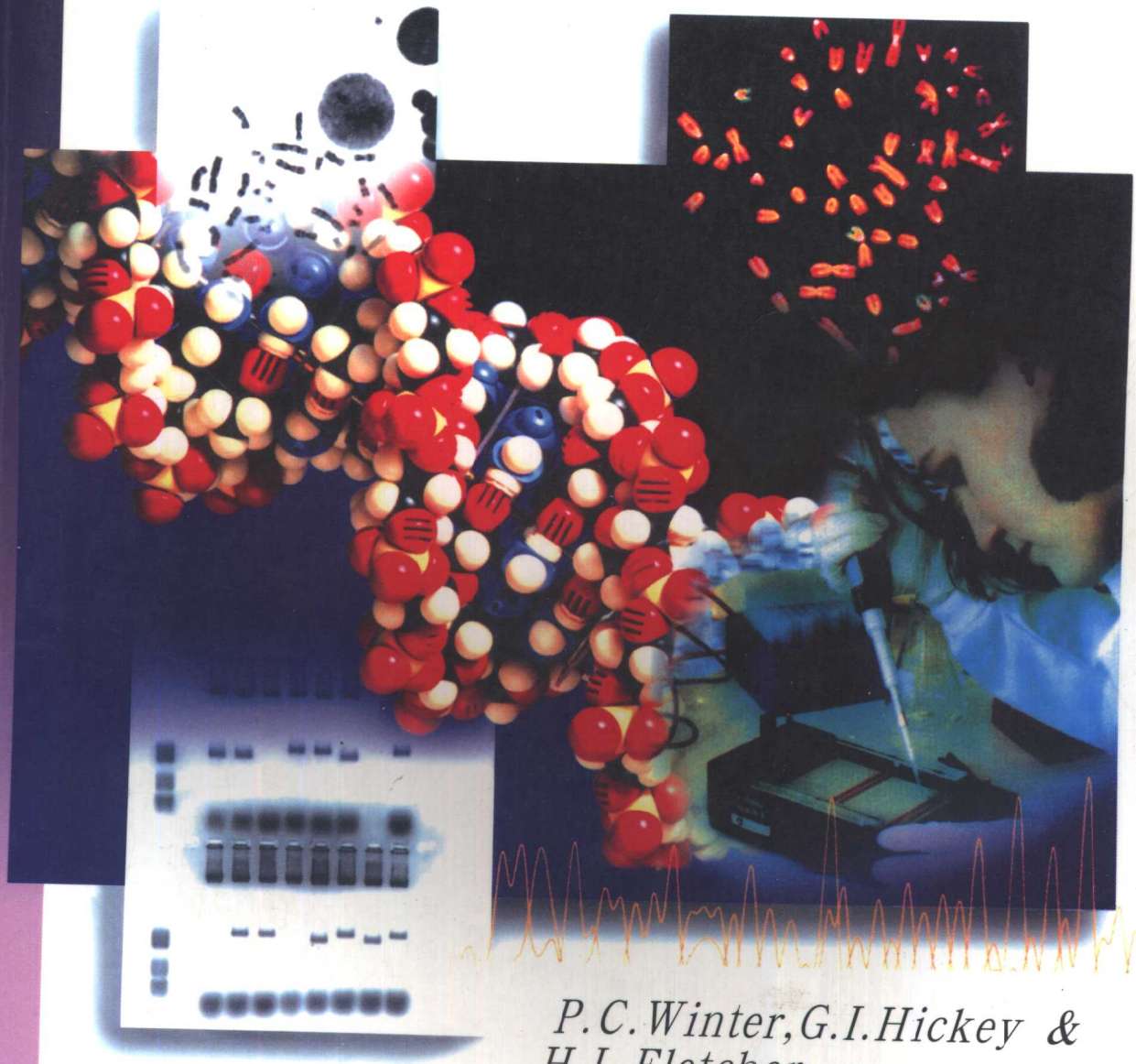
现代生物学精要速览

Instant Notes in GENETICS

遗传学

（影印版）



P.C.Winter, G.I.Hickey & H.L.Fletcher

科学出版社
BIOS SCIENTIFIC PUBLISHERS LIMITED

现代生物学精要速览

Insta■ Notes in

GENETICS

遗 传 学

(影印版)

P.C. Winter
Department of Haematology,
Royal Victoria Hospital, Belfast,
UK

and

G.I. Hickey & H.L. Fletcher
School of Biology & Biochemistry,
The Queen's University of Belfast,
Belfast, UK

科 学 出 版 社

BIOS SCIENTIFIC PUBLISHERS LIMITED

1999

内 容 简 介

本套丛书是国外优秀教材畅销榜的上榜教材，面向大学生，由英国著名大学具丰富教学经验的一流教授编写。它以一种风格独特的描述方式，全面、系统地概括了学科的核心内容和前沿动态，并以一种便于学习、利于复习的形式，使学生能快速、准确地掌握知识，很好地指导学习和考试。书中英文使用最为自然、易懂的语句，是提高专业外语的最佳套书。本书是该系列中的遗传学分册，共约 19 个章节。

P. C. Winter, G. I. Hickey and H. L. Fletcher

Instant Notes in Genetics

Original edition published in the United Kingdom under the title of Instant Notes in Genetics

图字：01-99-0194

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

中国科学院印刷厂 印刷

新华书店北京发行所发行 各地新华书店经售

*

1999 年 3 月第 一 版 开本：787×1092 1/16

1999 年 3 月第一次印刷 印张：22

印数：1—3 000 字数：507 000

ISBN 7-03-007307-X/Q・849

定价：45.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

ABBREVIATIONS

3D	three-dimensional
5BU	5-bromouracil
ADA	adenine deaminase
AIDS	acquired immune deficiency syndrome
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
bp	base pairs
BrdU	bromdeoxyuridine
bZIP	basic leucine zipper
CAP	catabolite activator protein
cAMP	cyclic adenosine monophosphate
cdk	cyclin dependent kinase
CFTR	cystic fibrosis transmembrane conductance regulator
cM	centiMorgan
DIG	digoxigenin
DNA	deoxyribonucleic acid
dATP	2′-deoxyadenosine 5′-triphosphate
dCTP	2′-deoxycytosine 5′-triphosphate
ddNTP	dideoxynucleotide triphosphate
dGTP	2′-deoxyguanosine 5′-triphosphate
dNTP	deoxynucleotide triphosphate
dTTP	2′-deoxythymidine 5′-triphosphate
ds	double-stranded
ES	embryonic stem
EST	expressed sequence tags
ETS	external transcribed spacers
F	fertility
FISH	fluorescent *in situ* hybridization
GDP	guanosine diphosphate
GTP	guanosine triphosphate
HFr	high frequency recombination
HIV	human immunodeficiency virus
HLH	helix-loop-helix
hnRNA	heterogeneous nuclear RNA
HUGO	human genome organization
ICR	internal control region
ITS	internal transcribed spacers
kb	kilo base
kbp	kilo base pairs
LINE	long interpersed nuclear elements
LTR	long terminal repeat
MCS	multiple cloning site
mDNA	mitrochondrial DNA
mRNA	messenger RNA
NOR	nuclear organizer region.
ORF	open reading frame
PAC	P1 artificial chromosome
PCR	polymerase chain reaction
PKU	phenylketonuria
pms	postmeiotic segregation
QTL	quantitative trait loci
R	resistance
RF	replicative form
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
ROS	reactive oxygen species
rRNA	ribosomal RNA
SAR	scaffold attachment regions
SCE	sister chromatid exchanges
SINE	short interspersed nuclear elements
snRNP	small nuclear ribonucleoproteins
ss	single-stranded
SSB	single-strand binding
STS	sequence tagged sites
TF	transcription factor
tRNA	transfer RNA
TIC	transcription initiation complex
VNTR	variable number tandem repeats
YAC	yeast artificial chromosome

PREFACE

This text aims to provide a comprehensive set of basic notes in genetics. Genetics is like riding a bike, easy when you know how, but impossible until you try it. Genetics is considered by some students to be the most difficult aspect of biology. This is often because you have to think about it. Insects have six legs. A simple observable fact of no further consequence, unless they start wearing shoes. The rules of segregation of genetic material, first discovered by Mendel, apply to most, perhaps all, of the 50 000 plus human genes. The rules of genetics are similar for all of the several million species on Earth and have consequences at all levels of life. Rules are conceptual but the facts of chromosome structure and genetic transmission which cause the rules are almost invisible. One understood, however, the behavior of all the genes is accessible. Thus, genetics is not suitable for rote learning, but like a new language, understanding the basics of genetics opens a book which gives an insight and coherence to all life.

Understanding requires a mental image. DNA is a long thin thing with a series of coded instructions on it. These are the plans for the molecular machines which make and run cells, and in turn read the DNA code and complete the cycle of interdependence. Groups of cells make organs and organisms, and the appearance (phenotype) of these reveals the instructions on their DNA (their genotype) to the outside world. These levels encompass all genetics from the structure of DNA and its mutation, through the workings and interactions of the proteins and RNAs it encodes, to evolutionary changes in the mixture of individual genotypes in populations.

The science of genetics has exploded since the development of techniques to manipulate and sequence DNA and RNA have enabled researchers to look at, and deliberately create, individual base changes in DNA. It is now possible to manufacture a DNA sequence to code for any interesting sequence of amino acids and so produce designer proteins. The gene can be given a selective promoter, which can be switched on or off as the researcher requires. Mutations which cause disease can be identified and the normal function of the gene investigated, with the intention of designing a drug to replace it. It may even be possible to replace the gene itself with the correct DNA sequence. An inhaler spray carrying the sequence of the gene whose absence causes cystic fibrosis is already in use. These exciting developments have led to genetics textbooks doubling in size. These things are interesting, but can make it difficult for the beginner to find the rules of the game amongst the advanced applications. Full appreciation of how genetic profiles can incriminate rapists needs an understanding of both the molecular genetic techniques used to produce the profile and the population genetics which tells how unlikely it is that the suspect is innocent.

We have attempted to cover the basics, without exhaustive detail or repetitive examples, and have used our experience of student's mistakes to draw attention to some common misconceptions and sources of confusion. The Keynotes at the start of each topic follow the series' successful format by presenting the absolute basics, but students should know why these brief statements are true, not just learn them. The breadth of the text was determined by attempting to include all the essential requirements for a foundation genetics

course, accommodating the fact that different courses will have different requirements and bias, some molecular, some medical, some biological, some ecological. Omitted topics were generally considered to be appropriate to a more advanced level. Depth was determined by the accessibility of the exciting advances. If they were not readily understandable, they were left out. We hope that the combination will provide broad support to most genetics lecturers and accesible information for genetics students.

The sequence of presentation is also arbitrary. Many texts run historically, starting with Mendel. We choose to run logically, starting with DNA and the genetic code, Understanding these gives the explanation for Mendels laws. The first section covers the chemistry of DNA, the genetic code and protein synthesis. The second covers the organisation of DNA into chromosomes at the cellular level. Thirdly the transmission of DNA between generations is discussed, with an exploration of the interaction between alleles, genes and their products, and how these determine the phenotype of the organism. The emphasis is on diploids, so is especially relevant to humans. Section D, population genetics and evolution, takes genetics to the next level, from examining the causes for the high level of cystic fibrosis in Europeans to the changes in allele frequencies which are evolution, and eventually cause speciation. Section E describes some of the techniques used in the molecular genetics revolution, and section F introduces some of the recent applications of genetics, and implications for the future. We hope this will encourage students to appreciate that genetics has an important position in society, and that the effort of understanding it is well worth while.

P.C. Winter, G.I. Hickey and H.L. Fletcher

CONTENTS

A1 DNA STRUCTURE

Key Notes

Nucleotides

DNA is a polymer containing chains of nucleotide monomers. Each nucleotide contains a sugar, a base and a phosphate group. The sugar is 2′-deoxyribose which has five carbons named 1′ (prime) 2′ etc. There are four types of base: adenine and guanine have two carbon–nitrogen rings and are purines; thymine and cytosine have a single ring and are pyrimidines. The bases are attached to the 1′ carbon of the deoxyribose. A sugar plus a base is termed a nucleoside. A nucleotide has one, two or three phosphate groups attached to the 5′ carbon of the sugar. Nucleotides occur as individual molecules or polymerized as DNA or RNA.

DNA polynucleotides

Nucleotide triphosphates of the four bases are joined to form DNA polynucleotide chains. Two phosphates are lost during polymerization and the nucleotides are joined by the remaining phosphate. A phosphodiester bond forms between the 5′ phosphate of one nucleotide and the 3′ hydroxyl of the next nucleotide. The polynucleotide has a free 5′ phosphate at one end (5′ end) and a free 3′ OH (3′ end) at the other end. The sequence of bases encodes the genetic information. It can be read 5′→3′ or 3′→5′. Polynucleotides are extremely long. It is possible to have 4^n different sequences.

The double helix

DNA molecules are composed of two polynucleotide strands wrapped around each other to form a double helix. The sugar–phosphate part of the molecule forms a backbone. The bases face inwards and are stacked on top of each other. The two polynucleotide chains run in opposite directions. The double helix is right-handed and executes a turn every 10 bases. The helix has a major groove which mediates interactions with proteins. Variant DNA structures have been identified including Z DNA which has a left-handed helix.

Complementary base pairing

Hydrogen bonds between bases on the two DNA strands stabilize the double helix. The available space between the strands restricts the bases that can interact such that a purine always interacts with a pyrimidine. Thus, A interacts only with T and G only with C. This is called complementary base pairing. The restriction on base pairing means that the sequence of bases on the two strands are related to each other, such that the sequence of one determines and predicts the sequence of the other. This allows genetic information to be preserved during replication of the DNA and expression of the genes. Disruption of the hydrogen bonds between the bases by heat or chemicals or by the action of enzymes causes the strands of the double helix to separate.

RNA structure

In RNA thymine is replaced by uracil and 2-deoxyribose by ribose. RNA normally exists as a single polynucleotide strand however, short stretches of base pairing may occur between complementary sequences.

Related topics

Gene transcription (A4)
DNA replication (A9)
DNA mutation (B5)

Nucleotides

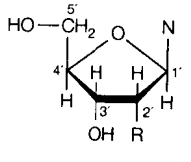
The ability of DNA to carry the genetic information required by a cell to reproduce itself is closely related to the structure of DNA molecules. DNA is a polymer and consists of a long chain of monomers called **nucleotides**. The DNA molecule is said to be a polynucleotide. Each nucleotide has three parts: a sugar, a nitrogen containing ring-structure called a **base**, and a phosphate group. The sugar present in DNA is a five carbon pentose called 2′-deoxyribose in which the –OH group on carbon 2 of ribose is replaced by hydrogen (*Fig. 1*). The carbon atoms in the sugar are numbered 1–5. The numbers are given a dash (′) referred to as **prime** to distinguish them from the numbers of the atoms in the base. The numbering is important because it indicates where other components of the nucleotide are attached to the sugar.

Nucleotides contain one of four bases: **adenine, guanine, thymine** or **cytosine** (*Fig. 2*). These are complex molecules containing carbon and nitrogen ring structures. Adenine and guanine contain two carbon–nitrogen rings and are known as **purines**. Cytosine and thymine contain a single ring and are called **pyrimidines**. The bases are attached to the sugar by a bond between the 1′ carbon of the sugar and a nitrogen at position 9 of the purines or position 1 of the pyrimidines. A sugar plus a base is called a **nucleoside** (*Fig. 3a*).

Nucleotides contain phosphate groups (PO_4) attached to the 5′ carbon of the sugar (*Fig. 3b*). A nucleoside is called a nucleotide when a phosphate group is attached, the attachment can consist of one, two or three phosphate groups joined together. The phosphate groups are called α, β and γ, with α directly attached to the sugar. Nucleotides may exist in cells as individual molecules (nucleotide triphosphates play an important role in cells as the carriers of energy used to power enzymatic reactions) or polymerized as nucleic acids (DNA or RNA).

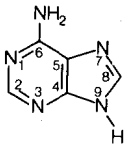
DNA polynucleotides

Nucleotide triphosphates are joined together to give polynucleotides. There are four used to synthesize DNA polynucleotides, 2′-deoxyadenosine 5′-triphosphate (dATP or A), 2′-deoxythymidine 5′-triphosphate (dTTP or T), 2′-deoxycytosine 5′-triphosphate (dCTP or C) and 2-deoxyguanosine 5′-triphosphate (dGTP or G). The β and γ phosphates are lost during polymerization and the nucleotide units are joined together by the remaining phosphate. The 5′ phosphate of one nucleotide forms a bond with the 3′ carbon of the next nucleotide eliminating the –OH group on the 3′ carbon during the reaction. The bond is called a **3′–5′ phosphodiester bond** (C–O–P) (*Fig. 4*). The polynucleotide chain has a free

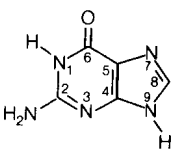


2′-Deoxyribose

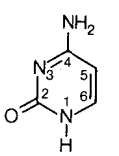
Fig. 1. Structure of 2′-deoxyribose.



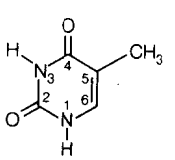
Adenine (A)



Guanine (G)



Cytosine (C)



Thymine (T)

Fig. 2. Bases in DNA.

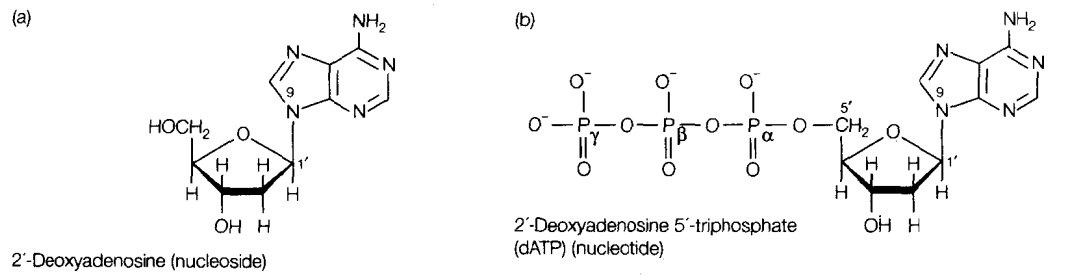


Fig. 3. Structure of (a) nucleosides, (b) nucleotides.

5′ triphosphate at one end known as the 5′ end and a free 3′ hydroxyl group at the other end called the 3′ end. This distinction gives the DNA polynucleotide polarity so that a DNA molecule can be described as running 5′→3′ or 3′→5′.

It is the sequence of the bases in the DNA polynucleotide that encodes the genetic information. This sequence is always written in the 5′→3′ direction (polymerase enzymes copy DNA molecules in this direction). Polynucleotides can be extremely long with no apparent limit to the number of nucleotides and no restrictions on the sequence of the nucleotides. The maximum number of possible base sequences for a polynucleotide is 4^n , where n is the number of nucleotides. This is an enormous number. For example, a polynucleotide containing just six bases could be arranged as $4^6 = 4096$ different sequences.

The double helix

DNA molecules have a very distinct and characteristic three-dimensional structure known as the double helix (*Fig. 5*). The structure of DNA was discovered

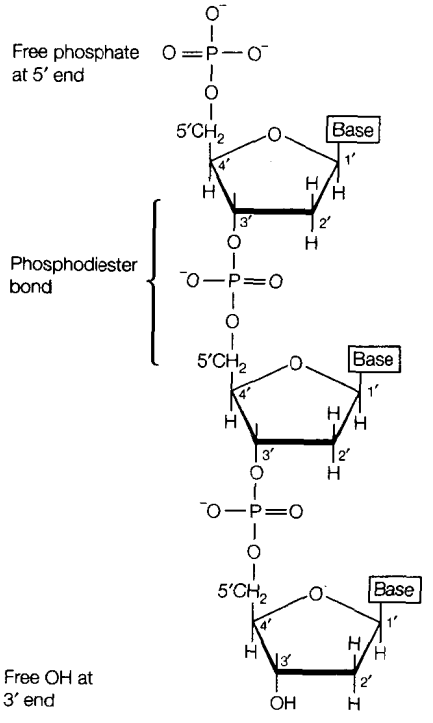


Fig. 4. Phosphodiester bonds join nucleotides in a DNA polynucleotide.

in 1953 by Watson and Crick working in Cambridge using X-ray diffraction pictures taken by Franklin and Wilkins. DNA exists as two polynucleotide chains wrapped around each other to form the double helix. The sugar–phosphate part of the molecule forms a spine or backbone which is on the outside of the helix. The bases, which are flat molecules, face inwards towards the center of the helix and are stacked on top of each other like a pile of plates.

X-ray diffraction pictures of the double helix show repeated patterns of bands that reflect the regularity of the structure of the DNA. The double helix executes a turn every 10 base pairs. The pitch of the helix is 34Å so the spacing between bases is 3.4Å. The diameter of the helix is 20Å. The double helix is said to be 3 **antiparallel**. One of the strands runs in the 5′→3′ direction and the other 3′→5′ direction. Only antiparallel polynucleotides form a stable helix. The double helix is not absolutely regular and when viewed from the outside a **major groove** and a **minor groove** can be seen. These are important for interaction with proteins, for replication of the DNA and for expression of the genetic information. The double helix is right-handed. This means that if the double helix were a spiral staircase and you were climbing up, the sugar–phosphate backbone would be on your right.

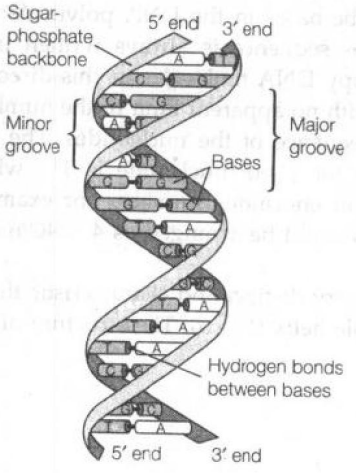


Fig. 5. The double helix.

A number of variant forms of DNA occur when crystals of the molecule are formed under different conditions. The form present in cells is called the **B form**. Another form called the A form has a slightly more compact structure. Other forms that exist are C, D, E and Z, which is striking because it exists as a left-handed helix. Regions in chromosomes containing nonstandard structures such as Z-DNA have recently been identified.

Complementary base pairing

The bases of the two polynucleotide chains interact with each other. The space between the polynucleotides is such that a two-ring purine interacts with a single-ring pyrimidine. Thus, thymine always interacts with adenine and guanine with cytosine. Hydrogen bonds form between the bases and help to stabilize the interaction. Two bonds form between A and T and three between G and C. Thus, G–C bonds are stronger than A–T bonds. The way in which the bases form pairs between the two DNA strands is known as **complemen-**

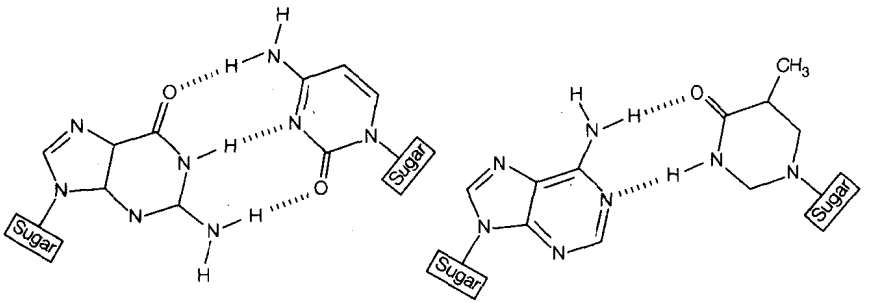


Fig. 6. Complementary base pairing. Hydrogen bonds are shown as dashed lines.

tary base pairing and is of fundamental importance (*Fig. 6*). Combinations other than G–C and A–T do not work because they are too large or too small to fit inside the helix or they do not align correctly to allow hydrogen bond formation. Because G must always bond to C and A to T the sequences of the two strands are related to each other and are said to be complementary with the sequence of one strand predicting and determining the sequence of the other. This means that one strand can be used to replicate the other. This is a vital mechanism for retaining genetic information and passing it on to other cells following cell division. Complementary base pairing is also essential for the expression of genetic information and is central to the way DNA sequences are transcribed into mRNA and translated into protein.

The double helix is stabilized by hydrogen bonds between the base pairs. These can be disrupted by heat and some chemicals. This results in separation of the double helix into two strands and the molecule is said to be denatured. In cells enzymes can separate the strands of the double helix for the purposes of copying the DNA and for expression of the genetic information.

RNA structure

The structure of RNA is similar to that of DNA but a number of important differences exist. In RNA ribose replaces 2′-deoxyribose and the base thymine is replaced by another base, uracil, which can also base pair with adenine (*Fig. 7*). In addition, RNA molecules normally exist as a single polynucleotide strand and do not form a double helix. However, it is possible for base pairing to occur between complementary parts of the same RNA strand resulting in short double-stranded regions.

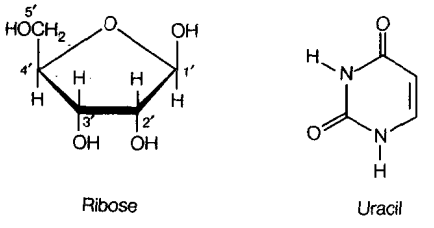


Fig. 7. Structures of ribose and uracil.

A2 GENES

Key Notes

Structure of genes

A gene is a unit of information and corresponds to a discrete segment of DNA that encodes the amino acid sequence of a polypeptide. Human cells contain 50–100 000 genes arranged on 23 chromosomes. The genes are dispersed and are separated by noncoding intergenic DNA. Information is encoded on the template strand which directs the synthesis of an RNA molecule. Both DNA strands can act as the template strand. DNA molecules have an enormous capacity to store genetic information.

Gene families

Some genes are arranged as clusters known as operons and multigene families. Operons occur in bacteria and contain coregulated genes with a related function. Multigene families occur in higher organisms and contain genes that are identical or similar that are not regulated coordinately. Simple multigene families contain identical genes whose product is required in large amounts. Complex multigene families contain genes that are very similar and encode proteins with a related function.

Gene expression

The biological information encoded in genes is made available by gene expression. In this process, an RNA copy of a gene is synthesized which then directs the synthesis of a protein. The central dogma states that information is always transferred from DNA to RNA to protein. The functioning of cells is dependent on the coordinated activity of many proteins. Gene expression ensures that proteins are synthesized in the correct place at the correct time.

Gene promoters

Gene expression is highly regulated. Not all of the genes present in a cell are active and different types of cell express different genes. The expression of a gene is regulated by a segment of DNA upstream of the coding sequence called the promoter, this binds RNA polymerase and associated transcription factor proteins and initiates synthesis of an RNA molecule.

Introns and exons

The coding sequence of a gene is split into a series of segments called exons which are separated by noncoding sequences called introns which usually account for most of the gene sequence. The number and sizes of the introns vary between genes. Introns are removed from RNA transcripts by a process called splicing prior to protein synthesis. Introns are not usually present in bacteria.

Pseudogenes

Copies of some genes exist which contain sequence errors acquired during evolution that prevent them from producing proteins. These are called pseudogenes and they represent evolutionary relics of original genes. Examples include the globin pseudogenes.

Related topics

Regulation of gene expression in prokaryotes (A10)
Regulation of gene expression in eukaryotes (A11)
The human genome (B4)

Structure of genes

The biological information needed by an organism to reproduce itself is contained in its DNA. The information is encoded in the base sequence of the DNA and is organized as a large number of genes, each of which contains the instructions for the synthesis of a polypeptide. In physical terms, a gene is a discrete segment of DNA with a base sequence that encodes the amino acid sequence of a polypeptide. Genes vary greatly in size from less than 100 base pairs to several million base pairs. In higher organisms the genes are present on a series of extremely long DNA molecules called **chromosomes**. In humans there are an estimated 50–100 000 genes arranged on 23 chromosomes. The genes are very dispersed and are separated from each other by sequences that do not appear to contain useful information; this is called **intergenic DNA**. The intergenic DNA is very long, such that in humans gene sequences account for less than about 30% of the total DNA. Only one of the two strands of the DNA double helix carries the biological information: this is called the **template strand** and it is used to produce an RNA molecule of complementary sequence which directs the synthesis of a polypeptide. The other strand is called the **nontemplate strand**. Both strands of the double helix have the potential to act as the template strand: individual genes may be encoded on different strands. Other terms are used to describe the strands of the double helix as alternatives to template and nontemplate. These include **sense/antisense** and **coding/noncoding**: the terms antisense and noncoding are equivalent to the template strand.

The capacity of DNA molecules to store information is enormous. For a DNA molecule n bases long, the number of different combinations of the four bases is 4^n. Even for very short DNA molecules the number of different sequences possible is very large. In practice, there are limitations to the sequences that can contain useful information. However the capacity to encode information remains vast.

Gene families

Most genes are spread out randomly along the chromosomes, however some are organized into groups or clusters. Two types of cluster occur: these are **operons** and **multigene families**.

Operons are gene clusters found in bacteria. They contain genes that are regulated in a coordinated way and encode proteins with closely related functions. An example is the *lac* operon in *E. coli* which contains three genes encoding enzymes required by the bacterium to break down lactose. When lactose is available as an energy source, the enzymes encoded by the *lac* operon are required together. The clustering of the genes within the operon allows them to be switched on or off at the same time allowing the organism to use its resources efficiently (*Fig. 1*).

In higher organisms, operons are absent and clustered genes exist as multigene families. Unlike operons, the genes in a multigene family are identical or are very similar and are not regulated coordinately. The clustering of genes in multigene families probably reflects a requirement for multiple copies of that

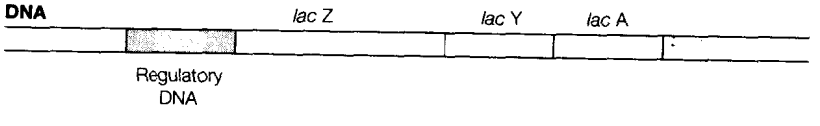


Fig. 1. The lac *operon. Three genes* (lac Z, Y and A) *are arranged and regulated together.*

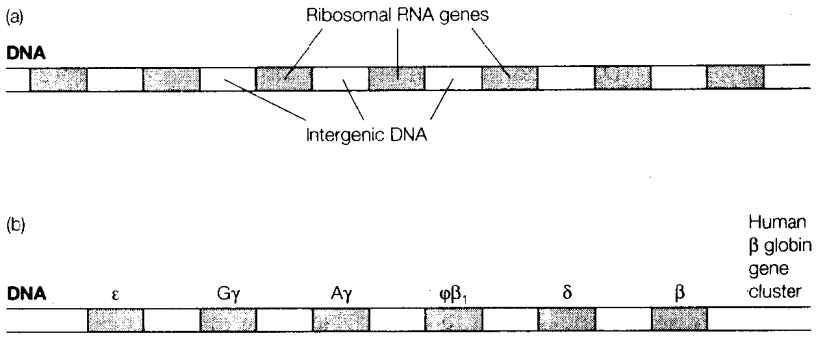


Fig. 2. (a) Simple multigene family; (b) complex multigene family.

gene which was fulfilled by duplication during evolution. Some multigene families exist as separate clusters on different chromosomes; this probably arose by rearrangements of the DNA during evolution which resulted in the breaking up of clusters. Multigene families may be simple or complex. In **simple multigene families** the genes are identical. An example is the gene for the 5S ribosomal RNA. In humans, there are about 2000 clustered copies of this gene reflecting the high demand of cells for the gene product (*Fig. 2a*). **Complex multigene families** contain genes that are very similar but not identical. An example is the globin gene family that encodes a series of polypeptides (α β γ ε ζ globins) that differ from each other by just a few amino acids. Globin polypeptides form complexes with each other and with a cofactor molecule called heme to give the adult and embryonic forms of the oxygen carrying blood protein, hemoglobin (*Fig. 2b*).

Gene expression

The biological information in a DNA molecule is contained in its base sequence. Gene expression is the process by which this information is made available to the cell. The use of the information is described by the **central dogma**, originally proposed by Crick, which states that information is transferred from DNA to RNA to protein (*Fig. 3*). During gene expression, DNA molecules copy their information by directing the synthesis of an RNA molecule of complementary sequence. This process is known as **transcription**. The RNA then directs the synthesis of a polypeptide whose amino acid sequence is determined by the base sequence of the RNA. This process is known as **translation**. The amino acid sequence of the protein determines its three-dimensional structure which in turn dictates its function. The central dogma states that the transfer of information can only occur in one direction – from DNA to RNA to protein – and cannot occur in reverse. An exception to this rule is found in retroviruses which have an enzyme called **reverse transcriptase** which can copy RNA into DNA. The functioning of cells, and in turn of living organisms, is dependent on the

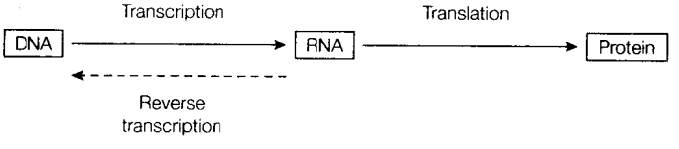


Fig. 3. The central dogma.

coordinated activity of many different proteins. The biological information contained within the genes acts as a set of instructions for synthesizing proteins at the correct time and in the correct place.

Gene promoters

The expression of the biological information present in genes is highly regulated. Not all the genes present in a cell's DNA are expressed and different genes are active in different cell types. The overall complement of genes that are active determines the characteristics of a cell and its function within the organism. Thus, for example, many of the genes that are active in muscle cells are different from those that are active in blood cells. Expression of genes is regulated by a segment of DNA sequence present upstream of the coding sequence known as the **promoter**. Conserved DNA sequences in the promoter are recognized and bound by the RNA polymerase and other associated proteins called **transcription factors** that bring about the synthesis of an RNA transcript of the gene. The expression of a gene in a cell is determined by the promoter sequence and its ability to bind RNA polymerase and transcription factors.

Introns and exons

One of the more surprising features of genes is that in higher organisms the coding information is usually split into a series of segments of DNA sequence called **exons**. These are separated by sequences that do not contain useful information called **introns** (*Fig. 4*). The number of introns varies greatly, from zero to more than 50 in some genes. The length of the exons and introns also varies but the introns are usually much longer and account for the majority of the sequence of the gene. Before the biological information in a gene can be used to synthesize a protein, the introns must be removed from RNA molecules by a process called **splicing** which leaves the exons and the coding information continuous. Introns are a feature of higher organisms only and are not usually found in bacteria.

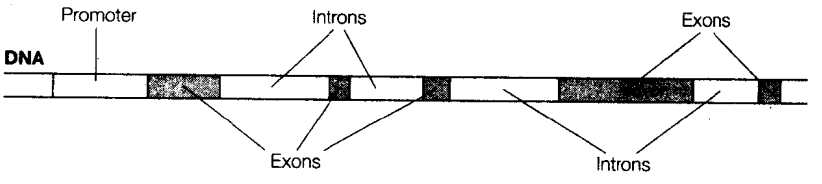


Fig. 4. Structure of a gene.

Pseudogenes

Some genes exist which resemble other genes but examination of their base sequence shows errors that make it impossible for them to contain useful biological information. These are called pseudogenes and they represent genes that have acquired errors or mutations in their DNA sequence during evolution causing their biological information to be scrambled so that they are no longer able to direct the synthesis of a protein. As such, pseudogenes are evolutionary relics. During evolution, the initial base changes causing loss of biological information are followed by more rapid changes so that the sequence of the pseudogene eventually deviates substantially from the original gene. Examples include several globin pseudogenes that are present in the globin gene clusters.