THE PRINCIPLES OF HUMAN BIOCHEMICAL GENETICS

Second Revised and Enlarged Edition

HARRY HARRIS

FRONTIERS OF BIOLOGY, Vol. 19*

THE PRINCIPLES OF HUMAN BIOCHEMICAL GENETICS

Second revised and enlarged edition

HARRY HARRIS

Galton Laboratory, University College, London

1975

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · OXFORD

Library of Congress Catalog Card Number: 74-83271
ISBN North-Holland for this series 0 7204 7100 1
ISBN North-Holland for this volume 0 7204 7119 2
ISBN American Elsevier 0 444 10012 1

PUBLISHERS:

NORTH-HOLLAND PUBLISHING COMPANY - AMSTERDAM
NORTH-HOLLAND PUBLISHING COMPANY, LTD. - OXFORD

SOLE DISTRIBUTORS FOR THE U.S.A. AND CANADA:

AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

This book is a completely revised edition of Volume 19, published by North-Holland Publishing Co in 1970

This book is also published in a paperback student edition

PRINTED IN THE NETHERLANDS

Editors' preface

The aim of the publication of this series of monographs, known under the collective title of '*Frontiers of Biology*', is to present coherent and up-to-date views of the fundamental concepts which dominate modern biology.

Biology in its widest sense has made very great advances during the past decade, and the rate of progress has been steadily accelerating. Undoubtedly important factors in this acceleration have been the effective use by biologists of new techniques, including electron microscopy, isotopic labels, and a great variety of physical and chemical techniques, especially those with varying degrees of automation. In addition, scientists with partly physical or chemical backgrounds have become interested in the great variety of problems presented by living organisms. Most significant, however, increasing interest in and understanding of the biology of the cell, especially in regard to the molecular events involved in genetic phenomena and in metabolism and its control, have led to the recognition of patterns common to all forms of life from bacteria to man. These factors and unifying concepts have led to a situation in which the sharp boundaries between the various classical biological disciplines are rapidly disappearing.

Thus, while scientists are becoming increasingly specialized in their techniques, to an increasing extent they need an intellectual and conceptual approach on a wide and non-specialized basis. It is with these considerations and needs in mind that this series of monographs, '*Frontiers of Biology*' has been conceived.

The advances in various areas of biology, including microbiology, biochemistry, genetics, cytology, and cell structure and function in general will be presented by authors who have themselves contributed significantly to these developments. They will have, in this series, the opportunity of bringing together, from diverse sources, theories and experimental data, and of integrating these into a more general conceptual framework. It is unavoidable, and probably even desirable, that the special bias of the individual authors will become evident in their contributions. Scope will also be given for presentation of new and challenging ideas and hypotheses for which

complete evidence is at present lacking. However, the main emphasis will be on fairly complete and objective presentation of the more important and more rapidly advancing aspects of biology. The level will be advanced, directed primarily to the needs of the graduate student and research worker.

Most monographs in this series will be in the range of 200–300 pages, but on occasion a collective work of major importance may be included exceeding this figure. The intent of the publishers is to bring out these books promptly and in fairly quick succession.

It is on the basis of all these various considerations that we welcome the opportunity of supporting the publication of the series '*Frontiers of Biology*' by North-Holland Publishing Company.

E. L. TATUM
A. NEUBERGER, *Editors*

Preface

This book has grown out of a course of lectures given at the Galton Laboratory, which was intended not only for students and research workers in human genetics, but also for biochemistry, biology and medical students as well as for interested research workers in related fields. I was concerned to explain the principal concepts which underlie modern ideas in human biochemical genetics, to present a picture of the extraordinary degree of inherited biochemical diversity which recent research has shown to be a characteristic feature of human populations, and to show how the detailed analysis of genetically determined biochemical differences between individual members of our species could throw new light on fundamental problems not only in genetics, but also in medicine and more generally in human biology.

Just over ten years ago I wrote an account of the subject (Human Biochemical Genetics, Cambridge University Press, 1959) covering most of the information which was then available, in what seemed at the time a logical order. Since then, however, research in this field has expanded almost explosively and in preparing the present book, it became very obvious that one could not be content with simply trying to update the earlier text. The many advances now called for a very different arrangement if present knowledge and concepts were to be presented in a coherent manner. This is not merely because a great deal more is now known about the particular topics that were dealt with in the earlier book, but because whole new areas of the subject have been opened up in a manner which could hardly have been envisaged only a few years ago. One of the important consequences of these developments has been the greater unity they have given to the subject as a whole. The interrelationships between what at one time seemed very different and unconnected types of phenomena such as the inborn errors of metabolism, the blood group antigenic differences, the haemoglobin diseases and the enzyme and protein polymorphisms, can now be thought about within a consistent theoretical framework in a way that was hardly possible previously. This of course gives one an opportunity to try and present the subject in a more systematic and analytical manner. So the present work differs consider-

ably in its approach and arrangement from my earlier one, and of course much of the material discussed is new.

One of the difficulties in writing this kind of text is deciding what examples should be used to illustrate the various points in the argument, and in how much detail they should be presented. Also one has to decide what must be left out, if the overall length is to be kept to a manageable size, and the argument not be obscured by an excessive amount of descriptive material. Since a book of this sort may also be useful as a source of reference, one is often placed in something of a quandary. I have tried to resolve this difficulty by giving key references to much material which is not described in detail in the text, and a great deal of this has been arranged in the form of tables or appendices so that the appropriate references can be extracted more easily. Nevertheless the scope of the subject is now so very extensive and the literature so vast and distributed over such a wide range of journals, that reference to many topics must inevitably have been omitted. I hope nevertheless that enough has been included so that the book may serve not only as an introductory text in which the main principles of the subject are formulated, but also as guide to further reading on specific topics.

Galton Laboratory,
University College, London

HARRY HARRIS, M.D., F.R.S.
July, 1969

Preface to 2nd edition

In the past five years, work in human biochemical genetics has continued to advance very rapidly. Consequently in revising the text for this edition much new material has been included, though the general arrangement of the book has been retained.

Galton Laboratory,
University College London

HARRY HARRIS
March 1974

Acknowledgements

It is a great pleasure to thank Mrs. N. Parry-Jones for preparing the figures, Mrs. J. Barrie for her secretarial assistance and Dr. D. A. Hopkinson for much helpful discussion and advice while the book was being written.

Many of the illustrations come from other publications and the permission for their reproduction is gratefully acknowledged. The original sources are given in the captions and bibliography.

Contents

CHAPTER 1

Gene mutations and single aminoacid substitutions

1.1. Introduction: genes, DNA and proteins

Human beings are exceedingly diverse. They differ from one another in their normal physical, physiological and mental attributes. They also differ in whether they suffer from particular diseases or other abnormalities. These variations are caused in part by differences in the environmental conditions in which they live. But they also depend on inborn differences. Indeed it is very probable that no two individuals with the exception of monozygotic twins are exactly alike in their inherited constitutions. The analysis in molecular terms of the nature and effects of such genetically determined differences forms the subject matter of human biochemical genetics.

Classical genetics led to the concept of the gene as the fundamental biological unit of heredity and postulated that it must possess three basic properties. It had to have a specific function in the cell, and hence in the organism as a whole. It had to be capable of exact self-replication so that its functional specificity would be preserved from one cell generation to the next. Finally, although usually an extremely stable entity, it had to be susceptible to occasional sudden change or mutation, which could result in the appearance of a new unit or allele differing functionally from the original one but self-replicating in its new form. It was shown how such units are arranged in linear order in the chromosomes, each gene having its own characteristic position or locus; how they are transmitted to an individual from his parents via the ovum and sperm, so that they are usually present in pairs, one member of a pair being derived from one parent and one from the other; and how because of mutational changes in previous generations multiple allelic forms of a gene can occupy a particular gene locus, so that individual members of a natural population may differ from one another in their characteristics according to the specific nature of the alleles that they happened to have received from their parents.

Four major advances made it possible to begin to understand the nature of

genetical diversity in molecular terms. The first was the discovery that the particular chemical substance which endows a gene with its characteristic properties is deoxyribosenucleic acid or DNA. The second was the elucidation of the molecular structure of this substance. The third was the recognition that the primary biochemical role of DNA in the cells of an organism is to direct the synthesis of enzymes and other proteins. The fourth was the unravelling of the genetic code. That is the details of the relationships between the structure of nucleic acids and the structure of proteins.

The main features of the molecular architecture of DNA were first formulated by Watson and Crick in 1953, who at the same time pointed out how the proposed structure would account for the three basic attributes of genetic material; gene specificity, gene replication and gene mutation.

The molecule is made up of two very long polynucleotide chains coiled round a common axis to form a double helix. The backbone of each chain consists of a regular alternation of phosphate and sugar (deoxyribose) groups. To each sugar group and projecting inwards from the chain is attached a nitrogenous base. This may be one of four different types; adenine or guanine which are purines, thymine or cytosine which are pyrimidines. The two chains are held together by hydrogen bonding between pairs of bases projecting at the same level from each chain, so that the whole structure may be likened to a spiral staircase, the pairs of bases representing the steps. There are certain restrictions on which bases can constitute a pair. In any one pair one base must be a purine and the other a pyrimidine, and of the possible combinations only two can occur; adenine with thymine, and guanine with cytosine. A given pair may be either way round. Thymine for example can occur in either chain, but when it does its partner on the other chain must be adenine.

A gene can be regarded as being represented by a length of DNA containing several hundred or thousand base pairs. While the phosphate–sugar backbones of the two chains which form the double helix are quite regular, the base pairs may occur in any sequence. A great many different permutations are therefore possible, and so each gene can have its own unique structure, from which is derived its functional specificity. The precise sequence of base pairs in a particular gene carries as it were in coded form a specific piece of genetical information.

Since the nature of one base fixes the nature of the other member of the pair, the two polynucleotide chains which make up the molecule, though qualitatively different, are exactly complementary. The sequence of bases in one chain fixes the sequence of bases in the other. Replication can occur by

the unwinding and separation of the chains and the reformation on each chain of its appropriate companion from an available pool of nucleotides. Each chain may thus act as a template for the formation of the other, so that from one molecule two precise replicates are produced each with exactly the same sequence of base pairs as the original.

A gene mutation can be envisaged as the consequence of some kind of event which results in an alteration of the base pair sequence of the particular gene. Many and perhaps most mutations probably represent no more than the change of one base for another at some point in the sequence. Others however involve more drastic changes such as the duplication or deletion of part of the sequence or some other kind of rearrangement. In general the new gene structure once formed will then be conserved in subsequent cell divisions by the ordinary process of DNA replication.

A great variety of different enzymes and other proteins are synthesised in the cells of a single organism. They each have their own distinctive properties and functions and together they define and control the complex pattern of metabolic and developmental processes which characterise the species and the individual. Proteins are composed of one or more polypeptide chains which are made up of long strings of aminoacids linked by peptide bonds in a specific linear order. Twenty different aminoacids may be present and typical polypeptide chains have sequences 100–500 aminoacids long, so as with DNA the number of possible structures is enormous. Furthermore the three-dimensional arrangement and hence the characteristic properties and functional activity of any given protein ultimately depends on the precise sequence of aminoacids in its constituent polypeptide chains.

The fundamental idea relating DNA structure to protein structure is that the sequence of base pairs in a given gene determines the sequence of amino-acids in a corresponding polypeptide chain. So the structures and hence the properties of all the enzymes and proteins an individual can make are thought to be defined by the base pair sequences of his genes.

The details of the genetic code – that is the relationship of base sequence to aminoacid sequence – have largely been worked out by experimental studies on microorganisms. But there is little doubt that in their main features they also apply in higher organisms including man. Each aminoacid is specified by a sequence of three bases. The base triplets occur consecutively and do not overlap. That is to say a triplet specifying one aminoacid is immediately followed by a separate triplet specifying the next aminoacid and so on. Thus the two sequences are colinear. The four characteristic bases of a DNA chain can occur in 64 different triplet sequences, and 61 of these

triplets each specify one of the twenty aminoacids, so that a particular aminoacid may be coded by two or more different base triplets (see fig. 1.5, p. 14). There are also three so-called 'nonsense' triplets which do not specify aminoacids but probably designate chain termination.

The series of processes by which the sequence of bases in the DNA of a gene is translated into the corresponding sequence of aminoacids in a polypeptide chain are complex and involve as intermediaries certain types of ribosenucleic acid (RNA) molecules. The first step involves the separation of the two polynucleotide chains of the DNA, so that one of them may serve as a template for the synthesis from available ribonucleotides of a complementary RNA chain. In this process the same base pairing rules as in DNA apply, except that uracil, which occurs in RNA instead of thymine, pairs with adenine. Thus a strand of RNA carrying the same genetic information as the DNA strand is formed, but it is coded in a complementary base sequence. This RNA strand, known as messenger or mRNA, then separates from the DNA and passes out of the cell nucleus to the ribosomes in the cytoplasm, which are the site of protein synthesis. The strands of mRNA attached to ribosomes then serve as templates for the formation of the polypeptide chains. Aminoacids come to the mRNA template attached to another species of RNA molecule, known as transfer RNA or tRNA. The tRNA molecules are relatively small (about 80 nucleotides) and occur as a series of distinct molecular types. Each is specific for a particular aminoacid which can be attached to one end of the tRNA molecule, and each contains within its polynucleotide sequence a characteristic base triplet complementary to a base triplet in mRNA which codes for that aminoacid. The attached aminoacid can thus be placed in the correct position for the synthesis of the polypeptide chain defined by the coded sequence of the mRNA.

The polypeptide chain is made sequentially, one aminoacid being added at a time, starting from the amino-terminal end. A key feature of the processes leading to polypeptide synthesis is that each aminoacid in the sequence is designated by a trinucleotide or base triplet in three types of molecule: DNA, mRNA and tRNA. The triplet in mRNA is complementary to that in DNA and also to that in tRNA, so that although the actual bases differ the same aminoacid is specified.

Thus the inherited information coded in the genes can be regarded as a kind of blueprint which defines the structures of all the enzymes and other proteins which an individual makes. But genes not only determine the structures of proteins, they are also apparently concerned in regulating their synthesis. The nature of the molecular relationships involved in these regula-