M. F. PERUTZ

PROTEINS AND NUCLEIC ACIDS

PROTEINS AND NUCLEIC ACIDS

STRUCTURE AND FUNCTION

Eighth Weizmann Memorial Lecture Series
April 1961

by

M. F. PERUTZ
Medical Research Council Laboratory
of Molecular Biology,
and
University of Cambridge, Great Britain

ELSEVIER PUBLISHING COMPANY
AMSTERDAM / LONDON / NEW YORK
1962

Preface

It was a great pleasure to go to the Weizmann Institute and to witness how the seed sown by its Founder has grown into a sturdy tree with many flourishing branches. Its fruit are the scientific achievements of the Institute which are known throughout the world for their high intellectual standard. I feel very honoured at having been asked to lecture at this leading centre of science, in memory of its great Founder.

I have tried to describe some of the spectacular progress in the field of molecular biology. My three lectures were concerned with the structure of proteins, the structure and replication of the genetic material, and the structure and function of the substances concerned with the biosynthesis of proteins. This book is a slightly expanded version of the lectures, and I hope that it will serve as an introduction for newcomers to the field of Molecular Biology. I have not attempted to be comprehensive, but have limited myself to some advances which seemed to me particularly significant. The experts will discover many omissions. These might be filled in by reference to the literature which I have listed as "Suggestions for further reading".

In conclusion I wish to thank the members of the Weizmann Institute for the warm reception given to my family and myself during our visit, and especially Professor Gerhard Schmidt, the Chairman of the Scientific Council, and Dr. Judith Bregman for their thoughtful and generous hospitality.

I also wish to thank all those who have helped me with the preparation of the book. These include D. M. Blow, S. Brenner, F. H. C. Crick, F. Jacob, J. C. Kendrew, M. Meselson, J. D. Watson, M. H. F. Wilkins and P. C. Zamecnik, who read and criticized parts or the whole of my manuscript; Miss Patricia White, who typed it; and my wife, who helped me with the references, the index, and the checking of the manuscript.

Finally, I should like to express my gratitude to the many colleagues and publishers who kindly allowed me to reproduce illustrations from books and papers.

ERRATUM

p. 119. It is stated that red cell reticulocytes still contain small quantities of DNA (Holloway and Ripley, 1952). I have since learnt that no DNA is found in reticulocyte preparations after careful removal of white blood cells (H. Borsook, private communication).

ACKNOWLEDGEMENTS

Thanks are due to the copyright holders and authors who kindly permitted the reproduction of published illustrations: to the Biochemical Society for Fig. 2; to Interscience Publishers for Fig. 3; to Nature for Figs. 4, 8, 9, 12, 40, 41 and 47; to Dr. J. C Kendrew and Dr. H. C. Watson for Figs. 5, 6, 10 and 11; to Princeton University Press for Fig. 7; to the Royal Society for Figs. 13–21 and 27; to Imperial Chemical Industries, Ltd. for Fig. 24; to the Journal of Molecular Biology for Figs. 26, 31, 36–38, 42, 56 and 57; to Scientific American, Inc. for Fig. 28; to the National Academy of Sciences of the U.S.A. for Figs. 29, 32, 48, 50 and 51; to Dr. M. Meselson for Fig. 30; to the American Society of Biological Chemists, Inc. for Figs. 33, 34 and 43; to Dr. J. Cairns for Fig. 35; to Academic Press, Inc. for Fig. 44; to Blood for Fig. 52. The references of the figures are included in the List of References.

Contents

Introduction

ENZYMES AND GENES

The bewildering variety of living forms conceals an underlying unity in the molecules that nature uses for its many purposes. Similar enzyme molecules catalyse the same reactions in unicellular organisms and in mammals; some of the same cofactors are probably at work in glycolysis and in photosynthesis; the same few simple molecules are used in the biosynthesis of the many complex organic compounds found in living organisms. Almost all metabolic processes are catalysed by enzymes and all enzymes are protein molecules. By teaching this creed 30 years ago Sir Frederick Gowland Hopkins created the scientific climate in which our research on the chemical structure and spatial architecture of proteins was begun.

In recent years biochemists have started to probe deeper, asking themselves what controls the synthesis of enzymes. It cannot be other enzymes, since they would have to be made by yet more enzymes and so on *ad infinitum*. We now know that enzyme synthesis is controlled by genes and we have good reason to believe that one gene controls the synthesis of one, or of part of one, enzyme. A gene must therefore possess a dual function: it must be able to replicate itself and to determine the specific structure of a protein molecule. This poses the question of the chemical nature of genes.

It would appear the simplest hypothesis, at any rate at first sight, to suppose that proteins were self-reproducing molecules in addition to being functional catalysts. However, recent research on viruses and bacterial transforming factors has given convincing proof that the essential part of a gene is made not of protein, but of nucleic acid. This may be of 2 different kinds; deoxyribonucleic acid (DNA) in cellular organisms and large viruses, or ribonucleic acid (RNA) in certain small viruses. Biological replication and metabolic activity therefore appear to depend primarily on the structure and interaction of 2 types of very large molecules: nucleic acids and proteins. The part played by lipoids and polysaccharides is not yet clear; at present it seems to be secondary.

The basic questions, therefore, are these. What structures do enzymes have and how do the structures determine their catalytic function? What is the structure of the genetic material? How does it replicate itself? How does it control the synthesis of enzymes? These are the questions to be discussed in the following chapters.

Chapter 1

The Structure of Proteins: Myoglobin and Haemoglobin

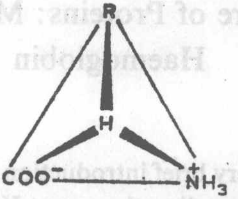
This chapter gives a very brief introduction to the general principles of protein structure, describes the recent X-ray analyses and chemical studies of myoglobin and haemoglobin and tries, as far as possible, to relate the results to biological function. The structural significance of the amino acid replacements in some of the abnormal haemoglobins is next considered. Finally, the nature of the forces determining the three-dimensional architecture of proteins, and the relationship between their structure in the crystal and in solution are discussed.

1. THE AMINO ACIDS

Proteins consist of amino acids linked together by peptide bonds. The commonly occurring amino acids are of 20 different kinds which all have the same dipolar ion group $\overset{+}{H_3N} \cdot \underset{|}{CH} \cdot COO^-$ in common (Fig. 1). The $—NH \cdot \underset{|}{CH} \cdot CO—$ group, derived from it by the elimination of H_2O, forms the backbone of the polypeptide chain.

Specificity is provided by the 20 different kinds of side chains R.

By convention, the carbon atoms of the amino acids are designated by Greek letters, the CH group which forms part of the main chain being called α, the first carbon of the side chain β, and so on. It will be noted that the α-carbon carries 4 different chemical substituents in all amino acids except glycine. Of the 2 possible enantiomorphs only the L form occurs in proteins. Its configuration is shown below.



Taking the hydrogen atom as the apex of a tetrahedron and looking down on it, the R, amino and carboxyl groups succeed each other in a clockwise direction.

The simplest amino acid is glycine where R is a hydrogen atom.

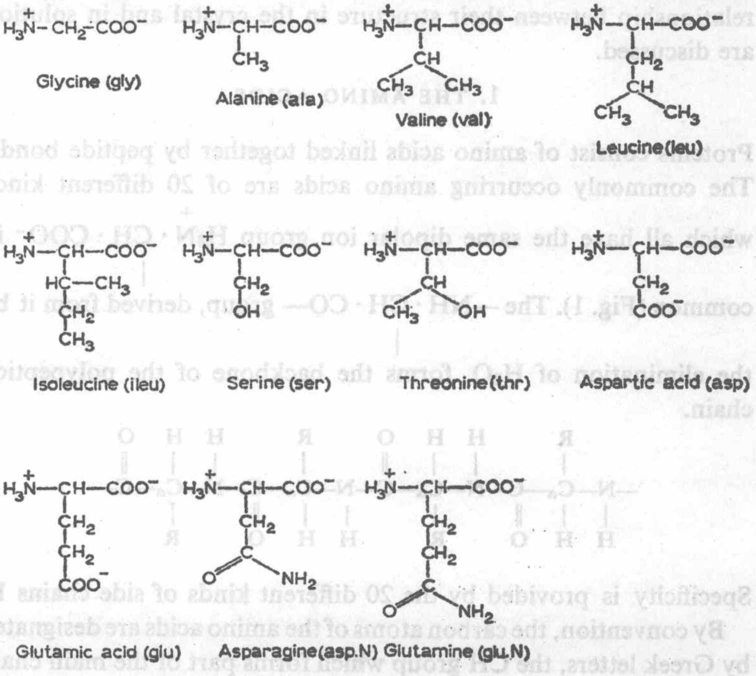
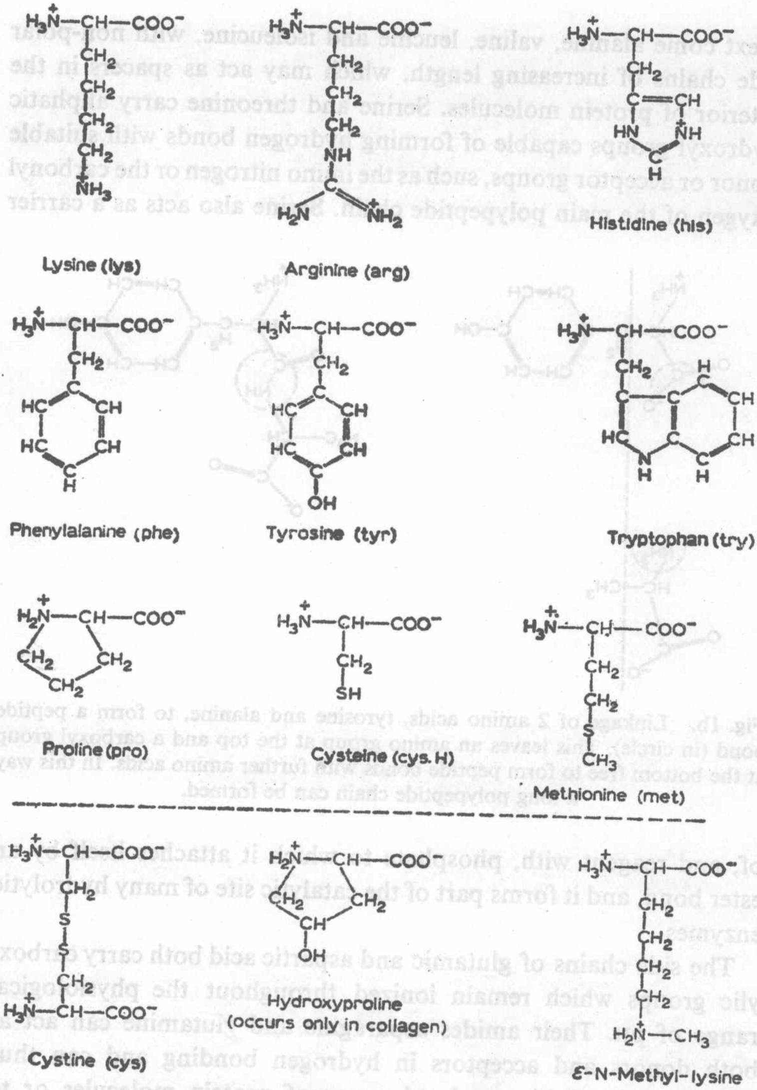

Glycine (gly)
Alanine (ala)
Valine (val)
Leucine (leu)
Isoleucine (ileu)
Serine (ser)
Threonine (thr)
Aspartic acid (asp)
Glutamic acid (glu)
Asparagine (asp.N)
Glutamine (glu.N)

Fig. 1a (1).



Lysine (lys)

Arginine (arg)

Histidine (his)

Phenylalanine (phe)

Tyrosine (tyr)

Tryptophan (try)

Proline (pro)

Cysteine (cys. H)

Methionine (met)

Cystine (cys)

Hydroxyproline (occurs only in collagen)

ε-N-Methyl-lysine

Fig. 1a (2).

Fig. 1a. The 20 commonly occurring amino acids (Abbreviations in brackets). The 3 amino acids below the broken line are formed by modification of one of the 20 after they have been assembled in the polypeptide chain; cystine is formed from cysteine, hydroxyproline from proline and ε-N-methyl-lysine from lysine.

Next come alanine, valine, leucine and isoleucine, with non-polar side chains of increasing length, which may act as spacers in the interior of protein molecules. Serine and threonine carry aliphatic hydroxyl groups capable of forming hydrogen bonds with suitable donor or acceptor groups, such as the imino nitrogen or the carbonyl oxygen of the main polypeptide chain. Serine also acts as a carrier of, and reagent with, phosphate to which it attaches itself by an ester bond, and it forms part of the catalytic site of many hydrolytic enzymes.

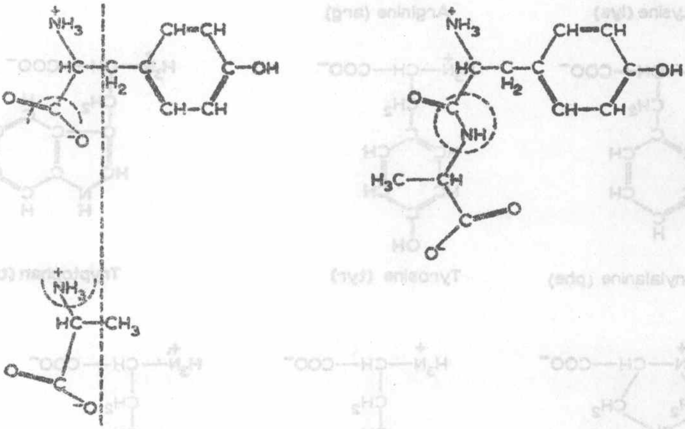


Fig. 1b. Linkage of 2 amino acids, tyrosine and alanine, to form a peptide bond (in circle). This leaves an amino group at the top and a carboxyl group at the bottom free to form peptide bonds with further amino acids. In this way a long polypeptide chain can be formed.

The side chains of glutamic and aspartic acid both carry carboxylic groups which remain ionized throughout the physiological range of pH. Their amides asparagine and glutamine can act as both donors and acceptors in hydrogen bonding and can thus contribute to the internal coherence of protein molecules or to their solubility in water.

Lysine and arginine carry cationic groups which also are ionized at physiological pH. They and the 2 carboxylic acids are the principal contributors to the electric charge of protein molecules and give them their properties of dipolar ions. Minor contributors are the terminal carboxyl and amino groups of the polypeptide chain, which are also ionized, and the histidines. These occupy a unique position, because they can change their state of ionization within the physiological range of pH. Their imidazole side chain carries a positive charge, provided by an extra proton, below about pH 6 (the exact pH varies with the environment) and is neutral above. For this reason, and because it forms co-ordination complexes with iron and other metals, histidine is often found at the catalytic site of enzymes.

Of the 3 amino acids with aromatic side chains, tyrosine and tryptophan carry hydrogen-bonding groups. Phenylalanine is non-polar. The structure of myoglobin suggests that aromatic rings may contribute to the stability of proteins by π bonding, or by acting as non-polar spacers, like the aliphatic side chains. Tryptophan appears to be a more specialized structure than the others, but its purpose is not yet clear.

The δ-carbon atom of the side chain of proline is linked to the imino nitrogen of the main chain, thereby inhibiting its participation in hydrogen bond formation and making proline a misfit in any helical polypeptide chain.

Cysteine carries the highly reactive sulphydryl group. This probably does not ionize nor form hydrogen bonds of significant strength, but 2 cysteines placed some distance apart along a polypeptide chain, or forming part of different chains, can be joined by oxidation to form the disulphide bridge of cystine which plays an important part in stabilizing protein structures. Cysteine may also be important in enzymic reactions involving the transfer of hydrogen, and in one enzyme, cytochrome *c*, it forms a thioester link between the protein and the non-protein part of the structure. The sulphur atom in methionine, on the other hand, is unreactive and may serve no function other than imposing a special configuration on the aliphatic side chain.

The stereochemical configuration of the amino acids, their interatomic distances and bond angles, and the absolute configuration of the L and D forms are known from crystal structures analyses.

Unfortunately there exists as yet no handbook in which all these data are collected, but tables with references to the original papers are included in the reviews by Kendrew and Perutz (1957) and by Rich and Green (1961). The physical chemistry of the amino acids in solution has also been studied in great detail, but in this case the data have been collected, correlated and discussed in relation to proteins in the classical monographs of Cohn and Edsall (1943) and of Edsall and Wyman (1958).

2. THE PRIMARY STRUCTURE OF PROTEINS

An enzyme molecule may consist of one or more polypeptide chains which, together, may contain between a hundred and several thousand amino acid residues, all arranged in a definite sequence. The number of chains and the sequence of residues within them constitute the primary structure of proteins. By convention the residues are numbered in sequence along each chain, beginning from the amino-terminal end. If there are several chains in one molecule they are designated by Roman or Greek letters which are added as suffixes under the residue numbers. Analysis of the primary structure of proteins presented insuperable difficulties until the development of chromatographic methods (Martin and Synge, 1945). These formed the experimental basis for Sanger's attack on the constitution of insulin which reached its climax with the complete elucidation of the sequence of its 51 residues in 2 closely linked chains (Ryle *et al.,* 1955; Fig. 2). Sanger's discovery was one of the milestones in protein chemistry. First of all it removed the last shadow of doubt from the polypeptide hypothesis enunciated by Hofmeister more than 50 years earlier. It established the fact that the amino acid residues really are arranged in a definite, genetically determined sequence, but disproved the widely held belief that this sequence was regular. It revealed the part played by cystine bridges in the architecture of protein molecules, and the chemical nature of species specificity. Most important of all, Sanger demonstrated that the complete formula of a protein can be determined by chemical methods, at least as far as the primary covalent bonds are concerned, and thereby stimulated a great volume of new research all over the world.

Since Sanger finished his work on insulin, he and others have tried to improve the methods of sequence study and to apply them to larger proteins (*e.g.* Hirs, Stein and Moore, 1956; Spackman, Stein and Moore, 1958; Braunitzer *et al.*, 1959). Sequences now completed include those of several large peptide hormones, of the enzymes ribonuclease and cytochrome *c*, of the 2 chains of human haemoglobin, and of the single long chain of the protein of tobacco mosaic virus. The extent of recent progress can be gauged from a comparison between the longer of the 2 chains of insulin analysed by Sanger and Tuppy in 1951, which contains 30 residues, and that of tobacco mosaic virus protein which contains 158. However, the difficulties of fractionation and analysis rise rapidly with the size of the protein and each analysis of even a small protein takes a team of research workers several years. Moreover, the number of known enzymes is over 1000 and most of them have molecular weights several times larger than the tobacco mosaic virus protein, which shows that the development of more rapid and sensitive methods is vitally important.

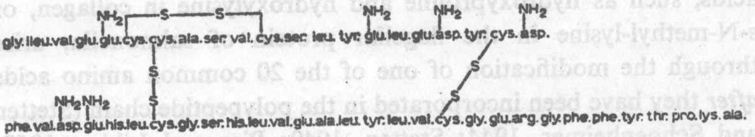

NH2
S
S
NH2
NH2
NH2
gly. ileu.val.glu.glu.cys.cys.ala.ser. val.cys.ser. leu. tyr. glu.leu.glu.asp.tyr. cys. asp.
S
S
S
S
NH2 NH2
phe.val.asp.glu.his.leu.cys.gly.ser.his.leu.val.glu.ala.leu. tyr. leu.val.cys.gly.glu.arg.gly.phe.phe.tyr. thr. pro.lys.ala.

Fig. 2. Chemical constitution of beef insulin (Ryle *et al.*, 1955).

Many enzymes contain one or more non-protein (prosthetic) groups which form the site of their catalytic activity. These may be metal-organic compounds as in the respiratory carriers and in certain respiratory enzymes, or pigments related to certain B vitamins, or nucleic acid derivatives, and others. It is a characteristic feature of these enzymes that the prosthetic groups by themselves generally do not possess the catalytic activity, and that this is conferred on them only by combination with a *specific* protein. Cases are known in which combination with different proteins confers different catalytic activity on the same metal-organic pigment.

It is an astonishing fact that the proteins of all species investigated contain the same 20 L-amino acids. It will be shown in later chapters that polypeptide chains are assembled by a template mechanism which apparently is so geared that no amino acids other than the 20 commonly occurring ones are accepted. It may be objected that other L-amino acids do occur in proteins and that small peptides with D-amino acids have been found in certain microorganisms. However, as far as is known the supernumerary amino acids, such as hydroxyproline and hydroxylysine in collagen, or ε-N-methyl-lysine in the flagellar protein of salmonella, arise through the modification of one of the 20 common amino acids *after* they have been incorporated in the polypeptide chain (Stetten and Schoenheimer, 1944; Stetten, 1949; Piez and Likins, 1957; Sinex *et al.*, 1959; Stocker *et al.*, 1961). D-Amino acids are probably produced, and incorporated into peptides, by special enzymic mechanisms.

3. THE CONFIGURATION OF THE POLYPEPTIDE CHAIN: SECONDARY STRUCTURE

Long-chain polymers in which the same atomic pattern repeats at regular intervals tend to have screw symmetry. This means that each unit of pattern, a methylene or an isoprene group say, is brought into congruence with its neighbours along the chain by a rotation about a common axis and a translation along it. Polypeptide chains are no exception, despite the diversity of their side chains. The nature of the side chains does, however, decide the most stable among several possible screw symmetries.

At first sight it would appear that a polypeptide chain might be able to assume a very large number of different configurations, but in fact rotation is restricted to a greater or lesser extent about each of the 3 different bonds making up the chain, and no configuration is stable unless it allows every imino group to be hydrogen-bonded to a carbonyl belonging either to the same chain or to a neighbouring one. The exact nature of these stereochemical restrictions was first pointed out by Pauling, Corey and Branson (1951) and by Pauling and Corey (1951), and has been summarized in Edsall and Wyman's monograph which gives an excellent introduction to the subject of polypeptide chain configuration. Briefly, the structures so far encountered can be described under 2 headings.

a. Structures with hydrogen bonds between chains

Three structures of this type are known. The simplest is a planar zigzag chain which occurs in synthetic polyamides (nylon). Neighbouring chains are joined to form sheets by hydrogen bonds between carbonyl and imino groups, while neighbouring sheets cohere through the residual forces provided by non-polar contacts. However, Huggins (1943), and Pauling and Corey (1951) discovered that side chains could not be accommodated if the polypeptide chains in the hydrogen-bonded sheets were fully extended, and that this difficulty could be overcome by pleating the sheets at right angles to the chain direction. This has the effect of making the C_α–C_β bond extend at right angles to the plane of the sheet and allows the methyl groups of alanine, for instance, to be neatly packed. The structure has been proved to occur in synthetic poly-L-alanine and poly-γ-methyl-L-glutamate, and in silk fibroin, where it is probably stabilized by the high content of glycine and alanine. It may also occur in stretched fibres of the keratin–myosin group, and possibly in feather keratin where its presence has been suggested, but is still controversial.

In the pleated sheet each residue is related to its neighbours along the chain by a rotation of 180° and a translation of between 6.5 and 7.0 Å. Cowan and McGavin (1955) discovered another type of structure in which residues are related by a rotation of 120° and a translation of 3.1 Å. This occurs as a left-handed helix in the synthetic polypeptides poly-L-proline and poly-L-hydroxyproline (Sasisekharan, 1959) and in one modification of polyglycine (Crick and Rich, 1955). It is important because it forms the structural basis of collagen (Rich and Crick, 1955 and 1961).

b. Structures with hydrogen bonds within the same chain

All structures of this type are helical. Sets of atoms form hy-