NEW STUDIES IN BIOLOGY

Genes and Gene Regulation

Norman MacLean

New Studies in Biology

Genes and Gene Regulation

Norman Maclean

Reader in Biology
Department of Biology, University of Southampton

Edward Arnold
A division of Hodder & Stoughton
LONDON NEW YORK MELBOURNE AUCKLAND

First published in Great Britain 1989
Distributed in the USA by Routledge, Chapman and Hall Inc.
29 West 23rd Street, New York, NY1001

British Library Cataloguing Publication Data

Maclean, Norman
Genes and gene regulation
1. Genetics
I. Title II. Series
575.1

ISBN 0-7131-2951-4

Text set in 10/11 pt Plantin
by Colset Private Limited, Singapore
Printed and bound in Great Britain
for Edward Arnold, the educational, academic and medical publishing
division of Hodder and Stoughton Limited, 41 Bedford Square, London
WC1B 3DQ by Whitstable Litho Printers Ltd, Whitstable, Kent

New Studies in Biology

Published in association with the Institute of Biology

The Institute of Biology aims to advance both the science and practice of biology. Besides providing the general editors for this series the Institute publishes two journals *Biologist* and *The Journal of Biological Education*, conducts examinations, arranges national and local meetings and represents the views of professional bodies to government and other bodies.
The emphasis of the *New Studies in Biology* will be on subjects covering major parts of first-year undergraduate courses. We will be publishing new editions of the 'bestsellers' as well as publishing additional mainstream titles. Each cover will be individually designed.

Already published in the New Studies in Biology.

Photosynthesis, *4th edition* D. O. Hall and K. K. Rao
Nitrogen Fixation, *2nd edition* John Postgate
Human Genetics and Medicine, *3rd edition* Cyril A. Clarke
Enzymes in Industry and Medicine Gordon F. Bickerstaff
Subtidal Ecology Elizabeth M. Wood
Control of Crop Diseases W. Carlile
Biotechnology, *2nd edition* John E. Smith
Phytoplankton, *2nd edition* A. D. Boney

In preparation for the New Studies in Biology

Plant Taxonomy, *3rd edition* V. H. Heywood
Pest Control and its Ecology, *2nd edition* M. F. van Emden

General Preface to the Series

Recent advances in biology have made it increasingly difficult for both students and teachers to keep abreast of all the new developments in so wide-ranging a subject. The New Studies in Biology, originating from an initiative of the Institute of Biology, are published to facilitate resolution of this problem. Each text provides a synthesis of a field and gives the reader an authoritative overview of the subject without unnecessary detail.

The Studies series originated 20 years ago but its vigour has been maintained by the regular production of new editions, and the introduction of additional titles as new themes become clearly identified. It is appropriate for the New Studies in their refined format to appear at a time when the public at large has become conscious of the beneficial applications of knowledge from the whole spectrum from molecular to environmental biology. The new series is set to provide as great a boon to the new generation of students as the original series did to their fathers.

1986

Institute of Biology
20 Queensberry Place
London SW7 2DZ

Preface

Genes have come to stay as part of our essential day to day vocabulary. Whether we are breeding sheep, elaborating educational strategies, or planning public health policy, we need to know about genes and their implications for our future.

Also, although less dramatic than setting foot on the moon, the scientific breakthrough of gene manipulation is hugely important both as a human achievement and in terms of its implications for economics and medicine.

This book sets out to explain what genes are, how they are organised as molecules, and how they are regulated in cells and organisms. It is not a primer in gene manipulation; other books in this series deal very competently with that topic. Rather, it considers genes in terms of their structure and function within biological systems, information which should be mastered before the student addresses topics such as, evolution, population genetics, or gene cloning.

Dr Trevor Beebee of the Department of Biochemistry University of Sussex has read and criticised all of the Chapters, and I am most grateful to him for his invaluable contribution.

Norman MacLean
1989

Contents

1

Genes and Genomes

1.1 The Gene Defined

Need for the word *gene* first arose out of Gregor Mendel's classical observations on peas, published in 1866, although the word was not coined until 1909 when Johannsen proposed its usage to denote an inherited factor in the genotype, following the discovery of Mendel's work in 1900 by de Vries. In order to appreciate its original significance and meaning, it is necessary to rid one's mind of subsequent genetical understanding and look again at the startling realisation of the distinction between phenotype and genotype, (see glossary for definition of these terms). By the late nineteenth century it was becoming clear in the minds of many experimentalists that the most remarkable observation concerning inheritance was that an organism could transmit to its progeny a potential for phenotypic expression not actually demonstrable in its own phenotype. Such an observation predated Mendel, since it had been clear for quite some time that parents, both possessing black hair, could produce a child with red hair. But it was not until after Mendel's period that it became evident that a genetic factor could exist unexpressed in both parents and yet be capable of transmission to, and expression in, some of their children.

This distinction between phenotype and genotype, implying that the phenotype represents only a partial expression of the total genotypic potential, still represents a foundation stone in genetics, and had created a need for a word which would designate the factor in the genotype, responsible for transmission and determination of a single phenotypic character. Such a factor came to be called a gene.

As genetics advanced, the gene was defined more precisely, especially in physical terms. Thus, when it was evident that chromosomes were the carriers of the genetic information, it became possible to define the position which an individual gene occupied on a chromosome arm. This position was designated its location, or *locus*, both within an individual chromosome, and, when the individual chromosomes could be distinguished and categorised, within the entire genome. Following this recognition that genes occupied distinct chromosomal loci, there followed the fundamental recognition of DNA as the genetic material. Although the manner of the precise arrangement of DNA in the chromosome remained obscure (and some of that obscurity persists even at the present time) it became evident that there was an approximately linear distribution of genes along a chromosome, and that the DNA formed a continuous thread running, albeit with many imposed orders of coiling and supercoiling, from one end of a chromosome to another. A gene could therefore be redefined with new precision as a length of DNA molecule, itself a sequence of nucleotides in a polynucleotide chain.

Because the gene was being actively investigated in a multi-disciplinary way in the 1940s and 1950s, it became desirable to further define the gene in terms of certain important parameters. Thus, in the 1950s Seymour Benzer proposed three further subdivisions of the basic word, the *cistron*, the *muton* and the *recon*; these were essentially operational definitions of the gene as determined or encountered by different approaches.

The cistron was defined in terms of gene function, following application of the cis/trans test, and represents the physical length of DNA that codes for a protein or RNA. A muton was defined as the smallest unit within DNA in which a change could result in mutation and it is now clear that this is in fact a single nucleotide. The third term, the recon, was defined as a unit of recombination, that is the smallest unit within the DNA capable of being independently involved in recombination, - this is now known to be the individual nucleotide. Although all of these words served useful functions in their time, they have largely fallen into disuse with increasing knowledge of DNA structure and nucleotide arrangement.

The situation now is that the word gene, although used more extensively than ever before, has been substantially overtaken by events and its use no longer carries the absolute clarity that is obviously desirable. How has this come about? Although the word gene continues to be used most frequently to designate a DNA molecule that carries in its nucleotide sequence the code for a protein or RNA, it is not absolutely restricted to this usage as we shall see in the following pages. For example, regulatory genes and pseudogenes may not necessarily subscribe to the above condition, nor is it always clear whether a gene sequence includes introns (see page 9) and even perhaps some flanking sequences which may be essential in transcription. A further area of uncertainty is the precise sequence referred to in using the word. Thus, in a population of diploid organisms, such as rabbits in a hayfield, a gene coding for the blood pigment globin might be referred to as the rabbit globin gene. But rabbits, in common with all other mammals, have multiple and differing types of globin, and so it is necessary to be more specific, referring to the rabbit beta globin gene. Since a population may have some variation due to stable polymorphism or a few mutant forms of the beta globin gene it becomes necessary either to refer broadly to the wild type, that is the most common allelic form of that gene sequence, or to actually designate a particular sequence. A further complication of this topic arises in that an individual may be heterozygous at the beta globin locus, carrying two dissimilar forms of this gene. How can it be made clear which allelic form is being designated?

It is likely that this introductory discussion has already introduced words and concepts which are unfamiliar to some readers, but hopefully all will be made clear in subsequent sections. It will be evident from this section, however, that as knowledge about genes has increased, so the precision of the term has diminished. A list of definitions of current terminology can be found in the Glossary (page 129), but some definitions are given below.

Gene

A sequence of DNA that carries the code for a protein or RNA molecule, and frequently includes regulatory regions at either or both ends. DNA sequences which are closely related by evolution or mutation to genes may retain the designation, even if no longer functional. Genes consist of DNA duplexes, only one strand of which carries the coding information. This is the sense or anti-coding strand, being effective in dictating the coding strand of messenger RNA. The other strand, the non-sense 'strand' of DNA, is not used genetically; but note that the strand acting as sense for one gene may well be in continuity with the non-sense strand of an adjacent gene. Both sense and non-sense strands together are commonly referred to as the gene. Most genes of higher animals and plants (eukaryotes) are interrupted by intron sequences, which are not represented in the messenger RNA (see discussion of introns on page 9 of this chapter). No genes in bacteria (prokaryotes) are interrupted by introns.

Structural Gene

A DNA sequence that codes for protein, thus excluding sequences such as regulatory genes. Some authors also describe genes coding for ribosomal and transfer RNA as structural genes, but more commonly the term refers purely to protein coding sequences.

Regulatory Gene (or Regulator Sequence)

A DNA sequence whose primary function is to control the rate of activity of other genes. The products of regulatory genes may be RNA or protein, or, in some usages, certain regulatory genes may have no products. Thus promoter or enhancer regions may be referred to as regulatory genes – in this book these regions will be termed promoter or enhancer 'sequences', thus avoiding such an ambiguous use of the term gene.

Pseudogene

A DNA sequence that closely resembles the sequence of a known gene at a different locus, but due to the insertion of stop signals or deletions, is unlikely to be translated or transcribed. Pseudogenes which lack introns are referred to as processed pseudogenes. It is assumed that pseudogenes have an evolutionary relationship to normal genes, and some may have arisen by reverse transcription from messenger RNA.

Many other classifications of genes exist in the literature, such as master genes, producer genes and so on, but these usually carry their own descriptions and will not be categorised here. What should be stressed is that unless the use of the word gene is clearly appropriate it is safest to refer to a stretch of DNA simply as a 'DNA sequence' or 'DNA region', and this terminology will be used elsewhere in this book. While the terminology is under discussion, it is useful to refer also to the word allele, itself a contraction for allelomorph. In theory any gene may exist in an almost infinite range of variants due to small base substitutions, but in reality, within a species or a population, the actual range is much smaller. Such variations of a gene are termed alleles of that gene, and if sufficiently distinctive, may earn the terminology of gene themselves. For example, the sickle-cell allele of the human beta globin gene is often referred to as the sickle cell gene.

1.2 Gene Structure

Although the double helical structure of DNA is very well known, and fully described in all basic texts of Biology and Biochemistry, a surprising number of possible pitfalls surround the topic, and it is as well to summarise the various physical characteristics and biomolecular properties of this remarkable molecule. These are presented below in an extended tabular form.

1.2.1 Summary of DNA structure

1. *It is a polynucleotide* of four different single nucleotides, each consisting of a purine or pyrmidine base, the pentose sugar deoxyribose, and a phosphoric acid group. The bases are adenine (A) and guanine (G), which are purines, and thymine (T) and cytosine (C) which are pyrimidines.
2. *It may exist as a single or double stranded molecule*, the latter being a double helix of two parallel single strands. When in the double helical form, the bases are connected by hydrogen bonds in specific and unique base pairing, two bonds linking the AT base pair and three bonds linking the GC base pair. Since the bases constitute the genetic code, it follows that the genetic information is actually on the *inside* of the double helix.
3. *The double helical form of DNA has strands in anti-parallel array.* This follows from the fact that each DNA molecule is polarised, one end having a phosphoryl radical on the 5′ carbon of its terminal nucleotide, the other possessing a free-OH on the 3′ carbon of the terminal nucleotide. In double helical array, one strand runs from 5′ to 3′ left to right, the complementary strand running 3′ to 5′ left to right. (DNA sequences are normally written and read running from 3′ on the left to 5′ on the right, and these may be designated as upstream and downstream respectively).
4. *The double helix comes in three alternative conformations.* These are known as the A, B and Z forms, of which B is the form most commonly found in nature. A fourth form C, is a variant of the B form. Both A and B forms are right-handed helices, the former with 11 bases per turn of the helix, the latter with 10 bases per turn. *Z* form DNA is a left-handed helix, with 12

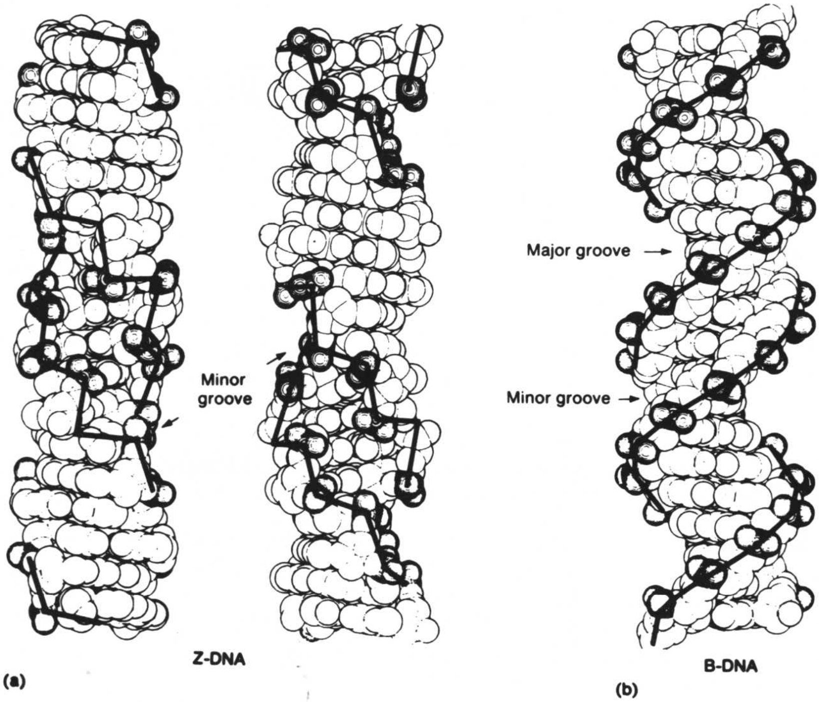


Fig. 1.1 Diagrams of space-filling models of DNA. In (a) the left handed Z DNA and in (b) the right handed B DNA. Heavy lines mark the course of the sugar-phosphate backbone, which can be seen to follow a somewhat zig-zag path in the Z form and a smooth path in the B form. (In Wang, A.H., *et al.* (1979) *Nature* **282**, 680. Kindly provided by Professor Alex Rich).

bases per turn, and adopts a zig-zag conformation. A form of DNA occurs under less hydrated conditions than the B form and has the bases tilted 20° away from the perpendicular axis and slightly laterally displaced (see Fig. 1.1).

5. *The helix of double stranded DNA may be further coiled to form positive or negative supercoils.* When DNA is not supercoiled it is said to be relaxed. Positive supercoiling implies futher coiling in the direction of rotation of the existing helix, and negative supercoiling indicates coiling in the opposite direction which tends to result in an untwisting of the molecule. Enzymes known as topoisomerases are necessary for the formation or removal of supercoiling. Class II topoisomerses (also termed gyrases), together with energy from ATP hydrolysis, can convert relaxed DNA to supercoiled form. They do so, not by physically turning the molecular screw, as it were, but by nicking and resealing overlying strands so that torsion is introduced (see Fig. 1.2). Topoisomerase class I enzymes resolve

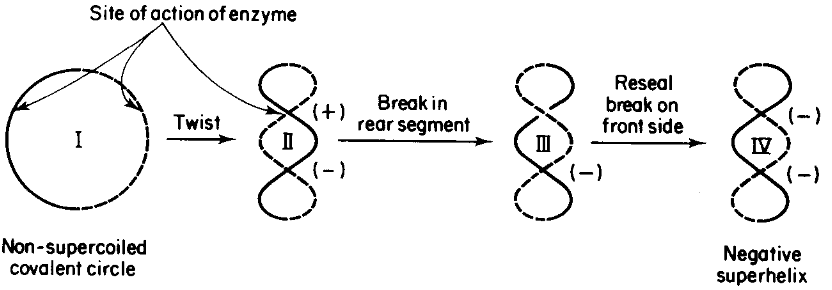


Fig. 1.2 Action of the topoisomerase II molecule of *E. coli* (also termed DNA gyrase) in introducing negative superhelix formation in a covalent circle of DNA. (From D. Freifelder, *Molecular Biology*, Van Nostrand Reinhold with permission).

supercoiling by nicking and resealing, thus establishing a return to relaxed DNA, but no energy input is required for this reaction.

6. *Some bases in DNA are modified by methylation after initial DNA synthesis.* Methylation is the only major modification undergone by DNA, and it occurs in most, but not all, organisms. In bacteria it involves chiefly the modification of adenine to 6-methyladenine, and in eukaryotes modification of cytosine to 6-methyl cytosine. It is enzyme mediated, but irreversible for a particular DNA molecule.

7. *In the DNA duplex only one strand functions as a template for messenger RNA.* The template strand is termed the sense or anticoding strand (since the mRNA message functions as the code for the protein sequence), while the sister strand with which it is base paired is the non-sense strand. However, it does not follow that the same strand of the DNA duplex is the sense strand along the length of an entire chromosomė. On the contrary, it is quite clear that both strands of DNA in an entire chromosome function as sense strands in certain loci. This implies that some are read in one direction and some in another, since the two strands lie in antiparallel array and a gene can only be read in its 3′ to 5′ orientation, giving message synthesised from 5′ to 3′. So if we imagine that a chromosome's worth of DNA is spread out before us a duplex running from left to right, some genes will be read from one strand running from left to right, others from the opposite strand running from right to left. But in eukaryotes no examples exist of genes actually overlapping so that opposite strands might serve as portions of the sense strand for two quite separate messages. It is the positioning of the promoter sequences that initially determines which parts of which strands are 'read' by the polymerase.

8. *DNA molecules can be readily dissociated into single strands and then recombined by specific base pairing under appropriate conditions.* The separation of a DNA duplex into single strands is known as denaturation or melting, and its reconstitution from single stands to duplex form as renaturation or

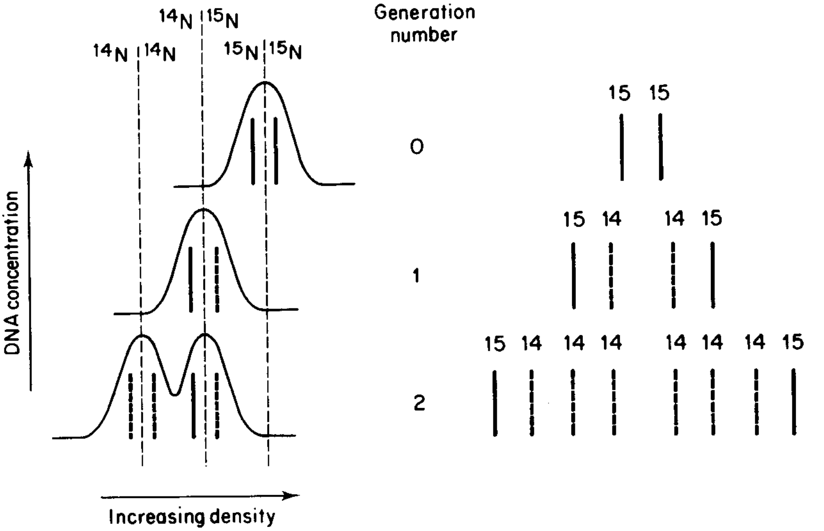


Fig. 1.3 The Meselson and Stahl experiment in which *E. coli*, grown for many generations in ^{15}N medium (black), were transferred at zero time to ^{14}N medium (hatched). After one generation, all DNA had adopted an intermediate density indicative of a hybrid molecule. In the second generation, two fractions were detected, one of ^{14}N duplex, and one of hybrid duplexes. Such an experiment confirms the semi-conservative mode of DNA replication. (From D. Freifelder, *Molecular Biology*, Van Nostrand Reinhold with permission).

annealling. If two strands from distinct sources are annealled, it is termed DNA hybridisation. The ability of the double helix to open out at least partially into a single stranded conformation is probably necessary for transcription, just as it is for replication (DNA synthesis). As we shall see later, this property of annealling makes it easy to compare sequence homologies in different strands in test tube experiments since most of the methodology of gene manipulation relies on DNA hybridisation.

9. *DNA replication is semi-conservative*; i.e. both single strands of the parent molecule serve as a template for the synthesis of new strands, and each new duplex molecule consists of one original and one newly synthesised strand. This was clearly demonstrated in the classic experiment of Meselson and Stahl (1958) outlined in Fig. 1.3. As in transcription, so in replication, the template strand is read from 3′ to 5′ and the newly synthesised strands therefore run from 5′ to 3′. As seen in Fig. 1.4, the opening up of the duplex during replication is gradual, which renders it necessary for one strand to be copied discontinuously (the so-called lagging strand), while the other can be read continuously (the leading strand). DNA replication is further complicated by the fact that a short RNA primer is synthesised first during ech synthetic initiation step, to be cut out by enzymatic means later

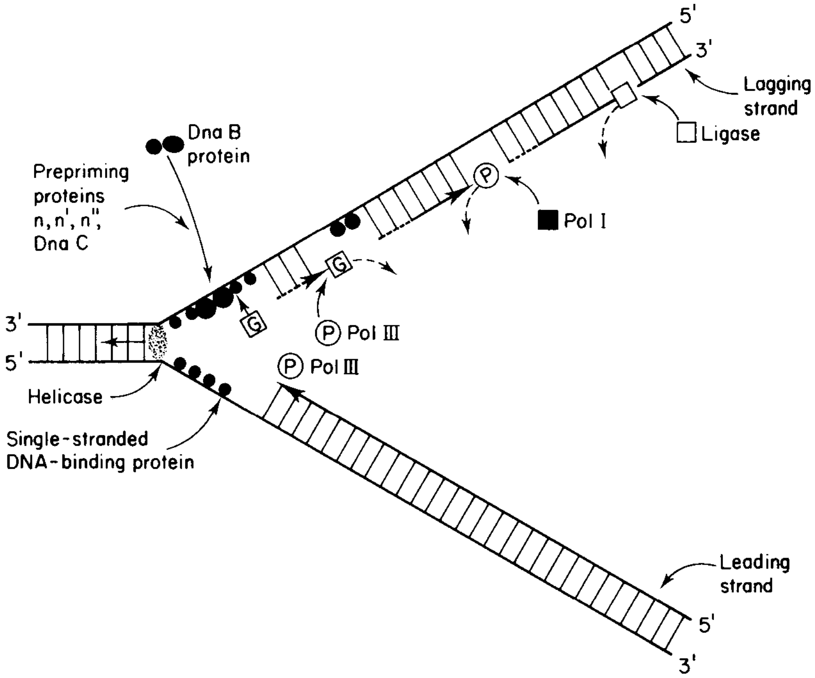
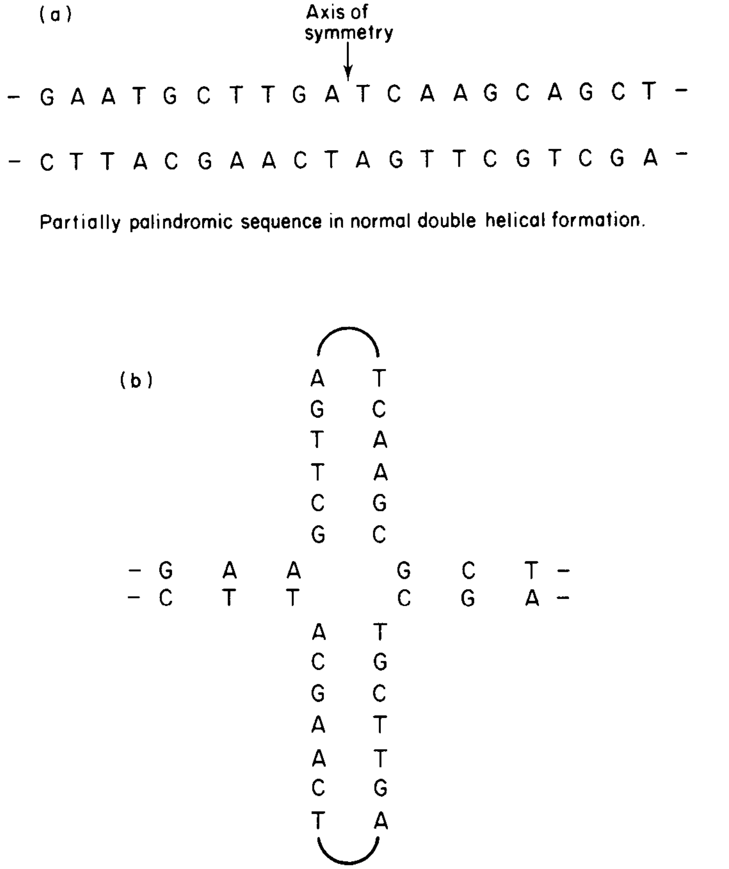


Fig. 1.4 Diagram illustrating the probable events at a bacterial DNA replication fork. Each section of DNA synthesised is headed by a short sequence of RNA primer which is removed prior to ligation of the fragments. Synthesis can only proceed in a 5′ to 3′ direction, reading the template strand from 3′ to 5′. This permits continuous replication to proceed with one strand (the leading strand) but requires discontinuous replication of the other strand (the lagging strand). (From D. Freifelder, *Molecular Biology*, Van Nostrand Reinhold with permission).

and replaced with DNA. Also the discontinuously synthesised strand produces short sequences called Okazaki fragments, and these have.to be enzymatically ligated together to form a complete strand.

10. *Although most DNA exists as a linear duplex, short palindromic stretches may adopt short alternative three-dimensional structures.* As seen in Fig. 1.5, a palindromic sequence, or inverted repeat, has a central axis of symmetry, and permits base paired loops to project out from the main duplex. Although palindromic sequences are relatively abundant in DNA, it is not clear how frequently these alternative structures do actually form. They might well provide easy recognition sites for other molecules, such as regulatory proteins, which interact specifically with DNA.

11. *Many genes contain introns.* One of the most astonishing discoveries of recent years, in the field of molecular genetics, is that most gene sequences



Partially palindromic sequence in normal double helical formation.

Partially palindromic sequence showing adoption of alternative conformation with specific three dimensional configuration.

Fig. 1.5 Adoption of a novel tertiary structure by a DNA sequence which includes a short palindrome.

of eukaryotic organisms (fungi, plants and animals) do not consist of a single continuous coding sequence which is simply transcribed into messenger RNA. Instead, it has become clear that most eukaryotic genes contain between one and fifty *introns*; these are lengths of sequence not represented in the message. The remaining parts of the gene which are represented in the message are known as *exons*. It is thus evident that cells must have a way of excluding the intron coding information from the message. This is done by initially transcribing the whole sequence, introns plus exons, into a large precursor RNA molecule (the heterogeneous RNA