The Principles of Human Biochemical Genetics

Third, revised edition

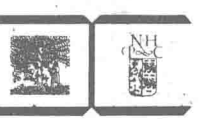
HARRY HARRIS

The Principles of Human Biochemical Genetics

Third, revised edition

HARRY HARRIS

Department of Human Genetics,
University of Pennsylvania,
Richards Building/G4
Philadelphia, PA 19174
U.S.A.



1980

ELSEVIER/NORTH-HOLLAND BIOMEDICAL PRESS
AMSTERDAM · NEW YORK · OXFORD

ISBN Paperback: 0 444 80256 8
ISBN Hardbound: 0 444 80264 9

Published by:
Elsevier/North-Holland Biomedical Press
335 Jan van Galenstraat, 1061 AZ Amsterdam
P.O. Box 211, 1000 AE Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:

Elsevier/North-Holland Inc.
52 Vanderbilt Avenue
New York, NY 10017, U.S.A.

First edition: 1970, second printing 1971
Second revised edition: 1975, second printing: 1977

This book is the third, revised edition of a student reprint edition of the completely revised Volume 19 (Volume 19) in the series Frontiers of Biology under the General Editorship of Professor A. Neuberger, London and Professor E.L. Tatum†, New York.*

PRINTED IN THE NETHERLANDS

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 a 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

Preface to 1st edition

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

Preface to 3rd edition

When I began to prepare this third edition it was already apparent that progress in the subject had become so rapid in the second half of the 1970's that several completely new sections were needed and much updating of the rest of the text was required. The new developments, however, which cover a very diverse range of topics, have nevertheless served to emphasize the unity of the subject as a whole and to reinforce the general line of argument set out in the earlier editions. So the general plan of the book has been conserved.

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.

Department of Human Genetics,
School of Medicine,
University of Pennsylvania
Philadelphia

HARRY HARRIS
June 1980

Acknowledgement

Several illustrations and tables in this volume have been obtained from other publications. The full sources can be found in the reference list at the end of the volume. The permission of the authors and the publishers for the reproduction of this material is gratefully acknowledged.

Fig. 1.1 – From Levitan and Montague, 1977, Oxford University Press.
Tables 1.1 and 6.4 – From Fredrickson et al., 1978, McGraw-Hill Co., Inc.
Fig. 1.11 – From Bunn and McDonough, 1974, American Chemical Society.
Figs. 2.8, 6.9 and 8.12 – From Deisseroth et al., 1978; Yen et al., 1978; and Kan and Dozy, 1978, resp., The National Academy of Sciences.
Fig. 2.14 – From Edwards and Hopkinson, 1977, Alan R. Liss, Inc.
Figs. 3.4, 7.4 and 8.8 – From Winter et al., 1979, Watkins, 1980, and Alper and Rosen, 1976, resp., Plenum Publishing Corp.
Figs. 3.7, 3.10, 8.7 and 8.16 – From Kan et al., 1975, Fritsch et al., 1979, Turner et al., 1975, and Goodman et al., 1975, resp., The Macmillan Press Ltd.
Figs. 3.8 and 5.10 – From Orkin et al., 1978, and Short et al., 1973, resp., Massachusetts Medical Society.
Figs. 4.5, 4.16 and 9.3 – From Muensch et al., 1978, Sensabaugh and Golden, 1978, and Willard et al., 1978, resp., The University of Chicago Press.
Fig. 5.9 – From Kennaway et al., 1975, Williams and Wilkins Co.
Figs. 6.3, 9.4 and Table 6.6 – From Goldstein and Brown, 1977, and Fenton and Rosenberg, 1978, resp., Annual Reviews Inc.
Figs. 6.4 and 8.3 – From Brown and Goldstein, 1976, and Scanlon et al., 1979, resp., American Association for the Advancement of Science.
Fig. 6.5 – From Goldstein and Brown, 1975, Dun-Donnelley Publishing Corp.
Fig. 8.5 – From Donald and Robson, 1978, Cambridge University Press.
Fig. 8.13 – From Jeffreys, 1979, M.I.T. Press.

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 deoxyribonucleic acid or DNA. The second was the elucidation of the molec-

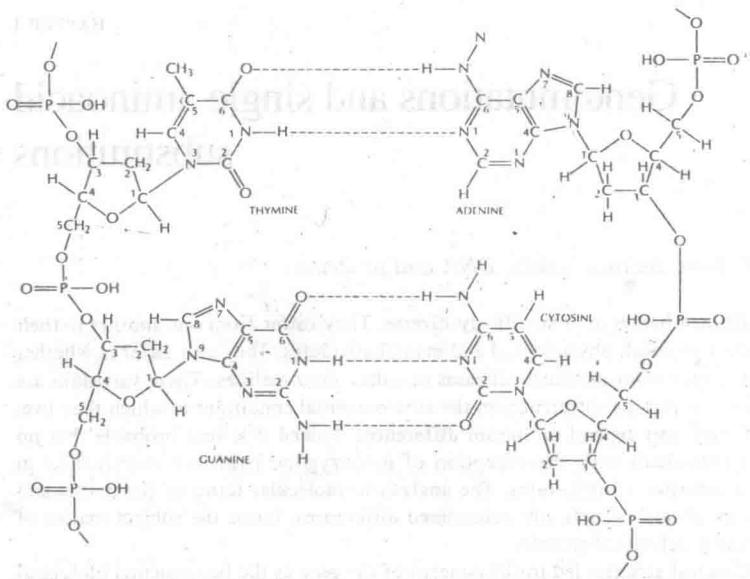


Fig. 1.1. Diagram illustrating the arrangement of two base pairs in DNA and the sugar-phosphate backbones. Note that the complementary chains run in opposite directions (i.e. are anti-parallel) as indicated by the opposite orientation of their sugar residues. The horizontal dashed lines indicate hydrogen bond connections between the complementary chains. (From Levitan and Montagu 1977.)

ular 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 unraveling of the genetic code. That is the details of the relationships between the structure of nucleic acids and the structure of proteins.

DNA: 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 (fig. 1.1). 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.

Proteins: A great variety of different enzymes and other proteins are synthesized 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 characterize 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

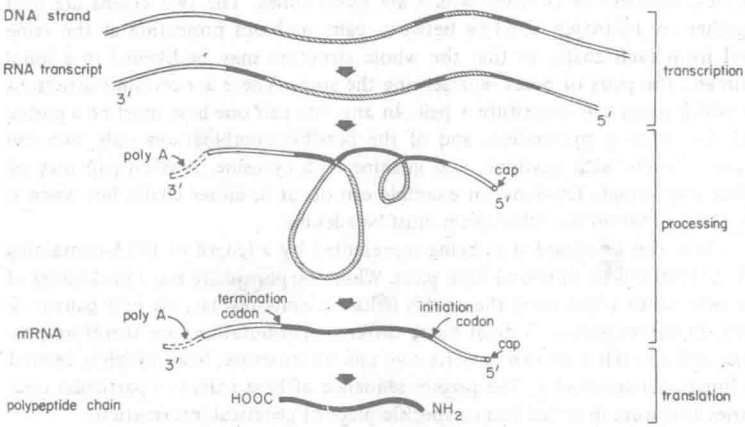


Fig. 1.2. Diagram illustrating transcription of the base sequence of the DNA of a gene into RNA; processing of the RNA transcript to give mRNA; and translation of the base sequence in the coding region of the mRNA into the aminoacid sequence of a polypeptide chain. *Key*: ▬ coding regions in DNA and RNA, and aminoacid sequence in polypeptide chain; ▨ intervening sequences.

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.

DNA and proteins: The fundamental idea relating DNA structure to protein structure is that the sequence of base pairs in a given gene determines the sequence of aminoacids 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.

Transcription: The series of processes by which the sequence of bases in the DNA of a gene is translated into a corresponding sequence of aminoacids in a polypeptide chain are complex and involve as intermediaries certain types of ribonucleic acid (RNA) molecules (fig. 1.2). 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, known as transcription, 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.

Processing: The RNA strand formed by transcription from a corresponding DNA strand is now subject to a series of alterations known as processing. Processing results in the formation of a so-called messenger RNA strand (mRNA) which is the form in which information about the sequence of aminoacids in a polypeptide chain is moved from the nucleus into the cytoplasm.

Processing of the primary RNA transcript involves a number of different changes. These include:

a) Removal of so-called 'intervening' sequences from within the strand and the reunion of the strand at the sites of cleavage. This is often referred to as 'splicing'. Removal of base sequences at the two ends of the primary RNA transcript may also occur.

b) Formation of a 'cap' structure at the 5′ end of the mRNA which is being produced. The cap consists of a methylated guanosine residue linked by a (5′-5′) triphosphate group to the first coded nucleotide of the mRNA. The latter and subsequent bases in the sequence may also be methylated giving rise to structures of the type m^7G (5′) ppp (5′) X^m $pY^m\cdots$, where X and Y are the first and second coded bases in the sequence.

c) The addition of a sequence of adenylate residues to the 3′ end of the mRNA being formed. This sequence is referred to as the poly-A tail.

The mRNA strand entering the cytoplasm from the nucleus carries in its base sequence all the information necessary to define the aminoacid sequence of the particular polypeptide chain and in fact after becoming attached to ribosomes in the cytoplasm, serves as the template for polypeptide synthesis.

The genetic code: The details of the genetic code – that is the relationship of base sequence to aminoacid sequence are given in table 1.1. Each aminoacid is specified by sequence of three bases, which are referred to as base triplets or codons. The four characteristic bases of a DNA chain can occur in 64 different base triplets and 61 of these triplets each specify one of the twenty aminoacids which occur in proteins, so that a particular aminoacid is usually coded by two or more different base triplets. There are also three so-called 'nonsense' triplets which do not specify aminoacids but designate polypeptide chain termination. In effect they provide 'stop' signals in polypeptide synthesis. In the process of transcription a particular base triplet in the DNA strand is transcribed into a complementary base triplet in the RNA strand. Thus the code may be written in two forms as shown in table 1.1, one form being appropriate for the DNA strand, and the other for the complementary RNA strand.

In mRNA the base triplets occur consecutively. That is to say, a triplet specifying one aminoacid is immediately followed by a separate triplet speci-