Lectures in Genetics for the first year students of Biotechnology

LECTURE NO. EIGHT

Genes

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 me 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.

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 50-100000 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 mat in humans gene sequences account for less than about 30% of me total DNA. Only one of the two strands of the DNA double helix carries the biological information: this is called me 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 me 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". Even for very short DNA molecules the number of different sequences possible is very large. In practice. Acre 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. Ah example is the toe operon in *E. coli* which contains three genes encoding enzymes required by me 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 multi-gene 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

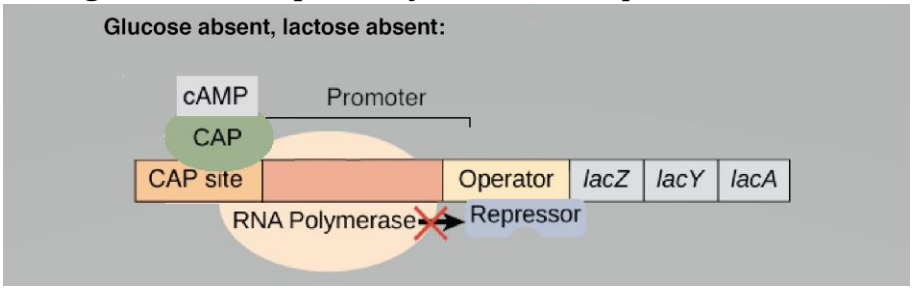
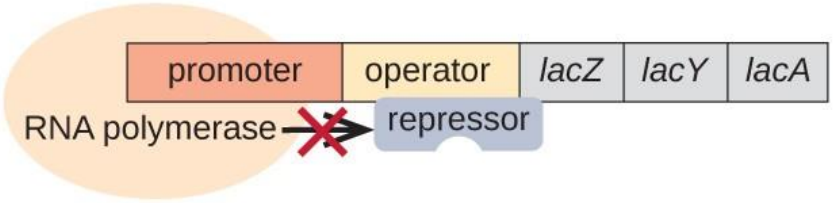


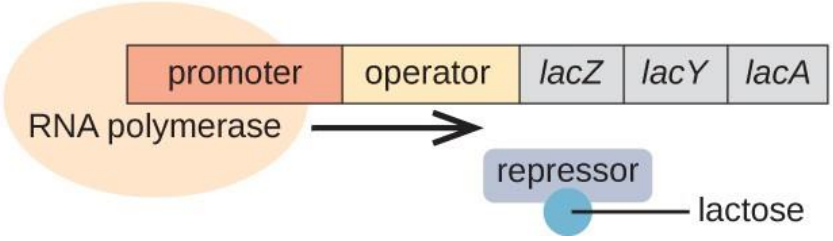
Fig. 1. The *lac* operon. *Three genes (lac Z.Y and A) are* arranged and regulated together.

multiple copies of that 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 dusters. Multigene families may be simple or complex. In **simple multi-gene 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 Ac adult and embryonic forms of the oxygen carrying blood protein, hemoglobin (Fig. 2b).

In the absence of lactose, the *lac* repressor binds the operator, and transcription is blocked.



In the presence of lactose, the *lac* repressor is released from the operator, and transcription proceeds at a slow rate.



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 me 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 add sequence is determined by the base sequence of the RNA. This process is known as translation. The amino add sequence of the protein determines its three-dimensional structure which in turn dictates its function. The central dogma states that die 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 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 promotor:

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 die promoter. Conserved DNA sequences in the promoter are recognized and bound

by die RNA polymerase and other associated proteins called transcription factor 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 mat 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 me 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.

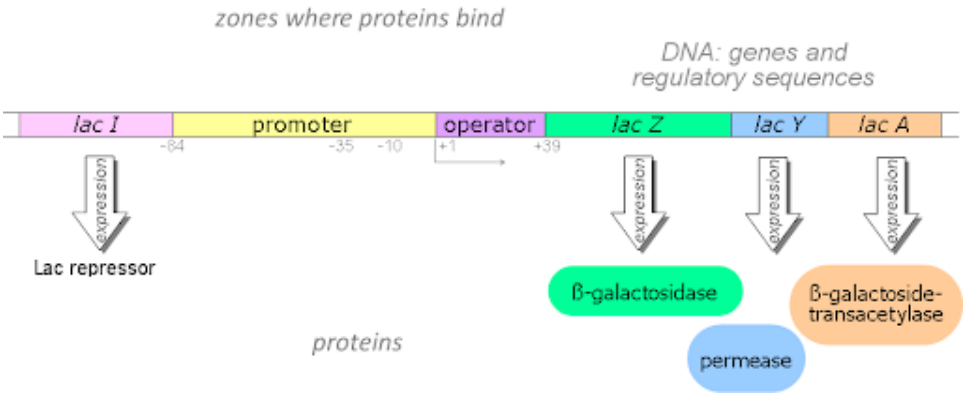


Fig. 4. Structure of a gene.

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 die original gene. Examples include several globin pseudogenes that are present in the globin gene clusters.