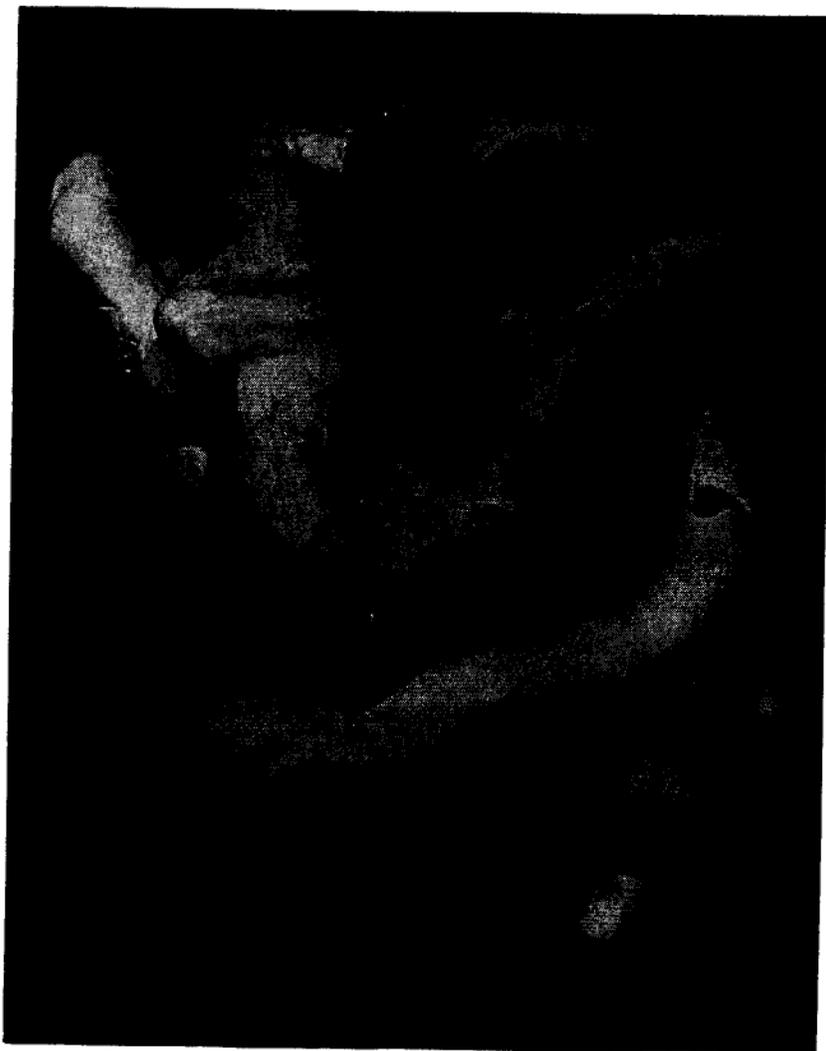
chapter 5

PROTEIN STRUCTURE

The study of the internal structure of proteins in the crystalline form by means of X-ray crystallography indicates a well-defined arrangement of atoms, more or less "frozen" into a definite pattern which is the same for each molecule in the crystal. The crystallographer thus obtains a picture of *the* structure of a protein which is rigid and essentially invariant (Figure 52). From the purely chemical point of view such a picture is completely satisfactory. In living cells, however, proteins are not in the solid state but are dissolved in the intracellular fluid and can be shown to be in a constant state of minor rearrangement in response to shifting hydrogen ion concentrations and salt levels and to reversible adsorption to intracellular surfaces.
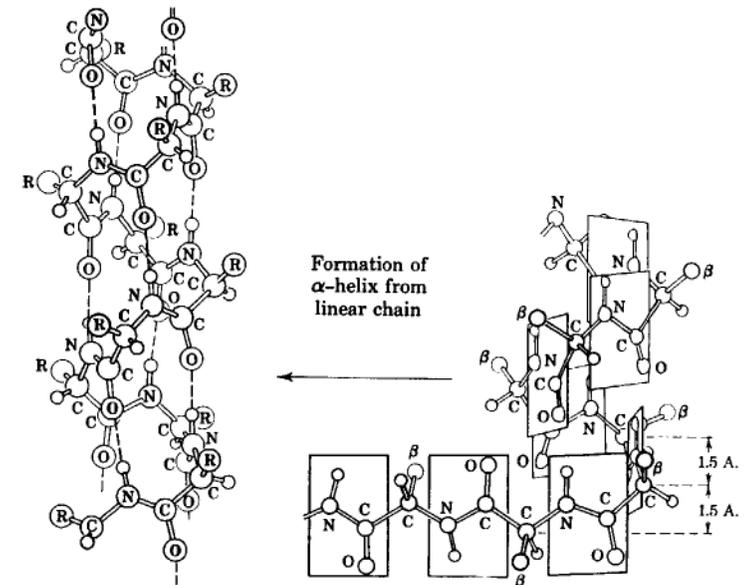
In considering such fluctuations it is convenient to think of the chemical fabric of proteins in terms of three distinct categories, christened by K. Linderstrøm-Lang—primary, secondary, and tertiary structure. The term primary structure refers to the fixed amino acid sequence of the polypeptide chain (or chains) making up the backbone of the molecule. This category also includes those covalent bonds that form fixed sites of cross-linkage such as the disulfide bonds between half-cystine residues and the phosphate diester linkages of certain phosphoproteins.

The concept of secondary structure is based, to a large extent, on
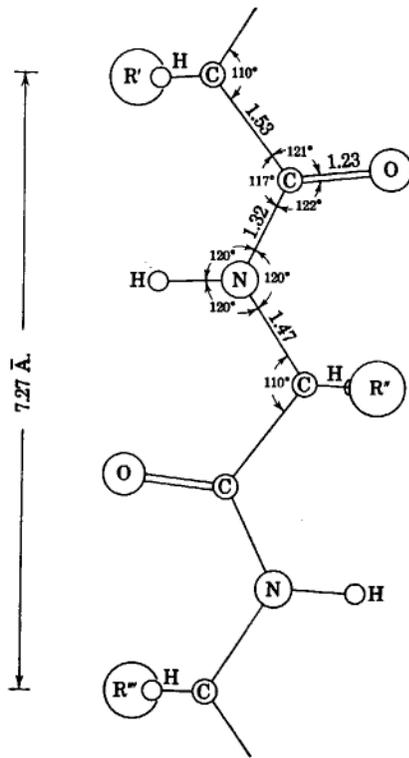
Figure 53. The dimensions of a fully extended polypeptide chain as derived from X-ray crystallographic data. From R. B. Corey and J. Donahue, *J. Am. Chem. Soc.*, **72**, 2899 (1950).

Dimensions of an α–helix

5th turn
4th turn
3rd turn
2nd turn
1st turn

18 residues
27 A.

5.1 A.

26°

5.4 A. pitch
3.6 residues

Rise per residue   1.5 A.

Figure 54. A schematic representation of the process of coiling an extended polypeptide chain into the α-helical configuration. Redrawn from L. Pauling and R. B. Corey, *Proc. Intern. Wool Textile Research Conf.*, **B**, 249 (1955).

Formation of α-helix from linear chain

the relatively recent discovery of helical coiling in proteins, stabilized by hydrogen bonds between the amide (CONH) linkages of the chain. A number of helical arrangements of the polypeptide chain may be constructed with atomic models which fit the geometrical restrictions imposed by the bond angles and interatomic distances that have been determined by the X-ray crystallographer on low molecular weight substances such as di- and tripeptides (Figure 53). These various models differ in respect to the number of amino acid residues per turn of helix, in the pitch of the helical screw, and in the amount of unfilled space left within their centers. Of all the helices that have been seriously considered as components of protein structure, the so-called α-helix appears to be the most probable. In this structure the maximum number of intrahelical hydrogen bonds are formed between the CONH linkages, and the structure is consequently a fairly stable one. Perhaps most important, the atoms comprising the peptide bond fall in one plane without strain. The likelihood of such a planar configuration is high as judged from crystal-

lographic studies of model peptides like that shown in Figure 53. Figure 54 illustrates the way in which the extended, "β-keratin" form of the polypeptide chain is coiled into the α-helical configuration. Each amide group is linked to the third amide group beyond by a hydrogen bond. A complete turn of the helix contains 3.7 amino acid residues, each residue contributing 1.47 A. of linear translation along the central axis. The pitch of the "screw" is thus 5.44 A.

Recent physicochemical studies on proteins and polypeptide model substances have shown that, in spite of the stabilization given by the internal hydrogen bonds, the helical structure is not sufficiently stable to exist in solution but unfolds into a random, disoriented strand. Stabilization of the helical coiling requires the presence of disulfide bridges and/or *tertiary structure*, and it is probable that only those parts of proteins which are properly anchored by such bonds can maintain the helical configuration. Tertiary bonds include such examples as the van der Waals interactions, the agglomeration of lyophobic side chains by mutual repulsion of solvent, and such special hydrogen bonds as might exist between the hydroxyl groups of tyrosine or the ε-amino groups of lysine and electronegative groups elsewhere along the chain (Figure 55). The variations, reversible and irreversible, that are possible in the secondary and tertiary structure of proteins are, basically, functions of the primary,
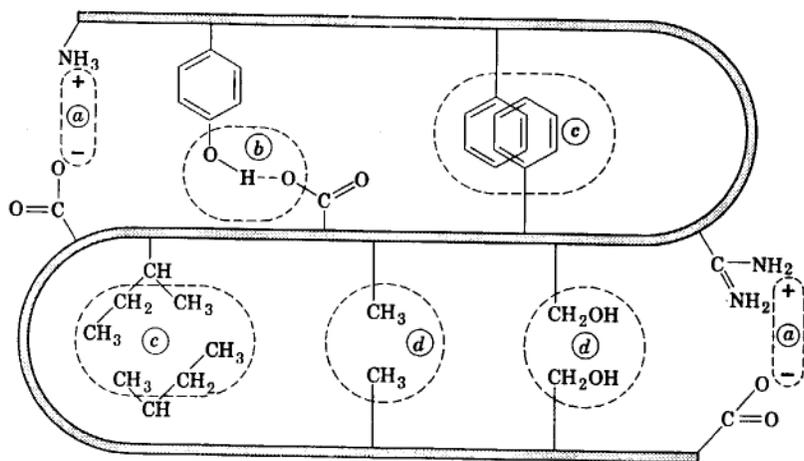

Figure 55. Some types of noncovalent bonds which stabilize protein structure: (a) Electrostatic interaction; (b) hydrogen bonding between tyrosine residues and carboxylate groups on side chains; (c) interaction of nonpolar side chains caused by the mutual repulsion of solvent; (d) van der Waals interactions.

covalent structure of the protein, which is fixed and invulnerable to modification by ordinary environmental fluctuations. Thus, a real understanding of the physicochemical behavior of a protein in solution must depend on a detailed knowledge of its structure in the organic chemical sense.

To appreciate our present state of knowledge we must attempt to view the problem from the position occupied by protein chemistry some fifteen years ago. In spite of a large body of rather precise data dealing with the behavior of proteins in solution, there existed a sort of mystical aura around macromolecular structures arising, for the most part, from the almost complete absence of accurate information on amino acid content and arrangement. The necessary background for the modern approach was provided when Fred Sanger and his colleagues[1] undertook the study of the insulin molecule and demonstrated that there existed, in the structural analysis of at least one protein, no insuperable chemical difficulties. The confidence engendered by Sanger's accomplishments and the techniques he employed have led to similar structural attacks on a number of other proteins during the past five years. This accumulated experience has now brought us to a point where we may, with good expectations of success, undertake the elucidation of the structure of nearly any protein whose molecular weight is within reasonable limits. At the present time, for example, complete sequential formulas are available for the insulins and adrenocorticotropic hormones of several species, for bovine glucagon, and for the melanocyte-stimulating hormones of pig and beef pituitary glands. The sequences of several enzymes, including ribonuclease, papain, and lysozyme, should be completed within a relatively short time, and many other biologically active proteins are under study in a number of laboratories.

Although it is outside the scope of this book to discuss, with any degree of completeness, the methodology involved in the determination of the sequence of amino acids in polypeptide chains, some consideration of this subject will be of interest. In the following summary we shall assume the availability of protein samples of proven homogeneity. A specific consideration of the heterogeneity of individual proteins, and of the implications of heterogeneity with regard to protein synthesis and genetics, will be made in subsequent chapters.

For reasons of personal familiarity with the protein, I have chosen as a model substance for the discussion of methodology the enzyme ribonuclease. This enzyme, which has been found in all cells tested so far, catalyzes a transphosphorylation reaction which leads to the
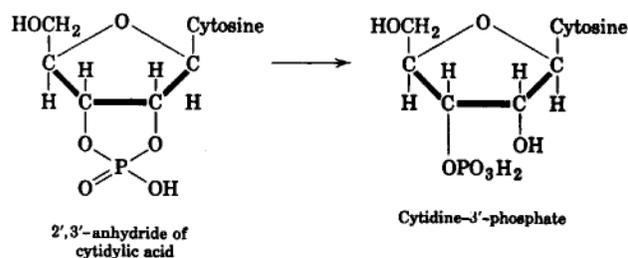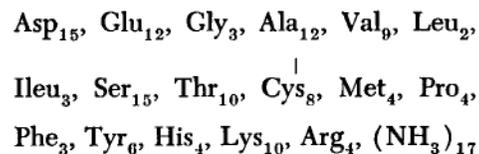
Figure 56. A synthetic material which is hydrolyzed by ribonuclease.

hydrolysis of certain phosphate diester linkages in ribonucleic acid involving pyrimidine nucleotide residues. The cleavage of a variety of synthetic substrates, of the sort shown in Figure 56, is also catalyzed by the enzyme and forms the basis for a convenient enzymatic assay. Although its specific role in cell metabolism is, as yet, unknown, it seems likely that ribonuclease plays an important part in some aspect of the process of protein biosynthesis (see Chapter 10). Large quantities of ribonuclease are formed and secreted by the pancreas, and the pancreatic enzyme must catalyze extensive hydrolysis of ribonucleic acid in the intestine.

The structure of bovine pancreatic ribonuclease has been under study for several years in two laboratories in which, fortunately for the present illustrative purposes, rather different methods have been employed toward the same end. As the first step in the study of the structure of any protein, it is necessary to obtain an accurate set of analytical data on the amino acid composition of the molecule. Such data were furnished for ribonuclease by Hirs, Moore, and Stein,[2] who established, by careful chromatographic techniques, that the protein contained 124 amino acid residues. Their results may be expressed in terms of the formula:

$$Asp_{15}, Glu_{12}, Gly_{3}, Ala_{12}, Val_{9}, Leu_{2},$$

$$Ileu_{3}, Ser_{15}, Thr_{10}, \overset{|}{Cys_{8}}, Met_{4}, Pro_{4},$$

$$Phe_{3}, Tyr_{6}, His_{4}, Lys_{10}, Arg_{4}, (NH_{3})_{17}$$

where each amino acid is indicated by an abbreviation derived from the first three letters of its name. Most readers of this book will be familiar with these abbreviated formulas, but for the benefit of those who are not a complete list is given in Table 3.

TABLE 3
Amino Acid Composition of Bovine, Pancreatic Ribonuclease[2]

Amino Acid	Abbreviations[a]	Number Residues per Molecule (mol. wt. 13,683)	Amino Acid	Abbreviations[a]	Number Residues per Molecule (mol. wt. 13,683)
Aspartic acid	(Asp)	15	Methionine	(Met)	4
Glutamic acid	(Glu)	12	Proline	(Pro)	4
Glycine	(Gly)	3	Phenylalanine	(Phe)	3
Alanine	(Ala)	12	Tyrosine	(Tyr)	6
Valine	(Val)	9	Histidine	(His)	4
Leucine	(Leu)	2	Lysine	(Lys)	10
Isoleucine	(Ileu)	3	Arginine	(Arg)	4
Serine	(Ser)	15	Amide NH_3		(17)
Threonine	(Thr)	10			
Half-cystine	(Cys)	8	Total number of residues		124

Not present in ribonuclease are tryptophan (Try) and cysteine (CySH). Cysteic acid ($CySO_3H$) is formed by oxidation of cysteine and cystine, and $MetO_2$ by oxidation of methionine. Amide nitrogens are derived from the side-chain carboxyl groups of glutamic acid and aspartic acid and are generally indicated as $Glu(NH_2)$ and $Asp(NH_2)$, although abbreviated forms, such as Gn and An, seem more convenient.

[a] As suggested by E. Brand and J. T. Edsall, *Ann. Rev. Biochem.*, **16**, 224 (1947).

To determine whether or not the 124 amino acids are arranged in a single, or in multiple, cross-linked polypeptide chains, various chemical and physical studies were carried out as summarized in Figure 57. The presence of as many as five cross-linked chains was possible in ribonuclease since the molecule contained four *cystine* residues. The protein was, therefore, oxidized with performic acid, which converts the sulfur atoms in disulfide bonds into sulfonic acid groups. Ultracentrifugation and viscosity measurements permitted calculations which indicated that the oxidized protein had essentially the same molecular weight as the native protein. The presence of more than one chain would have been reflected in a marked lowering of the average molecular weight as estimated by these physical measurements.
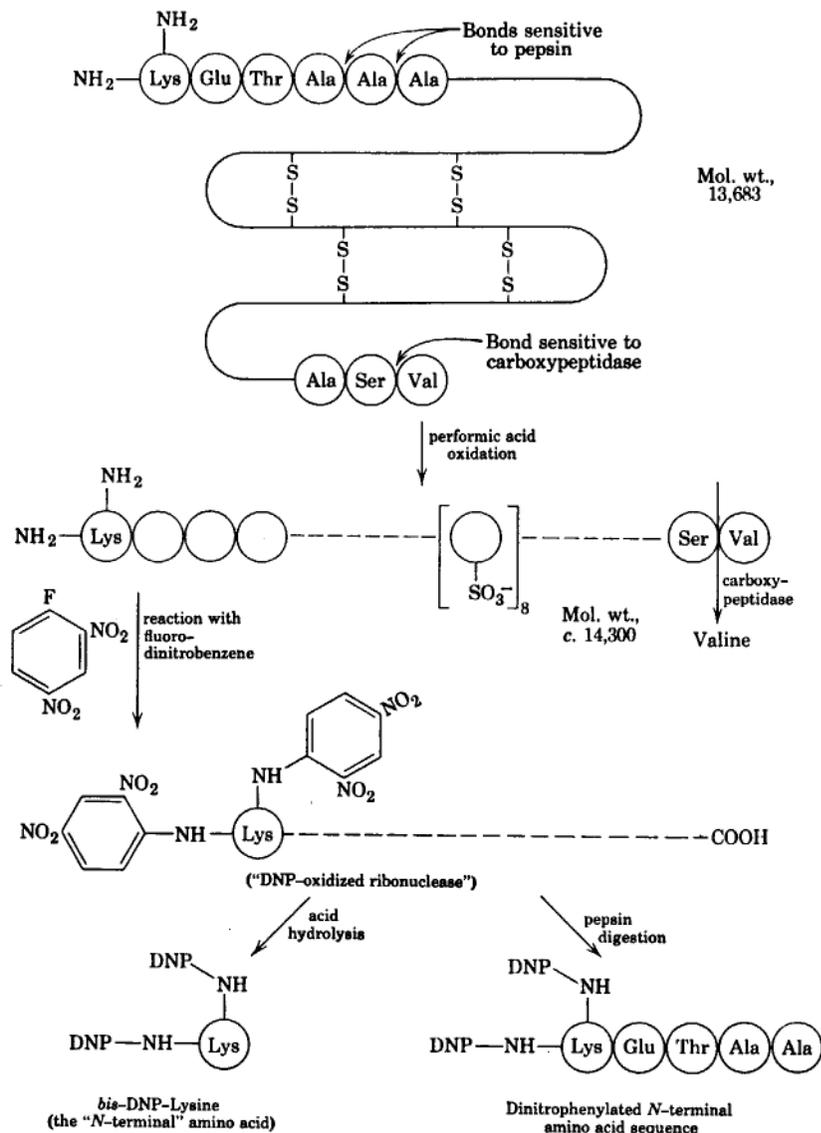
Figure 57. Demonstration of the presence of a single polypeptide chain in bovine pancreatic ribonuclease. The native molecule is oxidized with performic acid to yield an extended polypeptide chain. This derivative contains only a single N-terminal and a single C-terminal end group.
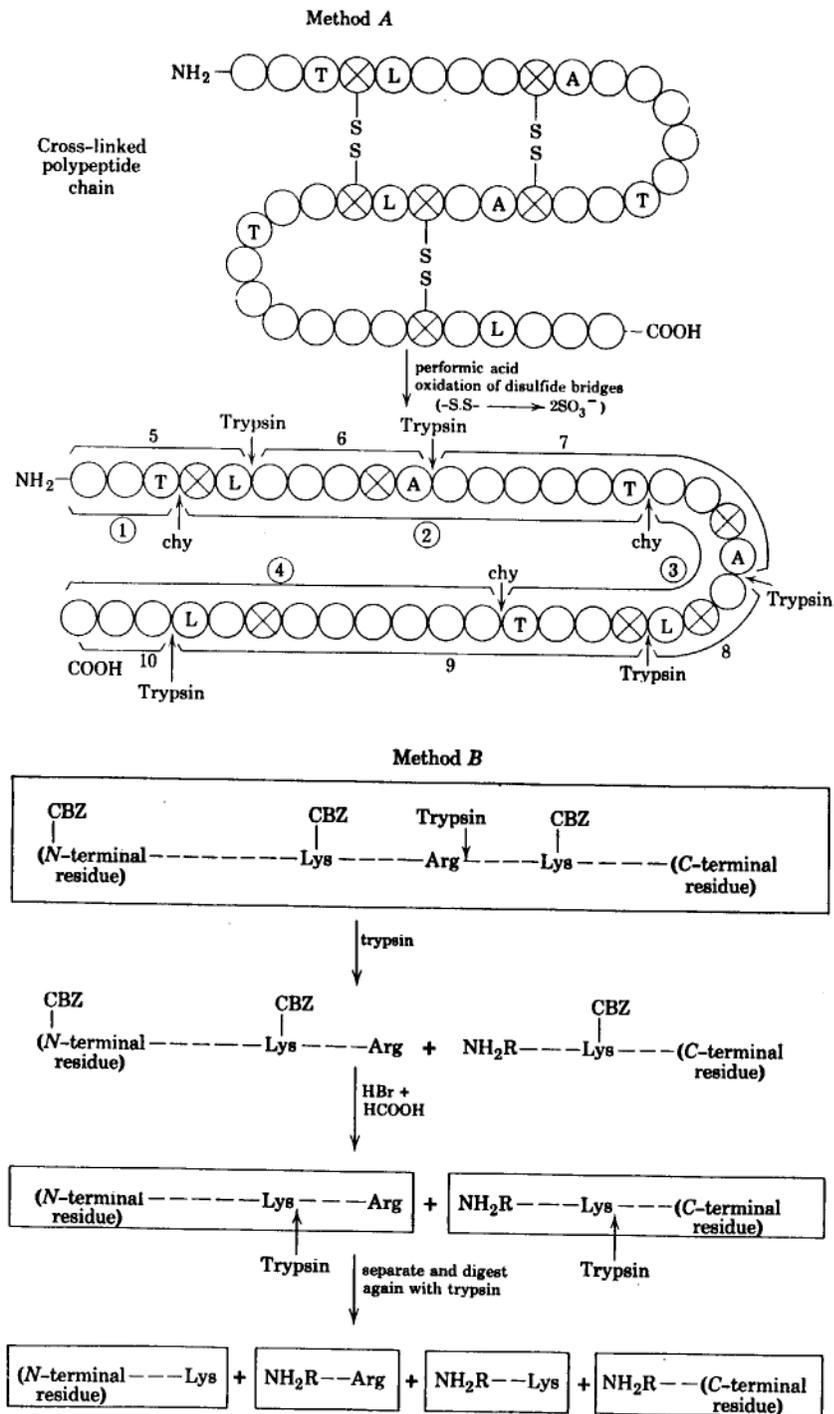
The single-chain character of the protein was further indicated by "end group analysis." The only free α-amino groups in proteins are those present on the amino acids at the termini of polypeptide chains. These amino groups may be reacted with reagents such as dinitrofluorobenzene, which forms a dinitrophenyl (DNP) derivative of the protein in which (Figure 57) the N-terminal (i.e., NH₂-terminal) amino acids have been converted to their DNP derivatives. Such a modified protein may be cleaved with acid or with proteolytic enzymes, and the DNP-N-terminal amino acid residue, or peptide, may be isolated and characterized chemically. By this method, evidence was obtained for only a single free α-amino group in the protein, present in the sequence Lys.Glu.Thr.Ala. For partial characterization of the C-terminal (i.e., COOH-terminal) end of the protein, the enzyme carboxypeptidase was used as one of several tools for the specific cleavage of the peptide bond at this end of the chain. This enzyme removed valine from ribonuclease, with only traces of other amino acids. The presence of single N-terminal and C-terminal residues in the protein confirmed the physical evidence for a single chain.

Hydrolysis of the oxidized chain has been carried out by a variety of proteolytic enzyme "reagents," since these tend to catalyze a much more limited number of peptide bond cleavages than do such agents as mineral acids. Indeed, it now appears that nonspecific hydrolysis by acids can, in general, be avoided at this stage and is of particular use only in the subsequent study of the sequence of the smaller peptide subunits produced by preliminary enzymatic attack.

Figure 58 is a schematic outline of two general methods of approach to the detailed structural analysis of polypeptide chains that have been applied to the ribonuclease molecule and are of general applicability. In both techniques the native structure of the protein is unfolded by cleavage of the disulfide bridges, either by oxidative methods to yield cysteic acid residues in place of half-cystine residues, or by reductive cleavage followed by chemical masking of the resulting sulfhydryl group by some suitable agent such as iodoacetic acid. (The latter method for cleavage has an intrinsic advantage over the oxidative method since it can be applied to proteins containing tryptophan, an amino acid residue which is extremely labile to the conditions of disulfide oxidation by performic acid.)

In method A one sample is hydrolyzed by trypsin which cleaves the chain at points following lysine or arginine residues, and another by chymotrypsin (or some other protease with different specificities from trypsin). As indicated, a completely different, but overlapping,

Method A

Cross-linked
polypeptide
chain

performic acid
oxidation of disulfide bridges
$(-S\text{-}S\text{-} \longrightarrow 2SO_3^-)$

Trypsin Trypsin

Method B

set of peptide fragments is obtained; this set may then be subjected to analysis and arranged in linear order on the basis of overlaps in composition.

In method B trypsin cleavage is restricted to positions following arginine only, by masking the ϵ-amino groups of the lysine residues with either dinitrophenyl groups or carbobenzoxy groups. In this method, therefore, we obtain a number of fragments one greater than the number of arginine residues in the protein. All the fragments contain C-terminal arginine except the one derived from the C-terminus of the chain (which fixes its position). Method B depends on the subsequent determination of the sequence of small peptides isolated from random hydrolysates prepared with acid or with some very nonspecific protease such as subtilisin (a bacterial protease) which yields short, easily manageable, peptides. These arginine-containing sequences are then used as connecting links between the larger fragments originally separated from the digest of the carbobenzoxylated protein.

The Rockefeller Institute group, following method A, have reacted separate samples of oxidized ribonuclease with trypsin, pepsin, and chymotrypsin. These hydrolysates have then been separated into component peptide fragments by column chromatography. A consideration of overlaps has then enabled these investigators to reconstruct large portions of the peptide chain. As an example of such a reconstruction, let us examine the experimental data which permitted Hirs, Moore, and Stein[3] to arrive at a partial reconstruction of the ribonuclease chain.

A typical analytical pattern is shown in Figure 59a, obtained by means of a Dowex-50 ion exchange column through which was passed a tryptic digest of oxidized ribonuclease under proper conditions of elution. The peptide peaks emerging from the column have then been hydrolyzed and rerun separately on subsequent Dowex-50 col-

Figure 58. Two approaches to the determination of the sequential arrangement of amino acids in a polypeptide chain. In method A the chain is digested with two or more proteolytic enzymes. The resulting fragments are then separated and analyzed for amino acid composition. A partial reconstruction of the molecule can then be made on the basis of overlapping amino acid compositions. The symbols T, L, and A represent tyrosine, lysine, and arginine respectively. In method B the number of cleavages caused by the proteolytic enzyme, trypsin, is limited to those bonds in which the carbonyl of the peptide bond is donated by an arginine residue. To arrange the fragments in their proper order we must first determine the nature of the amino acids in the immediate vicinity of each of the arginine residues.

PROTEIN STRUCTURE

Figure (a)

Dowex 50-X2
150 × 2cm.

Ninhydrin color value (leucine equivalents, mM. conc.)

(Glu,Gly,Ser$_2$,Thr)Lys[90,100%] 6

(Cys,AspNH$_2$,Ala,Val)Lys[95,95%] 5

Asp.Arg[70,90%] 7

NH$_3$ (from reagents)

1,2,3,4

Effluent liters 0.5   1.0   1.5   2.0

0.2N pH 3.1, 35° — Gradually increasing pH and [Na$^+$] —

(AspNH$_2$,GluNH$_2$,Ala,Thr$_2$) Lys[100,100%] 8

Lys.Glu.Thr.Ala.Ala.Ala.Lys[95,100%] 10

9

(AspNH$_2$,Leu,Thr)Lys[100,95%] 11

Ser.Arg[55,55%] 12

(GluNH$_2$,Phe)Arg[75,85%] 15

13+14

2.5   3.0   3.5   4.0   4.5

(0.2N pH 3.1 → 2.0N pH 5.1), 35° throughout

(a)

Figure (b)

Ninhydrin color value (leucine equivalents, mM. conc.)

4

(Cys,Asp(xNH$_2$),Glu(yNH$_2$), Gly,Ala,Val,Ileu$_2$,Pro)Tyr [70%]

(Ser,Thr,Met) [80%]

(AspNH$_2$,GluNH$_2$, Ala,Thr$_2$,Lys)[26%]

(Glu,Ser)Tyr[85%]

(14,15+16)

(11+12)

(13+NH$_3$)

2

5   6

1

3   7 8   9   10   17   18   19

Effluent liters 0.5   1.0   1.5   2.0   2.5

0.2N pH 3.1, 35° — Gradually increasing pH and [Na$^+$] —

(Asp(xNH$_2$),Glu$_2$(yNH$_2$)Thr, Ser$_3$,Met,His,Arg)[35%]

Lys.Glu.Thr.Ala.Ala.Ala.Lys.Phe[75%]

(Asp$_2$(xNH$_2$),Glu$_2$(yNH$_2$), Ala$_2$,Met,Thr,Ser$_6$, His,Arg)Tyr [25%]

(Cys,(AspNH$_2$)$_2$,GluNH$_2$,Leu,Met$_2$,Ser,Lys,Arg)[50%]

(Cys,Asp(xNH$_2$),Glu(yNH$_2$),Gly,Ala,Val,Ileu$_2$,Pro, Lys,His)Tyr[35%]

(Val$_2$,Pro,His)Phe[75%]

(Cys,Asp$_2$(xNH$_2$),Val,Thr$_2$,Pro,Lys$_2$, Arg)Phe[65%]

25

20   21   22 23   24   26   27   28 29   30   31   32

3.0   3.5   4.0   4.5   5.0   5.5

(0.2N pH 3.1 → 2.0N pH 5.1), 35° throughout

(b)

umns for amino acid analysis (the results are included in the figure). Similar analysis of a chymotryptic digest of oxidized ribonuclease yielded the results shown in Figure 59b.

The data in Figures 59a and 59b, when considered together with the known N-terminal sequence and with the knowledge that certain residues occur in ribonuclease in only limited numbers (e.g. phenylalanine, 3; glycine, 3; methionine and histidine, 4; etc.) permit the partial reconstruction of the sequence of the enzyme as shown in Figure 60. The parentheses surround the residues that may be grouped together and thus be given approximate linear assignments along the chain. Amide nitrogen atoms, unless specifically attributable to individual aspartic or glutamic acid residues, are distributed in an arbitrary fashion to the dicarboxylic amino acids within any given parenthesis.

The general approach described in Figure 58, method B, yields results that are in accord with the results obtained by method A. Since there exist four arginine residues in ribonuclease, we may expect five peptides of varying size by the application of this method. These peptides were, indeed, shown to be formed by trypsin digestion and, following isolation, were analyzed and partially characterized in terms of terminal sequences. Their composition is given in Table 4. Subsequent characterization of smaller arginine-containing sequences isolated from digests prepared with pepsin or partial acid hydrolysis permitted the alignment of these fragments in the order, A–B–C–D–E. Comparison of the partial reconstructions possible through the use of the two methods indicate how the different sets of information may be fitted together to give a final picture with greater detail than on the basis of either alone.

To construct a final picture of the covalent structure of the protein requires, in addition to the elucidation of the complete sequence of each peptide fragment, a study of the disulfide bridges. The location of these can be determined by application of the methods introduced by Sanger and his colleagues in the case of insulin. The na-

Figure 59. Separation of the peptides in proteolytic hydrolysates of oxidized ribonuclease. (a) A 20-hour tryptic hydrolysate of a 200-mg. sample of protein, chromatographed on a 150 x 1.8-cm. column of Dowex 50-X2. The points represent the ninhydrin color value given by aliquots of each of the fractions, expressed in leucine equivalents. Figures in brackets after each peptide after 3 and 20 hours of trypsin hydrolysis, respectively. From C. H. W. Hirs, S. Moore, and W. H. Stein, J. Biol. Chem., **219**, 623 (1956). (b) Peptides in a chymotryptic hydrolysate of oxidized ribonuclease. From C. H. W. Hirs, W. H. Stein, and S. Moore, J. Biol. Chem., **221**, 151 (1956).

Figure 60. A partial reconstruction of the polypeptide chain of ribonuclease based on procedures such as those described in Figure 58a. From C. H. W. Hirs, W. H. Stein, and S. Moore, *J. Biol. Chem.*, **221**, 151 (1956).

Lys.Thr.Ala.Ala.Ala.Lys.Phe.Glu.Arg.(Asp,Glu,Thr,Ser₃,Met,His).(Asp,Ala₂,Ser₃).Tyr.(Cys,Asp,Glu,Met₂).Lys.Ser.Arg.Asp.Leu.Thr.Lys.

Asp.Arg.(Cys,Asp,Val,Pro,Thr,Lys).Phe.(Glu,Val,Leu,Ser,His).(Asp,Glu,Glu,Ala,Val).(Ser,Cys,Glu,Ala,Val).Lys.Asp.(Cys,Ala,Val).Lys.

(Cys,Asp,Asp,Glu,Gly,Thr).Tyr.Glu.Ser.Tyr.(Thr,Ser,Met).(Cys,Asp,Ileu,Thr,Ser).Arg.(Glu,Gly,Thr,Ser₂).Lys.(Cys,Asp,Ala,Tyr,Pro).Tyr.Lys.

(Asp,Glu,Ala,Thr₂).Lys.His.(Val,Ileu₂).(Cys,Asp,Glu,Glu,Gly,Ala,Pro).Tyr.(Val₂,Pro,His).Phe.(Asp,Ala,Ser).Val

TABLE 4

Partial Structure of the Five Peptides Produced by Application of "Method B" to Oxidized Ribonuclease

Peptide	Structure
A	Lys.Glu.Thr.Ala.Ala.Ala.Lys.Phe.Glu.Arg. (cleavage markers: A ; P P ; T P ; T)
B	Glu.(Thr,Ser₆,Asp₃,His.Met₃,Ala₂,Tyr,Cys,Glu).Lys.Ser.Arg. (cleavage markers: T ; T)
C	Asp.Leu.Thr.Lys.Asp.Arg. (cleavage markers: A ; T ; T)
D	CySO₃.(Lys₃,Pro,Val₅,Asp₅,Thr₃,Phe,His,Glu₅,Ser₄,Leu,Al₁₃, Cys₃,Gly,Tyr₂,Met).Ser.Ileu.Thr.(Asp,Cys).Arg. (cleavage markers: C A ; T)
E	Glu.(Ser₂,Thr₃,Gly₂,Lys₃,Tyr₃,Pro₃,Asp₃,Ala₃,Cys₂,Glu₂, His₂,Ileu₂,Val₃).Phe.Asp.Ala.Ser.Val (cleavage markers: P ; A)

After digestion of carbobenzoxylated, oxidized ribonuclease with trypsin, the digest was decarbobenzoxylated by treatment with anhydrous HBr in glacial acetic acid. The five peptides were then separated by combined paper electrophoresis and paper chromatography and each one hydrolyzed further with acid (A), pepsin (P), trypsin (T), or chymotrypsin (C). A study of the composition of the fragments produced by these latter digestions permitted a partial reconstruction of the sequence as shown above.

tive protein, in which these bridges are intact, is digested with various proteases (e.g. trypsin plus chymotrypsin or subtilisin). From such digests can be isolated peptides of *cystine*, carved out from the structure of the cross-linked native molecule. The general approach to the determination of disulfide bridges which has been employed by both the group at the Rockefeller Institute and at the National Heart Institute is given in a disarmingly simplified form in Figure 61.

Following these general procedures, Spackman, Moore, and Stein[3] were able to isolate, for example, a *cystine*-containing peptide which,

Figure 61. Steps in the determination of how half-cystine residues are paired in disulfide linkage. The cysteic acid-containing peptides derived from each cystine peptide are separated by chromatography and/or electrophoresis. A consideration of their amino acid compositions in terms of the amino acid sequence of the protein will generally permit the location of each half-cystine residue along the polypeptide chain.

upon oxidation with performic acid, yielded the two *cysteic acid*-containing peptides

$$(CySO_3{}^-, An, Gn, Met_2, Lys) \quad and$$
$$(CySO_3{}^-, An, Ileu, Thr, Ser, Arg)r$$

The amino acid compositions of these peptides are such that the cysteic acid residues could only have been derived from half-cystines 1 and 6 respectively, establishing the presence of a disulfide bridge between these half-cystines in the native protein. (The half-cystines are numbered sequentially from the N-terminal end of the chain). In a similar fashion, evidence was obtained for a disulfide bond between half-cystines 4 and 5, and, with some reservations because of the possibility of chemical rearrangements during isolation, between 3 and 7, and 2 and 8. The results of the combined degradative studies of ribonuclease, together with the isolation and tentative characterization of the cystine bridges, permit us to construct a provisional formula for the enzyme as shown in Figure 62. The formula includes much of the total sequence within the parentheses in the reconstructions summarized in Figure 60 and Table 4.

One additional example of the determination of amino acid sequence is summarized in Figure 63. This figure shows, in a nutshell, the general approach used by Li and his colleagues in the study of the structure of the anterior pituitary hormone, adrenocorticotropin (ACTH).[4] The various fragments (see also Chapter 6), some of which were subjected to complete or partial sequence analysis by methods described in some of the references at the end of this chapter, are all consistent with only one unique sequence as shown.

Figure 62. The structure of bovine pancreatic ribonuclease. This figure has been constructed from the combined studies of a number of individuals at the Rockefeller Institute (including Drs. C. H. W. Hirs, S. Moore, S. H. Stein, D. Spackman, and L. Bailey) and at the National Heart Institute (including R. Redfield, J. Cooke, A. Ryle, and C. B. Anfinsen). The positions of the four disulfide bridges are probably as shown, although there is still some question about the bridge between half-cystine residues 2 and 8 and that between half-cystine residues 3 and 7, because of the poor yields obtained during the isolation of cystine peptides and because of the possibility of chemical rearrangements which might have occurred during certain steps in the sequence analysis of these peptides. At the present time some 95 residues may be assigned with reasonable certainty to definite positions in the polypeptide chain, which contains 124 residues in all. The residues that are shaded may be assigned to the general region of the molecule indicated, but their exact position in the sequence must await further double-checking by stepwise degradation.

Hydrolytic Agent	Amino Acid Sequence
Trypsin	Ser.Tyr.Ser(Met.Glu.His.Phe)Arg
Chymotrypsin	Arg.Try( )
Trypsin	Try.Gly.Lys.Pro.Val.Gly.Lys
Trypsin	Lys.Arg
Trypsin	Lys.Arg.Arg
Trypsin	Arg.Pro.Val.Lys
Acid	Pro(Val.Lys,Val,Tyr)
Trypsin	Val.Tyr.Pro.Ala.Gly.Glu(Asp,Asp,Glu,Ala,Ser,Glu,Ala,Phe,Pro,Leu,Glu,Phe)
Acid	Ala(Gly,Glu,Asp)
Pepsin	Asp.Glu
Pepsin	Asp(Glu,Ala)
Pepsin	Asp(Glu,Ala)Ser
Pepsin	Glu(Ala,Ser)
Pepsin	Ser.Glu
Pepsin	Ser(Glu,Ala)
Pepsin	Ser(Glu,Ala,Phe)
Pepsin	Glu.Ala.Phe
Pepsin	Phe(Pro,Leu,Glu)
Pepsin	Pro(Leu,Glu,Phe)

Complete sequence: Ser.Tyr.Ser.Met.Glu.His.Phe.Arg.Try.Gly.Lys.Pro.Val.Gly.Lys.Lys.Arg.Arg.Pro.Val.Lys.Val.Tyr.Pro.Ala.Gly.Glu.Asp.Asp.Glu.Ala.Ser.Glu.Ala.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34

Phe.Pro.Leu.Glu.Phe
35 36 37 38 39

Figure 63. A reconstruction of the structure of α-corticotropin from the structure of the peptides attained by proteolytic and partial acid hydrolysis.[1]

As we shall discuss subsequently, the knowledge of the complete covalent structure of a protein molecule is, unfortunately, not a magic key to the understanding of physicochemical and catalytic behavior in itself. A real appreciation of Nature's intent with respect to protein molecules must be sought through additional considerations of structure in three-dimensional terms.

Secondary and Tertiary Structure

Having established the details of the covalent structure of a protein, we may proceed to a comparison of these organochemical details with physicochemical behavior in solution. The total number of charges, positive and negative, along the polypeptide chains should agree with the values determined by titrating the protein with acid or base. A knowledge of the number of tyrosine, phenylalanine, and tryptophan residues in the molecule should make it possible to estimate, in advance, the shape and height of ultraviolet light absorption curves in the wavelengths from 250 to 290 mμ. Determinations of the content of sulfhydryl groups in the protein should confirm the analytical values for cysteine.

It was not surprising to the physical biochemists at work on such comparisons to find that these predictions seldom coincide with actuality. In general, many charges are masked or sequestered, absorption spectra are shifted from the theoretical, and sulfhydryl groups often appear in force only after serious denaturation of the protein molecule. The deviations are attributable to a number of known, and probably many unknown, parameters of secondary and tertiary structure which we shall attempt briefly to summarize.

We have already referred to the likelihood of helical coiling in proteins as the central theme of secondary structure. According to the dimensions of the α-helix and to the pitch of its screw, each residue in the polypeptide chain should contribute 1.5 A. to the length of the helical coil. Thus, in a helix with n residues, the total length of the coil should be $n \times 1.5$ A. (see Figure 54). By an extensive series of studies of the viscosity, light scattering, and sedimentation properties of the synthetic polyamino acid poly-γ-benzyl-L-glutamate in favorable solvents, Doty and his colleagues obtained data in full agreement with this theoretical calculation; that is, experimentally determined molecular lengths, in angstroms, were equal to the number of residues in the chain multiplied by 1.5.[5] It should be emphasized in the following that, although the word "helix" is

used freely for convenience, the presence of α-helical coiling in globular proteins is supported only by very circumstantial evidence. We may safely assume that there exists *some* sort of ordered folding within such proteins, but we must be wary of getting into the habit of thinking in terms of the α-helix exclusively until more data become available. It has not been established that all *proteins* contain such a structure, and the examination of this point, with special reference to globular proteins, is a particularly active area of research at the present time.

The study of the properties of helical systems in solution depends to a large degree on the use of optical rotatory methods. The mathematical theory of optical rotation by proteins is extremely complex and speculative. In spite of theoretical difficulties, however, there appears to be a high degree of consistency and predictability in connection with optical rotatory measurements, and we can gain an adequate appreciation of the usefulness of the method from a purely empirical point of view.

The optical rotation of a protein is made up of several components, one of which is the average rotation of the individual residues in the polypeptide chains. Superimposed on this rotation is the contribution of the ordered, repeating character of the helical configuration when this is present in the structure. It has been found experimentally, by studying the rotation of synthetic polyamino acids which can be made to assume varying degrees of helical coiling by modification of the solvent, that fully coiled chains possess an optical rotation approximately 90° (for the D line of the sodium lamp) greater than that to be expected as an average for the residues themselves. This completely empirical number may be used as a basis for calculating the content of helical coiling in a protein, according to the following equation:

$$\% \text{ helical coiling} = \frac{R_{\text{folded}} - R_{\text{unfolded}}}{90°} \times 100$$

The mean residue rotation of the folded protein, R_{folded}, is obtained by measurements on the native protein in water, and R_{unfolded} is estimated by measuring in the presence of such materials as urea or guanidine, or at elevated temperatures, under which conditions coiling, stabilized by hydrogen bonds, may be assumed to be destroyed. Some examples of optical rotatory measurements are given in Table 5 together with estimates of the amounts of helical coiling in these proteins. A measure of the rotation of the unfolded state may also be obtained on samples in which disulfide cross-linkages have been

TABLE 5

**Rotatory Properties of Native Proteins in the Folded and Unfolded States**[a]

Protein	R_{folded}	8 M urea R_{unfolded}	Calculated Per Cent of Protein in Helix
Ribonuclease	−90.7	−126.5	40
Insulin	−36.5	− 98.3	68
Ovalbumin	−33.9	−106.8	81
Lysozyme	−53.3	− 96.0	47
Chymotrypsin	−71.9	−115.9	49
Silk fibroin[b]	−40.	− 51.3	12

[a] C. Schellman and J. A. Schellman, *Compt. rend. trav. lab. Carlsberg, Ser. chim.*, **30**, 463–500 (1958).

[b] Silk fibroin is an example of a protein which is almost free of stabilizing cross-linkages and consequently exists in a form nearly free of secondary structure. Its mean residue optical rotation in urea (R_{unfolded}) is low owing to the high content of the optically inactive amino acid glycine.

broken by reductive or oxidative cleavage. In the absence of such stabilizing linkages, the helix is apparently unstable, and we observe rotations such as might be expected for the average amino acid residue, that is, about −100 to −120°. Disulfide bridges are not enough in themselves, however, to insure helical configuration in a protein molecule. We must take into account the fact that agents like urea can cause disorientation of secondary structure, even in the presence of intact disulfide linkages. It seems certain that a number of tertiary structural features also contribute to the determination of protein structure in solution. The presence of such structural interactions are indicated by the lack of agreement, mentioned earlier, between the experimentally determined values for ionizable groups, sulfhydryl groups, and the like and the values predicted from sequence data alone.

Functional groups which exhibit modified properties can be studied in a quantitative way. An excellent example comes from the spectrophotometric examination of proteins. It was first observed by Crammer and Neuberger[6] in 1943 that the spectrum of ovalbumin in the region attributable to tyrosine residue absorption was shifted to shorter wavelengths as compared with the spectrum of tyrosine itself. These results indicated that the absorption of light energy by the tyrosine chromophore had been modified in some way by the environ-
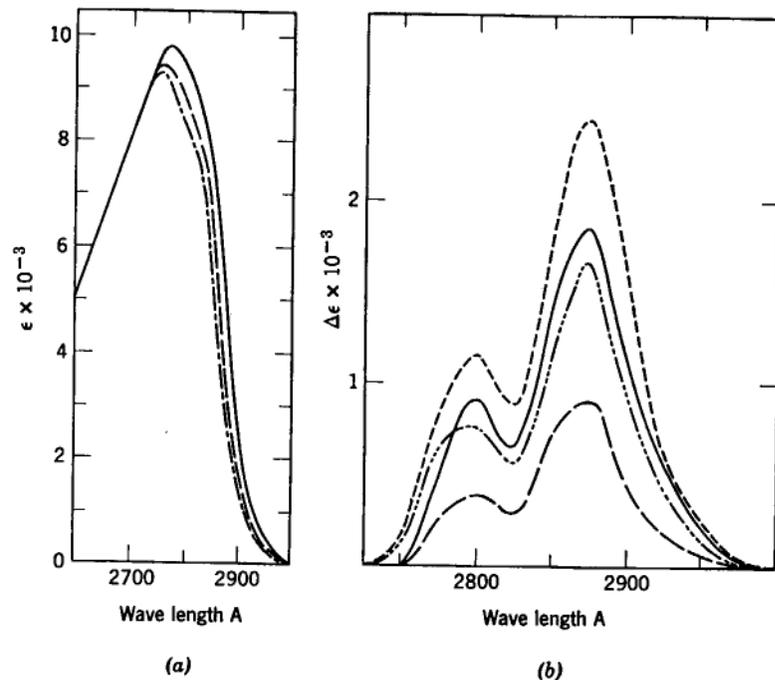
Figure 64. (a) Spectra of inactive ribonuclease (top curve), of inactive derivative of ribonuclease isolated from limited pepsin digests (middle curve), and of an exhaustive pepsin digest of ribonuclease (bottom curve). (b) Differences in extinction at various wavelengths between native ribonuclease and a complete pepsin digest of ribonuclease (upper curve), of the inactive pepsin derivative (second curve), of ribonuclease in 8 M urea (third curve), and of ribonuclease in 8 M urea in the presence of 0.003 M phosphate ions (bottom curve). From C. B. Anfinsen, *Federation Proc.*, **16**, 783 (1957).

ment of the tyrosine residues within the structure of the protein molecule. Similar observations were subsequently made for a number of other proteins including ribonuclease, lysozyme, and serum albumin. It was ultimately shown that the shift, characteristic of the abnormal spectrum (shown for ribonuclease in Figure 64), could be abolished by agents like urea which cause the rupture of hydrogen bonds. Simultaneously there were released, in a titratable form, carboxyl groups which had hitherto been masked and unavailable to titration by alkali. The number of such unmasked acid groups approximated the number of tyrosine residues which, it could be calculated, were responsible for the shift in extinction maximum. These findings furnished evidence for the presence of tyrosine hydroxyl-

carboxylate interactions of the type we have already shown schematically in Figure 55.[*] The stage was, of course, set for the detection of these particular hydrogen-bonded interactions since there is available here a simple method of observation, namely the spectrophotometer. Other hydrogen-bonded combinations, for example, between amino groups and carboxyl groups or between two carboxyl groups, may also be expected to exist in proteins, but they have so far escaped detection because of the lack of an adequate assay procedure for them.

Our consideration of the protein structure up to this point has been concerned with *static or equilibrium* properties. The covalent bonds in proteins can be relied on to resist most of the environmental fluctuations to which proteins are normally exposed. In a sense, they determine the three-dimensional potentialities by establishing a fixed distribution of charges and lyophobic groupings along the polypeptide chain and by acting as the basis for rigid and permanent cross-linkages. The covalent bonds do, however, have a certain degree of flexibility, and the slight deformations in bond angles between atoms, together with the rotation of atoms around single bonds, make it possible for a protein to assume a vast number of slightly different configurations in solution. When we measure the viscosity, optical rotation, or ultracentrifugal behavior of a protein preparation, the results obtained relate only to the *average* molecule in the solution. In general, the deviation from this average is probably fairly small because some arrangements have a much higher stability (i.e., a lower total free energy of configuration). The small fluctuations in configuration which do occur are, however, undoubtedly of considerable importance in connection with the biological activity of proteins and fortunately can be studied by means of a few specialized techniques which measure the *dynamic* aspects of protein structure.

As we have discussed earlier, it is possible to estimate the extent of helical coiling in proteins by measuring optical rotatory properties if we are willing to equate "repeating units of asymmetry" with "helix." The theoretical and experimental support for this assump-

[*] It will be noted that the spectral characteristics of tyrosine in native ribonuclease (and in other proteins in which similar effects have been observed) resemble to some extent those exhibited on ionization of the phenolic hydroxyl group of free tyrosine. In both instances a shift of the spectrum toward longer wavelengths as well as an increase in absorption occurs. It seems likely that formation of a hydrogen bond could bring about spectral changes analogous to those of ionization by increasing the phenolic O—H bond distance. Indeed a number of spectrophotometric studies on substituted phenols have demonstrated a shift of the spectrum toward the red under conditions favoring hydrogen bond formation.
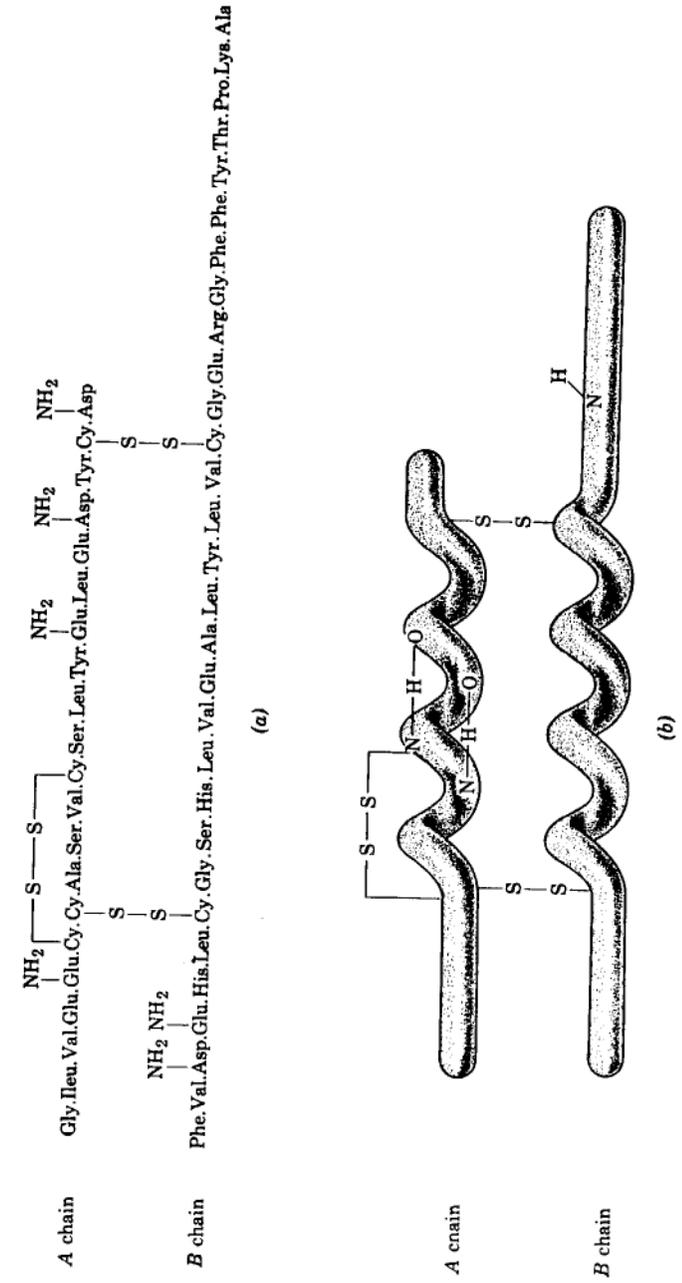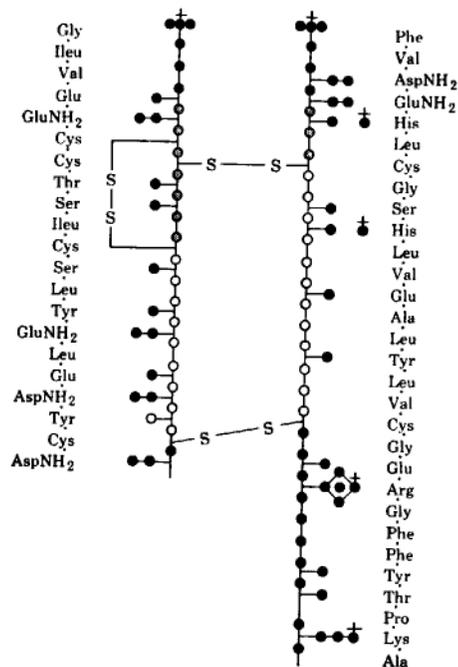
Figure 65. The structure of insulin (from A. P. Ryle, F. Sanger, L. F. Smith, and R. Kitai, *Biochem. J.*, **60**, 541 (1955), and (*b*) a schematic drawing of how insulin might be coiled within the portion of the molecule which is stabilized by disulfide bridges. Hydrogen atoms involved in bonding between amide groups within this region might be expected to be in much slower equilibrium with the hydrogen atoms of water molecules than those in the free "tails."

tion is good enough to satisfy most experts as a working hypothesis. By assuming, then, the presence of a certain amount of helical coiling in a protein like insulin, various models may be constructed based on the covalent structure derived by Sanger and his colleagues. In constructing such models it is necessary to guess where the helical portion belongs. Guessing is made somewhat easier by the observation that random polypeptide chains without internal cross-linkages do not spontaneously coil up into organized structures, as judged from viscosity studies and from the effect on optical rotatory properties when disulfide cross-linkages are ruptured by oxidation or reduction. Thus, in the case of insulin, whose structure is shown in Figure 65a, we may assume that the 60 per cent or so of the polypeptide chain which appears, polarimetrically, to be in the α-helix form is most likely to be located within the sterically stabilized region between the disulfide bridges joining the A chain to the B chain. The schematic drawing (Figure 65b) is intended to convey the idea of steric hindrance within the coiled portions of the chains. Within such a region of a protein molecule, we might expect to find that the CONH hydrogens involved in the hydrogen bonding of the helix are much less likely to exchange with the hydrogens of the water in which the protein is dissolved than are the hydrogens of side-chain groups or of CONH bonds outside the helical region. It is precisely this relative stability which the technique of *deuterium exchange* can measure, and this technique constitutes one of the most sensitive methods for the measurement of the *dynamic* state of protein structure.[7] Its particular value lies in the fact that, even when other physical measurements indicate that the average, equilibrium form of a protein contains some helical structure, it can measure the degree of deviation from the average resulting from the reversible opening and closing of hydrogen bonds with accordion-like flapping of the helix.

In practice, a protein is "loaded" with deuterium by solution in deuterium oxide for a time sufficient to allow all exchangeable hydrogens to come into equilibrium with the deuterium of the solvent. The protein is then carefully dried. Upon resolution in ordinary water, deuterium atoms equilibrate with the water, and the rate of appearance of deuterium, as measured by density change in the solvent, permits the division of total exchangeable atoms into those that are instantaneously, rapidly, and slowly exchangeable.

In insulin there are 91 theoretically exchangeable hydrogen atoms. These include the hydrogen atoms on the side chains of various amino acid residues and the CONH hydrogens forming the polypeptide

Gly
Ileu
Val
Glu
GluNH₂
Cys
Cys
Thr
Ser
Ileu
Cys
Ser
Leu
Tyr
GluNH₂
Leu
Glu
AspNH₂
Tyr
Cys
AspNH₂

Phe
Val
AspNH₂
GluNH₂
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

Figure 66. The relative degrees of "exchangeability" of various hydrogen atoms in insulin with deuterium atoms in the solvent: solid circles, instantaneous exchange; crosshatched circles, exchange at intermediate rates; open circles, restricted exchange. From K. Linderstrøm-Lang, *Symposium on Peptide Chemistry*, Special Publications 2, The Chemical Society, London, 1955.

chain. When deuterium-loaded insulin is dissolved in water at 0°C., 60 of these 91 exchange *instantaneously*. With further incubation additional exchange of other hydrogen atoms occurs, but at this temperature the rate of equilibration is relatively slow. In the presence of denaturing agents like urea, or at higher temperatures, the rate of equilibration is greatly increased, and all 91 of the deuterium atoms are ultimately exchanged.

These results have been interpreted by Linderstrøm-Lang[7] as shown in Figure 66. The *slowly* exchangeable hydrogen atoms (indicated by open circles) have been assigned to the region of insulin where the helical configuration is assumed to reside. The assignment of *"rapidly"* exchangeable atoms is somewhat more arbitrary, but can be made, in part, to that region of the A chain within the disulfide "loop" which, when forced into a helical arrangement in models, permits the formation of the intrachain disulfide bridge only with considerable bond deformation.

This interesting method furnishes perhaps the most clear-cut evidence for the existence of a continual flux of local minor rearrangement within the structure of proteins in solution. It brings out, in a particularly dramatic way, the intrinsic instability of the three-dimensional fabric of proteins which can lead to denaturation and loss of biological activity, and it emphasizes some of those structural components whose presence was not even suspected only a few years ago.

REFERENCES

1. F. Sanger in *Currents in Biochemical Research* (D. E. Green, editor), Interscience Publishers, New York, p. 434, 1956.
2. C. H. W. Hirs, W. G. Stein, and S. Moore, *J. Biol. Chem.*, 211, 941 (1954).
3. S. Moore and W. H. Stein in *The Harvey Lectures 1956–1957*, Academic Press, New York, p. 119, 1958.
4. C. H. Li, I. Geschwind, R. D. Cole, I. Raacke, J. Harris, and J. Dixon, *Nature*, 176, 687 (1955).
5. B. W. Low in *The Proteins*, volume 1, Part A (H. Neurath and K. Bailey, editors), Academic Press, New York, Chapter 4, 1957.
6. J. L. Crammer, and A. Neuberger, *Biochem. J.*, 37, 302 (1943).
7. K. V. Linderstrøm-Lang in *Symposium on Peptide Chemistry*, Special Publication 2, The Chemical Society, London, 1955.

SUGGESTIONS FOR FURTHER READING

Anfinsen, C. B., and R. R. Redfield, in *Advances in Protein Chemistry*, volume 11, Academic Press, New York, 1956.

Crick, F. H. C., and J. C. Kendrew, in *Advances in Protein Chemistry*, volume 12, Academic Press, New York, 1957.

Edsall, J. T., and J. Wyman, *Biophysical Chemistry*, volume 1, Academic Press, New York, 1958.

Linderstrøm-Lang, K. V., *Lane Medical Lectures*, volume 6, Stanford University Press, Stanford, 1952.

Low, B. W. and J. T. Edsall in *Currents in Biochemical Research*, (D. E. Green, editor), Interscience Publishers, New York, p. 379, 1976.

Pauling, L., and R. B. Corey, *Proc. Natl. Acad. Sci. U.S.*, 37, 235 (1951).

Pauling, L., R. B. Corey, and H. R. Branson, *Proc. Natl. Acad. Sci. U.S.*, 37, 205 (1951).