Page 1 of 30 Pages in this Full pdf

MZO-005: Genomics and Proteomics
Guess Paper-1

Q1. Describe the prokaryotic genomes and different types of plasmids found in the prokaryotes.

Ans. For a long time it was believed that most of the genomic information of prokaryotes is stored in a single, circular, double stranded DNA molecule called the chromosomal DNA. Some prokaryotes also contain additional genes in smaller, circular, autonomously replicating extra chromosomal DNA molecules called plasmids. Plasmids are much smaller in size than prokaryotic chromosomal genomes, and plasmid genes confer bacteria with special properties. Accordingly, plasmids can be distinguished into five types:

a) Resistance plasmids: They confer antibiotic resistance. For example, the plasmid pBR322 carries the genes for tetracycline and ampicillin resistance

b) Degradative plasmids: They are responsible for the degradation of unusual substances to which the bacteria are exposed, such as the TOL plasmid in Pseudomonas putida which contains genes encoding the enzymes required for toluene degradation)

c) Col or bacteriocinogenic plasmids: They direct the production of proteins which can kill other bacteria that do not carry these plasmids. For example, the Escherichia coli plasmid encodes the production of colicins.

d) Virulence plasmids: They make the bacteria pathogenic, such as the plasmids of the bacterium Yersinia pestis, which causes bubonic plague. These plasmids encode proteins called adhesins by means of which the bacteria bind to the cell membrane of animal cells, and then release their toxins. The toxins are also encoded by plasmid genes. Similarly, several E.coli strains contain the Ent plasmids which code for the production of enterotoxins that affects the intestines and causes bacterial diarrhoea.

e) Fertility plasmids: They are also called F-plasmids which contain transfer (tra) genes that allow the genetic material to pass directly between bacterial cells by conjugation.

Recent information indicates that this traditional view of the prokaryotic genome does not always hold true. Thus, the genomes of most prokaryotes, like E. coli is made up of a single circular DNA molecule which is supercoiled into a compact ball within the nucleiod. However, some prokaryotic genomes have been found to comprise of many linear or circular chromosomes. For example, Vibrio cholerae, the bacterium that causes cholera, contains two circular chromosomes, one of which contains the genes involved in metabolism and virulence, while the other contains the remaining essential genes. The bacterium Borrelia burgdorferi contains up to 11 copies of a single linear chromosome which are diffused as separate strands throughout the cell.

Prokaryotic genomes have a feature called an operon which is a co-regulated cluster of genes with related functions, such as genes involved in the same biochemical pathway. An operon consists of one or more structural genes, an operator gene, a promoter gene and a regulator gene. In addition, it requires an inducer, a repressor or a co-repressor to signal whether it has to be induced or repressed, thus distinguishing between an inducible operon and a repressible operon, respectively. The term 'operon' was coined by Jacob and Monod who first characterized an operon called the lac operon in Escherichia coli. The lac operon is an inducible operon which is expressed in response to the presence of lactose and absence of glucose in the growth medium of the bacterium. It consists of three structural genes that are required for lactose utilisation, namely, lacZ, lacY, and lacA which encode a β-galactosidase, a lac permease, and a transacetylase, respectively. The expression of all three genes is regulated by the common regulatory gene, lacI, which is usually located upstream of the structural genes. In fact, lac I, lacZ, lacY, and lacA are usually located together in a single cluster, which facilitates their inheritance together by vertical transfer, that is from parents to offsprings, or by horizontal gene transfer (HGT) between two organisms which are not related as parentoffspring (see unit 3). The 'selfish operon' hypothesis proposes that the clustering of genes in operons is advantageous to the constituent genes because it allows even non-essential genes which are functionally linked together in an operon to be transferred together along with the essential genes by HGT.

Prokaryotic genomes are compact and have only 6-14% of intergenic DNA interspersed between genes encoding proteins and structural RNAs. Their intergenic DNA mostly contains regulatory elements.

Q2. Write a short note on the organization and structure of eukaryotic genome.

Ans. Eukaryotic genomes are divided into multiple linear chromosomes, which are widely variable in number, ranging from 1 in males (which are haploid) or 2 in females of the Jack jumper ant, Myrmecia pilosula to 16000 macronuclear nanochromosomes in the ampliploid ciliated protozoan, Oxytricha trifallax. All aerobic eukaryotes have extrachromosomal DNA called the mitochondrial genome which exists in the form of multiple, double-stranded circular copies, in the mitochondria. In addition, photosynthetic eukaryotes also have the chloroplast genome which may be a single, or two or more circular DNAs.

Eukaryotic genomes have a high percentage of non-coding DNA, which ranges up to almost 99% in the human genome. Thus, only about 1% of the human genome codes for proteins, while the remaining genome has been referred to as 'junk' DNA or 'selfish' DNA. However, recent studies show that some of these non-coding DNA regions have several types of regulatory elements such as:

(a) Promoters usually precede a gene and provide binding sites for the protein machinery to carry out transcription of the gene.

Page 2 of 30 Pages in this Full pdf

(b) Enhancers provide binding sites for proteins that help activate transcription, and can be found before or after the gene they control, sometimes at large distances.
(c) Silencers provide binding sites for proteins that repress transcription, and can be found before or after the gene they control, sometimes at large distances.
(d) Insulators provide binding sites for proteins that control transcription by preventing enhancers from aiding in transcription (enhancer-blocker insulators), preventing structural changes in DNA that repress gene activity (barrier insulators), or by acting as both an enhancer blockers and barriers.
Other regions of non-coding DNA also code for various kinds of RNA molecules like transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), microRNAs (miRNAs), and long noncoding RNAs (lncRNAs), which either function in protein synthesis or the regulation of protein production. Eukaryotic genes are mostly split genes (see below) and have coding region or exons and noncoding DNA segments called introns located within the genes. Introns are removed from the mature mRNA by a process called splicing (see below) and some of them may contain enhancer elements.
The non-coding DNA of eukaryotic genomes contain high percentages of tandemly repetitive DNA sequences like satellite DNA, minisatellite DNA and microsatellite DNA or interspersed genome wide repeats (see Unit 2). Eukaryotic chromosomal genomes also vary widely in the number of proteincoding genes, ranging as widely as 6000 genes in the budding yeast, Saccharomyces cervesiae (genome size 12 million bp); 18000 genes in the roundworm, Caenorhabditis elegans (genome size 95.5 million bp); between 20000-25000 genes in humans, Homo sapiens (genome size 3.1 billion bp); 51000 genes in rice, Oryza sativa (genome size 470 million bp); and, 60000 genes in the single-celled protozoan, Trichomona vaginalis (genome size 160 million bp).
The mammalian mitochondrial genome is approximately 16.5 kb and it usually contains 13 protein coding genes, 2 ribosomal RNAs (one small 12S and one large 16S rRNA), 22 transfer RNAs (tRNAs), and a major non-coding region. In contrast, fungal genomes are somewhat larger (70-100 kb), while known plant mitochondrial genomes range in size from 187 kb to over 2400 kb. Plant mitochondrial genomes often have multiple large repeated regions, and also contain introns. The chloroplast genome in land plants ranges from 120 kb to 160 kb in size, and contains approximately 120 genes. The origin of mitochondria and chloroplasts in eukaryotes has been explained by the endosymbiotic hypothesis. According to this hypothesis, mitochondria and chloroplasts have descended from specialized bacteria that were endocytosed by another species of prokaryote (probably anaerobic) and became incorporated into their cytoplasm. This resulted in a symbiotic association wherein the symbiont bacteria was able to provide energy for the host cell by performing cellular respiration or photosynthesis, thus providing the host with a considerable evolutionary advantage over other prokaryotes.

Q3. How sequence complexity is determined? Explain.
Ans. The complexity of a genome is defined as the number of base pairs of unique or non-repetitive DNA in the genome or a segment of it. Unique or nonrepetitive DNA sequences occur only once in a genome and are not repeated, while repetitive DNA sequences such as satellite, minisatellite and microsatellite DNA (see unit 2) comprise of a sequence of variable length called the repeat sequence, which occurs multiple times in the genome. In the example depicted in Figure 1.1, S1 is a 30 bp long non-repetitive DNA sequence; S2 is a 30 bp long repetitive DNA sequence made up of the 10 bp repeat sequence, ATGATCTACG, repeated thrice; and, S2 is a 30 bp long repetitive DNA sequence made up of the 5 bp repeat sequence, ATGAT, repeated six times. Thus, the complexity (c) of S1 is 30, that of S2 is 10 while that of S3 is 5.

Sequence complexity	L	c
S1 5′–ATGATCTACGTAGTATGGCCTATGCCGGCT–3′	30	30
S2 5′–ATGATCTACGATGATCTACGATGATCTACG–3′	30	10
S3 5′–ATGATATGATATGATATGATATGATATGAT–3′	30	05

Fig.: Complexity of DNA

The complexity of a genome is related to its renaturation kinetics. A DNA double helix (duplex) comprises of two antiparallel and complementary DNA strands held together in a helical structure by inter-chain hydrogen bonds between pairs of bases. When a DNA solution is heated, the two strands of the double helix unwind and finally separate into individual strands by breaking of the inter-chain hydrogen bonds. This process is called denaturation. Cooling the solution eventually causes the hydrogen bonds between the two complementary strands to reform thus bringing about the renaturation of the two strands into the double helix.
The two-step renaturation model states that DNA renaturation occurs in two steps: nucleation and zipping. During nucleation a short duplex region forms between the complementary regions of the two strands of DNA. Once the strands are perfectly aligned by nucleation, the two strands are zipped together by rapid hybridization leading to complete renaturation of DNA. The zipping rate is directly proportional to the length of single stranded DNA (L) and inversely proportional to the sequence complexity (c). DNA renaturation is a second order kinetic reaction whose rate depends upon the concentration of the two strands of single stranded DNA. It can be analysed by Cot curve analysis which gives an estimate of the complexity of the genome.

Page 3 of 30 Pages in this Full pdf

If C is concentration of single stranded DNA at time, t, the rate of renaturation is given by:

$$-\frac{dC}{dt} = kC^2$$ where k is the rate constant

or $$\frac{dC}{C^2} = -kdt$$

Integrating between the limits $C = Co$ at initial time, $t = to$ and $C = C$ at time t, we have,

$$\frac{C}{Co} = \frac{1}{1 + k\text{Cot}}$$

When renaturation is half complete, $C = \frac{1}{2}Co$ at time $t = t_{1/2}$

Thus,

$$Cot_{1/2} = \frac{1}{k}$$

The rate constant, k, is given by the equation:

$$K = \frac{\sqrt{L}}{c}$$

Thus, $$Cot_{1/2} = \frac{c}{\sqrt{L}}$$

Since the length of any given DNA sequence (L) is constant, it follows that:

$$Cot_{1/2} = c$$

Thus, in the above given example, $Cot_{1/2}S1 > Cot_{1/2}S2 > Cot_{1/2}S3$. Genomes of higher complexity with a large percentage of unique and non-repetitive DNA sequences will take a longer time to renature than genomes of lower complexity with a low percentage of unique and, conversely, high percentage of repetitive DNA sequences.

Q4. Describe the model nuclear genomes of Drosophila and mammals.

Ans. Drosophila Genome: The fruit fly, Drosophila melanogaster is an important model organism, and its genome sequence was released in 2002. The euchromatic DNA is less condensed and more easily transcribed into gene expression. In contrast, heterochromatin is the relatively highly condensed and more often gene-poor region and transcriptionally silent in gene expression. The genome of Drosophila is 180 Mb in size, of which 120 Mb is euchromatic while the remaining is heterochromatic. The genome encodes 13676 genes, of which the euchromatin contains about 98% of the protein-coding genes. The heterochromatin is largely composed of simple sequence repeats. The genome comprises four chromosomes - the sex chromosomes X and Y, and two autosomes referred to as chromosomes 2 and 3. In addition, there is a small chromosome 4, which is negligible.

Transposable elements or jumping genes make up nearly 4% of the euchromatic Drosophila genome and vary in copy number from one to 146. They mostly occur in intragenic regions, often nested within other transposons. Transposable elements were also found inserted into the introns of 431 genes.

The Drosophila genome exhibits complex gene structures. Approximately 15% of the genes (2,054) are overlapping genes, involving the overlap of messenger RNAs (mRNAs) on opposite strands. For example, the 3' end of CG9455 overlaps the 5' end of Serine protease inhibitor 1 (Spn1:CG9456), sharing 408 base pairs of sequence. The translation start site of Spn1 is 219 bp beyond the translation stop of CG9455 such that the genes encode distinct proteins (Fig. 2.3A). A nested gene is a gene whose entire coding sequence lies between the start codon and the stop codon of a larger external, or surrounding gene. Approximately 7.6% of the Drosophila genes (1038) are nested genes, such that the coding sequences of one gene are included within the introns of other genes. For example, two genes, CG31049 and CG33204, map to the second and third introns, respectively, of gene Darkener of Apricot (Doa) (Fig. 2.3B). The majority (574) of the nested protein-coding genes are transcribed from the opposite strand of the surrounding gene.

Page 4 of 30 Pages in this Full pdf

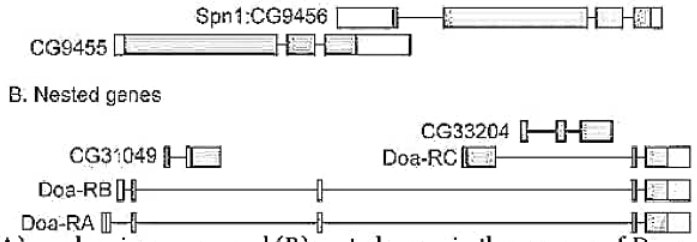

Fig.: Examples of (A) overlapping genes, and (B) nested genes in the genome of Drosophila melanogaster

A segment of DNA that structurally resembles a gene but is not capable of coding for a protein or a functional RNA product, is called a pseudogene. The Drosophila genome contains 17 pseudogenes. Of these, 15 originated by recombination and contain intact introns, one originated by retro transposition and lacks introns, and one could not be classified definitively.

The Drosophila genome comprises of about 30% heterochromatin DNA. The centric region of chromosomes X, 2 and 3, most part of chromosome 4 and all of chromosome Y is heterochromatic in nature. Approximately two thirds of heterochromatin is made up of satellite sequences. It also contains repetitive elements (e.g., transposons), and some single copy DNA. For example, approximately 80% of heterochromatic Y chromosome is composed of nine simple repeated sequences, including (AAGAC)n (8 Mb), (AAGAG)n (7 Mb), and (AATAT)n (6 Mb). In addition, the Y chromosome also contains one of two ribosomal DNA loci (bobbed) and six genes essential for male fertility (kl-1, kl2, kl-3, kl-5, ks-1, and ks-2).

The heterochromatic genes of Drosophila encode products with diverse cellular functions. Examples include light (functions in post-Golgi protein trafficking), concertina (encodes a G-protein subunit), Nipped-B (involved in morphogenesis), rolled (functions as a MAP kinase), poly ADP-ribose polymerase (regulates chromatin structure by poly-ADP ribosylation of proteins), and bobbed (ribosomal RNA). These genes differ significantly from euchromatic genes in that introns and regulatory regions contain clusters of transposable elements and some introns are hundreds of kilobases in length.

Mammalian Genome: The genomes of several mammals have been sequenced over the years. The genome of the mouse (Mus musculus) is 2.6 Gb in size while that of rat (Rattusnorvegicus) is 2.75 Gb. The domestic dog (Canis lupus familiaris) has a genome size of 2.5 Gb, while primates such as chimpanzee (Pan troglodytes) and gorilla (Gorilla gorilla), have genomes of nearly 3 Gb, like that of humans (Homo sapiens). The mouse genome has approximately 40 % repetitive sequences while the human genome has about 45 % repetitive sequences. Both genomes contain a total of 20000-25000 genes.

The genes of mammalian genomes appear to be organized in distinct 'gene rich' clusters which are separated by gene-poor 'deserts'. The human and mouse genomes have large-scale functional clustering of genes with some of them clustering together in specific chromosomal regions. Both human and mouse genomes have the gene deserts which are devoid of genes and may be up to 3 Mb long. Some of these gene deserts have been found to contain enhancers to distant flanking genes.

Mammalian genomes contain "ultra-conserved elements" (UCEs) that are highly conserved between evolutionarily distant species. The human genome contains 481 UCEs that are >200 bp in length and are 100% invariant between the human, rat, and mouse genomes. Some of these UCEs code for absolutely conserved protein but some UCEs are enhancer elements of nearby genes which contain multiple transcription-factor-binding sites, such that variation of a base would disrupt the function of one or more transcription factors, making the UCEs highly resistant to evolutionary change.

The organization of the human genome is shown in Figure 4. Genes make up only about 1 % of the human genome, with the remaining 99 % of the genome comprising intergenic DNA. Out of this, gene-related non-coding sequences such as pseudogenes, introns and gene fragments make up about 39 % of the genome. A large percentage of the latter consists of interspersed genome wide repeats namely LINEs, SINEs, transposons and long terminal repeats (LTRs), followed by tandemly repeated DNA regions, namely the satellite, minisatellite and microsatellite regions. It has been proposed that genomes having a large percentage of intergenic DNA may have an evolutionary advantage over those with a lower percentage of intergenic DNA because the flanking non-coding DNA protects the coding regions from sequence change by spontaneous mutation or recombination.

Q5. Define Pseudogenes and explain their types.

Ans. The term 'pseudogene' was coined in 1977 to refer to a non-functional 5S rRNA gene in Xenopus laevis. They are important constituents of non-coding DNA found in plant, insect, mammalian and bacterial genomes. A DNA segment that structurally resembles a gene but is unable to code for a protein is known as a pseudogene. Most often, pseudogenes are derived from genes that have accumulated mutations during evolution to the stage that they can no longer code for proteins. They have high sequence homology to known functional genes, but do not code for proteins. They are inheritable genetic elements and are often members of gene families containing functional genes. For example, the beta-globin gene family located on chromosome 11 consists of five functional genes and two pseudogenes (Fig.).

Page 5 of 30 Pages in this Full pdf

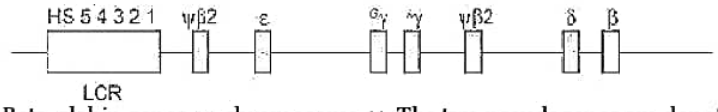

Fig.: The Beta-globin genes on chromosome 11. The two pseudogenes are denoted by Ψ

Pseudogenes are of two types:

(a) Unprocessed pseudogenes: The unprocessed pseudogenes are formed when genes are duplicated by one or more copies underwent deleterious genetic alterations such as frameshift mutations and creation of a premature stop codon which prevents the gene from encoding a functional protein. Some pseudogenes such as the OCT4-pseudogene 1 (OCT4-pg 1), OCT4-pseudogene 2 (OCT4-pg 2), OCT4-pseudogene 1 (OCT4-pg 3) and OCT4-pseudogene 4 (OCT4-pg 4) are transcribed to produce unstable proteins in various human cancer cell lines and are thought to have a role in tumor cell growth and tumor progression. Unprocessed pseudogenes are found in proximity to the genes from which they are derived, such as the pseudogenes in the previous example of the beta-globin gene family.

(b) Processed pseudogenes: Processed pseudogenes are produced by a retrotransposition event involving the synthesis of complementary DNA (cDNA) from an RNA product of a gene by reverse transcription, followed by insertion of the cDNA in a different region of the genome. Unlike the functional gene, the cDNA copy lacks a promoter, since the promoter sequence itself is not transcribed during RNA synthesis. Thus, the cDNA copy forms a pseudogene which is not expressed. A processed pseudogene also often lacks introns, since these are absent from the mature mRNA template, and instead contain a poly A-tail. For example, a processed pseudogene, Mgst1-psi, is found in the GST supergene family in Drosophila melanogaster. This pseudogene contains a poly-A tail towards the 3′ end, has no introns, a premature stop codon, and a direct repeat at both ends. It is presumed to have been derived from the reverse transcription and subsequent integration of Mgst1 mRNA into the Drosophila melanogaster genome.

Pseudogenes were once referred to as 'genomic fossils' and were regarded as junk DNA because they are not able to produce functional proteins. However, approximately 20% of pseudogenes are transcribed into RNAs that control expression of other genes by producing natural small interfering RNA (siRNAs) and long non-coding RNAs (lncRNAs). A few pseudogenes occurring in the human genome have been found to have oncosuppressive functions, that is, they can prevent the growth of cancer cells and the progression of cancer. The PTEN pseudogene, PTEN1, is a processed pseudogene of PTEN. It contains a mutated start codon and is unable to produce a protein. However its mRNA acts as a 'sponge' by complementary binding to a number of miRNAs such as miR-17, miR-19, miR-21, miR-26, and miR-214 which normally target and inhibit PTEN. The sponging effect of PTEN1 relieves the inhibitory effects of the miRNAs on PTEN, restoring its normal tumor suppressor activities and thus leading to a decrease in the growth of many cancers.

Q6. How the location of genes in a genome can be identified by computational methods or experimental analysis? Explain.

Ans. Eukaryotic genomes are less compact than prokaryotic genomes, and the genes are organized into 'gene rich' or 'gene poor' regions, separated by stretches of intergenic DNA of variable length. The location of genes in a genome can be identified by computational methods or experimental analysis.

Computational methods for locating gene: There are two approaches commonly used for locating genes in a genome sequence by computational analysis:

(a) Locating genes by genome analysis: Computer-based gene prediction softwares like GENSCAN and GrailEXP are used to locate genes in a genome sequence by identifying the distinctive features of a gene. These softwares search open reading frame (ORF) of a proteincoding gene, codon bias, identifying exon-intron boundaries, upstream regulatory elements and CpG islands.

- **ORF Scanning:** This involves scanning a genome sequence for ORFs. An ORF is the sequence of codons which specify the amino acids of a protein. It begins with the start codon ATG (GTG and ATA in the mitochondrial genome of some species), and ends with a stop codon (TAA, TAG, TGA). The length of the ORF is crucial in correct identification, as the average length of a gene is 317 codons for genes of Escherichia coli, 483 codons for genes of Saccharomyces cerevisiae, and approximately 450 codons for genes of humans.
- **Codon bias:** The genetic code is degenerate hence multiple codons can code for particular amino acid, not all codons are used at an equal rate in the genes of an organism, and some codons are found to be more favoured than others. For example, while the amino acid may be coded by four different codons- GCA, GCC, GCU and GCG, the preferred codon for alanine in human genes is GCC. This feature should be kept in mind while locating the genes of a species in a genome.
- **Exon-intron boundaries:** Exon-intron boundaries are the sequences which signify the end of an exon and the beginning of an intron. Locating them in series can indicate the presence of a gene in a eukaryotic genome. Exon-intron boundaries are indicated by the sequence 5'-AG↓GT/A-3' at the exon-intron junction.
- **CpG islands:** CpG islands are 500-1500 bp long stretches of DNA rich in CpG dinucleotides (p indicates a phosphodiester bond), and are characterised by a CG: GC ratio of more than 0.6.They are frequently associated with the promoters of vertebrate genes, and are located in the vicinity of about 40% of promoters of mammalian genes. Thus, CpG islands are good indicators of the location of genes in vertebrate genomes.

(b) Locating genes by homology with known proteins: Computer-based gene prediction softwares may also be used to search databases for sequences of known proteins or segments of proteins which match with the protein predicted

IGNOU MOTHER PUBLICATION - ASSIGNMENTS & GUESS PAPERS & PYQ SOLVED PDF

INDIA'S NUMBER 1 STUDY MATERIAL PROVIDER

यह केवल एक सैंपल पीडीएफ है।

यह केवल Sample PDF है। पूरी Solved Assignment या Guess Paper PDF खरीदने के लिए हमारी वेबसाइट पर जाएँ

या

WhatsApp पर मैसेज करें।

WHATSAPP ONLY: +91 9485943709

WHATSAPP ONLY:: +91 9992290905

1. Solved Assignment Visit -:: https://motherpublication.in/

2. Pyq Solved Guess paper Visit:-: https://mothernotes.in/

Mother Publication • IGNOU Study Resources