would be favoured energetically. The presence of a catalyst would enhance the rate of peptide bond formation.

The cellular factory responsible for synthesising proteins is the ribosome. The ribosome consists of structural RNAs and about 80 different proteins. In its inactive state, it exists as two subunits; a large subunit and a small subunit. When the small subunit encounters an mRNA, the process of translation of the mRNA to protein begins. There are two sites in the large subunit, for subsequent amino acids to bind to and thus, be close enough to each other for the formation of a peptide bond. The ribosome also acts as a catalyst (23S rRNA in bacteria is the enzyme- ribozyme) for the formation of peptide bond.

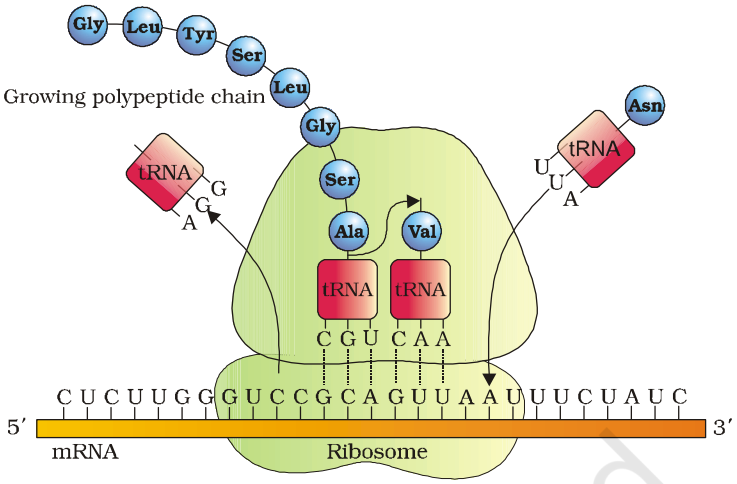

Figure 6.13 Translation

A translational unit in mRNA is the sequence of RNA that is flanked by the start codon (AUG) and the stop codon and codes for a polypeptide. An mRNA also has some additional sequences that are not translated and are referred as **untranslated regions** (**UTR**). The UTRs are present at both 5′-end (before start codon) and at 3′-end (after stop codon). They are required for efficient translation process.

For initiation, the ribosome binds to the mRNA at the start codon (AUG) that is recognised only by the initiator tRNA. The ribosome proceeds to the elongation phase of protein synthesis. During this stage, complexes composed of an amino acid linked to tRNA, sequentially bind to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. The ribosome moves from codon to codon along the mRNA. Amino acids are added one by one, translated into Polypeptide sequences dictated by DNA and represented by mRNA. At the end, a **release factor** binds to the stop codon, terminating translation and releasing the complete polypeptide from the ribosome.

6.8 Regulation of Gene Expression

Regulation of gene expression refers to a very broad term that may occur at various levels. Considering that gene expression results in the formation of a polypeptide, it can be regulated at several levels. In eukaryotes, the regulation could be exerted at

(i) transcriptional level (formation of primary transcript),
(ii) processing level (regulation of splicing),
(iii) transport of mRNA from nucleus to the cytoplasm,
(iv) translational level.

The genes in a cell are expressed to perform a particular function or a set of functions. For example, if an enzyme called beta-galactosidase is synthesised by *E. coli*, it is used to catalyse the hydrolysis of a disaccharide, lactose into galactose and glucose; the bacteria use them as a source of energy. Hence, if the bacteria do not have lactose around them to be utilised for energy source, they would no longer require the synthesis of the enzyme beta-galactosidase. Therefore, in simple terms, it is the metabolic, physiological or environmental conditions that regulate the expression of genes. The development and differentiation of embryo into adult organisms are also a result of the coordinated regulation of expression of several sets of genes.

In prokaryotes, control of the rate of transcriptional initiation is the predominant site for control of gene expression. In a transcription unit, the activity of RNA polymerase at a given promoter is in turn regulated by interaction with accessory proteins, which affect its ability to recognise start sites. These regulatory proteins can act both positively (activators) and negatively (repressors). The accessibility of promoter regions of prokaryotic DNA is in many cases regulated by the interaction of proteins with sequences termed **operators**. The operator region is adjacent to the promoter elements in most operons and in most cases the sequences of the operator bind a repressor protein. Each operon has its specific operator and specific repressor. For example, *lac* operator is present only in the *lac* operon and it interacts specifically with *lac* repressor only.

6.8.1 The *Lac* operon

The elucidation of the *lac* operon was also a result of a close association between a geneticist, Francois Jacob and a biochemist, Jacque Monod. They were the first to elucidate a transcriptionally regulated system. In *lac* operon (here *lac* refers to lactose), a polycistronic structural gene is regulated by a common promoter and regulatory genes. Such arrangement is very common in bacteria and is referred to as **operon**. To name few such examples, *lac* operon, *trp* operon, *ara* operon, *his* operon, *val* operon, etc.

The *lac* operon consists of one regulatory gene (the *i* gene – here the term *i* does not refer to inducer, rather it is derived from the word inhibitor) and three structural genes (*z*, *y*, and *a*). The *i* gene codes for the repressor of the *lac* operon. The *z* gene codes for beta-galactosidase (β-gal), which is primarily responsible for the hydrolysis of the disaccharide, lactose into its monomeric units, galactose and glucose. The *y* gene codes for permease, which increases permeability of the cell to β-galactosides. The *a* gene encodes a transacetylase. Hence, all the three gene products in *lac* operon are required for metabolism of lactose. In most other operons as well, the genes present in the operon are needed together to function in the same or related metabolic pathway (Figure 6.14).

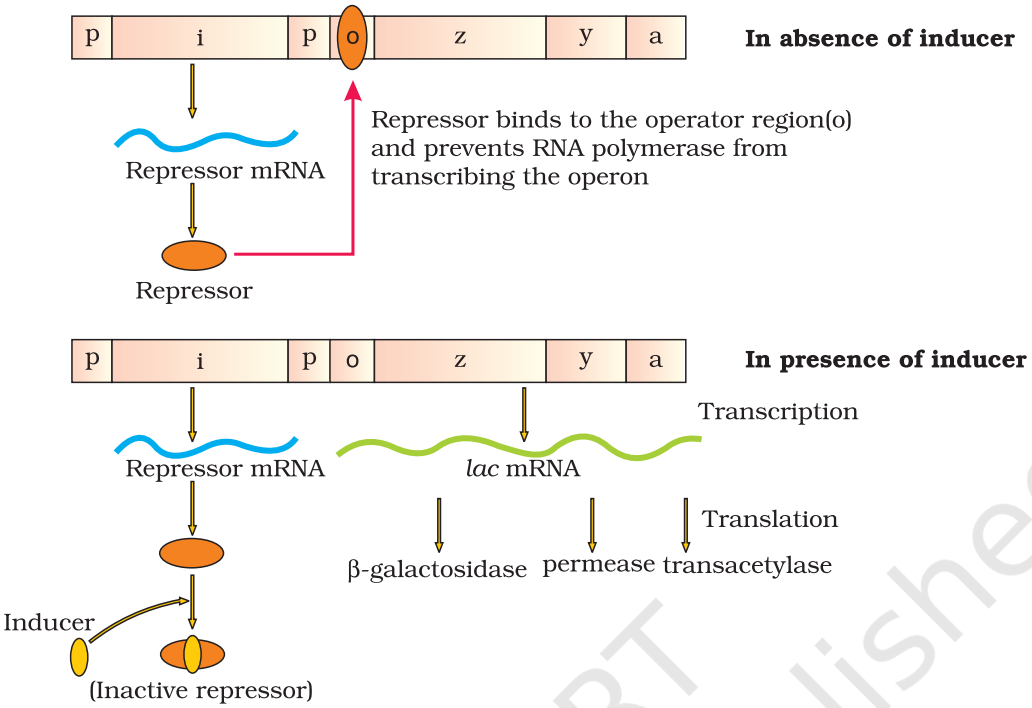

Figure 6.14 The *lac* Operon

Lactose is the substrate for the enzyme beta-galactosidase and it regulates switching on and off of the operon. Hence, it is termed as **inducer**. In the absence of a preferred carbon source such as glucose, if lactose is provided in the growth medium of the bacteria, the lactose is transported into the cells through the action of permease (Remember, a very low level of expression of *lac* operon has to be present in the cell all the time, otherwise lactose cannot enter the cells). The lactose then induces the operon in the following manner.

The repressor of the operon is synthesised (all-the-time – constitutively) from the *i* gene. The repressor protein binds to the operator region of the operon and prevents RNA polymerase from transcribing the operon. In the presence of an inducer, such as lactose or allolactose, the repressor is inactivated by interaction with the inducer. This allows RNA polymerase access to the promoter and transcription proceeds (Figure 6.14). Essentially, regulation of *lac* operon can also be visualised as regulation of enzyme synthesis by its substrate.

Remember, glucose or galactose cannot act as inducers for lac operon. Can you think for how long the lac operon would be expressed in the presence of lactose?

Regulation of *lac* operon by repressor is referred to as **negative regulation**. *Lac* operon is under control of positive regulation as well, but it is beyond the scope of discussion at this level.

6.9 Human Genome Project

In the preceding sections you have learnt that it is the sequence of bases in DNA that determines the genetic information of a given organism. In other words, genetic make-up of an organism or an individual lies in the DNA sequences. If two individuals differ, then their DNA sequences should also be different, at least at some places. These assumptions led to the quest of finding out the complete DNA sequence of human genome. With the establishment of genetic engineering techniques where it was possible to isolate and clone any piece of DNA and availability of simple and fast techniques for determining DNA sequences, a very ambitious project of sequencing human genome was launched in the year 1990.

Human Genome Project (HGP) was called a mega project. You can imagine the magnitude and the requirements for the project if we simply define the aims of the project as follows:

Human genome is said to have approximately 3 x 10^9 bp, and if the cost of sequencing required is US $ 3 per bp (the estimated cost in the beginning), the total estimated cost of the project would be approximately 9 billion US dollars. Further, if the obtained sequences were to be stored in typed form in books, and if each page of the book contained 1000 letters and each book contained 1000 pages, then 3300 such books would be required to store the information of DNA sequence from a single human cell. The enormous amount of data expected to be generated also necessitated the use of high speed computational devices for data storage and retrieval, and analysis. HGP was closely associated with the rapid development of a new area in biology called **Bioinformatics.**

Goals of HGP

Some of the important goals of HGP were as follows:

(i) Identify all the approximately 20,000-25,000 genes in human DNA;
(ii) Determine the sequences of the 3 billion chemical base pairs that make up human DNA;
(iiii) Store this information in databases;
(iv) Improve tools for data analysis;
(v) Transfer related technologies to other sectors, such as industries;
(vi) Address the ethical, legal, and social issues (ELSI) that may arise from the project.

The Human Genome Project was a 13-year project coordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the HGP, the Wellcome Trust (U.K.) became a major partner; additional contributions came from Japan, France, Germany, China and others. The project was completed in 2003. Knowledge about the effects of DNA variations among individuals can lead to revolutionary new ways to diagnose, treat and someday prevent the thousands of

disorders that affect human beings. Besides providing clues to understanding human biology, learning about non-human organisms DNA sequences can lead to an understanding of their natural capabilities that can be applied toward solving challenges in health care, agriculture, energy production, environmental remediation. Many non-human model organisms, such as bacteria, yeast, *Caenorhabditis elegans* (a free living non-pathogenic nematode), *Drosophila* (the fruit fly), plants (rice and *Arabidopsis*), etc., have also been sequenced.

Methodologies : The methods involved two major approaches. One approach focused on identifying all the genes that are expressed as RNA (referred to as **Expressed Sequence Tags** (ESTs). The other took the blind approach of simply sequencing the whole set of genome that contained all the coding and non-coding sequence, and later assigning different regions in the sequence with functions (a term referred to as **Sequence Annotation**). For sequencing, the total DNA from a cell is isolated and converted into random fragments of relatively smaller sizes (recall DNA is a very long polymer, and there are technical limitations in sequencing very long pieces of DNA) and cloned in suitable host using specialised vectors. The cloning resulted into amplification of each piece of DNA fragment so that it subsequently could be sequenced with ease. The commonly used hosts were bacteria and yeast, and the vectors were called as **BAC** (bacterial artificial chromosomes), and **YAC** (yeast artificial chromosomes).

The fragments were sequenced using automated DNA sequencers that worked on the principle of a method developed by Frederick Sanger. (Remember, Sanger is also credited for developing method for determination of amino acid sequences in proteins). These sequences were then arranged based on some overlapping regions present in them. This required generation of overlapping fragments for sequencing. Alignment of these sequences was humanly not possible. Therefore, specialised computer based programs were developed (Figure 6.15). These sequences were subsequently annotated and were assigned to each chromosome. The sequence of chromosome 1 was completed only in May 2006 (this was the last of the 24 human chromosomes – 22 autosomes and X and Y – to be

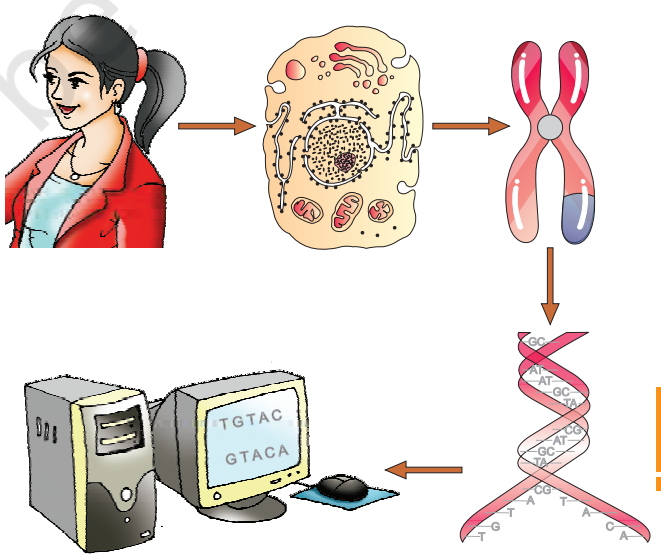

Figure 6.15 A representative diagram of human genome project

sequenced). Another challenging task was assigning the genetic and physical maps on the genome. This was generated using information on polymorphism of restriction endonuclease recognition sites, and some repetitive DNA sequences known as microsatellites (one of the applications of polymorphism in repetitive DNA sequences shall be explained in next section of DNA fingerprinting).

6.9.1 Salient Features of Human Genome

Some of the salient observations drawn from human genome project are as follows:

(i) The human genome contains 3164.7 million bp.

(ii) The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.

(iii) The total number of genes is estimated at 30,000–much lower than previous estimates of 80,000 to 1,40,000 genes. Almost all (99.9 per cent) nucleotide bases are exactly the same in all people.

(iv) The functions are unknown for over 50 per cent of the discovered genes.

(v) Less than 2 per cent of the genome codes for proteins.

(vi) Repeated sequences make up very large portion of the human genome.

(vii) Repetitive sequences are stretches of DNA sequences that are repeated many times, sometimes hundred to thousand times. They are thought to have no direct coding functions, but they shed light on chromosome structure, dynamics and evolution.

(viii) Chromosome 1 has most genes (2968), and the Y has the fewest (231).

(ix) Scientists have identified about 1.4 million locations where single-base DNA differences (**SNPs** – **single nucleotide polymorphism**, pronounced as 'snips') occur in humans. This information promises to revolutionise the processes of finding chromosomal locations for disease-associated sequences and tracing human history.

6.9.2 Applications and Future Challenges

Deriving meaningful knowledge from the DNA sequences will define research through the coming decades leading to our understanding of biological systems. This enormous task will require the expertise and creativity of tens of thousands of scientists from varied disciplines in both the public and private sectors worldwide. One of the greatest impacts of having the HG sequence may well be enabling a radically new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences and new high-throughput technologies, we can approach questions systematically and on a much

broader scale. They can study all the genes in a genome, for example, all the transcripts in a particular tissue or organ or tumor, or how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life.

6.10 DNA Fingerprinting

As stated in the preceding section, 99.9 per cent of base sequence among humans is the same. *Assuming human genome as 3 × 10^9 bp, in how many base sequences would there be differences?* It is these differences in sequence of DNA which make every individual unique in their phenotypic appearance. If one aims to find out genetic differences between two individuals or among individuals of a population, sequencing the DNA every time would be a daunting and expensive task. Imagine trying to compare two sets of 3 × 10^6 base pairs. DNA fingerprinting is a very quick way to compare the DNA sequences of any two individuals.

DNA fingerprinting involves identifying differences in some specific regions in DNA sequence called as **repetitive DNA**, because in these sequences, a small stretch of DNA is repeated many times. These repetitive DNA are separated from bulk genomic DNA as different peaks during density gradient centrifugation. The bulk DNA forms a major peak and the other small peaks are referred to as **satellite DNA**. Depending on base composition (A : T rich or G:C rich), length of segment, and number of repetitive units, the satellite DNA is classified into many categories, such as micro-satellites, mini-satellites etc. These sequences normally do not code for any proteins, but they form a large portion of human genome. These sequence show high degree of polymorphism and form the basis of DNA fingerprinting. Since DNA from every tissue (such as blood, hair-follicle, skin, bone, saliva, sperm etc.), from an individual show the same degree of polymorphism, they become very useful identification tool in forensic applications. Further, as the polymorphisms are inheritable from parents to children, DNA fingerprinting is the basis of paternity testing, in case of disputes.

As polymorphism in DNA sequence is the basis of genetic mapping of human genome as well as of DNA fingerprinting, it is essential that we understand what DNA polymorphism means in simple terms. **Polymorphism** (variation at genetic level) arises due to mutations. (Recall different kind of mutations and their effects that you have already studied in Chapter 5, and in the preceding sections in this chapter.) New mutations may arise in an individual either in somatic cells or in the germ cells (cells that generate gametes in sexually reproducing organisms). If a germ cell mutation does not seriously impair individual's ability to have offspring who can transmit the mutation, it can spread to

the other members of population (through sexual reproduction). Allelic (again recall the definition of alleles from Chapter 5) sequence variation has traditionally been described as a DNA polymorphism if more than one variant (allele) at a locus occurs in human population with a frequency greater than 0.01. In simple terms, if an **inheritable mutation** is observed in a population at high frequency, it is referred to as **DNA polymorphism**. The probability of such variation to be observed in non-coding DNA sequence would be higher as mutations in these sequences may not have any immediate effect/impact in an individual's reproductive ability. These mutations keep on accumulating generation after generation, and form one of the basis of variability/polymorphism. There is a variety of different types of polymorphisms ranging from single nucleotide change to very large scale changes. For evolution and speciation, such polymorphisms play very important role, and you will study these in details at higher classes.

The technique of DNA Fingerprinting was initially developed by Alec Jeffreys. He used a satellite DNA as probe that shows very high degree of polymorphism. It was called as **Variable Number of Tandem Repeats** (VNTR). The technique, as used earlier, involved Southern blot hybridisation using radiolabelled VNTR as a probe. It included

(i) isolation of DNA,

(ii) digestion of DNA by restriction endonucleases,

(iii) separation of DNA fragments by electrophoresis,

(iv) transferring (blotting) of separated DNA fragments to synthetic membranes, such as nitrocellulose or nylon,

(v) hybridisation using labelled VNTR probe, and

(vi) detection of hybridised DNA fragments by autoradiography. A schematic representation of DNA fingerprinting is shown in Figure 6.16.

The VNTR belongs to a class of satellite DNA referred to as mini-satellite. A small DNA sequence is arranged tandemly in many copy numbers. The copy number varies from chromosome to chromosome in an individual. The numbers of repeat show very high degree of polymorphism. As a result the size of VNTR varies in size from 0.1 to 20 kb. Consequently, after hybridisation with VNTR probe, the autoradiogram gives many bands of differing sizes. These bands give a characteristic pattern for an individual DNA (Figure 6.16). It differs from individual to individual in a population except in the case of monozygotic (identical) twins. The sensitivity of the technique has been increased by use of polymerase chain reaction (PCR–you will study about it in Chapter 11). Consequently, DNA from a single cell is enough to perform DNA fingerprinting analysis. In addition to application in forensic

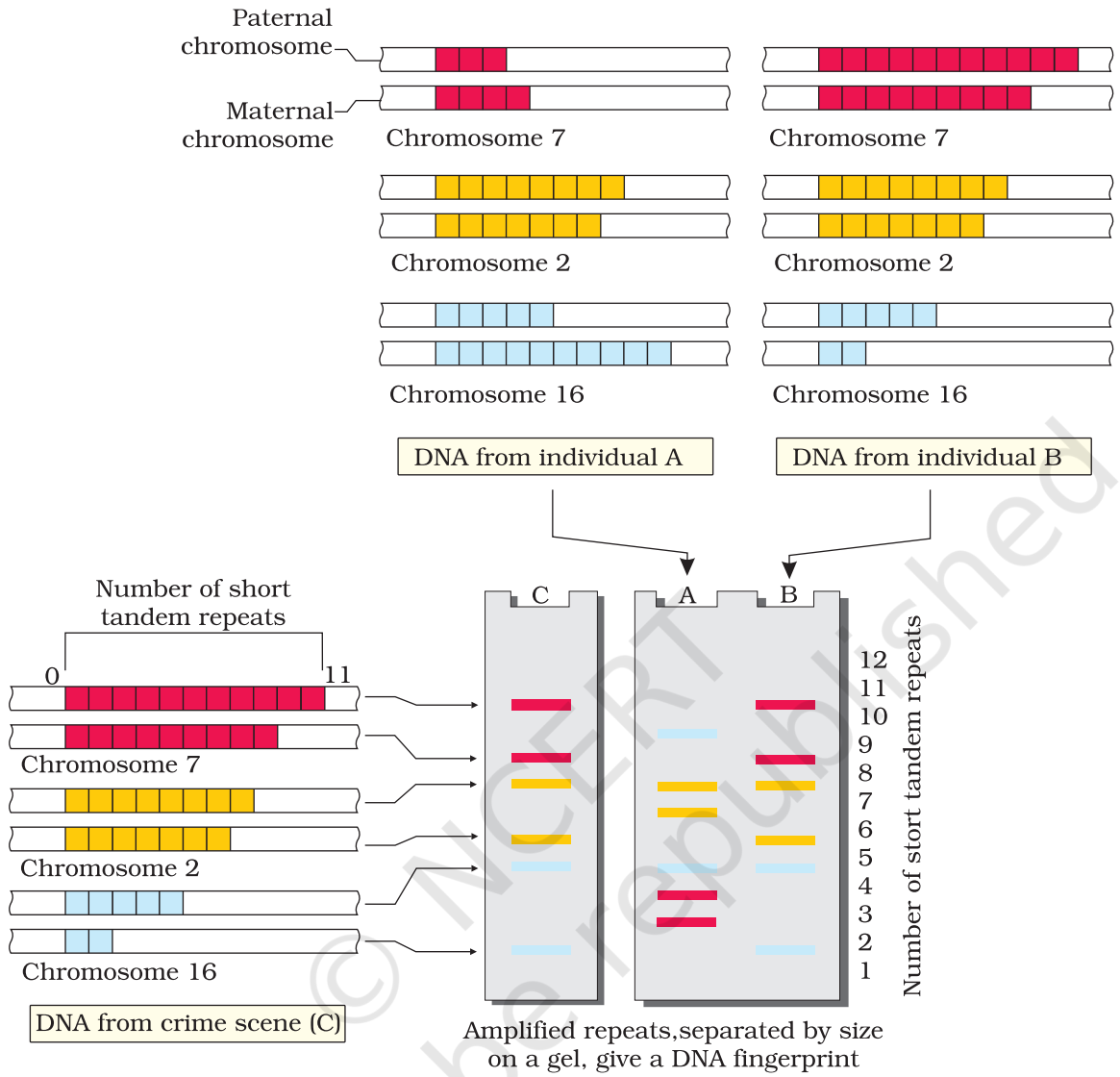

Figure 6.16 Schematic representation of DNA fingerprinting: Few representative chromosomes have been shown to contain different copy number of VNTR. For the sake of understanding different colour schemes have been used to trace the origin of each band in the gel. The two alleles (paternal and maternal) of a chromosome also contain different copy numbers of VNTR. It is clear that the banding pattern of DNA from crime scene matches with individual B, and not with A.

science, it has much wider application, such as in determining population and genetic diversities. Currently, many different probes are used to generate DNA fingerprints.

SUMMARY

Nucleic acids are long polymers of nucleotides. While DNA stores genetic information, RNA mostly helps in transfer and expression of information. Though DNA and RNA both function as genetic material, but DNA being chemically and structurally more stable is a better genetic material. However, RNA is the first to evolve and DNA was derived from RNA. The hallmark of the double stranded helical structure of DNA is the hydrogen bonding between the bases from opposite strands. The rule is that Adenine pairs with Thymine through two H-bonds, and Guanine with Cytosine through three H-bonds. This makes one strand complementary to the other. The DNA replicates semiconservatively, the process is guided by the complementary H-bonding. A segment of DNA that codes for RNA may in a simplistic term can be referred as gene. During transcription also, one of the strands of DNA acts a template to direct the synthesis of complementary RNA. In bacteria, the transcribed mRNA is functional, hence can directly be translated. In eukaryotes, the gene is split. The coding sequences, exons, are interrupted by non-coding sequences, introns. Introns are removed and exons are joined to produce functional RNA by splicing. The messenger RNA contains the base sequences that are read in a combination of three (to make triplet genetic code) to code for an amino acid. The genetic code is read again on the principle of complementarity by tRNA that acts as an adapter molecule. There are specific tRNAs for every amino acid. The tRNA binds to specific amino acid at one end and pairs through H-bonding with codes on mRNA through its anticodons. The site of translation (protein synthesis) is ribosomes, which bind to mRNA and provide platform for joining of amino acids. One of the rRNA acts as a catalyst for peptide bond formation, which is an example of RNA enzyme (ribozyme). Translation is a process that has evolved around RNA, indicating that life began around RNA. Since, transcription and translation are energetically very expensive processes, these have to be tightly regulated. Regulation of transcription is the primary step for regulation of gene expression. In bacteria, more than one gene is arranged together and regulated in units called as operons. *Lac* operon is the prototype operon in bacteria, which codes for genes responsible for metabolism of lactose. The operon is regulated by the amount of lactose in the medium where the bacteria are grown. Therefore, this regulation can also be viewed as regulation of enzyme synthesis by its substrate.

Human genome project was a mega project that aimed to sequence every base in human genome. This project has yielded much new information. Many new areas and avenues have opened up as a consequence of the project. DNA Fingerprinting is a technique to find out variations in individuals of a population at DNA level. It works on the principle of polymorphism in DNA sequences. It has immense applications in the field of forensic science, genetic biodiversity and evolutionary biology.

EXERCISES

1. Group the following as nitrogenous bases and nucleosides:
   Adenine, Cytidine, Thymine, Guanosine, Uracil and Cytosine.
2. If a double stranded DNA has 20 per cent of cytosine, calculate the per cent of adenine in the DNA.
3. If the sequence of one strand of DNA is written as follows:

   5′-ATGCATGCATGCATGCATGCATGCATGC-3′

   Write down the sequence of complementary strand in 5′→3′ direction.
4. If the sequence of the coding strand in a transcription unit is written as follows:

   5′-ATGCATGCATGCATGCATGCATGCATGC-3′

   Write down the sequence of mRNA.
5. Which property of DNA double helix led Watson and Crick to hypothesise semi-conservative mode of DNA replication? Explain.
6. Depending upon the chemical nature of the template (DNA or RNA) and the nature of nucleic acids synthesised from it (DNA or RNA), list the types of nucleic acid polymerases.
7. How did Hershey and Chase differentiate between DNA and protein in their experiment while proving that DNA is the genetic material?
8. Differentiate between the followings:
   (a) Repetitive DNA and Satellite DNA
   (b) mRNA and tRNA
   (c) Template strand and Coding strand
9. List two essential roles of ribosome during translation.
10. In the medium where *E. coli* was growing, lactose was added, which induced the *lac* operon. Then, why does lac operon shut down some time after addition of lactose in the medium?
11. Explain (in one or two lines) the function of the followings:
    (a) Promoter
    (b) tRNA
    (c) Exons
12. Why is the Human Genome project called a mega project?
13. What is DNA fingerprinting? Mention its application.
14. Briefly describe the following:
    (a) Transcription
    (b) Polymorphism
    (c) Translation
    (d) Bioinformatics