MICROBIOLOGY AND IMMUNOLOGY (DBT01) (PG DIPLOMA)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

Lesson 1.1.1

DIVERSITY OF CELL SIZE AND CELL SHAPE

Objective

Objective

The objective of this lesson is to know the properties of the cell it's functions ,the strcture of prokaryotic and eukaryotic cells, and the differences between prokaryotic and eukaryotic cells

1.1.1.1 Introduction

The cell is the basic unit of organization or structure of all living matter within a selective and retentive semipermeable membrane, it contains a complete set of different kinds of units necessary to permit its own growth and reproduction from simple nutrients. Different cell biologists defined cell in different ways as.

- "A unit of Biological activity delimited by a semi permeable membrane and capable of self reproduction in a medium free of other living systems"
- "The simplest integrated organization in living systems capable of independent survival".

All these definitions have excluded viruses. A virus is neither an organism nor a cell, yet it consists of a core of nucleicacid (DNA or RNA) enclosed in an external mantle of protein. In the free state viruses are quite inert. They become activated only when they infect a living host cell. In a way, thus viruses are cellular parasites that can not reproduce by themselves. But because viruses are primitive and simpler units of life, they should be discussed prior to other cells.

Cells and the structures they comprise are too small to be directly seen, heared or touched. Yet, inspite of this tremendous handicap, cells are the subject of thousands of publications each year, with virtually every aspect of their mimsale structure coming under scrutinity. In many ways, the study of cell biology stands as a tribute to human curiosity for seeking to discover, and to human creative intelligence for devising the complex instruments and elaborate techniques by which these discoveries can be made.

It wasn't until the 1830's that the wide spread importance of cell was realized. In 1838 Matthias Schleiden, a German Lawyer Tusmed botanist, concluded that despite differences in the structure of various tissues, plants were made of cells and that the plant embryo arose from a single cell. In 1839 Theodar Schwann, a German Zoologist, and Colleague of Schleiden, Published a comprehensive report on the cellular basis of animal life.

Schwann concluded that the cells of plants and animals are similar structures and proposed these two tenets of "cell theory".

1. All organisms are composed of one or more cells.

2. The cell is the structural unit of life. By 1855 Rudolf Virchow, A German Pathologist, had made a convincing case for the third tenet of the cell theory.

3. Cells can arise only by division from pre existing cells.

1.1.1.2 Basic properties of cells

Just as plants and animals are alive, so too are cells. Life, infact is the most basic property of cells and cells are the smallest units to exhibit this property. Unlike the parts of a cell, which simply deteriorate, if isolated, whole cells can be removed from a plant or animal and cultured in a laboratory where they will grow and reproduce for extended periods of time. The first culture of human cells bearn in 1951. The cells were obtained from a malignant tumor and named HeLa cells after the donor, Henrietta Lacks. HeLa cells descended by cell division from this first cell sample – are still being grown in laboratories around the world today. Because they are so simple to study than cells situated with in the body, cells grown *invitro* (i.e. in culture outside the body) have become an essential tool of cell and molecular biologists.

1) Cells are highly complex and organized

Complexity is a property that can be described in terms of order and consistency. The more complex the structure is, the greater the number of parts that must be in their proper place, the less tolerance of errors in the nature and interactions of parts, and the more regulation or control that must be exerted to maintain the system.

2) Cells posses a genetic program and two means to use it

Organisms are built according to the information encoded in a collection of genes. The human genetic program contains enough information. Remarkbly, this vast amount of information is packed into a set of chromosomes that occupy the space of cell nucleus that is of few nanometers size.

Genes are more than storage lockers for information they constitute the blue prints for constructing cellular structures, the directions for running cellular activities and the program for making more of themselves.

3) Cells are capable to reproduce

Cells reproduce by division, a process in which the contents of a “Mother” cell are distributed into two “daughter” cells prior to division, the genetic material is faithfully duplicated and each daughter cell receives a complete and equal share of genetic information.

4) Cells acquire and utilize energy

Developing and maintaining complexity requires the constant input of energy. Virtually all of the energy required by life on the earth’s surface arrives in the form of electromagnetic radiation from the sun. The energy of light is trapped by light absorbing pigments present in the membranes of photosynthetic cells. It is converted by photosynthesis into chemical energy that is stored in energy-rich carbohydrates such as sucrose or starch. The energy trapped in these molecules during photosynthesis provided the fuel that runs the activities of nearly all the organisms on earth. For most animal cells energy arrives prepacked, usually in the form of sugar, glucose.

5) Cells carry out variety of chemical reactions

Cells function like miniaturized chemical plants. Even the simplest bacterial cell is capable of performing hundreds of different chemical transformations; virtually all chemical changes that take place in the cells require enzymes – molecules that greatly increase the rate at which a chemical reactions occur. The sum of the total chemical reactions that occur in a cell represents that cell's metabolism.

6) Cells engage in numerous mechanical activities

Cells are sites of bustling activity. Materials are transported from place to place, structures are assembled and then rapidly disassembled and in many cases the cell moves it self from one site to another. These types of activities are based on dynamic mechanical changes within the cells, most of which are initiated by changes in the shape of certain "Motor proteins".

7) Cells are able to respond to stimuli

Some cells respond to stimuli in obvious ways; a single celled protista, moves away from an object in its path or moves towards a source of nutrients. Most cells are covered with receptors that interact with substances in the environment in highly specific ways. Cells possess receptors to hormones, growth factors, extracellular materials, as well as to the substances on the surfaces of another cells. Cells may respond to specific stimuli by altering their metabolic activities, preparing for cell division, moving from one place to another or even committing suicide.

8) Cells are capable of self regulation

In addition to requiring energy, maintaining a complex ordered state requires constant regulation. As in the body as a whole, many different control mechanisms operate with in each living cell. The importance of cells regulatory mechanisms becomes evident when they break down. For example, failure of a cell to correct a mistake when it duplicates its DNA may result in a debilitating mutation or a break down in a cells growth control can transform the cell into a cancer cell with the capability of destroying the entire organism.

1.1.1.3 Different classes of cells

Once the electron microscope become widely available, biologists were able to examine the internal structure of a wide variety of cells. It became apparent from these studies that there were two basic classes of cells- prokaryotic and eukaryotic – distinguished by their size and the types of internal structures or organelles, they contain.

The structurally simple, prokaryotic cells are found only among bacteria, and conversely all bacteria consist of prokaryotic cells. All other types of organisms – protists, fungi, plants and animals consist of structurally more complex eukaryotic cells.

The comparison of both prokaryotic and eukaryotic cell reveals many basic differences as well as similarities. The similarities reflect the fact that eukaryotic cells almost are certainly evolved from prokaryotic ancestors. Because of their common ancestry, both types of cells share an identical genetic language, a common set of metabolic pathways, and many common structural features. For example both types of cells are bounded by plasma membrane of similar construction that serve as a selectively permeable barrier between the living and nonliving worlds.

Internally eukaryotes are much more complex both structurally and functionally than prokaryotic cells.

The genetic material of a prokaryotic cell is present in a nucleoid; a poorly demarcated region of the cell that lacks a boundary membrane to separate it from the surrounding cytoplasm. In contrast, the eukaryotic cells possess a nucleus, a region bounded by a complex membranous structure called the nuclear envelope. This difference in nuclear structure is the basis for the terms.

Prokaryotic: (pro = before; karyon = nucleus)

Eukaryotic: (eu = true; karyon = nucleus)

Prokaryotic cells contain relatively small amounts of DNA; a total length of DNA of a bacterium ranges from about 0.25 mm to about 3mm which is sufficient to encode between several hundred and several thousand proteins.

The cytoplasm of the two types of cells is also very different. The cytoplasm of a eukaryotic cell is filled with a great diversity of structures. Most notably, eukaryotic cells contain an array of membranous and membrane bound organelles.

For example: Plant and animal cells typically contain mitochondria. Where chemical energy is made available to fuel cellular activities, an endoplasmic reticulum, where many of cell proteins and lipids are manufactured; Golgi complexes where materials are stored, modified and transported to specific cellular destinations and a variety of simple membrane bound vesicles of varying dimensions.

Taken as a group the membranes of the eukaryotic cell serve to divide the cytoplasm into compartments with in which specialized activities can take place. In contrast cytoplasm of prokaryotic cells is essentially devoid of membranous structures. Exceptions to this generalization include mesosomes, which are derived from simple infoldings of the plasma membrane and the complex photosynthetic membranes of cyanobacteria.

The cytoplasmic membranes of eukaryotic cells form a system of inter connecting channels and vesicles that function in the direct transport of substances from one part of the cell to another, as well as between the inside of the cell and its environment. Because of their small size, directed intra cytoplasm communication is less important in prokaryotic cells, where the necessary movement of materials can be accomplished by simple diffusion. Both types of cells may be surrounded by a rigid, non-living cell wall that protects the delicate life form within. Although the cell walls of prokaryotes and eukaryotes may have similar functions, their chemical composition is very different.

Another major difference between eukaryotic and prokaryotic cells is that the eukaryotic cells divide by a complex process of mitosis in which duplicated chromosomes condense into compact structures that are reaggregated by a elaborate mitotic spindle that allows each daughter cell to receive an equivalent array of genetic material. In prokaryotes there is no condensation of the chromosome and no mitotic spindle. The DNA is duplicated and the two copies are separated simply and accurately by the growth of an intervening cell membrane. This simpler mechanism of division allows prokaryotic cells to proliferate much more rapidly than eukaryotic cells; a malt fed population of bacteria can double in number after every 20-40 minutes. For the most part, prokaryotes are non-sexual organisms. They contain only one copy of their single chromosome and have no processes comparable to meiosis, gamete formation or true fertilization. Even though true sexual reproduction is lacking among prokaryotes, some are capable of “conjugation”, in which a piece of DNA is passed from one cell to another.

Eukaryotic cells posses a variety of complex locomotary mechanisms, whereas those of prokaryotes are quite simple.

1.1.1.4. Types of prokaryotic cells

Prokaryotes are divided into two major groups or domains

1. The archaea (the archaeons)
2. The bacteria (or eubacteria)

The living archaeons include the methanogens (prokaryotes capable of converting CO_2 and H_2 gases into Methane (CH_4) gas.

The halophiles – prokaryotes that live in extremely salty environments.

The thermophiles: prokaryotes that live at very high temperatures. Included in this latter group are hyper thermophiles such as *pyrolobus fumarii*, which lives in the hydrothermal vents of the ocean floor and is capable of reproducing in super heated water at temperatures above 109^{o}C.

All other types of prokaryotes are classified in the domain II bacteria. This domain includes the smallest living cells, the mycoplasma (0.2 mm in diameter) which are also the only prokaryotes lacking a cell wall. The most complex prokaryotes are the cyanobacteria (formerly known as blue-green algae because of the bluish-green scum they can form on the surface of lakes & ponds) cyanobacteria contain elaborate arrays of cytoplasmic membranes, which serve as sites of photosynthesis.

1.1.1.5 Types of eukaryotic cells

In many regards the most complex cells are not found inside of plants or animals but rather are present in single celled (unicellular) protists where all of the machinery required for the complex activities in which this organism engages like sensing the environment, trapping of food, expelling excess fluid, evading the predators are present with in the single cell.

Complex unicellular organisms represent one evolutionary pathway has led to the evolution of multicellular organisms in which different activities are conducted by different types of specialized cells. The advantages provided by the division of labour among cells makes the life simple by the process of “differentiation”.

As a result of differentiation, different types of cells acquire a distinctive appearance and contain unique materials.

The pathway of differentiation followed by each embryonic cell depends primarily on the signals it receives from the surrounding environment, these signals in turn depends on the position of that cell with in the embryo.

Skeletal muscle cells contain a network of precisely aligned filaments composed of unique contractile proteins; cartilage cells become surrounded by a characteristic matrix containing polysaccharides and the protein collagen, which together provide mechanical support; Red blood cells become disk shaped sacks filled with a single protein haemoglobin.

Despite their many differences, all the cells of a multicellular plant or animal are composed of similar organelles, though in different shapes according to the necessity.

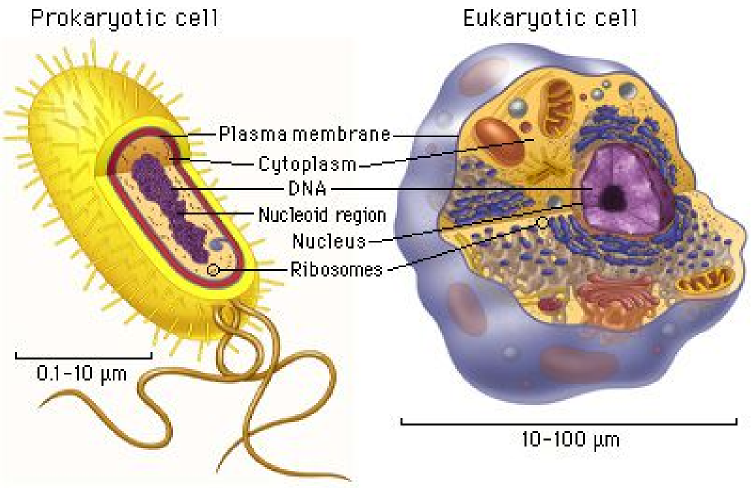


Fig: Structure of prokaryotic and eukaryotic

The size units

Nearly all cells are microscopic, thus the units most commonly used describe very small linear dimensions.

Two units of linear measure are most commonly used to describe structures with in a cell. The "Micrometer" (µm) and the "Nanometer" (nm). One µm is equal to 10^{-6} meters and one nm is equal to 10^{-9} meters.

Also Angstrom (^{o}A) which is no longer accepted in this metric nomenclature which is equal to one-tenth of a nanometer (nm), is still employed for atomic dimensions.

Bacteria typically range in length from about 1-5 µm. Eukaryotic cells range from about 10-30 µm.

1.1.1.6 Comparison of prokaryotic & Eukaryotic cells features that held in common by two types of cells

1. Plasma membrane of similar construction.
2. Genetic information encoded in DNA using identical genetic code.
3. Similar mechanisms for transcription and translation of genetic information including similar ribosomes.

4. Shared metabolic pathways (Ex: Glycolysis and TCA cycle).
5. Similar apparatus or complex for conversion of chemical energy as ATP (located in the plasma membrane of prokaryotes and the mitochondrial membrane of eukaryotes)
6. Similar mechanism of photosynthesis (between cyanobacteria and green plants).
7. Similar mechanism for synthesizing and inserting membrane proteins.
8. Proteasomes (protein digesting structures) of similar construction (between archebacteria and eukaryotes).

Features of eukaryotic cells not found in prokaryotes

1. Division of cells into nucleus and cytoplasm, separated by a nuclear envelope containing complex pore structures.
2. Complex chromosomes composed of DNA and associated proteins that are capable of compacting into mitotic structures.
3. Complex membranous cytoplasmic organelles (includes endoplasmic reticulum, golgicomplex, lysosomes, endosomes, peroxisomes and glyoxisomes).
4. Complex cytoskeletal system (including microfilaments, intermediate filaments, and microtubules).
5. Complex flagella and cilia.
6. Capable of ingesting fluid and particulate material by enclosure with in plasma membrane vesicles (endocytosis and phagocytosis).
7. Cellulose – containing cell walls (in plants).
8. Cell division utilizing a microtubule. Containing mitotic spindle that separates chromosomes.
9. Presence of two copies of genes per cell (diploidy) one from each parent.
10. Sexual reproduction requiring meiosis and fertilization. Summary

The cell is the basic unit of organization or structure of all living matter. Cells are highly complex and organized they can able to reproduce, acquire and utilize energy to carryout various chemical reactions, mechnical activities. The cells are mainly classified in to two types –prokaryotic and eukaryotic cells. the prokaryotes lacks the nuclear membrane that separates the nucleus from cytoplasm but it is present in eukaryotes. the cytoplasm of both cells is also different .the eukaryotic cell contains different membrane bound cytoplasmic organelles such as ER chloroplast ,mitochondria etc. so there is compartmentalization in eukaryotes but not in prokaryotes.

Model Questions

1. Compare a prokaryotic and eukaryotic cell on the basis of structural, functional and metabolic differences?
2. Describe the basic properties of a cell?
3. What is the importance of cell differentiation?

Reference Books

1. Cell and molecular biology by Gerald Karp.
2. Cell and molecular biology by De. Robertis.
3. Cell & molecular biology by P.K. Gupta.

K.Haritha

Lesson 1.1.2

CELL THEORY

Contents

Objective

The objective of the lesson is to know how the cell as in tented and by whom, and its characteristic features and various properties exhibited by the cell. This lesson also deals with the functions of a typical cell and different types of cells and these differentiations.

1.1.2.1. INTRODUCTION:

The cell is the fundamental unit of life, the under lying building block from which all organisms are constructed. The properties of cells which are the smallest units exhibit the characteristics of life; define both the potential capabilities and the inherent limitations of all living organisms. In the past several decades a wide variety of powerful new experimental approaches have been utilized to investigate the intricate of cell organization and function. As a result a revolution has occurred in our understanding of how cells are constructed and how they carryout the activities required for maintaining life.

1.1.2.2. The Discovery of Cells Led to the Formulation of Cell Theory

Most cells are Invisible to the naked eye, so scientists did not know of their existence prior to the invention of the Microscope. The perfection in the ability to cast and grind magnifying lenses in the early 17th century triggered a scientific and Intellectual revolution. In 1665 Reobert Hooke used one of the microscopes to examine thin slices of cork, leading him to observe tiny little boxes which he named " Cells".

Later Antoine Van Leeuwenhoek developed better instruments with much more magnifications of almost 300 fold. There superior Microscopes allowed him to discover blood cells, sperm cells and the one-celled organisms present in pond water.

During the 150 years that followed the pioneering observations, many other cell types were discovered using light microscopy, but the Biological significance of cells remained unclear. Then in 1839 the German Biologists Mathias Schleiden and Theodor Schwann integrated the growing body of information on the universal occurrence of cells into one of the first great unifying theories of Biology the "Cell Theory".

1.1.2.3. Tenets:

This theory has two important facts:

1. First, it stated that in spite of the enormous diversity of living organisms on Earth, all organisms are composed of cells.
2. Second, it proposed that all living cells are structurally similar to one another

(The cells Retains a dual existence as a distinct entity and a building block in the construction of organisms).

In 1855 the German physiologist Rudolf Virchow added a third principle to the cell theory, where he concluded that all cells arise from the division of preexisting cells. (An idea summarized in Virchow's famous Latin phase. "Omnis cellula e cellula" which means that all cells arise only from pre-existing cells".

The Modern Tenets of the Cell Theory Include:

1. All known living things are made up of cells.

2. The cell is the structural and functional unit of all living things.

3. All cells come from pre existing cells by division.
   (Spontaneous Generation does not occur)
4. Cells contain hereditary information which is passed from cell to cell during cell Division.

5. All cells are basically the same in chemical composition

6. All energy flow (Metabolism and Biochemistry) of life occurs with in cells.

1.1.2.4. Basic Properties of Cells:

Just as plants and animals are alive so too are cells. Life is the most basic property of cells and cells are the smallest units to exhibit this property. The exploration of cells begins with the examination of few of the most fundamental properties of the cells.

1) Cells are Highly Complex and Organized:

Complexity is a property that is evident but difficult to describe. It is seen in terms of order and consistency. The more complex the structure is, greater the number of parts that must be in their proper place and less tolerance of errors in the nature and interactions of the parts, and the more regulation or control that must be existed to maintain the system. Each type of cell has a consistent appearance in the electron microscope that is; its organ cells have a particular shape and location, from one individual of species to another. Similarly, each type of organelle has a consistent composition of Macromolecules, which are arranged in a predictable pattern.

2) Cells Possess a Genetic Program and the means to use it:

Organisms are building according to information encoded in a collection of genes. The human genetic program contains enough information, if converted to words, can fill millions of pages of text. Remarkably, this vast amount of information is packed into a set of chromosomes that occupy the space of a cell nucleus –thousands of times smaller than the dot on this i. Genes constitute the blue prints for constructing cellular structure, directions for running cellular activities and the program for making more of themselves.

3) Cells are Capable of producing more of themselves:

Cells reproduce by division, a process in which the contents of a “mother” cell are distributed into two “daughter” cells. Prior to division the genetic material is faithfully duplicated and each daughter cell receives a complete and equal share of genetic information.

4) Cells Acquire and Utilize Energy:

Developing and maintaining complexity requires the constant input of energy. Virtually all of the energy required by life on Earth arrives in the form of electromagnetic radiation from the Sun. The energy of light is trapped by light absorbing pigments present in the membranes of photosynthetic cells.

Diagram: Typical Eukaryotic cell structure from Cell Molecular Biology by Gerald Karp.

Light energy is converted by photosynthesis into chemical energy that is stored in energy – rich Carbohydrates such as sucrose or starch. The energy trapped in these molecules during photosynthesis provides the fuel that runs activities of nearly all the organisms on the Earth.

5) Cells Carry out a Variety of Chemical reactions:

Cells function like miniaturized chemical plants. Even the simplest bacterial cell is capable of hundreds of different chemical transformation, none of which occurs at any significant rate in the inanimate world. Virtually all chemical changes that take place in cells require enzymes – molecules that greatly increase the rate at which a chemical reaction occurs. The sum total of the chemical reaction that occurs in a cell represents that cells metabolism.

6) Cells Engage in Numerous Mechanical Activities:

Cells are sites of bustling activity. Materials are transported from place to place, structures are assembled and then rapidly dissembled and in many cases, the entire cell moves itself from one site to another. These types of activities are based on dynamic, mechanical changes with in cells most of which are initiated by chances in shape of certain 'motor' proteins.

7) Cells are able to respond to stimuli:

Some cells respond to stimuli in obvious ways, a single celled protist for example moves way from an object in its path or moves toward a source of nutrients. Cells with in a multicellular plant or animal respond to stimuli less obviously, but they respond none the less. Most cells are covered with receptors that interact with substance in the environment in highly specific ways. Cells possess receptors to hormones, growth factors, extra cellular materials, as well as to substances on the surfaces of other cells. Cells may respond to specific stimuli by altering their metabolic activities, preparing for cell division, moving from one place to another or even committing suicide.

8) Cells are Capable of Self Regulation:

In addition to requiring energy, maintaining a complex ordered state requires constant regulation. As in the body as a whole, many different control mechanisms operate with in each living cell. The importance of cells regulatory mechanisms becomes most evident when they break down. For example, failure of a cell to correct a mistake when it duplicates its DNA may result in a difilitating mutation, or a break down in a cells growth control can transform the cell into a cancer cell with the capability of destroying the entire organisms.

1.1.2.5. Cells Perform Four Basic Functions:

Although the cell theory was originally based on observation that different kinds of cells resemble one another when viewed microscopically the functional similarities between cells have turned out to be even more pronounced than their structural similarities. It the properties of many different cell types were examined its becomes apparent that cells share the following functional characteristics.

1) Cells maintain a selective barrier called the Plasma Membrane, which separates the inside of the cell from the external environment. By regulating the passage of materials into and out of cells the plasma membrane ensures that optimum conditions for living processes prevail within the cell interior. Membrane barriers are also employed to subdivide the cell into multiple compartments specialised for different activities.

2) Cells utilise genetic information to guide the synthesis of most of the cells components. This genetic information is stored in molecules of DNA and is duplicated prior to cell division so that each newly formed cell inherits a complete set of genetic instructions.

3) Cells contain Catalysts called enzymes which speed up chemical reactions involved in the synthesis and breakdown of organic molecules. The sum of all these reaction is referred to as metabolism. Metabolism converts food stuffs into molecules that are needed by the cell, breaks down molecules that are no longer required, and traps energy in useful chemical forms. Which in turn provide the power for energy requiring activities.

4) Cells almost always exhibit some type of motility mechanisms for moving components from one location to another with in the cell are virtually universal. In many cell types such mechanisms also permit movement of the cell as a whole.

   The preceding functions are performed by specialised structures with in the cell called as the organelles.

1.1.2.6. Two Fundamentally Different Classes of Cells:

With the availability of electron microscope, biologists were able to examine the internal structure of a wide variety of cells. It came apparent from these studies that there are two basic classes of cells.

1) Prokaryotic

2) Eukaryotic

Distinguished by their size and the types of internal structures or organelles they contain.

The structurally simpler, prokaryotic cells are found only among bacteria and conversely all bacteria consists of prokaryotic cells.

All other types of organisms, protists, fungi, plants and animals consists of structurally more complex Eukaryotic cells.

Characteristics that Distinguish Prokaryotic and Eukaryotic Cells:

The following brief comparison between prokaryotic and eukaryotic cells reveals many basic differences, as well as similarities. The similarities and differences between the two types of cells are tabulated in table 1.

The similarities reflect the fact that Eukaryotic cells almost certainly evolved from Prokaryotic ancestors. Because of their common ancestry, both types of cell share an identical genetic language, a common set of metabolic pathways, and many common structural features.

Internally Eukaryotic cells are much more complex both structurally and functionally than Prokaryotic cells.

TABLE – 1

A Comparison of Prokaryotic and Eukaryotic Cells:

Features held in common by the two types of cells:

1. Plasma membrane of similar constructions.

2. Genetic information encoded in DNA using identical genetic code.

3. Similar mechanisms for transcription and translation of genetic information, including similar ribosome's (functionally).

4. Shared metabolic pathways (eg. Glycolysis and TCA cycle)

5. Similar apparatus for conservation of chemical energy as ATP (located in the plasma membrane of Prokaryotes and the mitochondria membrane of Eukaryotes).

6. Similar mechanism of photosynthesis (between cyano bacteria and green plants).

7. Similar mechanism for synthesizing and inserting membrane proteins.

8. Proteasoms (protein digesting structures) of similar construction (between Archea bacteria and Eukaryotes).

TABLE 1 (A)

Features of Eukaryotic Cell not found in Prokaryotes:

1. Division of cells into mules and cytoplasm, separate by a nuclear envelope containing complex pore structures.

2. Complex Chromosomes composed of DNA & Associated proteins that are capable of compacting into mitotic structures.

3. Complex membranes Cytoplasmic organelles (includes Endoplasmic Reticulum,
Golgi complex, Lysosomes, Endosomes, Peroisomes & Glyoxysomes)

4. Specialised cytoplasmic organelles for Aerobic respiration (Mitochondria) and photosynthesis(Chloroplasts).

5. Complex Cytoskeletal system (including Micro filaments Intermediate filaments and
Micro tubules).

7. Complex Flagella and Cilia.

8. Capable of ingesting fluid and particulate material by enclosure with in plasma membrane vesicles (endocytosis and phagocytosis).

9. Cellulose containing cell walls (in plants).

10. Cell division utilising a microtubule, containing Mitotic spindle that separates chromosomes.

11. Presence of two copies of genes per cell (diploid) one from each parent.
12. Sexual reproduction requiring meiosis and fertilisation.

Summary :

The Summary of the lesson is as follows :

- ➔ Cell is the basic structural & functional unit of life.
- ➔ All organisms are composed of cells.
- ➔ Life is the most basic property of cells and they are highly complex in their arrangement.
- ➔ Cells are basically two types based on this complexity.
- ➔ Prokaryotes.
- ➔ Eukaryotes.

Model questions :

1) What is cells Theory? Give the tenets of Cell theory in detail?

2) Describe the various, properties exhibited by the Cells?

3) Differentiate Prokaryotes from Eukaryotes in detail?

4) Describe the basic functions of a Cell?

Diagrams: Typical structure of a Eukaryotic Cell.

Table : 1) Features held in common between Prokaryotes and Eukaryotes.
1) (A) Features of Eukaryotic Cells not found in Prokaryotes.

References:

1. Cell and Molecular Biology by Gerald Karp.
2. Cell by Bruce Alberts.
3. Cell & Molecular Biology by Baltimore.

Lesson 1.1.3

GENERAL ACCOUNT OF CELL MEMBRANE & FUNCTIONS

Objective

Objective

The structure and function of cells are critically dependent on membranes, which not only separate the interiors from its environment but also define the internal compartments of the Eukaryotic cells, including the nucleus and cytoplasmic organelles. The formation of biological membranes is based on the properties of lipids and all the cell membranes share a common structural organization layers of phospholipids with associated proteins. The membrane proteins are responsible for many specialized functions. Some act as receptors that allow to respond to external signals, some are responsible for selective transport of molecules across the membrane, others participate in electron transport and oxidative phosphorylation. In addition, membrane proteins control the interactions between cells of multicellular organisms. The common structural organization of membranes under lies a variety of biological processes and specialized membrane functions, thus known as biomembranes.

1.1.3.1 General Account of Cell Membrane & Functions

1. Compartamentalisation

Membranes are continuous unbroken sheets and, as such, inevitably enclose compartments. The plasma membrane encloses the contents of the entire cell, while the nuclear and cytoplasmic membranes enclose various internal cellular spaces. Because of compartmentalization, different types of activities proceed with a minimum of outside interference and can be regulated independently.

2. Selective permeable barriers :

Membranes prevent the unrestricted exchange of molecules from one side to the other. At the same time membranes provide the means of communication between the compartments they separate.

3. Transportation

The plasma membrane contains the machinery for physically transporting substances from one side of the membrane to another, often from a region where the solute is present at low concentration into a region where the solute is present at much higher concentration, also allows the cell to accumulate substances such as sugars and amino acids in the cell metabolism.

4. Response to external signals

By the process of signal transduction, plasma membrane plays a critical role in the response of cell to external stimuli with the aid of structures called receptors. The interaction of the plasma membrane receptors with an external ligand may cause the membrane to generate a new signal that stimulates or inhibits internal activities.

5. Intercellular Interaction

Plasma membrane allows the cells to recognize one another, and adhere when appropriate information is available.

6. Energy Transduction

Membranes are intimately involved in the process by which one type of energy is converted to another type (Energy Transduction). The most fundamental, energy transduction occurs during photosynthesis when energy in sun light is absorbed by membrane bound pigments and converted into chemical energy contained in carbohydrates.

1.1.3.2 History of Plasma Membrane Structure

The first information on the chemical nature of the outer boundary layer of a cell was obtained by *E.ouerton* in 1990's. Basing on the principle that non polar solutes dissolve more readily in non polar solvents than in polar solvents, he reasoned that more could be learnt by determine the rate at which different types of chemicals diffuse through that boundary. By placing plant root hairs into solutions of different composition overton discovered that, the more lipid soluble compound, the more rapidly it would enter the root hair cells. He concluded that the dissolving power of the outer boundary of the cell matched that of a fatty oil.

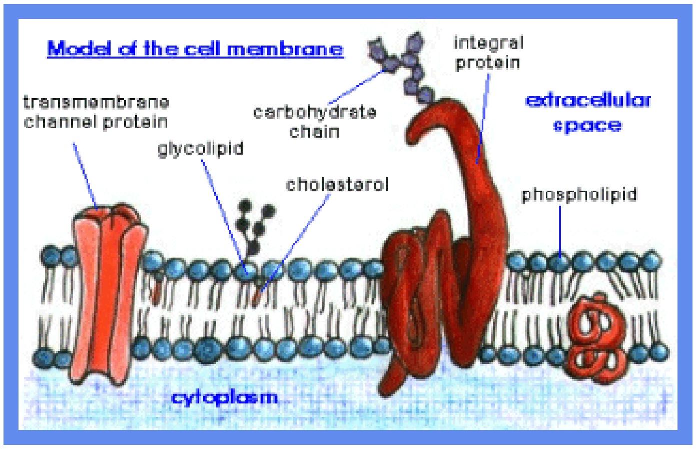
The first proposal that cellular membranes might contain a lipid bilayer was made in 1925 by two Dutch Scientists, E.Gorter and F. Grendel. They extracted the lipid from human red blood cells and measured the amount of the surface area the lipid would cover when spread over the surface of water. Getting an actual ratio of 2:1 (lipid: water), they concluded that the plasmamembrane contained a bimolecular layer of lipids or simple a lipid bilayer. They also suggested that the polar groups of each molecular layer were directed toward the outside of the bilayer. This would be the thermodynamically favoured arrangement because the polar head groups of lipid could interact the surrounding water molecules, while the hydrophobic fattyacyl chains would be protected from the aqueous environment.

In 1920's & 1930's cell physiologists proposed that lipid solubility was not the sole determining factor as to whether or not a substance could penetrate the plasma membrane. Similarly the surface tensions of the membranes were declined to be much lower than those of pure lipid structures and it was shown that the presence of a protein film over an artificial lipid layer greatly lowered its surface tension.

In 1935 Hugh Daveson & James Danielli proposed that the plasmamembrane was composed of a lipid bilayer that was lined on both its inner and outer surface by a layers of globular proteins Later in their revised model, in addition to inner and outer protein layers, the bilayer was also found to have protein lined-pores which could provide the entry and exit of polar solutes in the cell.

Fluid mosaic model

Experiments conducted in late 1960 are led to the proposal of Fluid Mosaic model in 1972 by S. Jonathan singer and Garth Nicolson. In this, attention had been focused on the physical state of the lipid rather than existing in a frozen, immobile bilayer; the lipid molecules are present in a fluid state and can move laterally with in the plane of the membrane. The proteins of the fluid mosaic model occur as a "Mosaic" of discontinuous particles that penetrate deeply into and even completely through the lipid sheet. This model presents cellular membranes as dynamic structures in which the components are mobile and capable of coming together to engage in various types of transient or semi permanent interactions.



1.1.3.3. The chemical composition of membranes

All membranes are lipid protein assemblies in which the components are held together in a thin sheet by non covalent bonds.

The lipid bilayer serves primarily as a structural framework for the membrane and as barriers preventing the indiscriminate movements of water – soluble materials into and out of the cell. The proteins on the other hand carryout specific transport channel functions.

In addition to lipid and protein, membranes also contain carbohydrates. The ratio of lipid to protein varies considerably depending on the type of cellular membrane (Plasma vs Endoplasmic reticulum vs Golgi), depending on the type of organism (Prokaryote Vs Eukaryote Vs plant) and the type of cell (Cartilage Vs Muscle Vs Liver). These differences can be correlated with the particular functions of the membranes. For instance, the inner mitochondrial membrane contains protein carriers of the electron transport chain with relatively less lipid than other membranes, where as the myelin sheath a electrical insulation for the neuron enclose has a thick lipid layer of high electrical resistance with minimum content of protein.

1) Membrane lipids

Membranes contain a wide diversity of lipids all of which are amphipathic (both hydrophilic and hydrophobic). There are three main types of membrane lipids. Phosphoglycerides, sphingolipids and cholesterol.

a) Phosphoglycerides

Most membrane lipids contain a phosphate group, which makes them phospholipids. As most membrane phospholipids are built on a glycerol backbone, they are called Phosphoglycerides. Unlike triglycerides membrane glycerides are diglycerides only two of the hydroxyl groups of glycerol are esterified to fatty acids; the third is esterified to a phosphate group. All the membrane Phosphoglycerides have an additional group linked to the phosphate, most commonly either choline (forming phosphotidyl choline), or exthanolamine (phosphotidyl ethanolamine) or serine (phosphotidyl serine) or Inositol (phosphotidyl inositol). These small and hydrophilic groups along with the charged phophate to which it is attached forms a highly water soluble domain at one end of the molecule called the head group. In contrast the fattyacyl chains are long, unbranched, hydrophobic hydrocarbons. A membrane fatty acid may be fully saturated. (i.e lack double bonds), Monounsaturated (i.e. possess one double bond) or polyunsaturated (possess more than one double bond). Phosphoglycerides often contain one unsaturated and one saturated fattyacyl chain.

b) Sphingolipids

A less abundant class of membrane lipids called sphingolipids, are derivatives of sphingosine an amino alcohol that contains a long hydrocarbon chain, sphingo lipids contain a sphingosine linked to a fattyacid by its aminogroup. This molecule is a ceramide. The various sphingosine based lipids have additional groups esterified to the terminal alcohol of the sphingosine moiety. If the substitution is phosphotidylcholine, the molecule is sphingomyelin, which is the only phospholipid of the membrane that is not built with a glycerol backbone. If the molecule is a carbohydrate, the molecule is glycolipid. If the carbohydrate is a simple sugar, the glycolipid is called the cerebroside. If it is a oligosaccharide, the glycolipid is called a Ganglioside. The nervous system is particularly rich in Glycolipids. Glycolipids playa role in certain infectious diseases; the cholera toxin and the influenza viruses both enter their target cell by first binding to cell-surface gangliosides.

c) Cholesterol

Another lipid component of certain membranes is the sterol cholesterol. It is absent in the plasmamembranes of most plant and all bacterial cells. Cholesterol is smaller than the other lipids of the membrane and less amphipathic. They are oriented with their hydrophilic hydroxyl groups towards the membrane surface and their hydrophobic tails embedded in the lipid bilayer.

2) Membrane carbohydrates

The plasma membranes of eukaryotic cells contain carbohydrates that are covalently linked to both lipid and protein components. Depending on the species and cell type, the carbohydrate content of the plasma membrane ranges between 2-10% by weight. The plasma membrane of redblood cells contains approximately 52% protein, 40% lipid and 8% carbohydrate of the 8% ; 7% is covalently linked to lipids to form glycolipids and the remaining 93% is covalently linked to proteins to form glycoproteins. All of the carbohydrate of the plasmamembrane faces outward into the extracellular space. The carbohydrate of internal cellular membranes also faces away from the cytosol.

The carbohydrte of glycoproteins is present as short, branched oligosaccharides, typically having fewer than 15 sugars per chain. These carbohydrate projections are thought to play a role in mediating the interaction of a cell with other cells as well as its non living environment. The carbohydrates of the glycolipids of the red blood cell plasmamembrane determine whether a persons blood type is A, B, AB or O. The ABO determinants are short branched oligosacharide chains.

In contrast to most high-molecular weight carbohydrates (such as glycogen, starch or cellulose), which are polymers of a single sugar, the oligosaccharides attached to membrane proteins and lipids can display considerable variability in composition and structure. They provide specificity in their interactions with one another and with other types of molecules.

3) Membrane proteins

Depending on the cell type and the particular organelle with in that cell, a membrane may contain from a dozen to more than 50 different proteins. These proteins are not randomly arranged with in the membrane, but each is located and oriented in a particular position relative to the cytoplasm. All the membrane proteins are asymmetrically situated so that the properties of one surface of a membrane are very different from those of the other surface. This property is referred to as membrane “sidedness”. Membrane proteins can be grouped into three distinct classes distinguished by the intimacy of their relationship to the lipid bilayer. These are

a). Integral proteins

These penetrate into the lipid bilayer. In fact, nearly all integral proteins are transmembrane proteins, that i.e. is, they pass entirely through the lipid bilayer and thus have domains that protrude from both the extra cellular and cytoplasmic sides of the membrane.

b) Peripheral proteins

These are located entirely outside of the lipid bilayer, on the cytoplasmic surface, yet are associated with the surface of the membrane by non covalent bonds.

c) Lipid anchored proteins

These are located – out side the lipid bilayer, on either the extracellular or cytoplasmic surface, but are covalently linked to a lipid molecule that is situated with in the bilayer.

a) Integral membrane proteins

These are also amphipathic having both hydrophilic and hydrophobic portions. The regions of the protein having non-polar surfaces are embedded with in the lipid bilayer where they form hydrophobic interactions with the fatty acyl chains, sealing the protein into the lipid “wall” of the membrane preserving the permeability barriers of the membrane and brings the proteins into direct contact with surrounding lipid molecules, which give rise to many of the membrane dynamic properties. The other parts of the protein composed of largely ionic and polar amino acids protrude beyond the edge of the bilayer on one or both sides of the hydrophilic regions serve as the parts of an integral protein that interact with water soluble substances (ions, low molecular weight substances, hormones and other proteins) at the membrane surface or within a central channel.

As a result of their hydrophobic surfaces, integral membrane proteins are difficult to isolate in a soluble form. Removal of these proteins from the membrane normally requires the use of a detergent. Such as the ionic (charged) detergent SDS (Sodium dodecyl sulphate) which denatures proteins or the nonionic (uncharged) detergent, Triton x-100 (which generally does not alter proteins tertiary structure.

Detergents are amphipathic, being composed of a polar end and a non polar hydrocarbon chain. As a consequence of their structure, detergents can substitute for phospholipids in stabilizing integral proteins while rendering them soluble in aqueous solution. Once the proteins have been solubilised by the detergents, various analytical procedures can be carried out to determine the proteins amino acid composition molecular weight, amino acid sequence and so forth.

b) Peripheral membrane proteins

Peripheral proteins are associated with the membrane by weak electrostatic bonds either to the hydrophilic head groups of the lipids or to the hydrophilic portions of integral proteins protruding from the bilayer. Peripheral proteins can usually be solubilised by extraction with aqueous solutions of high salt or alkaline pH.

The best studied peripheral proteins (most notably members of the spectrin family) are located on the inner surface of the plasma membrane, where they form a fibrillar network that acts as a membrane skeleton. These proteins provide mechanical support for the membrane and function as an anchor for integral membrane proteins. Other peripheral proteins on the inner membrane surface function as enzymes or factors that transmit transmembrane signals.

c) Lipid anchored membrane proteins

Two types of lipid anchored membrane proteins are distinguished by the types of lipid anchors and the membrane surface on which they are exposed.

A variety of proteins present on the external face of the plasma membrane are bound to the membrane by a short oligosaccharide linked to a molecule of Glyco phosphatidyl inositol (GPI)that was embedded in the outer leaflet of the lipid bilayer. A rare type of Anemia, paroxysmal moctugnal haemoglobinuria, results from a deficiency in GPI synthesis that makes red blood cells susceptible to lysis. Another group of proteins present on the cytoplasmic side of the plasma membrane is anchored to the membrane by long hydrocarbon chains embedded in the inner leaflet of the lipid bilayer. At least two proteins associated with the plasma membrane in this way (Sre and Ras) have been implicated in the transformation of a normal cell to a malignant state.

1.1.3.4 Mode of transport across plasma membrane

The plasma membrane acts as a semipermeable barrier between the cell and the extracellular environment. The selective permeability of the plasma membrane allows the cell to maintain a constant internal environment. In consequence, in all types of cells there exists a difference in ionic concentration with the extracellular space. Transport across the membrane may be passive (or) active. It may occur via the phospholipid bilayer (or) by the help of specific integral membrane proteins, called permeases (or) transport proteins.

Passive transport

It is a type of diffusion in which an ion (or) molecule crossing a membrane moves down its electrochemical (or) concentration gradient. No metabolic energy is consumed in passive transport. Passive transport is of following three types:

The concentration gradient

The difference in the levels of the two concentrations is called the concentration gradient. The two levels are greater concentration and lesser concentration.

1. Osmosis

The plasma membrane is permeable to water molecules. The to and fro movement of water molecules through the plasma membrane occurs due to the differences in the concentration of the solution either sides. "The process by which the water molecules pass through a membrane from a region of higher water concentration to the region of lower water concentration is known as osmosis (Gr: osmos = pushing)". The process in which water molecules enter into the cell is known as endosmosis and the process which involves the exit of water molecules from the cell is known as exosmosis. In plant cells due to excessive exosmosis, the cytoplasm along with the plasma membrane shrinks away from the cell-wall. This is known as plasmolysis. (Gr. Plasma = molded; lysis=loosing). Due to endosmosis (or) exosmosis, the water molecules come in (or) go out of the cell. The amount of the water inside the cell causes a pressure which is caused by the osmosis is known as osmotic pressure. The plasma membrane maintains a balance between the osmotic pressure of the intra-cellular and inter-cellular fluids.

2. Simple diffusion

Transport of metabolites across the membrane along the concentration gradient and without the use of a carrier molecule is called simple diffusion. Movement of substances takes place from a high concentration to a low concentration region and the concentration gradient disappears as diffusion occurs. Simple diffusion does not involve any stereo-specificity (i.e both L and D isomers move across at equal rates), and is a slow process. This transport is not believed to be an important mechanism for transport across cell membranes.

In the Danielli-Davson model it was assumed that passage of substances took place through small (7A^O) rigid protein-lined pores in membrane. According to Singer-Nicolson fluid mosaic model the pores may not be stable, but may be constantly appearing and disappearing (statistical pore concept). They are believed to be formed by the appearance of gaps in the highly fluid lipid bilayer because of random movement of membrane phospholipids. Small polar molecules could cross the membrane through the gaps (pores) which arise in a random manner and are transitory.

3. Facilitated diffusion

It resembles simple diffusion in that it does not require energy and takes place along the concentration gradient. It differs in certain respects. Firstly, the process is stereospecific, i.e only one of the two possible isomers, L and D is transported. Secondly, it shows saturation kinetics i.e. increase in the concentration of the substance to be transported results in an increase in the rate of transfer up to an asymptotic value. Thirdly, a carrier is required for transport across the membrane.

Experimental evidence indicates that the proteins are highly selective as carriers. Carrier proteins specific for individual sugars and amino acids, phosphate, Ca^{++}, Na^{+} and K^{+} have been isolated. The carrier protein molecules move to and fro across the membrane by thermal diffusion. The metabolite binds to the carrier protein at the outer surface of the membrane to form a carrier-metabolite complex. This diffuses along the concentration gradient, i.e from high concentration to low concentration regions. The metabolite is set free at the inner surface of the membrane because of the relatively low concentration of metabolite on the inner side of the membrane. Transportation of metabolite continues as long as there is a concentration gradient.

Example : The entry of glucose into erythrocytes is an example of facilitated transport.

Active transport

Active transport uses specific transport proteins, called pumps, which use metabolic energy (ATP) to move ions (or) molecules against their concentration gradient. For example in both vertebrates and invertebrates, the concentration of sodium ion is about 10-20 times higher in the blood than within the cell. The concentration of the potassium is 20-40 times higher inside the cell. Such a low sodium concentration inside the cell is maintained by the sodium-potassium pump. There are different types of ions (or) molecules such as calcium pump, proton pump etc.

Examples :

1. Na^{+} - K^{+} - ATPase
2. Calcium ATPase

3. Proton pump

Let, us consider Na^+ - K^+ - ATPase;

Na^+ - K^+ - ATPase

It is an ion pump (or) cation exchange pump which is driven by energy of one ATP molecule to export three Na^+ ions outside the cell in exchange of the import of two K^+ ions inside the cell. Na^+ - K^+ - ATPase is a transmembrane protein which is a dimer having two sub units: one smaller unit, is a glycoprotein of 50,000 daltons M.wt; having a unknown function and another larger unit having 1,20,000 daltons M.Wt. The larger subunit of Na^+-K^+ - ATPase performs the actual function of cation transport. It has three sites on its extra cytoplasmic surface: two sites for K^+ ions and one site for inhibitor ouabain on it's cytosolic side, the larger subunit contains three sites for three Na^+ ions and also has one catalytic site for a ATP molecule. It is believed that the hydrolysis of one ATP molecule some how drives confirmational changes in the Na^+ - K^+ - ATPase that allows the pump to transport three Na^+ ions out and two K^+ ions inside the cell.

In the similar process, the above quoted examples too carry the same mechanism of transport based on electrochemical gradient.

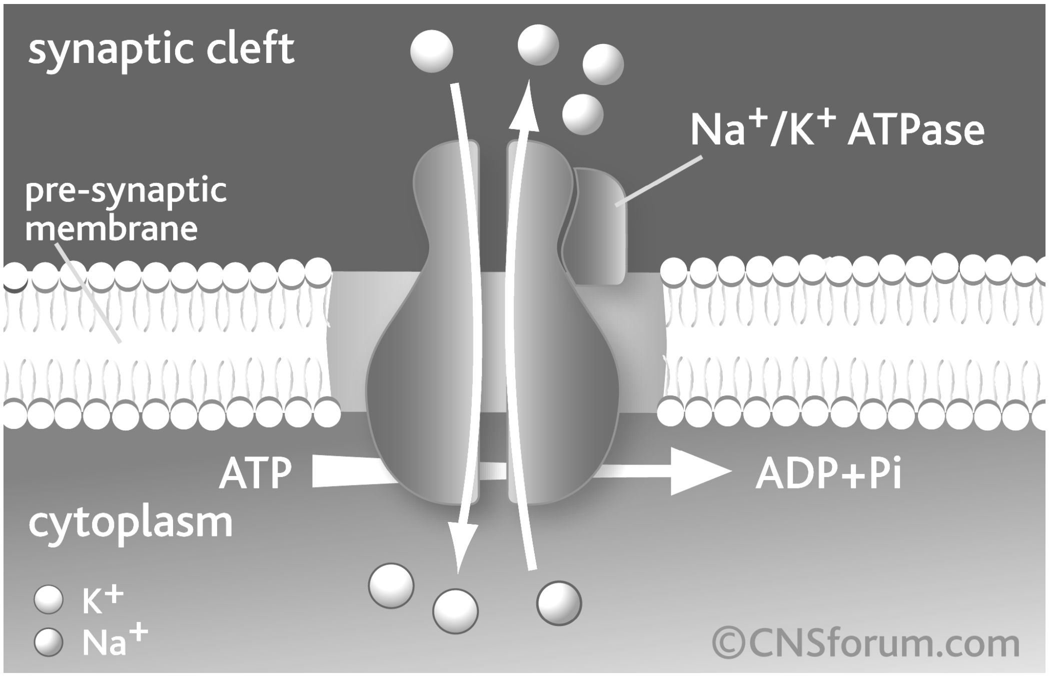


Fig. There are various types of transport events; such as:

1. Uniport

The proteins that transport a single molecule in a unidirectional fashion across the membrane are called uniporters and the type of transport is called uniport.

2. Symport

The proteins that transport a substance along with another substance in the same direction are called symporters and the type of transport is called symport.

3. Antiport

The proteins that transport a substance across the membrane in one direction while at the same time transport a substance in the opposite direction are called antiproters and the type of transport is called antiport.

Bulk transport

Cells routinely import and export large molecules across the plasma membrane. Macromolecules are secreted out from the cell by exocytosis and are ingested into the cell from outside through phagocytosis and endocytosis.

1. Exocytosis

It is also called emeiocytosis and cell vomiting. In all eukaryotic cells, secretory vesilces are continually carrying new plasma membrane and cellular secretions such as proteins, lipids and carbohydrates from the Golgi apparatus to the plasma membrane (or) to cell exterior by the process of exocytosis. During exocytosis the vesicle membrane is incorporated into the plasma membrane. The amount of secretory vesicle membrane that is temperorily added to the plasma membrane can be enormous.

2. Phagocytosis

Some times the large –sized solid food (or) foreign particles are taken in by the cell through the plasma membrane. The process of ingestion of large – sized solid substances by the cell is known as phagocytosis.

The process of phagocytosis involves the process of adsorption, formation of phagosome and phagolysome and the process of egestion.

Types of phagocytosis

The process in which phagocytosis is expressed is of 2 types:

1. Colloidal

The process in which plasma membrane ingests smaller colloidal particles is known as colloidopexy (or) Ultraphagocytosis. Eg: leucocytes and the macrophagic cells of mammals.

When the cell ingests colloidal chromogen particles phagocytically, this process is known as chromopexy. Eg: some mesoblastic cells.

2. Endocytosis

In endocytosis, small regions of plasma membrane fold inwards (or) invaginate, until it has formed new intra cellular membrane limited vesicles. In eukaryotes, endocytosis is of two types.

1. Pinocytosis

It is non-specific uptake of small droplets of extracellular fluid by endocytic vesicles (or) pinosomes. The pinocytosis which occurs at sub-microscopic level is known as micropinocytosis. Pinocytosis (Gr: pinein = to drink; `cell drinking').

2. Receptor mediated endocytosis

In this type of endocytosis, a specific receptor on the surface of the plasma membrane "recognizes" an extracellular macromolecule and binds with it. The substance bound with the receptor is called the ligand. Examples of ligands may include viruses, small proteins etc.. The region of plasma membrane containing the receptor ligand complex undergoes endocytosis.

Therefore, all cells, both prokaryotic and eukaryotic are surrounded by a plasma membrane, which defines the boundary of the cell and separate its internal content from the environment. The passage of ions and most biological molecules are therefore mediated by proteins responsible for the selective traffic of molecules into and out of the cell. Other proteins serve as sensors through which the cell receives signals from the environment. The plasma membrane thus plays a dual role: (1) It isolates the cytoplasm and (2) Mediates interactions between the cell and environment.

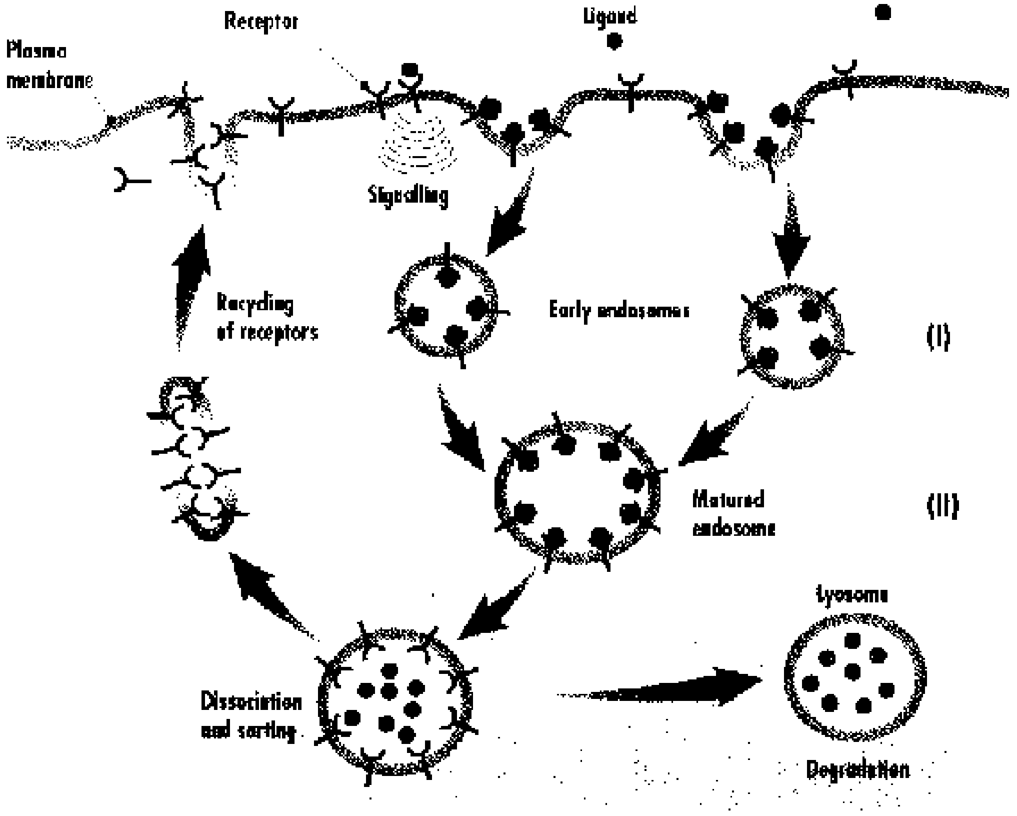


Fig.

1.1.3.5. The importance of membrane fluidity

Membrane fluidity provides a perfect compromise between a rigid, ordered structure in which mobility would be absent and a completely fluid, nonviscous liquid in which the components of the membrane could not be oriented and structural organization and mechanical support would be lacking. In addition, fluidity allows for interactions to take place within the membrane. For example, membrane-fluidity makes it possible for clusters of membrane proteins to assemble at particular sites with in the membrane and form specialized structures, such as intercellular junctions, light capturing complexes and synapses. Because of membrane fluidity molecules that interact can come together, carry out the necessary reaction and move apart.

1.1.3.6. Functions

The plasma membrane has many common functions in all the cells which are essential. These include

→ Transporting nutrients into and metabolic wastes out of the cell.

→ Preventing unwanted materials in the extra cellular matrix from entering the cell.

→ Preventing the loss of needed metabolites and maintaining the proper ionic composition (pH ≅ 7.2) and osmotic pressure of the cytosol.

→ To carryout these functions, the plasma membrane contains specific transport proteins that permit the passage of certain small molecules but not others. Several of these proteins use the energy released from the hydrolysis of ATP to pump ions and other molecules into and out of the cell against the concentration gradient. Small charged molecules such as ATP and amino acids can diffuse freely within the cytosol but are restricted in their ability to leave or enter across the plasma membrane.

→ Specialised areas of the plasma membrane contain proteins and glycolipids that form specific contacts and junctions between cells to strengthen tissues and to allow the exchange of metabolites between cells.

→ Other proteins in the plasma membrane act as anchoring points for many of the cytoskeletal fibers that permeate the cytosol, imparting shape and strength to the cell.

→ The plasma membrane of many types of eukaryotes also contains receptor proteins that bind specific signaling molecules (Eg. Hosmones, growth factors, neurotransmiters) leading to various cellular responses.

→ Unlike animal cells, plant cells are surrounded by a cell wall and lack the extracellular matrix found in animal tissues.

→ The walls are primarily built of cellulose, a rod like polysaccharide formed from β (1→4) linked glucose monomers.

→ In plants, the cell wall, which is built mainly of cellulose is the major determinant of cell shape and imparts rigidity to cells.

→ Animal cells which lack a wall are surrounded by an extracellular matrix consisting of collagen, glycoproteis and other components that give strength and rigidity to tissues and organs.

Summary

Membranes are continuous unbroken sheets. They act as selectively permeable barriers and maintain the cell environment. To explain the structure of membrane S. Jonathan Singer and Garth Nicolson proposed fluid mosaic model. Membranes contain lipids carbohydrates and proteins. The proportion of the three components vary from one membrane to the other. The plasma membrane has many common essential functions in all the cells like transportation, response to external signals, intracellular interaction and energy transduction etc.

Model questions

1. Describe the functions of membranes ?
2. Give a detailed account on the chemical composition of the membranes.
3. Describe the mode of transport across plasma membrane.

Reference books

1. Cell & Molecular Biology by Geralad Karp.
2. Molecular Cell Biology, Ladish et al.

K.Haritha

Lesson 1.1.4

GENERAL ACCOUNT OF CELL ORGANELLES & FUNCTION

Contents

Objective

- The objective of this lesson is to know the cell organelles and their functions.

1.1.4.1 Introduction

The body of all living organisms (bacteria, blue green algae, plants and animals) except viruses has cellular organization and many contain one or more cells. The organisms with only one cell in their body are called unicellular organisms (Eg. Bacteria, Blue green algae, Protozoa etc). The organisms having many cells in their body are called multicellular organisms (Eg. Most plants and Animals). Any cellular organism must contain only one type of cell from the following types.

A. Prokaryotic cells

B. Eukaryotic cells

The prokaryotic (pro=primitive; karyon = nucleus (Greek)) are small, simple and most primitive organisms. They are probably the first to come into existence. The eukaryotic (Eu=true or well; karyon = nucleus) cells have evolved from the prokaryotic cells, that are highly evolved.

1.1.4.2. Prokaryotic cell organization

The prokaryotic cells are the most primitive cells from the morphological point of view. A prokaryotic cell is essentially a one envelope system organized in depth. It consists of central nuclear components (i.e. DNA, RNA & nuclear proteins) surrounded by cytoplasmic ground substance, with the whole enveloped by plasma membrane. Neither the nuclear apparatus nor the respiratory enzyme system are separately enclosed by membranes, although the inner surface of the plasmamembrane itself may serve for enzyme attachment. The cytoplasm of the prokaryotic cell lacks in well defined cytoplasmic organelles such as Endoplasmic reticulum, Golgi Apparatus, mitochondria, centrioles etc. In the nutshell, the prokaryotic cells are distinguished from the eukaryotic cells primarily on the basis of what they lack i.e. they also don’t contain nucleoli, cytoskeleton and centrioles & basal bodies.

1.1.4.3. Organelles of the eukaryotic cell

The various techniques available have led to an appreciation of the highly organized internal structure of eukaryotic cells, marked by the presence of many different organelles.

Unique proteins in the interior and membranes of each type of organelle largely determine its specific functional characteristics.

1. Lysosomes

Lysosomes are acidic organelles that contain a battery of degradative enzymes. These provide an excellent example of the ability of intra cellular membranes to form closed compartments in which the composition of the lumen differs substantially from that of the surrounding cytosol.

→ Found in the animal cells, Lysosomes are bounded by a single membrane and are responsible for degrading certain components that have become obsolete for the cell or organism.

In some cases materials taken into a cell by endocytosis or phagocytosis are degraded by lysosomes. Endocytosis refers to the process by which extra cellular materials are taken up by invagination of a segment of the plasmamembrane to form a small membrane bounded vesicle (Endosome).

In phagocytosis relatively large particles are enveloped by the plasmamembrane and digested.

→ Lysosomes contain a group of enzymes that degrade polymers into their monomeric subunits.

For example: nucleases degrade RNA & DNA into their mono nucleotide building blocks.

→ Proteases degrade a variety of proteins and peptides.

→ Phosphatases remove phosphate groups from mono nucleotide phospholipids and other compounds;

All the lysosomal enzymes work most efficiently at acid pH values and collectively are termed acid hydrolases. A hydrogen ion pump and a Cl^- channel protein in the lysosomal membrane maintain the pH of the interior at ~ 4.8. Together they transport the HCl. The acid pH helps to denature proteins, making them accessible to the action of the lysosomal hydrolases. Which themselves are resistant to acid denaturation.

Lysosomal enzymes are little active or poorly active at the neutral pH values of the cell. Thus if a lysosome releases its contents into the cytosol, where the pH is between 7.0 & 7.3, little degradation of cytosolic components takes place.

Lysosomes vary in size and shape and several hundred may be present in a typical animal cell. They function at sites where various materials to be degraded collectively.

Primary lysosomes are roughly spherical and do not contain obvious particulate or membrane debris.

Secondary lysosomes which are larger and irregularly shaped, appear to result from the fusion of primary lysosomes with other membrane organelles; they contain particles or membranes in the process of being digested. The process by which an aged organelle is degraded in a lysosome is called autophagy. (eating onself).

Taysachs disease is caused by a defect in one enzyme catalyzing a step in the lysosomal breakdown of certain glycolipids called gangliosides, which are abundant in nerve cells protection of with devastating consequences. The symptoms of this inherited disease usually are evident before the age one. Affected children commonly become demented and blind by age 2, and die before the age of three. Nerve cells from such children are greatly enlarged with swollen lipid-filled lysosomes.

Vacuoles

Most plant cells contain at least one membrane limited internal vacuole. The number and size of vacuoles depend on both the type of cells and its stage of development. Plant cells store water, ions and nutrients such as sucrose, and amino acids with in these vacuoles. Vacuoles also act as receptacles for waste products and excess salts taken up by the plants and may function similarly to lysosomes in animal cells.

Like lysosomes, vaculoses have an acidic pH, maintained by a proton pump and a Cl^- channel protein and contain a battery of degradative enzymes. Similar vacuoles are found in green algae and many microorganisms such as yeast.

2. Peroxisomes

All animal cells (except erythrocytes) and many plant cells contain peroxisomes, a class of small organelles (0.2 - 1μm) in diameter bounded by a single membrane. Peroisomes contain several oxidase – enzymes that use molecular oxygen to oxidize organic substances, in the process forming hydrogen peroxide (H_2O_2) a corrosive substance. Peroxisomes also contain copious amounts of enzyme catalase which degrades hydrogen peroxide to yield water and oxygen.

$$2H_2O_2 \xrightarrow{\text{catalase}} 2H_2O + O_2$$

→ Glyoxisomes are similar organelles found in plant seeds that oxidize stored lipids as a source of carbon and energy for growth. They contain many of the same types of enzymes as peroisomes as well as additional ones used to convert fatty acids to glucose precursors.

In contrast to oxidation of fatty acids in mitochondria, which produces CO_2 and is coupled to generation of ATP, peroxisomal oxidation of fatty acids yields acetyl groups and is not linked to ATP formation. The energy released during peroxisomal oxidation is converted into heat and the acetyl groups are transported into the cytosol, where they are used in the synthesis of cholesterol and other metabolites. In most eukaryotes, the peroxisomes are the principal organelles in which fatty acids are oxidized, there by generating precursors for important biosynthetic pathways. Particularly in liver and kidney cells various toxic molecules that enter the blood stream also are degraded in peroxisomes producing harmless products.

In the human genetic disease x-linked adreno leuko dystrophy (ADL), peroxisomal oxidation of very long chain fatty acids is defective. Individuals with severe form of ADL are unaffected until mid-childhood, then severe neurological disorders appear, followed by death with in a few years.

3. Mitochondria

Mitochondria are the principal sites of ATP production in aerobic cells. Most eukaryotic cells contain many mitochondria, which occupy up to 25% of the volume of the cytoplasm. Mitochondria contain two very different membranes, an outer one and an inner one, separated by the intermembrane space. The outer membrane, composed of about half lipid and half protein, contains proteins that render the membrane permeable to molecules having molecular weights as high as 10,000 D.

The surface area of the innermembrane is greatly increased by a large number of infoldings, or cristae that protrude into the matrix or central space.

In non photosynthetic cells, the principal fuels for ATP synthesis are fatty acids and glucose. The complete aerobic degradation of glucose to CO_2 and H_2O is coupled to synthesis of as many as 36 molecules. In eukaryotic cells, the initial stages of glucose degradation occur in cytosol, where two ATP molecules per glucose molecule are generated. The terminal stages, including those involving phosphorylation coupled to final oxidation by oxygen are carried out by enzymes in the mitochondrial matrix and cristae. As many as 34 ATP molecules per glucose molecule are generated in mitochondria, although this value can vary because much of the energy released in mitochondrial oxidation can be used for other purposes (eg; heat generation and transport of molecules into or out of the mitochondria), making less energy available for ATP synthesis. Similarly, all the ATP formed during the oxidation of fattyacids to CO_2 is generated in the mitochondrion. Thus the mitochondrion can be regarded as the “power plant” of the cell.

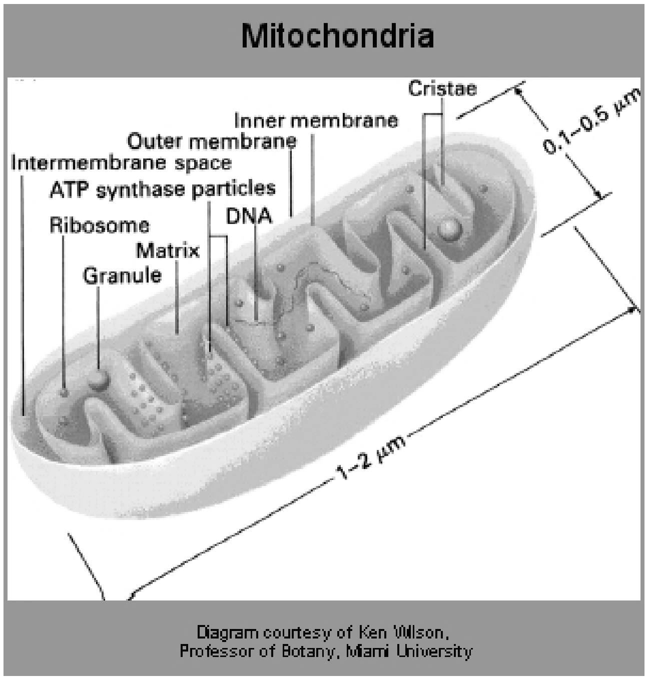

Mitochondria

Diagram courtesy of Ken Wilson,
Professor of Botany, Miami University

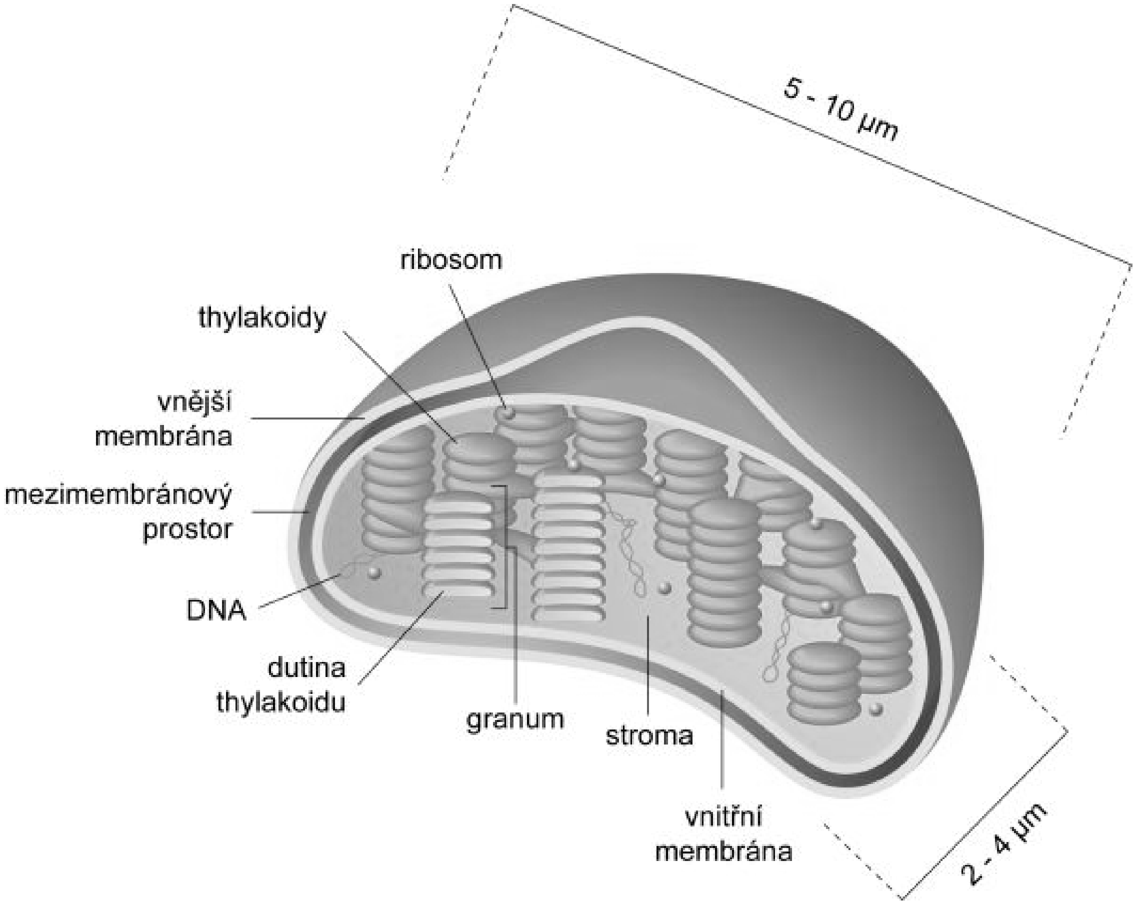
5. Chloroplast

Chloroplasts are the sites of photosynthesis, containing three membrane-limited compartments. Except for vacuoles, chloroplasts are the largest and most characteristic organelles in the cells of plants and green algae. They can be as long as 10µm and are typically 0.5-2 µm thick, but they vary in size and shape in different cells. Like the mitochondrion, the chloroplast is surrounded by an outer and an inner membrane. Chloroplasts also contain an extensive internal system of inter connected membrane – limited sacs called thylakoids, which are flattened to form disks; there often are grouped in stacks called Grana and embedded in matrix, the stroma. The thylakoid membranes contain green pigment (chlorophyll) and other pigments and enzymes that absorb light and generate ATP during photosynthesis. Part of this ATP is used by enzymes located in the stroma to convert CO_2 into three-carbon intermediates; there are then exported to the cytosol and converted into sugars.

The molecular the mechanism by which ATP is formed in mitochondria and chloroplasts are very similar. These also share other features like

- → Both often migrate from place to place with in the cells and also contain their own DNA, which encodes some of the key organellar proteins.

The proteins encoded by mitochondrial and chloroplast DNA are synthesized on ribosomes with in the organelles. However most of the proteins in each organelle are encoded in nuclear DNA and are synthesized in the cytosol; and are transported into the organelles.



5. Endoplasmic reticulum

The cytoplasmic matrix is transversed by a complex network of inter-connecting membrane bound vacuoles or cavities. These vacuoles or cavities often remain concentrated in the endoplasmic portion of the cytoplasm, therefore known as endoplasmic reticulum a name derived from the fact that in the light microscope it looks like a “mal in the cytoplasm”.

The name “Endoplasmic reticulum” was coined in 1953 by Porter.

The largest membrane in a eukaryotic cell encloses the Endoplasmic Reticulum (ER) - A compartment comprising a network of inter connected, closed, membrane – bounded vesicles.

The ER is particularly important in the synthesis of many membrane lipids and proteins.

The Endoplasmic Reticulum is of two forms.

1. Smooth Endoplasmic Reticulum
2. Rough Endoplasmic Reticulum.

Occurrence

The occurrence of endoplasmic reticulum varies from cell to cell. The Erythrocytes, egg and embryonic cells lack in ER. The adipose tissue, brown fat cells, adrenocortical cells, retinal pigment cells contain Smooth Endoplasmic Reticulum (SER). The cells of those organs which are actively engaged in the synthesis of proteins such as acinar cells of pancreas, plasma cells, globlet cells, and some endocrine glands are found to contain rough endoplasmic reticulum which is highly developed. The presence of both SER and RER in the hepatocytes (liver cells) is reflective of the variety of roles played by the liver in metabolism.

Morphology

Morphologically, the ER occur in the following three forms.

1. Lamellar form or cisternae (A closed, fluid filled sac. Vesicle or cavity is called cisternae).
2. Vesicular form or vesicle
3. Tubular form or tubules.

1. Cisternae

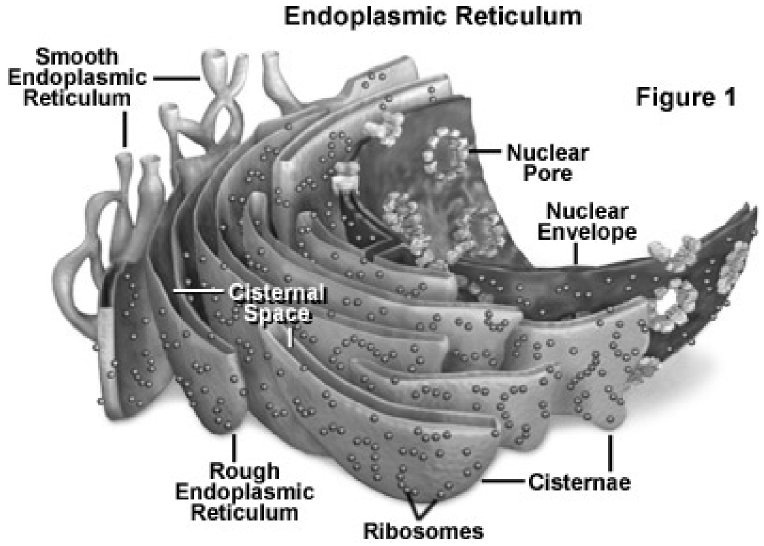
The cisternae are long, flattened, sac like, unbranched tubules, having the diameter of 40-50μm. They remain arranged parallely in bundles or stakes. RER usually exists as cisternae which occur in those cells which have synthetic roles as pancreas, notochord and brain.

2. Vesicles

The vesicles are oval, membrane bound vacuolar structures having a diameter of 25-500 μm. They often remain isolated in the cytoplasm and occur in most cells but especially abundant in the SER.

3. Tubules

The tubules are branched structures forming the reticular system along with the cisternae and vesicles. They usually have the diameter of 50-190µm and occur in almost all cells. Tubular form of ER is found in SER and is dynamic in nature i.e, it is associated with membrane movements, fission and fusion between membranes of cytocavity networks.



Smooth Endoplasmic Reticulum

The Smooth Endoplasmic Reticulum is smooth because it lacks ribosomes.

The synthesis of fattyacids and phospholipids occurs in the smooth ER. Although many cells have verylittle smooth ER, this organelle is abundant in hepatocytes.

Enzymes in the smooth ER of the liver modify or detoxify hydrophobic chemicals such as pesticides and carcinogens by chemically converting them into more water soluble, conjugated products that can be secreted from the body. High doses of such compounds result in a large proliferation of the smooth ER in liver cells.

The Rough Endoplasmic Reticulum

The Endoplasmic Reticulum is termed as Rough as it is studded with Ribosomes. Ribosomes bound to the Rough ER synthesize certain membrane and organelle proteins and virtually all proteins to be secreted from the cell. The ribosomes that fabricate secretory proteins are bounded to the rough ER by the nascent polypeptide chain of the polypeptide. As the growing secretory polypeptide emerges from the ribosome, it passes through the rough ER membrane with the help of certain specific proteins in the membrane. The newly made secretory proteins accumulate in the lumin of the rough ER before being transported to their next destination.

All eukaryotic cells contain a considerable amount of rough ER, because it is needed for the synthesis of plasmamembrane proteins and proteins of the extracellular matrix. Rough ER is particularly abundant in cells that are specialized to produce secreted proteins.

For example: Plasma cells produce antibodies, which circulate in the blood stream and pancreatic acinar cells synthesize digestive enzymes, which are transported to the intestine via a series of progressively larger ducts. In both types of cells, a larger part of the cytosol is filled with Rough ER.

Functions of endoplasmic reticulum

The endoplasmic reticulum acts as secretory, storage, circulatory and nervous system of the cell.

It performs following important functions.

1. The endoplasmic reticulum provides an ultra structural skeletal frame work to the cell and gives mechanical support to the colloidal cytoplasmic matrix.

2. The exchange of molecules by the process of osmosis, diffusion and active transport also occur through the membranes of ER like plasmamembrane. The ER membranes has permeases and carriers.

3. The ER membranes contain many enzymes which perform various synthetic and metabolic activities. Further it provides increased surface area for various enzymatic reactions.

4. The ER acts as an intracellular circulatory or transporting system. Various secretory products of granular ER are transported to various organelles as follows.

Granular ER → Agranular ER → Golgimembranes → Lysosomes, transport vesicles or secretory granules.

5. The ER membranes are found to conduct intracellular impules.

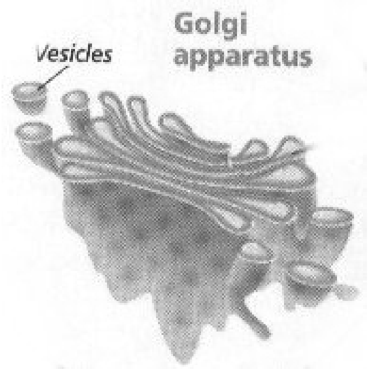
For example : The sarcoplasmic reticulum transmits impulses from the surface membrane into the deep region of the muscle fibres.

6. The ER membranes form the new nuclear envelope after each nuclear division.

7. The sarcoplasmic reticulum plays a role in releasing calcium when the muscle is stimulated and actively transporting back into the sarcoplasmic reticulum when the stimulation stops and the muscle must be relaxed.

6. Golgiapparatus

For the performance of certain important cellular functions such as biosynthesis of polysaccharides, packaging (compartmentalizing) of cellular synthetic products (proteins), production of exocytotic (secretory) vesicles and differentiation of cellular membranes, there occurs a complex organelle called Golgicomplex, or Golgiapparatus in the cytoplasm of animal & plant cells. The Golgi apparatus like the endoplasmic reticulum is a canalicular system with sacs, but unlike the ER it has parallely arranged, membrane bound vesicles which lack ribosomes and stainable by osmium tetraoxide and silver salts.



Occurrence

The Golgi apparatus occurs in all cells except the prokaryotic cells (i.e. Mycoplasma, bacteria, and blue green algae) and eukaryotic cells of certain fungi and red blood cells of animals.

Morphology

The golgi apparatus is morphologically very similar in both plant and animal cells. However it is extremely pleomorphic. Its shape and form vary depending on cell type. Typically, Golgi Apparatus appears as a complex array of inter connecting tubules, vesicles and cisternae. The detailed structure of three basic components of the Golgi apparatus can be studied as follows.

1. Flatenned sac or cisternae

Cisternae (about 1μm in diameter) are central, flattened, plate-like or saucer like closed compartments which are held in parallel bundles or stacks one above the other. In each stack, cisternae are separated by a space of 20 to 30nm which may contain rod-like elements of fibres. Each stack of cisternae forms a dictyosome which may contain 5 or 6 Golgi cisternae in animal cells or 20 or more in plant cells. Each cisternae is bounded by a smooth unit membrane.

2. Tubules

A complex array of associated vesicles and anastomosing tubules (30.50nm diameter) surround the dictyosome and radiate from it. The peripheral area of dictyosome is fenestrated (lace-like) in structure.

3. Vesicles

The vesicles 60nm in diameter are of 3 types.

1. Transitional vesicles

Transitional vesicles are small membrane limited vesicles which discharge from margins of the transitional ER (SER) to migrate and converge to cisface of Golgi, where they coalasce to form new cisternae.

2. Secretory vesicles are varied sized membrane limited vesicles which discharge from margins of cisternae of Golgi. They, often occur between the maturing face of Golgi and the plasmamembrane.

3. Clathrin coated vesicles are spherical protuberances about 50μm in diameter and with a Rough surface. They are found at the periphery of the organelle, usually at the ends of single tubules and are morphologically very distinct coated vesicles are known to play a role in intra-cellular traffic of membranes and of secretory products i.e between Endoplasmic reticulum as well as between GERL region and the Endosomal and lysosomal compartments.

The GERL region

The Golgi apparatus is a differentiated portion of the endomembrane system found in both animals and plants. This membraneous compound is spatially and temporally related to ER on one side and by way of secretory vesicles, may fuse with specific portions of the plasmamembrane. To the trans face of the Golgi is associated the trans-reticular Golgi network or GERL (Golgi + Smooth ER + Lysosomal) in which acid phosphatase enzyme a charecteristic lyzosomal enzyme makes its first appearance. GERL is found to be involved in the origin of primary lyzosome and of reelanin granules; in the processing, condensing and packaging of secretory material in endocrine and exocrine cells and in lipid metabolism. GERL is also a region of sorting of cellular secretory proteins.

Functions

1. Golgi functions inplants

Inplants, Golgi apparatus is mainly involved in the secretion of materials of primary and secondary cell walls (Eg: formation and export of glycoproteins, lipids, pectins and monomess of hemicellulose, cellulose, lignin etc). During cytokinesis of mitosis or meiosis, the vesicles originating from the periphery of Golgi apparatus, coalesce in the phragmoplast area to form a semi-solid layer, called cell-plate. The unit membrane of Golgi vesicles guses during cell plateformation and becomes part of plasmamembrane of daughter cells.

Golgi complex

A series of flattened sacs like structures located near the nucleus in many cells is known to be the Golgi complex whose main function is to process and sort the secretory and membrane proteins. Several minutes after proteins are synthesized in the Rough ER, most of them leave the organelle with in small membrane bounded transport vesicles. The vesicles which bud off from regions of the rough ER not coated with ribosomes, carry the proteins to the luminal cavity of Golgi complex.

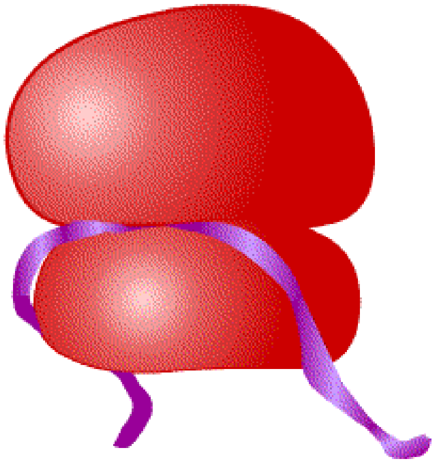
Three-dimensional reconstructions from serial sections of a Golgi complex reveal a series of flattened membrane vesicles or sacs, surrounded by a number of more or less spherical membrane vesicles. The stalk of flattened Golgi sacs has three defined regions – the cis, the medial and the trans.

Transfer vesicles from the rough ER fuse with the is region of the Golgi complex, where they deposit their proteins. These proteins then progress from the cis to the middle to the trans region with in each region are different enzymes that modify secretory and membrane proteins differently, depending on their structures and their final destinations.

After secretory proteins are modified in the Golgi sacs, they are transported out of the complex by a second set of transport vesicles, which seem to bud off the trans side of the Golgi complex. Some of these transport vesicles termed coated vesicles are surrounded by an outer protein cage composed primarily of the fibrous protein clathrin, some vesicles contain membrane proteins destined for the plasma membrane. Others, proteins for lyzosomes or for other organelles.

7. Ribosomes

The ribosomes are small, dense, rounded and granular particles of the ribonucleo protein. They occur either freely in the matrix of mitochondria, chloroplast and cytoplasm or remain attached with the membranes of the endoplasmic reticulum and nucleus. They occur in most prokaryotic and eukaryotic cells and are known to provide a scaffold for the ordered interaction of all molecules involved in protein synthesis.



Occurrence

The ribosomes occur in cells, both prokaryotic and eukaryotic cells. In prokaryotic cells the ribosomes occur freely in the cytoplasm. In eukaryotic cells the ribosomes either occur freely in the cytoplasm or remain attached to the outer surface of ER.

The cells in which active protein synthesis takes place, the ribosomes remain attached with the membranes of ER.

Methods of isolation

The ribosomes are usually isolated from the cell by differential centrifugation method in which an analytical centrifuge is employed. The sedimentation coefficient is determined by the various optical and electronic techniques. It is expressed in the “svedberg” unit `s’. The coefficient is related with the size and molecular weight of the ribosomal particle.

Types of ribosomes

1. 70s ribosomes

The 70 s ribosomes are comparatively smaller in size and have sedimentation coefficient 70 s. they occur in the prokaryotic cells of the Blue green algae and bacteria and also in mitochondria and chloroplasts of eukaryotic cells.

2. 80 s Ribosomes

The 80 s are larger than 70 s type ribosomes and have a sedimentation coefficient of 80 s. There occur in eukaryotic cells of the plants and animals.

Count of ribosomes

An E.coli cell contains 10,000 ribosomes, forming 25% of the total mass of the bacterial cell. In contrast mammalian cultured cells contain 10 million ribosomes per cell, each of which is about twice as large as prokaryotic ribosomes.

Structure of ribosomes

The ribosomes are oblate, spheroid structures of 150 to 250 ^{o}A in diameter. Each ribosome is porous, hydrated and composed of two sub units. One ribosomal sub unit is large in size and has a dome like structure, while the other subunit is smaller in size and occurring above the larger subunit and forming a caplike structure.

The 70s ribosome consists of two subunits i.e. 50s and 30s. The 50s subunit is larger and 30s subunit is smaller.

The 80s ribosome consists of 60s and 40s subunits 60s subunit is the larger one and the 40s is the smaller one. Both the sub units remain attached by a narrow cleft.

The two ribosomal subunits remain united with each other due to high concentrations of Mg^{++} ions. When the concentration of Mg^{++} ions reduces in the matrix, both the subunits get separated.

When they are attached they forma dimer. Further during protein synthesis many ribosomes are aggregated due to common messenger RNA and form the structure called poly ribosomes or polyzomes.

Chemical composition

The ribosomes are chemically composed of RNA and proteins as their major constituents: There is noipid content in ribosomes.

Ribosomal RNA's

The 70s ribosomes contain three types or rRNA i.e 23srRNA, 16s rRNA, 5s rRNA. The 23s and 5s rRNA occur in the larger 50s ribosomal subunit while the 16s rRNA occur in the smaller 30s ribosomal subunit.

The 80s ribosomes contain four types of rRNA i.e 28s rRNA, 18s rRNA, 5s rRNA, 5.8s rRNA. The 28s, 5s, 5.8s rRNA's occur in the larger 60s ribosomal subunit, while the 18s rRNA occurs in the smaller 40s ribosomal sub unit.

Ribosomal proteins

70s ribosomes of E.coli are composed of about 55 ribosomal proteins. Out of which 21 different molecules have been isolated from the 30s ribosomal subunit and 32-34 proteins from the 50s subunit. 80s ribosomes of eukaryotes are composed of 70 different proteins of which 30 types are present in the smaller 40s subunit and nearly 40 types are present in the larger 60s subunit.

8. Nucleus

The nucleus, the largest organelle in the eukaryotic cells is surrounded by two membranes, each one a phospholipid bilayer containing many different types of proteins. The inner nuclear membrane defines the nucleus itself. In many cells the outer nuclear membrane is continuous with the rough ER and the space between the inner and outer nuclear membranes is continuous with the lumen of the Rough Endoplasmic reticulum.

Robert brown for the first time in 1883 discovered a prominent body with in the cell and termed it as nucleus. A synonymous term of this organelle is the Greek Word Karyon on the basis of presence or absence of well defined nucleus, living organisms have been classified into two groups. There groups are (1) prokaryotes the organisms which do not have a well organized nucleus and will therefore include viruses, bacteria and blue-green algae and

2. Eukaryotes which would include the remaining types which have a well organized nucleus.

The nucleus is found in all the eukaryotic cells of the plants and animals. However certain eukaryotic cells such as the mature sieve tubes of higher plants and mammalian erythrocytes contain no nucleus.

Morphology

Usually the cells contain single nucleus but the number of the nucleus may vary from cell to cell. According to the number of nuclei following types of cells have been recognized.

1. Mononucleate cells

Most plant and animal cells contain single nucleus, such cells are known as mononucleate cells.

2. Binucleate cells

Cells containing two nuclei

Example: protozoans such as paramecium and cells of cartilage and liver.

3. Polynucleate cells

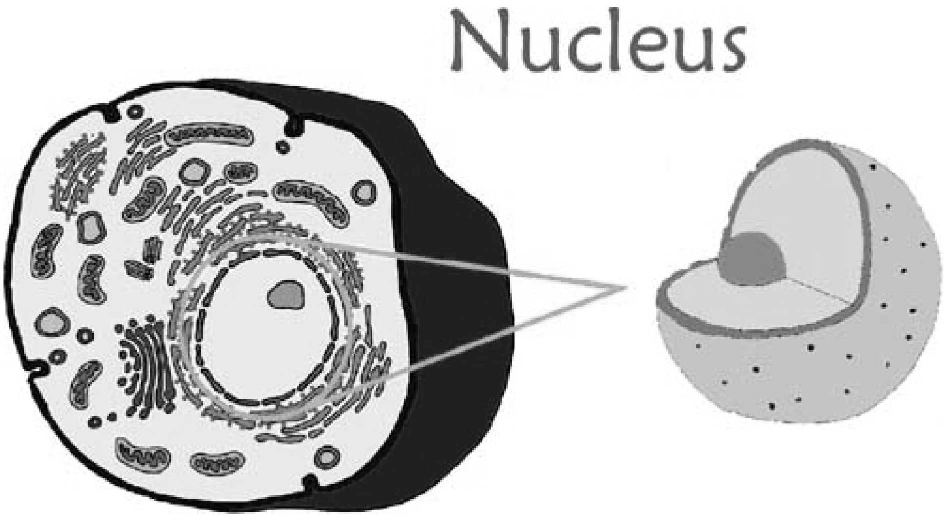
Cells containing many nuclei (3-100) are known as polynucleate cells. The polynucleate cells of animals are called as syncytial cells and polynucleate cells of plants are known as coenocytes. Example for syncytial cells are the osteo blast (polykaryocytes of bone marrow) which contain about 100 nucleiper cell and the example for coenocytes are the siphonal algae vaucheria.

Shape

The shape of the nucleus normally remains related with the shape of the cell, but certain nuclei are almost is regular in shape. The shape of the nucleus varies from spheroidal to ellipsoid and also found as discoid in some cells like the cells of the squamous epithelium.

Size

Generally nucleus occupies about 10 percent of the total cell volume. Nuclei vary in size from about 3µm in volume.



Ultra structure

The nucleus is composed of following structures:

1. The nuclear membrane or envelope.
2. The nucleoplasm.
3. The chromatin fibres.
4. The nucleous

The nuclear envelope

The nuclear envelope encloses the DNA and defines the nuclear compartment of interphase and prophase nuclei. It is formed from two concentric unitmembrane contains specific proteins that act as binding sites for the supporting fibrous sheath of intermediate filaments called nuclear lamina. Nuclear lamina has contact with the chromatin (chromosomes) and nuclear RNA's. The inner nuclear membrane is surrounded by the outer nuclear membrane, which closely resembles the membrane of the endoplasmic reticulum that is continuous with it like the membrane of the Rough ER, the outer surface of the outer nuclear membrane is studded with ribosomes engaged in protein synthesis. The proteins made on there are transported into the space between the inner and outer nuclear membrane called perinulcear space. The perinuclear space is a 10-50 nm wide fluid filled compartment which is continuous with the ER lumen.

Nuclear lamina

Nuclear lamina is a protein mesh work which is 50-80 nm thick, that lines the inner surface of the inner membrane, except the areas of nucleopores and consists of a square lattice of intermediate filaments forming a very dynamic structure.

Nuclear pore

The nuclear envelope in all eukaryotic cells is perforated by nuclear pores which appear circular in surface view and have a diameter between 10nm to 100nm. These pores aid in the transport of materials. These nuclear pores are evenly or randomly distributed over the surface of nuclear envelope.

Nucleoplasm

The space with in the nuclear envelope is filled with a transparent, semisolid granular slightly acidophilic ground substance or the matrix known as nuclear sapor nucleoplasm. It has a complex chemical composition, composed mainly of nucleoproteins along with other organic & inorganic substances i.e. Nucleic acids, proteins, enzymes, minerals.

Chromatin fibres

The nucleoplasm consists many thread like, coiled much elongated structures which take readily the basic stains such as the basic fuschin. These thread like structures are known as the chromatin fibres. These are observed only in the interphase nucleus. During the cell division (Mitosis & Meiosis) chromatin fibres become thick ribbon – like structures which are known as the chromosomes. The fibres of the chromatin are twisted, finely anatomized and uniformly distributed in the nucleoplasm. Two types of chromatin can be recognized basing on the staining properties.

A. Hetero chromatin

The darkly stained condensed region of the chromatin is known as hetero chromatin. The hetero chromatin occurs around the nucleolus and at the periphery. It is supposed to be metabolically and genitically inert.

B. Euchromatin

The lightly stained and diffused region of the chromatin is known as the euchromatin. The euchromatin is genetically active.

Nucleolus

Most cells contain in their nuclei one or more prominent spherical colloidal acidophilic bodies, called nucleolus. However bacteria and yeast lack nucleolus. The size of the nucleolus is found to be related with the synthetic activity of the cell. The number of the nucleoli in the nucleus depends on the species and the number of chromosomes. The number may be one, two, or four. The position of the nucleolus is eccentric.

Nucleolus is not bounded by any limiting membrane; calcium ions are supposed to maintain its intact organization. Nucleolus are the sites where biogenesis of ribosomal subunits.

Chromosomes

The chromosomes are the nuclear components of the special organization, individuality and function. They are capable of self replication and play a vital role in heredity, mutation, variation and evolutionary development of the species. The number of chromosomes is constant for a particular species. Therefore, there are of great importance in the determination of the phylogeny and taxonomy of the species.

Chromosome number of some organisms

Common name	Chromosome number
Hydra	32
Housefly	12
Mosquito	6
Rabbit	44
Gorilla	48
Man	46

Summary

A prokaryotic cell is essentially a one envelope system consists of central nuclear components (i.e. DNA, RNA & nuclear proteins) surrounded by cytoplasmic ground substance, with the whole enveloped by plasma membrane. whereas trhe eukaryotic cell contains several membrane bound organelles such as Lysosomes , mitochondria ,chloroplast,ribosomes ,peroxisomes etc that perform different functions.

Model Questions

1. Write a detailed account on eukaryotic organellar organization.
2. Give an account on structure & function of endoplasmic reticulum.
3. Give an account an structure & function of ribosomes.

References Books

1. Cell & Molecular Biology by Gerald Karp.
2. Cell & Molecular Biology by P.K.Gupta.
3. Cell & Molecular Biology by De. Robertis.

K.Haritha

Lesson 1.2.1

STRUCTURE AND FUNCTION OF PROKARYOTIC GENOME ORGNIZATION

Objective

This chapter of prokaryotic genome organization deals with the prokaryotes. The main difference between the prokaryotes and eukaryotes is that eukaryotic cells have a complex internal cellular organization with membranous organelles and a separate nucleus. Prokaryotic cells lack this complex internal organization and have no distinct nucleus. As far as genetics and molecular biology studies are concerned the well studied prokaryote is the E.coli, that we are going to see in this chapter.

1.2.1.1. Different Genome Organizations

A. E.coli Genome

Among the bacteria the well studied genome organization is in E.coli. The E.coli genome is 4639kb in length and comprises for known protein functions are studied. About one third of the protein – coding genes are organized into 75 different operons, with the remainder scattered around the rest of the genome at random. Operon is a set of genes under one regulatory system. Most of genes present as a single copy in the chromosome. The exception to single copy is rRNA gene cluster, as seven copies of these rRNA genes present in a single E.coli chromosome.

The 2400 genes of E.coli will be about 80% of its genetic material. The functions of the remaining 20% are

i. mainly acting as spacer or intergenic region between genes.

ii. One of these regions act as the origin of replication for the E.Coli DNA molecule.

iii. Some regions interact with DNA binding proteins to package the DNA molecule into the nucleoid.

B. Genomes of the bacteria

The basic features of gene organization, with the operons and few repeated genes appear to be same in all the bacteria. The bacteria related to E.coli have their genes arranged in a similar order, where as the map for a more distant species is more different. There are no absolute requirements for particular genes to be adjacent to one another, except when those genes are clustered in an operon. Any similarities that are seen presumably arise from evolutionary relationship rather than functional requirements.

Depending on the functions the requirements of cell varies, that will be proportional to the genome complexity. For example the 30,000kb genome of *Bacillus megaterium* is probably taken up by extra genes that enable this bacterium to synthesize heat-resistant spores. The extra DNA cannot be entirely accounted for only one function and the larger molecules must also have more extensive intergenic regions.

As the bacteria have the smallest genomes, they are mostly obligate parasites, depend on the host for many of its requirements.

C. Genomes of archaea

The first indications that the archaea are very different from other prokaryotes identified, when their rRNA genes were compared with those of bacteria. It was shown that the archaeal rRNA take up a base paired secondary structure that is significantly different from that of E.coli. *Methanococcus Jannaschi* genome complete sequencing of provides the information of archaea bacterial complete organization. Although the genome as a whole has a close resemblance to a typical bacterial molecule, being circular with groups of genes arranged in operons, the molecules involved in expressing the genes are much more closely related to the eukaryotic equivalents. For example, the M.Jannaschii RNA polymerase has 11 subunits, rather than four subunits in bacterial enzyme. This RNA polymerase resembles RNA polymerase II subunits in eukaryotes rather than the E.Coli proteins. The translation initiation factors of M. Jannaschii appear to comprise a mixture of bacterial and eukaryotic proteins.

These genetic differences are supported by a number of biochemical differences between the archaean bacteria in particular regarding the structure and chemical composition of the cell wall. Overall, it is becoming clear that the archaea make up a distinct group of organisms.

1.2.1.2. Physical Organization of Bacterial Genomes

The prokaryotic cells are not divided into membrane bound compartments; their internal structure is not entirely featureless. The electron microscope reveals two distinct regions with in an E.Coli cell-a central area called the nucleoid taking up about one third volume of the cell, surrounded by a peripheral region that is usually referred to as the cytoplasm.

A. Structure of Bacterial Nucleoid

The nucleoid is made up of DNA and protein. The DNA is a single, circular molecule that carries bacterial genes. In E.coli the DNA molecule is of approximately 1.5 mm, and has to be placed in a cell of about 1um to 2 um. This is not impossible as a DNA molecule is very thin and so does not take up much space when it is folded up tight by supercoiling. The problem is that the DNA molecule has to be accessible, as the genes it carries must be transcribed, and it must also be possible to replicate the DNA and separate the daughter molecule must be folded up in a very precise way, probably with the protein component of the nucleoid helping to build and maintain the ordered structure.

The nucleoid contains a protein at the center. From the protein core upto 100 supercoiled loops of DNA radiate. Several DNA binding proteins, thought to be involved in packing the DNA molecule, have been isolated from E.Coli nucleoids. The protein present at the center are HU, H, H_1, H_2P_1, etc., the HU of E.Coli forms a tetramer around which 58 bp of DNA becomes wound. There are some 60,000 HU proteins per cell.

B. Replication and partitioning of bacterial genome

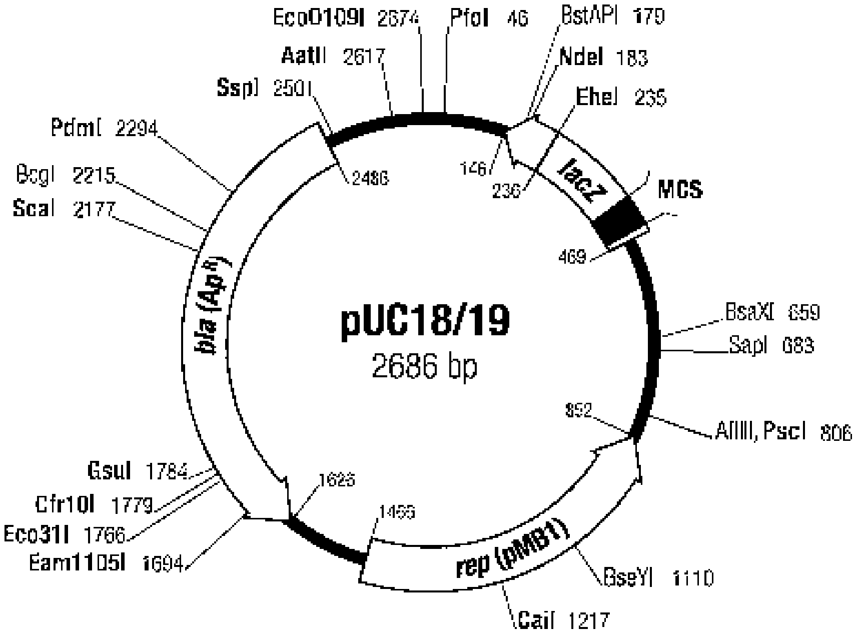
The E.coli genome has a single replication origin, from which two replication forks progress in opposite directions around the molecule. The termination sites are located on opposite side of DNA origin site.

With such a great length of DNA in such a small volume, partitioning of cells after replication is a difficult task. In 1963 Jacob and Bernner proposed that each daughter molecule has an attachment point on the cell membrane, and that partitioning at cell division occurs by synthesis of new membrane between two attachment points. A protein, MuKB, which has weak structural similarities with eukaryotic force generating proteins such as Myosin. (It has been proposed that a limited filamentons system is synthesized with in the bacterium during cell division and that MuKB generated forces to move DNA molecules a long. These filaments to opposite parts of the dividing cells.)

1.2.1.3. Plasmids

The nucleoid DNA is not the only repository for genes in the bacterial cell. Bacteria often contain small independent DNA molecules called plasmids, which carry genes not found on the main DNA molecule.

Plasmids are circular DNA molecules that lead in independent existence in the bacterial cell. Plasmids almost always carry one or more genes and often these genes are responsible for a useful characteristic. For example, the ability to survive in normally toxic concentrations of antibiotics such as chloramphenicol or ampicillin is often due to the presence in the bacterium of a plasmid that carries antibiotic-resistance genes. Now the plasmids playing important role in rDNA technology. We will see the example of plasmid PUC 18 map in Fig.1.



All plasmids possess at least one DNA sequence that can also act as an origin of replication, so they are able to multiply in the cell independently of the main DNA molecule. The smaller plasmids make use of the cells own DNA replicative enzymes to make copies of themselves, where as some of the larger ones carry genes that code for special enzymes that are specific for plasmid replication.

Few types of plasmids are also able to replicate by inserting themselves into the main DNA molecule. These integrative plasmids or episomes may be stably maintained in this form through numerous cell divisions, but always at some stage exists as independent elements.

A. Types of Plasmids

All species of bacteria possess plasmids and a large number of different types are known. These are classified according to the genes that they carry and the characteristics they confer to the bacterium.

a. Fertility (F) plasmids: are able to direct conjugation between different bacteria.

b. Resistance (R) plasmids: Carry genes conferring resistance to one or more antibacterial agents such as ampicillin, tetracycline.

c. Col plasmids: Carry genes coding colicins. Colicins are proteins that kill other bacteria.

d. Degradative plasmids : Allow the bacterium to metabolize unusual molecules such as toluene and salicylic acid.

e. Virulence plasmids : Carry pathogenicity on the bacterium.

B. Copy number and incompatibility

The copy number refers to the number of molecules of a plasmid that are present in a single cell and is a characteristic value for each plasmid type.

Depending on copy number plasmids are:

i. Stringent plasmids – have a low copy number of perhaps just one or two per cell.

ii. Relaxed plasmids – present in multiple copies of 10 or more per cell.

Incompatibility refers to the fact that certain plasmid types cannot coexist in the same cell. An E.coli cell can contain upto seven or more types of plasmid at the same time, but these must all belong to different incompatibility groups.

The size of the plasmids vary from approximately 1kb for the smallest to over 250kb for the larger ones. Some plasmids are found in no other bacteria. Some plasmids have broad host range and can exists in numerous species.

1.2.1.4. Bacterial transposons

Transposons are DNA sequence elements that are capable of moving around in the genome. They occur in both eukaryotes and prokaryotes, where they are present on bacterial chromosomes and on plasmids. The process of movement is called transposition and depends on recombination between DNA sequences. Transposons are autonomous units and each encodes an enzyme called transponase that catalyzes its own transposition. Transposons are of different types like Insertion sequences, composite transposons, T_n3-type transposons.

Summary

Prokaryotic cells have no distinct nucleus for locating the chromosome. Operon arrangement is the specific feature of prokaryotic organization. rRNA's secondary organization is different in eubacteria and archaeabacteria.

Technical terms

1. Copy number – the number of molecules of a plasmid in a single cell.

2. Incompatibility – certain plasmid types can’t coexist in the same cell.

Model Questions

1. What are plasmids ?
2. How the genome of bacteria is organized in a cell ?

Reference Books

1. Genes VI – Lewin
2. Principles of Biochemistry – Lehninger, Nelson & Cox.

Mrs. G.V. Padmavathi,

Lesson 1.2.2

STRUCTURE AND FUNCTION OF EUKARYOTIC GENOME ORGANIZATION

Objective

The genome of eukaryotes has more complex organization than that of prokaryotes. The prokaryotes contain haploid number of chromosomes. Haploid cell is the cell which contain only one set of genes. But the eukaryotic cells mostly diploid, the cell having two complete sets of genes, one from each parent. Some eukaryotic cells are polyploid, the cells carrying several copies of the genome. All the DNA of a cell does not contain genes, only some areas of DNA contain genes.

The eukaryotes conatains many times the amount of DNA than in prokaryotes. The DNA of eukaryotes is packaged into several chromosomes. Each chromosome present only one time in case of haploid cells, two times in diploid and many times in polyploids. The chromosome length of E.Coli is 1.5 mm, which is very small when compared with 100nm of human haploid chromosomal length. It is very surprising how this large polynucleotide DNA arranged in the chromosome to fit in the nucleus of a cell.

We will consider the organizational aspects of complex eukaryotic chromosome in the nucleus.

1.2.2.1. Chemical Composition of Eukaryotic Chromosomes

The chromatin form of chromosome, present at the interphase of cell cycle, is the well studied form of chromosome. The chromatin contain 50% DNA and 50% proteins, and very little amount of RNA is supposed. The proteins present in the chromatin are of two types.

A. Histone proteins

B. Nonhistone proteins

A. Histone proteins

Histone proteins are positively charged basic proteins at body pH. Histones play major role in chromatin. Five types of histones present in almost all eukaryotic cells. Few exceptions are there like in sperm cells instead of histones protamines present. The histone proteins complex with DNA forming small, ellipsoidal, beads called nucleosomes. H1 attaches on the outersurface of bead.

B. Nonhisotne Proteins

Contains large number of different proteins. The composition of the nonhistone chromosomal protein fraction varies widely among different cell types of the same organism. The non histone protein play much role in regulating gene expression than in organizing the chromosomes.

1.2.2.2. Packing of Chromosomes in Nucleus

Each chromosome contains one large DNA molecule of 1-20cm. During metaphase this is further packaged to 1 to 10 um. For this high complex, compact organization three levels of DNA packing is established.

A. Packing DNA as 10 nm diameter nucleosome.
B. Additional folding to form 30 nm chromatin fiber.
C. Nonhistone proteins condense the 30 nm fiber to form scaffold chromosome.

A. 10nm diameter nucleosome

When isolated chromatin is examined by electron microscopy, it is found to consists of a series of beads joined by thin threads. The beads are of 10 to 11nm in diameter and 6nm length. The beaded area of the DNA is packaged in a nuclease resistant form using the histone proteins, so called nucleosome. The DNA sequence between two successive nucleosomes is called linker DNA.

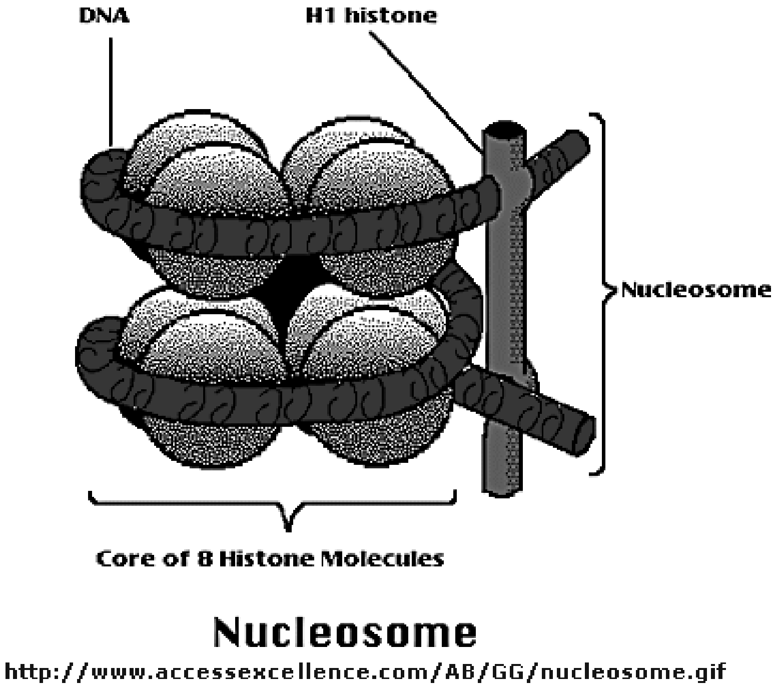

DNA
H1 histone
Nucleosome
Core of 8 Histone Molecules
Nucleosome
http://www.accessexcellence.com/AB/GG/nucleosome.gif

Figure.1: (lehninger 923 Fig 24-23) nucleosomes.

When the chromatin is partially hydrolysed by the endonucleases 200 nucleotides pair length DNA associated with nucleosome forms are produced. After extensive nuclease digestion of chromatin with endonuclease 146 nucleotides pair long segment of DNA remains present in each nucleosome. These nuclease resistant structure is called nucleosome core. The nuclosome core consists of a 146 nucleotide pair length of DNA and two molecules each of H_2a, H_2b, H_3 and H_4. The histones protect the segment of DNA in the nucleosome core from cleavage by endonucleases.

The complete chromatin subunit consists of the nucleosome core, the liner DNA and the associated nonhistone chromosomal proteins, all stabilized by the binding of one molecule of histone H_1 to the outside of the structure. The size of linker DNA varies from species to species and from cell type to cell type. Linkers of as short as 8 nucleotide pairs and as long as 114 nucleotide pairs are observed.

B. 30nm chromatin fiber

The nucelosome organization compacts DNA length about seven folds. But the overall compaction in chromosome is greater than 10,000. This much compaction is with the high orders of structural organization ----- The 30nm fiber provides an approximately 100 fold compaction of the DNA. This involves additional folding or supercoiling of the 10nm nucleosome to produce the 30 nm chromatin fiber, characteristic of mitotic and meiotic chromosomes. Histone H1 is involved in this supercoiling of 10nm fiber to produce 30nm fiber.

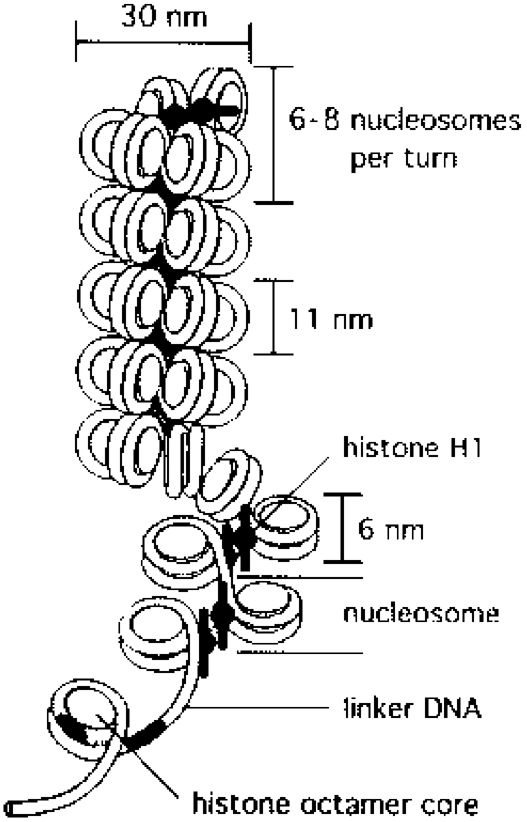
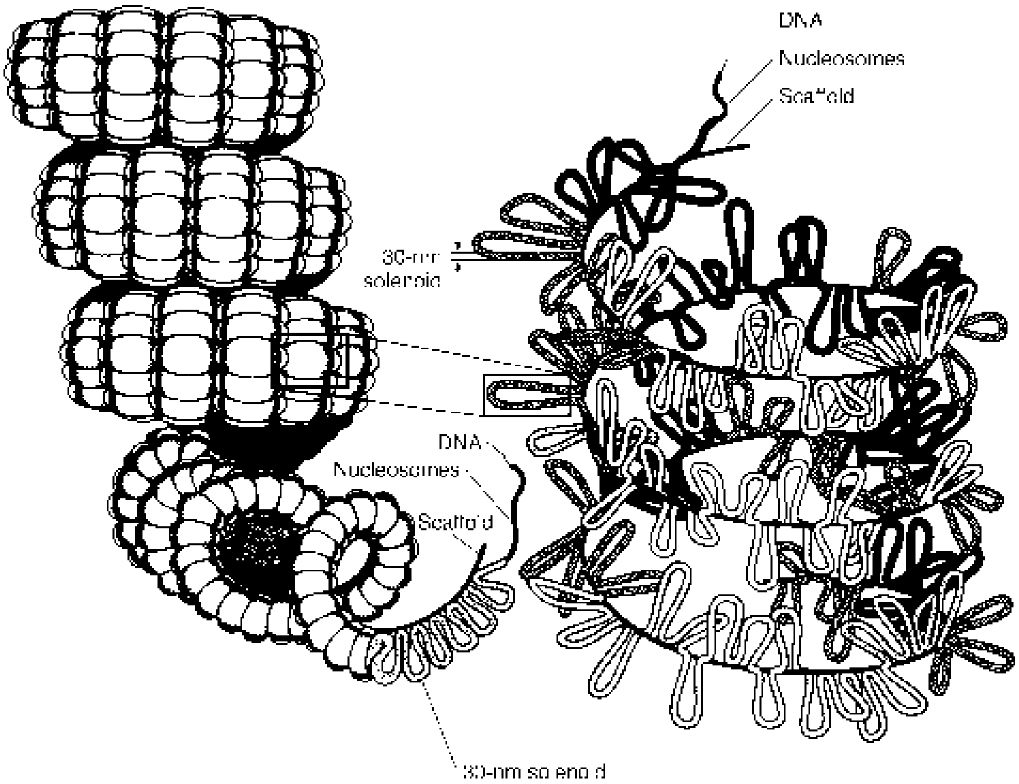


Figure 2. Scaffold structure (Lehninger 926, 24-29)

C. Scaffold structure

The final organization is scaffold structure chromosome formation. In this nonhistone chromosomal proteins form a scaffold that is involved in condensing the 30nm chromatin fiber into the tightly packaged metaphase chromosomes. This third level of condensation appears to involve the separation of segments of the giant DNA molecules present in eukaryotic chromosomes into independently supercoiled domains or loops.



1.2.2.3. Single Copy and Repetitive DNA

The large variations in the genome size of organisms and its renaturation & denaturation studies shows to have similar gene contents. Some part of most eukaryotic nuclear genomes is made up of repetitive DNA. The individual sequence elements that are repeated many times over, either in tandem arrays or interspersed throughout the genome are called repetitive DNA. Some of these sequences are repeated several thousand to several million times in the genome, others only as 10 to several hundred copies on contrast to this most genes are made up of sequences that are not repeated alese where, called as single copy DNA. In humans the repetitive DNA makes up about 15% of the genome. Some repetitive DNA sequences code for some functions, but most of the repetitive DNA sequences present in eukaryotes are commonly grouped into three classes:

1. Unique or single copy DNA sequences – 1 to 10 copies per genome
2. 2. Moderately repetitive DNA sequences – 10 to 10^5 copies per genome.
3. Highly repetitive DNA sequences – more than 10^5 copies per genome.

1.2.2.4. Chromosome Morphology

The overall structure of chromosome contains centromer, telomere and chromatids. The two arms present in a chromosome on two sides of centromer are called chromatids.

a. Centromere

This is the important region in chromosomal division. The position of centromere is characteristice for a particular chromosome and is one of the features used to distinguish individual members of the karyotype. This also the point of chromosome attachment to the microtubules that draw the daughters into their respective nuclei during cell division. The centromeric region of the chromosome contains special proteins and is characterized by specific centromeric DNA sequences.

b. Telomeres

The ends of chromosome are called the telomeres. The telomeres protect chromosomal ends. The telomeres three important roles are:

i. The telomeres must protect the ends of the chromosomes from attack by nuclease enzymes.

ii. They must prevent chromosomes from joining together

iii. Overcome the problems rised by replication.

Telomeres, like centromeres, contain special DNA sequences that appear to underlie their functional specialization.

1.2.2.5. Extrachromosomal Genes

The genetic material is generally supposed to be present only in nucleus of eukaryotes. But later studies established that DNA is present in chloroplasts of algae such as *chlamydomonas reinhardtii* and in mitochondria of fungi such as

Neurospora crossa. It was then realized that the unusual genes must be carried by the DNA molecules in these organelles.

Mitochondrial and chloroplast DNA molecules are generally circular and double stranded, although a few linear mitochondrial genomes are known.

1.2.2.6. Human Genome

The human genome is approximately 3 x 10^9 b with 65,000 to 80,000 genes per 23 chromosomes. The human genome is a single, linear, double stranded DNA molecule. About 30% of the genome consists of genes, the remainder being extrageni and having no known function.

Summary

1. Interphose chromatin structure is well studied chromosomal form.

2. The chromosome apart from DNA contain histone and non histone proteins also, in its organization.

3. The three levels of packing are 10nm fiber, 30nm fiber, scaffold form.

Technical Terms

1. Nucleosome – The beaded area of the DNA is packaged in a nuclease resistant form using the ristone proteins.

2. Repetitive DNA – The individual sequence elements that are repeated many times over, either in tandem arrays or interspersed throughout the genome.

3. Single copy DNA – The sequence on the DNA that are not repeated elsewhere in chromosome.

Model Questions

1. What are the proteins participating in chromosomal organization
2. Explain the morphology of DNA.

Reference Books

1. Principles of Biochemistry – Lehninger, Nelson & Cox
2. Gentics – Weaver

Mrs. G.V.Padmavathi

Lesson 1.2.3

CELL CYCLE

Objective

The basic objective to learn about cell cycle is, cell is the unit of biological continuity.

Introduction

The cell cycling is important for three biological aspects.

1.Reproduction; 2.Growth; 3. Replacement.

The cell cycle is defined as the sum total of the division related events that occur between the time a cell completes one cell division and the time it completes the next one.

The basic requirements of cell cycle are :

1. The genetic material in both the nucleus and organelles must replicate itself completely, and one copy of it must end up in each of the two cells that are formed, this is called nuclear division.
2. The cytoplasm materials and membranes must arrange themselves so that there will be two complete cells to receive this DNA, this is termed as cytokinesis.

The growth rates of a human infant compared to an adult, or a plant seedling compared to a mature tree, indicates that the rate of cell division is not constant. There must be a set of controls that determine when a cell is to divide and how quickly. These controls are necessary for the normal cell cycle otherwise leads to complications like cancer.

1.2.3.1. Cell Cycle Parameters

Most cells of an adult multicellular organism divide slowly, if at all; that is, their cell cycle times are quite long. For example, neurons in adult mammals and cortical cells in the plant stem are generally non dividing, once they are differentiated. In contrast certain cell types divide rapidly. In higher plants, where most growth occurs in cell elongation, cell division takes place in localized regions called meristems, which are located at the tips of both roots and shoots. This allows for growth in length.

In human adults there are also regions of constant cell division, mostly to replace cells that are lost. For example, about 200 million erythrocytes are destroyed each day. In normal circumstances the demand of the organism activities result in the occurrence of 25 million cell divisions at any time in the human body.

Typical cell cycle times for these cells in both plants and animals range from 15 to 40 hrs. (For organs such as the liver or for tumor cells the cells often cycling). But generally we can say that embryonic and regenerating tissues have shorter cycle times than adult differentiated tissues.

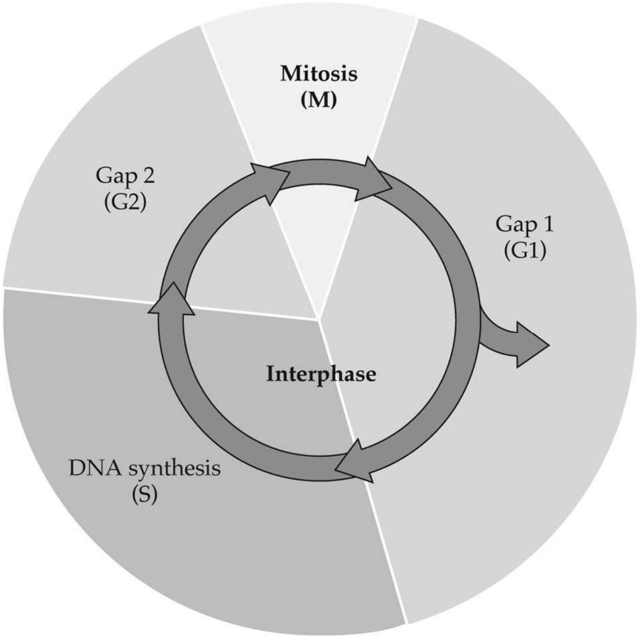
The concept of cancer as an uncontrolled growth leads to the assumption that cell cycle times of cancer cells are shorter than those of normal cells.

However, the actual reason is not that individual cells divide more quickly but that a greater fraction of the cells is cycling at a perceptible rate and fewer of the cancer cells die.

1.2.3.2. Specific Events in the Cell Cycle

The cell cycle is divided into four distinct phases:

1. This is followed by the G_1 phase, which covers the longest part of the cell cycle. Some cells assume a quiescent state known as the G_o phase.
2. G_1 gives way to the S phase which, in contrast to events in prokaryotes, is the only period in the cell cycle when DNA is synthesized.
3. During the relatively short G_2 phase, the now tetraploid cell prepares for next phase.
4. Mitosis or Meiosis and cell division occur during the M phase. It then enters G_1 phase once again and thereby commences a new round of the cell cycle.



LIFE: THE SCIENCE OF BIOLOGY, Seventh Edition, **Figure 9.3 The Eukaryotic Cell Cycle**
© 2004 Sinauer Associates, Inc. and W. H. Freeman & Co.

Fig.1. eukaryotic cell cycle (Voet 1039 Fig. 31-28)

1. G_1 phase (Gap_1 phase)

The duration of the G_1 phase of the cell cycle is the most variable of the four phases. In synchronous cells, it can be shown by the use of inhibitors and mutant strains that G_1 events are necessary for entry into S phase and the DNA synthesis.

Nuclei and cells grow throughout the G_1 period. Among the macromolecules that are synthesized in greater amounts during G_1 of cultured animal cells is fibronectin, In addition, a number of enzymes involved in DNA synthesis are made late in this phase, in preparation for S phase.

During G_1, cell commits itself to one of the three life-styles:

a. The cell can continue on the cycle and divide.

b. The cell can permanently stop division, such as neurons, which are terminally differentiated and do not divide.

c. The cell can become reversibly quiescent. Such nondividing cells are called G_o cells. They differs from neurons in that they can be induced to proliferate by appropriate stimuli memory cells of immune system are of this nature.

The G_1 and G_o cells differ in nonhistone proteins and RNA synthesis profiles. Hydorxy urea, a DNA synthesis inhibitor, is more potent in cells that were in G_1 than G_o cells stimulated to divide. A glycoprotein inhibitor acts as control point in releasing a cell from the quiescent G_o state. The inhibitor disappears rapidly from nuclei that have been reactivated from G_o to enter G_1.

The events in G_1 phase are :

a. There is an early increase in transport of various nutrients across the plasma membrane, as well as some changes in DNase sensitivity of chromatin. These events and progression into G_1, do not occur unless the cell is given platelet – derived growth factor.

b. The Epidermal growth factor is needed to stimulate the next series of events, which involve increases in polyribosomes and in some glycolytic enzymes.

c. A third growth factor, insulin like growth factor I, is required for the events late in G_1, the synthesis of proteins needed for DNA replication.

d. All the other proteins must be made for cells to enter S phase.

2. S phase (Synthetic phase)

The initiation of DNA replication requires that the cells have sufficient "division potential", which may or may not be acquired during G_1. This potential is a diffusible cytoplasmic factor that acts at the G_1-S transition.

The replication of chromosomal DNA can be studied by fiber autoradiography. In these experiments, cells in S phase are given a short pulse of radioactive thymidine, which is incorporated into DNA as it is synthesized. After the pulse, the cells are incubated for a time in non radioactive thymidine, and then the DNA is carefully isolated to preserve as long strands a possible. The location of radioactive elements then monitored by autoradiography.

The results of the experiment are :

a. The replication rate is 3000 base pairs per minute.

b. More than one origin of replication per chromosomal DNA molecule.

The number of origins varies depending on how short a time the S phase lasts. The origins must be on nucleosome – free DNA.

A number of proteins are involved in replication including DNA polymerase, RNA primase, nucleases, helicases, topoisomerases, ligases, single strand binding proteins.

A number of anticancer drugs are specific for S phase cells, like Thiogunine, Methotrexate, etc,. These agents generally inhibit DNA synthesis by interfering with nucleotide pools. Because cancer cells are not necessarily rapidly dividing, the drugs also affect those normal tissues that are likely to be in the S phase. Thus bone marrow, epithelia, and hair follicles are likely to be inhibited, as well as the targeted tumor cells.

3. G_2 phase (Gap$_2$ phase)

The G_2 phase, a period for preparation for chromosome condensation and mitosis, is typically one to four hours long. The exposure of cycling cells to sublethal doses of ionizing radiation of DNA – damaging chemicals causes a lengthening of the G_2 phase. A number of anticancer drugs act to cause cells to accumulate in G_2. The include mitomycin C, cyclophosphamide.

The events occurring during G_2 phase are :

a. Histone phosphorylation – especially H1, is apparently a prerequisite for the interaction of chromosomal fibers into higher order structures. This event is catalyzed by a protein kinase that is activated at the G_2-S transition point.

b. Tubulin synthesis – Tubulin is a major component of the mitotic spindle.

4. M phase

The M phase is of mitotic type or meiotic type. These two different types of nuclear divisions along the cytoplasmic division. We will see in detail in the cell division (1.2.4) chapter. Here we will see about the similarities and differences between Mitosis and Meiosis many aspects of mitosis and meiosis are similar:

a. There is a G_1 phase, followed by an S phase during interphase.

b. Chromosomes condens during prophase, are at an equatorial plate during metaphase, migrate to the pole during anaphase, and decondense during telophase.

c. Spindles form during prometaphase

d. Migration of chromosomes occurs along spindle fibers attached to kinetochores.

e. Centromeres and centrosomes may be involved in spindle orientation.

Differences between mitosis and meiosis :

a. While mitosis results in the exact duplication of the genetic material, with an equal partitioning into the two resulting cells, meiosis is a reduction division. If an organism has two chromosomes in its cells, the result of a mitotic cell cycle is two cells with two chromosomes each. But after meiotic cycle, the result is four cells with one chromosome each.

b. The mitosis occurs in many cell types and tissues, meiosis in higher animals occurs only in germ cells that form gametes.

c. While the mitotic cell cycle has one period of DNA replication followed by a single cell division, the meiotic cycle has one period of DNA replication followed by two cell divisions. This if a mitotic cell in G_1 has a one c content of DNA, after it has 2 c and after division each cell has one c. But if meiotic cell has 2c in G_1 and 4c after S phase after the two divisions each of the four progeny cells has one c.

d. In meiosis as crossing over occurs results in some gene variabilities. In mitosis no crossing over, a very rare event.

e. Mitotic cell division takes from hours to days time, but meiosis takes much longer, from weeks to months to years. For weeks example is human testis, for years example is oocytes in females.

1.2.3.3. Control of the Cell Cycle

Control of the Sequence of Events

The cell cycle is an orderly series of events, as is the development of an organism from a fertilized egg. Determining the genetic switches during the cell cycle has been an intensive area of research.

Some proteins of the cell become active and / or are synthesized at certain times in the cycle. Enzymes involved in DNA synthesis are usually induced in G_1 and S and tubulin for the spindle is made during G_2. In budding yeast, about 400 genes are directly involved in the cell cycle.

The microtubule inhibitor colchicines destroys the spindle and therefore blocks movement along the fibers. In dividing cells in a root meristem are exposed to this drug for several cell cycles, there is an increase in chromosome number. All events of the cycle except anaphase and cytokinesis occur, showing that cytokinesis depends on an intact spindle and chromosome migration but that the other events are not dependent on spindle integrity. This semi-independence has practical applications: colchicines is used to increase chromosome numbers and cell size in crop plant breeding.

The study of mutations has established that during the cell cycle, at least three major checkpoints exist, where the cell is monitered or checked before it can proceed to the next stage of the cycle.

The products of many of these genes are enzymes called kinases that can add phosphates to other proteins. They serve as master control molecules that work in conjugation with proteins called cyclin. These kinases phosphorylate cyclins and influence their activity at the cell cycle checkpoints. These activities thus regulate the cell cycle. When such a kinase works in conjugation with a cyclin, it is called a cdk protein, for cyclin – dependent kinase protein.

The three check points of a cell cycle are :

1. G_1 / S check point
2. G_2 / M check point
3. M check point

1. G$_1$ / S check point

This monitors the size that the cell has achieved following the previous mitosis and whether the DNA has been damaged. If the cell has not achieved an adequate size or if the DNA has been damaged, further progress through the cycle is arrested until these conditions are "corrected". If both conditions are initially normal, then the checkpoint is traversed and the cell proceeds to the S phase of the cycle.

2. G$_2$ / M check point

At this checkpoint the physiological conditions in the cell are monitored prior to entering mitosis. If DNA replication or repair to any DNA damage has not been completed, the cell cycle is arrested until these processes are completed.

3. M Check point

At this checkpoint both the successful formation of the spindle fiber system and the attachment of spindle fibers to the kinetochores associated with the centromeres are monitored. If spindle fibers are not properly formed on attachment is inadequate, mitosis is arrested.

The importance of cell cycle control and the checkpoints can be demonstrated by considering what happens when this regulatory system is impaired. For example, a cell that has incurred damage to its DNA is allowed to proceed through the cell cycle, the damage may lead to uncontrolled cell division, precisely the definition of a cancerous cell. Such a damaged cell would normally be arrested at either the G$_1$ / S or the G$_2$ / M checkpoints.

An interested related finding involves the protein product of the P^{53} gene in humans and its involvement during scrutiny at the G$_1$ / S checkpoint. This protein functions during the regulation of apoptosis the genetic process whereby programmed cell death occurs. When the normal p^{53} gene product is present, a proliferative cell that has incurred severe damage to its DNA will be targeted for programmed cell death at the G$_1$ / S checkpoint and this, effectively removed from the cell population. This surveillance is dependent on the product of the p^{53} gene. However, if the gene has mutated, resulting in abnormal function of the p^{53} gene product, the damaged cell may proceed through the checkpoint and continue to proliferate in an uncontrolled manner. Infact, a high percentage of human cancers, including colon, breast, lung and bladder malignancies. The p^{53} is referred to as a tumor suppressor gene.

Much of the basic work on the cell cycle has been conducted mainly by two groups of geneticists. They worked with yeasts, especially *Saccharomyces* *pombe*. The developmental biologist studying newly fertilized eggs of organisms such as frogs, sea wrching and newts. Both groups have succeeded in identifying and characterizing genes involved in the cell cycle and their role. In 2001 Nobel prize for physiology and medicine was awarded to lee Hartwell, Tim Hunt, and Paul Nurse for their work on the cell cycle.

Summary

The cell cycle is important for three biological aspects – Reproduction, growth and replacement. The four phases of cell cycle are G_1, S, G_2 and M. DNA replication takes place in the S phase of cell cycle. The cell cycle will be regulated at the checkpoints with inducer and inhibitor factors.

Technical Terms

1. G indicates Gap phase
2. S is the Synthetic phase of DNA.
3. Check points – the points where the cell cycle is check for its perfect maintenance.

Model Questions

1. How can you regulate the cell cycle
2. What are the different phases of cell cycle
3. What the way of G_1 phase commited life styles

Reference Books

1. Genetics – P.K. Gupta
2. Cell Biology – David E. Sadava

Mrs. G.V. Padmavathi

Lesson 1.2.4

CELL DIVISION

Objective

The primary activity concerned with cells is physical growth of primitive organisms without cell division, growth occurs through an increase in volume and an enlargement of the outer surface membrane. In case of spherical cells an increase in size produces a relatively smaller increase in surface area than in volume. As the surface area is proportionately less and increasingly demand of cells to obtain more food, oxygen and the various metabolic necessities and to excrete wastes and metabolic products imbalances the cell. This lead to cell death. So some form of cell division have been a primary necessity for the maintenance of life. How this cell division takes place in the cells is the objective of the lesson. In this lesson we will see the different types of divisions and how they occur.

1.2.4.1 Introduction

In case of eukaryotes cell division is more complex than prokaryotes because of variation in cell structure and composition. For a multicellular organism, such as humans countless divisions of a single-cell called zygote produce an organism of very high cellular complexity and organization. The cell division does not stop with the formation of the mature organisms, but continues throughout life. It is estimated that more than 25 million cells are undergoing division each second in an adult human to replace aged cells.

As the cell division not only involved in formation of more cells, but also in connecting link between parent and offspring, it plays vital role.

Cell division is basically of two types

1. Mitotic cell division
2. Meiotic cell division

1.2.4.2. Mitotic Cell Division

The name mitosis comes from the Greek word “mitos” meaning “thread”, used to describe thread like chromosomes. Mitosis is a process of nuclear division in which duplicated chromosomes are faithfully separated from one another, producing two nuclei, each with one copy of chromosome. Nuclear division is called karyokinesis.

The karyokinesis of a cell is usually accompanied by cytokinesis, a process by which the cell splits into two, partitioning the cytoplasm into two roughly equal compartments. The two daughter cells formed by mitosis and cytokinesis possess a genetic content identical to each other and to their mother cell.

The functions of mitosis are, maintaining the chromosome number and generates the new cells for the growth, maintenance and repair of an organism. In single called organisms that reproduce by cell divisions such as protozoans, algae and some fungi, mitosis provides the mechanism for asexual reproduction. Multicellular organisms begin life as zygotes. The mitotic activityof zygote and the daughter cells is the basic for development and growth of the organism.

In adult organisms, mitotic activity associated with cell division is prominent in wound healing and other forms of cell replacement in certain tissues like epidermal skin cells. In abnormal situations, somatic cells may exhibit uncontrolled cell divisions, resulting in cancer.

At the time of mitosis of a cell in the cell cycle except separating duplicated chromosomes all other activities of the cell are switched off and the cell becomes comparatively unresponsive to external stimuli. Mitosis takes place in somatic cells.

Interphase is the interval time between two cell divisions. Mitosis mainly involves – prophase, metaphase, anaphase and telophase. The mitotic karyokinesis follows cytokinesis.

A. Interphase

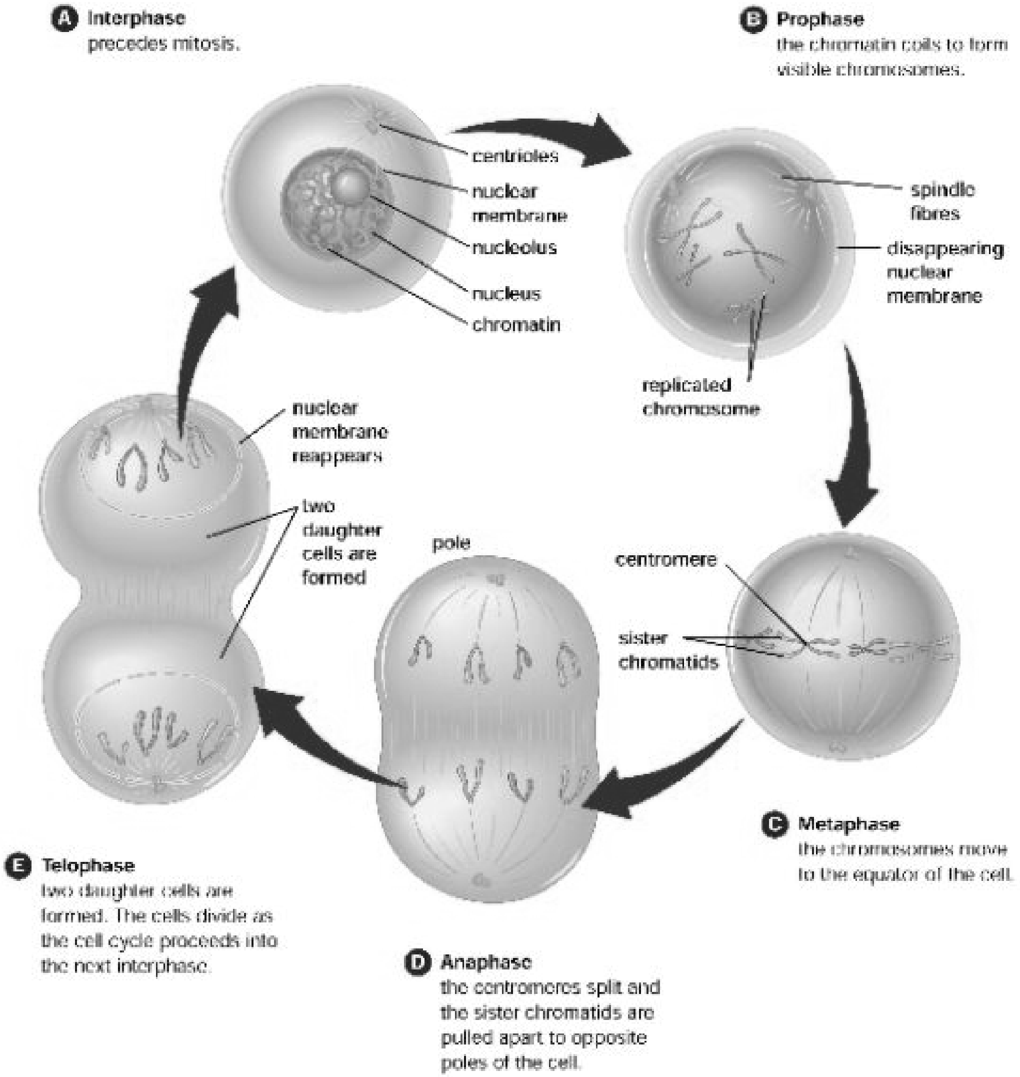
Many cells undergo a continuous alteration between division and non division. The period occurring from the competition of one cell division to the beginning of the next division is called interphase. In this phase cell growth and normal functions of cell will be continued. In the interphase along with normal biochemical activities replication of the DNA of each chromosome take place.

The period of DNA synthesis in interphase is called “S” phase and is separated in time from the previous cell division by a gap called G_1. After DNA synthesis a further gap called G_2 occurs before the next cell division.

Cytologically, Interphase is characterized by the absence of visible chromosomes. Instead of chromosomal form the unfolded and uncoiled chromatin form. Once G_1, S and G_2 are completed, mitosis is initiated. The mitosis is divided into different phase.

If the total cell cycle time is considered as for 16 hrs, within this 16 hrs, 15 hrs will be spent in interphase and only one hour in mitotic phase. With in one hour mitotic phase 36 minutes in prophase, 3 minutes in metaphase, 3 minutes in anaphase, and 18 mminutes in telophase.

Fig.1. Mitotic division in animal cells


A Interphase
precedes mitosis.
centrioles
nuclear membrane
nucleolus
nucleus
chromatin
B Prophase
the chromatin coils to form visible chromosomes.
spindle fibres
disappearing nuclear membrane
replicated chromosome
centromere
sister chromatids
C Metaphase
the chromosomes move to the equator of the cell.
pole
D Anaphase
the centromeres split and the sister chromatids are pulled apart to opposite poles of the cell.
nuclear membrane reappears
two daughter cells are formed
E Telophase
two daughter cells are formed. The cells divide as the cell cycle proceeds into the next interphase.

B. Prophase

Most of the mitotic phase time spent by the dividing cell in prophase. In the prophase the chromosomes are prepared for separation and gathers all the proteins and other biomolecules necessary for mitosis.

In the prophase of mitosis the main steps are :

1. Chromosomal material condenses to form compact mitotic chromosomes.

2. Chromosomes composed of two chromatids attached together at the centromere.

3. Mitotic spindle is assembled.

4. Cytoskeleton and nuclear envelope disappear.

5. Golgi complex and endoplasmic reticulum will be fragmented.

To separate the duplicated chromosomes, a cell makes them as shorter and thick fibers. At this stage chromosomes appear to be split longitudinally into two different strands, called as chromatids. The two chromatids of each chromosome are connected at the centromeres.

C. Prometaphase

The disappearance of nuclear envelope in prophase indicates the starting of next phase, prometaphase. During this phase mitotic spindle formation will be finished and the chromosomes are moved into the center of the cell.

The steps of mitosis involved in prometaphase:

1. Chromosomal microtubules attach to kinetochores of chromosome.
2. Chromosomes are moved to spindle equator.

D. Metaphase

The cell enters into metaphase when all the chromosomes are positioned at spindle equator. At this stage one chromatid of each chromosome connected to one pole and its sister chromatid connected to the opposite pole. Microtubules participates in separating two chromatids of a chromosome.

The steps in metaphase of mitosis:

1. The chromosomes are placed in the cell center by attaching chromosomal microtubules to both poles.

E. Anaphase

The mitotic anaphase begins when the centromere splits into two, allowing sister chromatids to separate and move towards the opposite poles. The arms of each chromosome drags behind their centromers, giving the characteristic metacentric, submetacentric and telocentric shape to each chromosome according to the location of centromere in it. The migration of chromatids to opposite poles of the cell is achieved by the contraction of chromosomal fibers and certain other cytoplasmic activities.

The step involves in Anaphase are

1. Centromeres split and chromatids separate
2. Chromosomes move to opposite spindle poles.
3. Spindle poles move for apart.

F. Telophase

The last phase in mitosis, telophase involves in assemble of identical set of chromosomes at each pole of the cell. The chromosomes begin to uncoil and return to the interphase condition. The mitotic spindle degenerates, the endoplasmic reticulum forms new nuclear envelops around both chromosomal sets. The nucleolus appears in each daughter nuclei.

The step of telophase are

1. Chromosomes cluster to opposite spindle poles
2. Chromosomes become dispersed
3. Nuclear envelop assembles around chromosome cluster
4. Golgi complex and Endoplasmic reticulum reforms
5. Cytokinesis

G. Cytokinesis

In animals the cytokinesis is accomplished by the formation of a cleavage furrow in the cytoplasm. The cleavage furrow is formed by the cyclosis movement of cytoplasm, formation of a contracting ring in the equator region, the expansion of plasma membrane and interaction of microtubules and the cell surface. The furrow deepens and pinches the cell into daughter cells.

H. Significance of Mitosis

Mitosis results in the precise, equal distribution of chromosomes from a parent nucleus to the daughter nuclei. It maintains an equilibrium in the amount of DNA contents in the cell. It replaces the aged cells.

1.2.4.3 Meiotic Cell Division

The type of cell division in which the diploid cell is divided into four haploid daughter cells is called meiosis. Meiosis must be highly specific since haploid gametes or spores must highly specific since haploid gametes or spores must contain precisely one member of each homologous pair of chromosomes. Successfully completed meiosis ensures genetic continuity from generation to generation. The general conception is, meiosis takes place in germ cells. In mitosis each paternally and maternally derived member of any given homologous pair of chromosomes behaves autonomously during division. But in meiosis homologous chromosomes pair up. Since each paired up structure contains two chromosomes it is called "bivalent". As the bivalent condenses both chromosomes making up it have been duplicated. This results in a tetrad consisting of two pairs of chromatids. In order to achieve haploidy two divisions in a single meiotic cell division occur – Meiosis I and Meiosis II. Meiosis I is reductional division, as the number of centromeres reduces to half in this division. Meiosis II is an equational division as the number of centromeres remains same after this division.

Fig 3. Meiotic cell division (Genetics – weaver)

A.Meiosis I

Before entering meiosis each cell remains in the interphase, where the genetic materials are duplicated due to replication. Meiosis I includes

1. Prophase I
2. Metaphase II
3. Anaphase I
4. Telophase I

1. Prophase I

In meiosis I prophase is longer than the mitotic prophase. Synapsis and crossing over occurs during prophase I. Depending on the morphology or activity of chromosomes with in the nuclear membrane the prophase I is divided into (a) leptonema, (b) zygonema, (c) pachynema and (d) diplonema and (e)diakinesis.

a. Leptonema (leptotene)

During this stage the chromosomes appear as long single threads, unassociated with one another. The centrioles move towards the opposite poles of the cell and a definite type of orientation and polarization of chromosomes towards the centrioles takes place.

b. Zygonema (zygotene or Synaptenes)

During this stage the homologous chromosomes pair with one another over the entire length of the chromosomes. The pairing of homologous chromosomes is called synapsis. Each pair of homologous chromosome is called bivalent.

c. Pachynema (Pachytene)

During this stage the paired chromosomes becomes shorter and thicker than the earlier stages and splits into two sister chromatids except at the region of centromere. This results in tetrad formation.

During this stage exchange of sister chromatids takes place called crossing over. This crossing over results in hereditary variations.

d. Diplonema (Diplotene)

During diplonema the points of interchange called chiasmata, move towards the ends of the synapsed chromosomes.

e. Diakinesis

During diakinesis the chromosomes begin to coil and so become shorter and thicker. The nucleolus detaches from the nuclear organization and disappears completely. The nuclear envelope starts to degenerate and spindle formation starts.

2. Metaphase I

The centromere of each chromosome of a tetrad is directed towards the opposite poles. The homologous microtubular spindle fibres remain attached with the centromeres and homologous chromosomes become ready to separate.

3. Anaphase I

In this stage the separation of whole chromosomes of a tetrad take place, so that each pole of the dividing cell receives either a parental or maternal longitudinally double chromosome of each tetrad.

4. Telophase I

During telophase I, the chromosome may persist for a time in the condensed state. The nucleolus and nuclear membrane will be reconstituted and sometimes cytokinesis may also occur to produce two cells.

B. Meiosis II

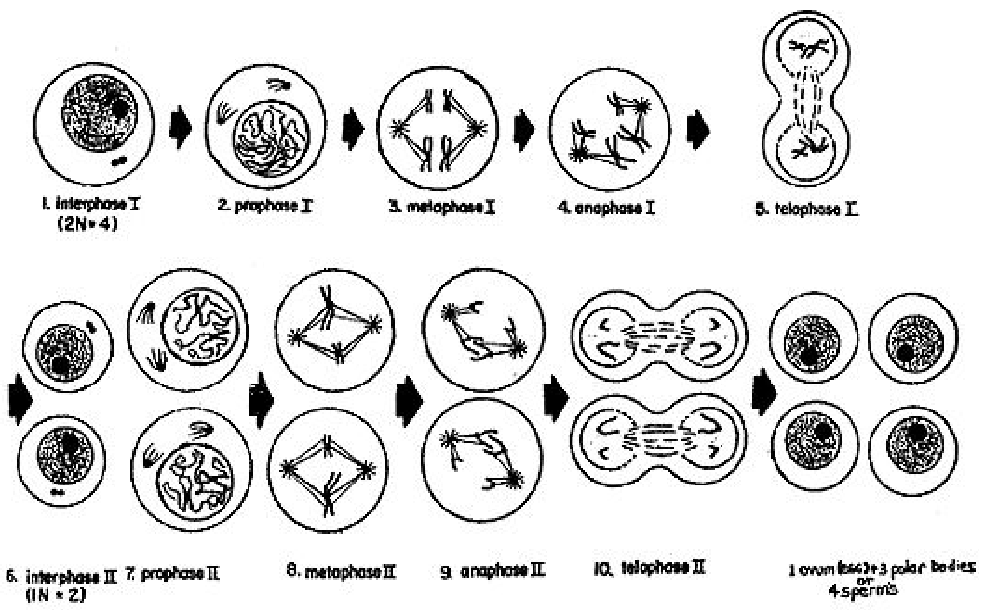
A second division is essential if each gamete is to receive only one chromatid from each original tetrad. This meiosis II also involves prophase II, metaphase II, anaphase II and telophase II.

During prophase II each diad is composed of one pair of sister chromatids attached by a common centromere. During metaphase II, the centromere is positioned on the equatorial plate, when they divide, anaphase II is initiated and the sister chromatids of each diad are pulled to opposite poles. Because the number of diads is equal to the haploid number telophase II separates one member of each pair of homologous chromosomes present at each pole. Each chromosome is referred to as a monod. Following cytokinesis in Telophase II, four haploid gametes may result from a single meiotic event.

D. Significance of

E. meiosis

Generalized Meiosis in an Animal Cell



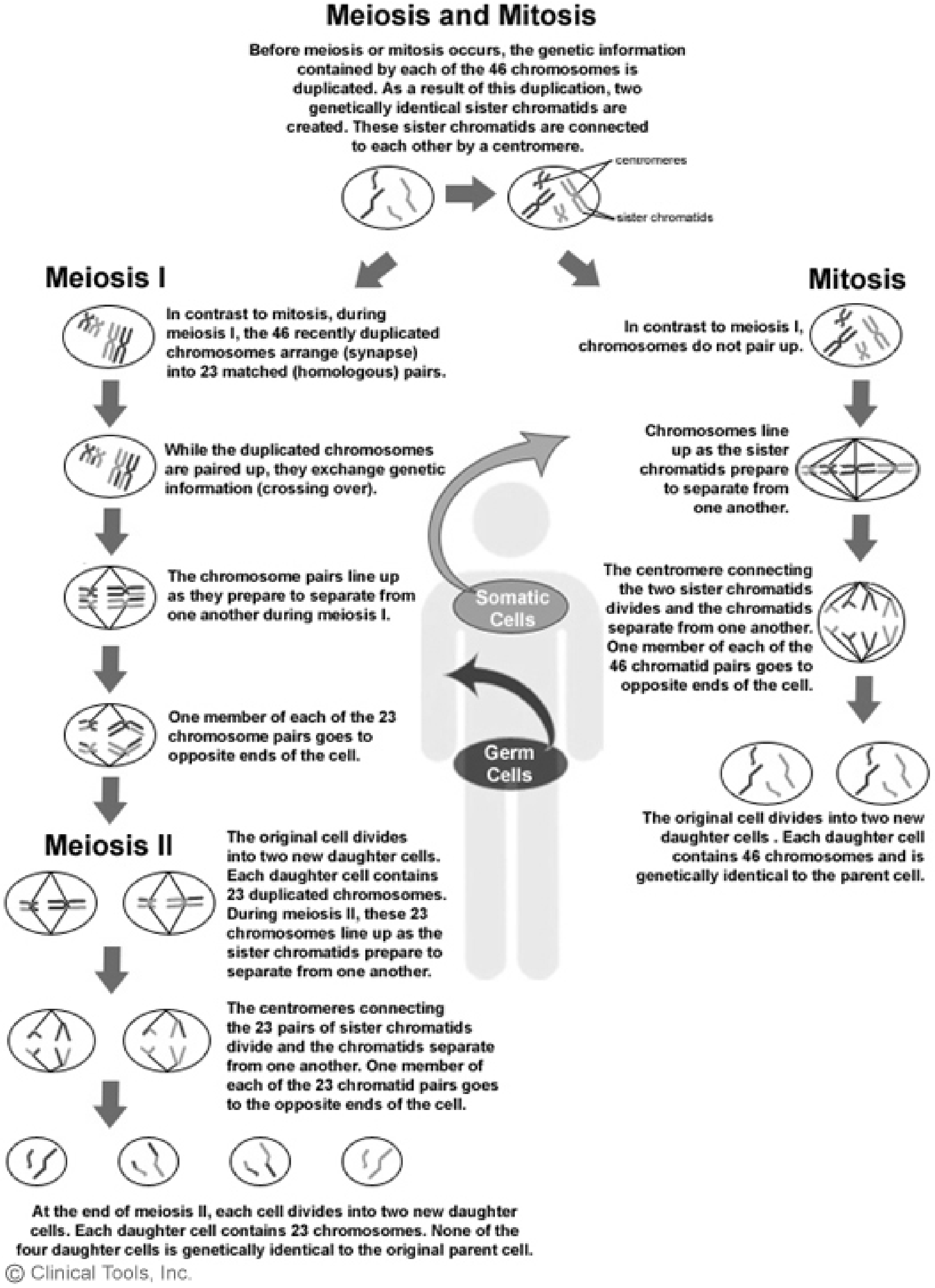
1. The formation of four haploid nuclei from a single diploid cell for balancing of doubling of chromosome number that results during fertilization.
2. The crossing over occurs in its prophase I, provides new combinations of genetic substance.

3. The two members of a homologous pair of chromosomes pass on two different daughter cells. This results in different combinations of chromosomes, so different characters in both daughter cells.

1.2.4.4 Comparision of Mitosis and Meiosis

1. While mitosis gives rise to two daughter cells, which are identical to each other as to the parent cell, meiosis gives rise to four daughter cells. These four cells resemble each other with respect to chromosome number they differ, since paternal and maternal chromosomes would reassort during first division and would also undergo exchange of chromosome segments during crossing over. The four daughter cells will also differ from the parent cell in having half the chromosome number.

2. In meiosis, first division is reductional and second division is equational. The mitotic division is purely equational.



Meiosis and Mitosis

Before meiosis or mitosis occurs, the genetic information contained by each of the 46 chromosomes is duplicated. As a result of this duplication, two genetically identical sister chromatids are created. These sister chromatids are connected to each other by a centromere.

Meiosis I

In contrast to mitosis, during meiosis I, the 46 recently duplicated chromosomes arrange (synapse) into 23 matched (homologous) pairs.

While the duplicated chromosomes are paired up, they exchange genetic information (crossing over).

The chromosome pairs line up as they prepare to separate from one another during meiosis I.

One member of each of the 23 chromosome pairs goes to opposite ends of the cell.

Meiosis II

The original cell divides into two new daughter cells. Each daughter cell contains 23 duplicated chromosomes. During meiosis II, these 23 chromosomes line up as the sister chromatids prepare to separate from one another.

The centromeres connecting the 23 pairs of sister chromatids divide and the chromatids separate from one another. One member of each of the 23 chromatid pairs goes to the opposite ends of the cell.

At the end of meiosis II, each cell divides into two new daughter cells. Each daughter cell contains 23 chromosomes. None of the four daughter cells is genetically identical to the original parent cell.

Mitosis

In contrast to meiosis I, chromosomes do not pair up.

Chromosomes line up as the sister chromatids prepare to separate from one another.

The centromere connecting the two sister chromatids divides and the chromatids separate from one another. One member of each of the 46 chromatid pairs goes to opposite ends of the cell.

The original cell divides into two new daughter cells . Each daughter cell contains 46 chromosomes and is genetically identical to the parent cell.

© Clinical Tools, Inc.

Fig.4. Comparision of mitosis and meiosis

3. In meiosis, homologous chromosomes undergo pairing. In mitosis no chromosome pairing takes place.

4. In meiosis, with each pair of homologous chromosomes forming a bivalent, a tripartite structure called synaptonemal complex. No such structure is observed in mitosis.

5. In meiosis, synthesis of small fraction of DNA is not completed in S phase, but is delayed till zygotene. No such delayed synthesis of DNA is observed in mitosis.

Summary

For the maintenance of life cell division is the compulsory process. In mitosis duplicated chromosomes are faithfully separated from one another, producing two nuclei, each with one copy of chromosome. In meiosis the diploid cell is divided into four haploid daughter cells. Crossing over in meiosis provides new combinations of genetic substance.

Technical terms

1. Synapsis – Pairing of homologous chromosomes.

2. Crossing over – Exchange of nucleotide sequences between sister chromatids.

3. Karyokinesis – Nuclear division of a cell

4. Cytokinesis – Splitting of a cell into two

Model questions

1. How the somatic cell divides – explain.

2. Write down the differences between meiotic and mitotic cell divisions.

Reference books

1. Genetics - P.K. Gupta

2. Principles of Genetics – Smustad and Simmons.

G.V.Padmavathi

Lesson 1.3.1

Introduction To Molecular Biology

1.3.1.1 Objective

The objective of the lesson is to explain in detail the molecular biology and its importance

1.3.1.2 Introduction

1.3.1.2.1Goals of Molecular Biology

The ultimate goal of molecular biology is ambitious: to understand the five basic cell behavior patterns (growth, division, specialization, movement, and interaction) in terms of the various molecules that are responsible for them. That is, molecular biology wants to generate a complete description of the structure, function, and interrelationships of the cell's macromolecules, and thereby to understand why living cells behave the way they do.

This goal might appear overly zealous. Yet the rate at which progress is being made often astonishes even the most optimistic scientists. One might, in fact, consider these years to represent a golden era for biology, with the field of molecular biology providing the main driving force.

Significant discoveries are emerging from research laboratories nearly every day and the front pages of national newspapers frequently herald exciting announcements of the identification of disease-causing genes, or promising biotechnology products, or new agricultural processes.

A few decades ago the most important discoveries in molecular biology were made using the simplest organisms (e.g., viruses and bacteria). Nowadays, however, equally important findings are regularly reported for both plants and mammals. A few key discoveries and the efforts of a small group of pioneering scientists have set the stage for the present era.

1.3.1.2.2 The Early Years

The term molecular biology was first used in 1938 by Warren Weaver. As Director of the Natural Sciences Section of the Rockefeller Foundation, he advocated that financial support be given to this "new branch of science—a new biology— **Molecular Biology.**" By that time, biochemists began to discover many fundamental intracellular chemical reactions, and to appreciate the importance of specific reactions and of protein structure in defining the numerous properties of cells. However, the development of molecular biology itself could not begin until this realization was reached:

The most productive advances would be made by studying "simple" systems such as bacteria and bacteriophages (bacterial viruses). Although bacteria and bacteriophages are still quite complicated, they are far simpler than animal cells. In fact, they enabled scientists to identify DNA as the molecule that contains most, if not all, of the genetic information of a cell. Although DNA was first described in 1869 by F. Miescher, its significance to cell function and the definitive proof that it is responsible for inherited traits did not come until almost a century later. The experimental evidence that led to the assignment of genes to DNA depended heavily on the use of bacteria and their viruses.

Once it became clear that DNA contained the chemical basis of heredity, it was not long before J. D. Watson and F. H. C. Crick offered a model for the physical structure of DNA. That model also proposed a mechanism for DNA replication and the spontaneous origin of mutations. Shortly thereafter, RNA was revealed to be an intermediate for the synthesis of enzymes and other proteins.

Following these discoveries, the new field of molecular genetics progressed rapidly in the late 1950s and early 1960s. It provided new concepts at a rate matched only by the development of quantum mechanics in the 1920s. The initial success and the accumulation of an enormous body of information enabled researchers to apply the techniques and powerful logical methods of molecular genetics to a variety of subjects: muscle and nerve function, membrane structure, the mode of action of antibiotics, cellular differentiation and development, immunology, and others. Faith in the basic uniformity of life processes was an important factor in this rapid growth. That is, it was believed that the fundamental biological principles that govern the activity of simple organisms, such as bacteria and viruses (organisms that lack an organized nucleus), must apply to more complex cells; only the details should vary. This faith has been amply justified by experimental results. In this book, prokaryotes and eukaryotes will often be discussed separately and compared and contrasted. Usually prokaryotes will be discussed first, because they are simpler. In keeping with this practice, we begin by briefly reviewing the properties of several model living systems, including bacteria and their viruses.

1.3.1.3 Model Biological Systems

The simplest living organism is, of course, the virus. As a minimalist life form, it consists of a DNA (or, in some cases, an RNA) inner core surrounded by a protein coat. The key to the virus's simplicity is its parasitic nature. It borrows functions from its host cell. The host is, for some kinds of viruses, a bacterial cell, while for others it is a plant cell, and for yet others it is an animal cell. Of these hosts, the bacterial cell is the simplest. We will review its general features here briefly, then return to viruses. Bacteria are free-living unicellular organisms. They have a single chromosome, which is not enclosed in a nucleus (they are prokaryotes), and, compared to eukaryotes, they are simple in their physical organization. For all practical purposes, a bacterium can be thought of as consisting of several thousand chemicals and a few organized particles, all in liquid solution, enclosed in a rigid cell wall.

Bacteria have many features that make them suitable objects for the study of fundamental biological processes. For example, they can be grown easily and rapidly and, compared to cells in multicellular organisms, they are relatively simple in their needs. The bacterium that has served the field of molecular biology best is *Escherichia coli* (usually referred to as *E. coli*), which divides every 20 minutes at 37^{O}C under optimal conditions. Thus, a single cell becomes 109 bacteria in about 20 hours!

Bacteria can be grown in a liquid growth medium or on a solid surface. A population growing in a liquid medium is called a bacterial culture. If the liquid is a complex extract of biological material, it is called a broth. If the growth medium is a simple mixture containing no organic compounds other than a carbon source, such as a sugar, it is called a minimal medium. A typical minimal medium contains each of the ions Na+, K+, Mg2+, Ca2+, NH+ 4, Cl-, HPO2- 4, SO2- 4, and a source of carbon (such as glucose, glycerol, or lactate). If a bacterium can grow in a minimal medium—that is, if it can synthesize *all* necessary organic substances, such as amino acids, vitamins, and lipids—the bacterium is said to be a prototroph. If any organic substances other than a carbon source must be added for growth to occur, the bacterium is termed an auxotroph.

Bacteria are commonly grown on solid surfaces. The earliest surface used for growing bacteria was a slice of raw potato. This was eventually replaced by agar, a jelling agent obtained from seaweed. Agar is resistant to the action of bacterial enzymes and, hence, is considered biologically inert. Figure 1-1 illustrates growth of bacteria on an agar surface.

Metabolism in bacteria is precisely regulated. Thus, bacteria represent the most efficient free-living organisms yet discovered. They rarely synthesize substances that are not needed. For example, the amino acid tryptophan is not formed if tryptophan is present in the growth medium, but when the tryptophan in the medium is used up, the tryptophan-synthesizing enzymatic system will be quickly activated. The systems responsible for the utilization of various energy sources are also efficiently regulated. A well-studied example is the metabolism of the sugar lactose as an alternate carbon source to glucose. Control of both tryptophan

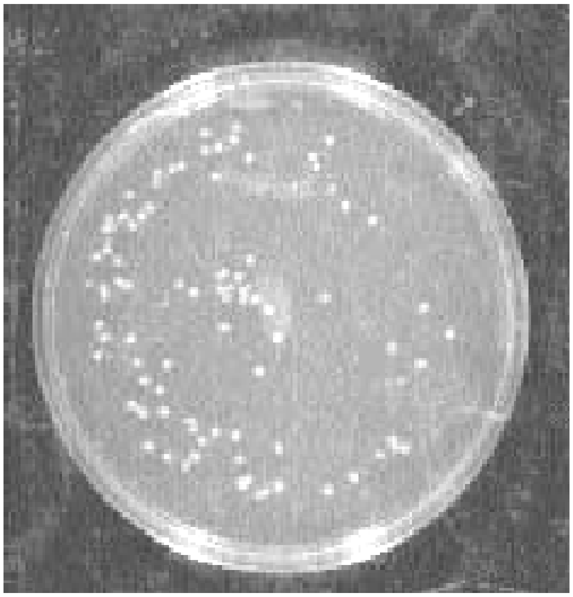


Figure 1-1 A glass dish displaying colonies of *E. coli* grown on the surface of an agar-containing growth medium. Each colony contains a cluster, or clone, of cells that represents the progeny of a single cell that divided many times. For instance, if 100 cells were spread on the agar surface, 100 colonies would appear the next day.

Synthesis and lactose degradation are two examples of metabolic regulation. This very general phenomenon will be explored extensively throughout the book, especially in Part III. Both simple and complex regulatory systems will be described, all of which act to determine how much of a particular compound is utilized and how much of each intracellular compound is synthesized at different times and in different circumstances. We will learn about the lengths to which the so-called simple cells go to utilize limited resources efficiently and to optimize their metabolic pathways for efficient growth. The ease with which bacteria can be grown in the laboratory on agar surfaces or even in liquid growth medium in large, industrial-scale tanks has facilitated rapid progress in the physical and chemical characterization of macromolecules. For example, the original descriptions of nucleic acids and proteins were largely made with components fractionated and isolated from mass quantities of bacterial cells. In addition, knowledge of metabolic regulation (Part III) can be traced back to early studies of bacteria, and more recently to plant and animal cells, grown under conditions similar to those illustrated in Figure 1-1.

1.3.1.3.1 Bacteriophage

Once the culture conditions for growing bacteria were established and many of the metabolic processes of normal bacterial life were known, bacterial viruses (bacteriophage, or the shortened form phage) were studied in earnest. Being much simpler than bacteria (unlike bacteria, they are not "free-living"), they could be studied in a relatively straightforward fashion. Since they indeed represent the simplest form of life, as their biochemical features became known, several physicists began to study them in the hopes of discovering new laws or first principles of physics! Figure 1-2 illustrates a bacteriophage that is relatively complex. It contains a protein coat, or phage head, to which is attached a tail. Some phages are simpler yet, and lack well-defined tail structures. Molecular biologists have utilized viruses as simple model systems for many kinds of studies. One of the most significant studies used the minimalist protein coat/DNA core composition of phage to establish whether protein or DNA carries heredity information. Chapter 6 will explain the life cycle of a typical bacteriophage and review how phage, like the one illustrated in Figure 1-2, were used in pioneering experiments that helped prove that DNA rather than protein contains genetic information.

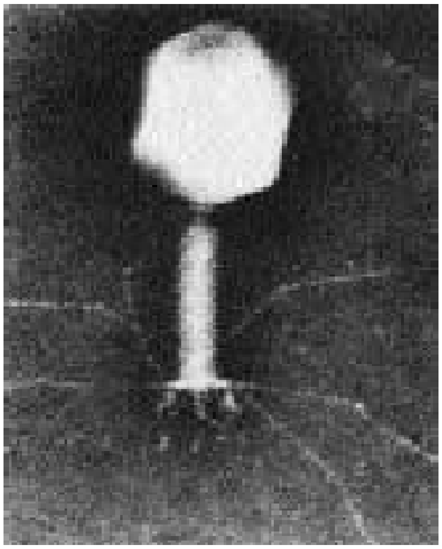


Figure 1-2 An *E. coli* T4 phage. The DNA is contained in the head. Tail fibers come from the pronged plate at the tip of the tail and serve to attach the virus to the host bacterium's surface.

1.3.1.3.2 Archaebacteria

In addition to those two broad categories of organisms, another group of prokaryotes has recently been designated—the Archaebacteria. They are probably modern descendants of a very ancient type of prokaryote. Features that distinguish them include gene expression and cell division mechanisms, which are intermediate between typical prokaryotes and eukaryotes. Microbiology texts should be consulted to gain an understanding of the ways in which they differ from both prokaryotes and eukaryotes, such as energy metabolism, environmental tolerance, and ribosomal RNA sequence.

1.3.1.3.3 Yeasts

Another favorable model system that shares many of the advantages of bacteria is yeast [Figure 1-3(a, b)]. They are, however, eukaryotes. Having a true nuclear membrane surrounding their chromosomes, they represent a higher level of organization than bacteria. That level approaches the complexity of animal (e.g., human) cells. Yet being a microorganism, yeast can be grown and manipulated much like *E. coli.* Yeasts have been used for millennia for producing wine and beer. A great deal of early biochemical research was carried out with yeasts rather than bacteria, work stimulated mainly by interest in understanding and improving beer. In contemporary molecular biology, mutant strains of yeast are often employed to discover genes that control growth, division, and cell behavior patterns. In addition, yeasts are presently employed as tools for producing large numbers of copies of human chromosome fragments. So-called “yeast artificial chromosomes” represent miniature chromosomes that contain foreign (e.g., human) DNA. They are propagated within yeast cells, providing the molecular biologist with opportunities for genetic engineering.

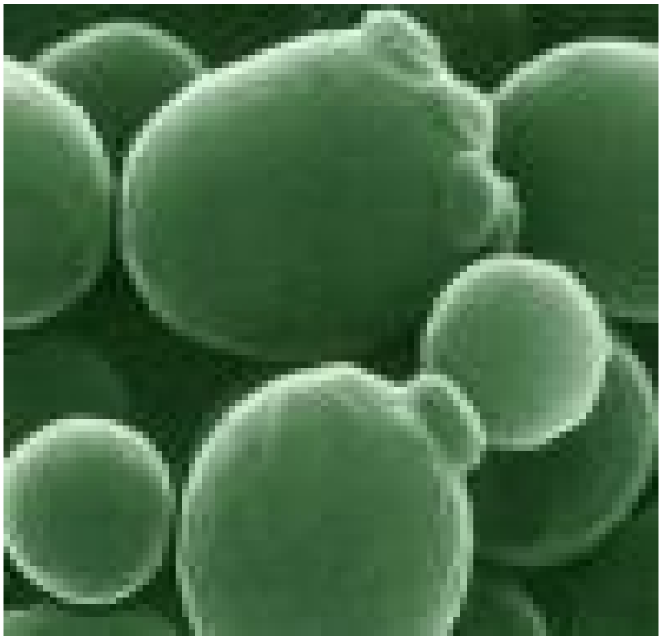


Figure 1-3 (a) A light micrograph of the yeast *Saccharomyces cerevisiae*. Many cells are budding by outgrowth from the cell wall of the mother.

1.3.1.3.4 Animal Cells (and Embryos)

Many types of animal cells, including several kinds of human cells, can be cultured using methods based on bacteria and yeast culture techniques. Primary cell cultures represent normal animal tissue that is usually derived from either skin cells (Figure 1-4) or embryos. They grow well initially, but eventually die off. Tumor cells, however, grow indefinitely, and are often easier to propagate in culture. Such tumor (or “transformed”) cells are therefore frequently used for routine laboratory experimentation.

Cells from early-stage mammalian (e.g., mouse) embryos provide an especially versatile model system. When cultured in an appropriate medium, they specialize in various ways that can be conveniently studied by the molecular biologist. These so-called embryonal stem cells can, in some instances, be grown in the presence of various culture medium supplements and thereby specialize into one or another cell type (e.g., muscle, nerve, liver, etc.). As it holds great promise as a starting point for the production of tissues and organs as human replacement parts, embryonal stem cell research is attracting great interest among molecular biologists. Molecular biologists also can introduce foreign genes into isolated embryonal stem cells and then reintroduce them into a growing (mouse) embryo. Often, the foreign genes will be expressed as the embryo develops, providing novel insights into the ways gene expression is regulated

In addition, molecular biologists have succeeded in injecting foreign genes directly into animal eggs and thereby generating transgenic animals. In the case of some animals, improved agricultural productivity, such as enhanced milk production from dairy cows, has been achieved. Some scientists speculate that similar procedures will one day be applicable to humans. Attempts will be made to inject foreign genes into human cells in order to correct genetic defects.

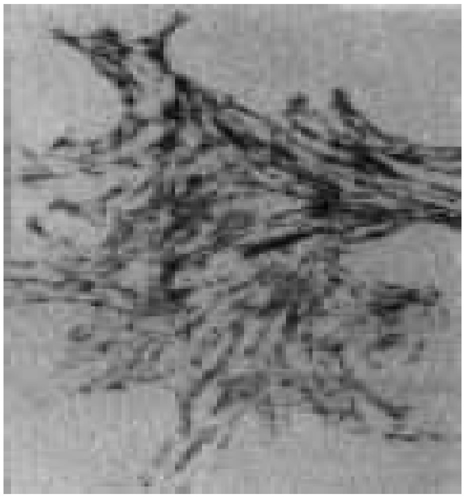


Figure 1-4 A microcolony of Chinese hamster fibroblasts that have been growing for a few days on a glass surface. (Courtesy of Theodore Puck.)

1.3.1.3.5 Plant Cells

The discovery that the cells of many plants can be nurtured and cultivated into whole, complete plants expanded the horizons of molecular biologists. Studies on the mechanism of action of plant hormones such as auxin were given a tremendous boost when cultivation methods for plant cells were perfected. More recently, a small simple plant, the common wall cress (*Arabidopsis*), which can be grown in a test tube, is being established as a tool for molecular genetics. It has a small amount of nuclear DNA, and exhibits several advantageous features (such as the ease with which genetic variants can be produced) that will be exploited by molecular biologists to learn about gene function in higher plants.

1.3.1.4 Methodology of Molecular Biology Methodology of Molecular Biology

For certain, the choice of appropriate research organisms was important for advancing the knowledge base of molecular biology. Especially advantageous was the fact that, despite the vast array of bacteria, viruses, and animal cells that were potentially available for laboratory experimentation, a consensus among researchers quickly emerged. A limited selection, which included *E. coli,* its bacteriophages, and only a small number of animal cell lines, were chosen as focal points. Thus, during the early years, researchers were able to conveniently exchange information and collaborate on achieving common goals, such as preparing genetic maps and understanding metabolic regulation

1.3.1.4.1 Physical and chemical characterization methods

The development of a vast array of laboratory methods was, however, equally important. In several cases, the application of those methods provided great leaps in the research scientist's ability to discover new features and functions of macromolecules. The originators of several of those breakthrough methods were awarded Nobel Prizes.

In the review of those methods, it should quickly become apparent that they provide ways to learn mainly about physical and chemical features of macromolecules. Most of them represent procedures for measuring or visualizing ever smaller quantities or tinier features of molecules, or parts of macromolecules.

Especially when applied in combination, the application of these methods has led to enormous advances in understanding the structural features of proteins and nucleic acids. For example, Electrophoresis of nucleic acids and hybridization have been coupled (e.g., Southern blotting) to learn about gene expression patterns in cells that are undergoing specialization. The polymerase chain reaction and DNA sequencing are often coupled to gain knowledge about the structure of a particular gene. From this knowledge, researchers can investigate the way cells regulate genes during growth.

1.3.1.4.2 Genetic Methods for Molecular Biology

Many of the most significant advances in molecular biology have come about through the application of genetic analysis to molecular phenomena. Initially, genetics provided an abstract, logical system for deducing relationships between specific genes and the metabolic processes for which they are responsible. Mutant organisms have traditionally provided the tools that allowed geneticists to develop models or schemes to explain how a metabolic process works, or how a structural component contributes to cell function.

In some cases, mutants have allowed the molecular biologist to recognize processes previously not known to exist (e.g., steps in DNA replication). In other instances, mutants have helped scientists unravel a complex metabolic pathway (e.g., carbohydrate metabolism,11). Perhaps most importantly, mutants often allow scientists to establish a direct correlation between an identified gene and its previously unknown function (e.g., termination step in protein synthesis,]

More recently, with the availability of pure genes prepared using methods, nucleotide sequences can be altered. Those alterations can be directed to specific nucleotides, either through simple pinpoint substitutions or through deletion. In these ways, the role particular nucleotide sequences play in determining protein structure and/or function can be directly tested. No longer is it necessary to rely exclusively on naturally occurring or laboratory-induced mutations in whole organisms. Beginning with a pure gene, a researcher can work back to a function in the intact organism (“reverse genetics”).

1.3.1.5 Rapid Progress in Molecular Biology

Direct experimentation using the model systems and laboratory methods just described has proven to be a powerful driving force in molecular biology. A variety of thought processes has emerged as the discipline has matured, and they guide today’s molecular biologist. Although certainly not exclusive to molecular biology, these thought processes are employed by most molecular biologists on an almost daily basis as data are collected, interpretations are formulated, conclusions are drawn, and additional experiments are designed. The following are examples of what constitutes the logic of molecular biology.

1.3.1.6 The Efficiency Argument

During the hundreds of millions of years that living organisms have been evolving, the rigors of competition for survival in an environment where resources are limited has led to a natural selection for efficiency. Accordingly, when attempting to understand how a mechanism or process works, molecular biologists intuitively favor the simplest, most efficient scheme over more complex, contrived, or cumbersome schemes. This approach has been especially useful for understanding the molecular mechanisms involved in regulating bacterial metabolism, RNA, DNA, and protein synthesis, as well as phage life cycles. However, analogous processes in the cells of higher plants and animals lack this intense competitive pressure between individual cells for survival. Simple, efficient mechanisms, therefore, do not always exist in the cells of higher organisms.

1.3.1.7 Phylogenetic History

When attempting to explain why a particular mechanism works the way it does, researchers often make comparisons with similar mechanisms in organisms that are either more or less advanced on the evolutionary scale. Species divergence is often due to changes in molecular processes. By comparing these processes among evolutionarily diverse species, biologists frequently gain a deeper nderstanding of molecular mechanisms.

1.3.1.8 Quantitative Assessment

The molecular biologist is constantly searching for quantitative relationships. In the process of unraveling a molecular mechanism, an assessment of the timing, chemical balance, and flow of components in space is routinely made in order to establish whether a proposed mechanism or hypothesis is feasible. Often, proposed pathways or schemes can be either approved or disregarded based on quantitative assessments.

1.3.1.9 Conceptual Model Building

Models attempt to explain complex hypotheses by casting them in a readily understandable form (often in the form of a diagram or cartoon). In this context, they often represent conceptual entities rather than the three-dimensional hardware associated with molecular model kits for constructing larger-than-life representations of the structure of one or another macromolecule (e.g., a protein). For example, models have been developed to explain how an antigen stimulates the immune system to produce antibodies that react specifically with the challenge antigen. Conceptual models are generally regarded as being good only if they can be subjected to a direct experimental test. That situation generally applies to the usefulness of almost any hypothesis. Molecular biologists are, therefore, constantly revising and refining their models as the interpretations of new experimental results call for modification.

1.3.1.10 Parallelism

Molecular biologists have great faith in the universality of basic biological processes. As a starting point in an attempt to understand a molecular mechanism or process, explanations that are satisfactory for simple systems (e.g., bacteriophages) are often scaled up to fit similar phenomena in more complex systems (e.g., mammalian cells). This approach does not always work, as has been demonstrated for messenger RNA processing, which is inherently different in eukaryotic cells. Nevertheless, newly discovered phenomena in molecular biology are often initially explained in terms of similar processes that had been analyzed earlier.

1.3.1.11 Strong Inference

Like all scientific enterprises, molecular biology is a human endeavor. Accordingly, intuition, which we might define here as a sort of comprehension based more on experience and common sense than on a conscious and elaborate line of logic argument, is often employed to develop a set of explanations or possible models for a particular molecular event. Strong inference is based on the exclusion of all but one final alternative explanation for a phenomenon. Initially, one states all the reasonable explanations for a particular phenomenon. Then, by direct experimentation, various possible explanations are ruled out one by one. The final alternative, which cannot be ruled out, and which satisfies our common sense, is considered most likely to be correct. This strong inference approach accounts for the relatively common use of the phrase “it is likely that” when experimental data are being interpreted. Despite the importance of intuition in “strong inference,” do not forget the critical role played by experimental evidence!

1.3.1.12 Optimism

Molecular biology thought processes enthusiastically endorse “optimism” as a key feature. That the use of the reductionist approach, which breaks a process down to its simplest components (e.g., molecules), will continue to be productive is a basic tenet of the discipline. Indeed, many molecular biologists believe that not only cellular functions but also increasingly higher-order biological phenomena (including perhaps even human behavior) will be comprehended in terms of molecular processes.

Rapid progress in molecular biology has generated literally thousands of volumes of detailed information about the structure and function of cellular components. Organizing those details in a comprehensive way represents a challenge for both the beginning student and the accomplished professional scientist. In order to develop a working knowledge of the discipline, details about molecular structure and function require attention, and generalizations need to be made

1.3.1.13 Putting the Details of Molecular Biology in Perspective

Molecular biology has profited greatly from the use of the so-called “reductionist” approach to the study of biological phenomena. That is, many molecular researchers believe the whole (organism) can only be fully understood when it is dissected and analyzed. Eventually, the lowest common denominators, such as the nature of the bonding between DNA and proteins, are revealed and general theories are constructed. Often, the analysis of one or another aspect of a single molecule (e.g., the enzyme lysozyme) has provided the focus of a large number of research laboratories, each following a similar reductionist theme

That approach represents a sharp contrast to the "holistic" approach, which gives special emphasis to interactions between components. The old adage—"the whole is greater than the sum of its parts"—applies to the holistic approach. This approach emphasizes the integration of various biological phenomena into higher order systems. In some cases, those systems are represented by individual types of organisms (e.g., the physiology of amphibia; the human embryo), while in other cases several organisms are integrated into even higher orders of organization, such as the disciplines of animal or plant ecology.

Because of the spectacular success of the reductionist approach in molecular biology, modern researchers instinctively favor reductionist strategies. However, beginning students need to appreciate the fact that individual molecules do not function in a vacuum. They function rather as components of complex gene expression mechanisms, in metabolic pathways, or as structural elements, always in
concert with other molecules.

Those various mechanisms, pathways, and structural elements themselves also operate in the context of other systems, which are themselves usually organized into cells. In eukaryotic organisms, those cells cluster to form tissues or organs that in turn comprise a single whole organism. Finally, whole organisms interact with one another and their environment to form an ecosystem. Individual molecules are, of course, the basis of all living activities. Nevertheless, in order to fully comprehend the function and significance of any single molecule, an excursion beyond the test tube and perhaps even out of the laboratory and on to biological field stations will eventually be necessary. In several instances, such a "vertical" integration of knowledge has generated spectacular results. One classic example is the protein hormone prolactin found in virtually all vertebrates. It was first discovered to stimulate the growth of the cells that line the crop in the pigeon. Later it was found to stimulate milk secretion in mammals. Then it was discovered to promote tail growth in aquatic amphibia. More recently, it was discovered to regulate salt balance in fish.

The boundaries of the holistic approach are even being extended. For example, data generated in the molecular biology laboratory from isolation and characterization procedures have been integrated into the research activities of other disciplines. Mitochondrial DNA research provides an interesting case— it has been discovered that mitochondrial DNA is inherited separately from chromosomal (nuclear) DNA. It also evolves quickly, since mitochondria lack DNA repair enzymes, and mutations, therefore, arise relatively often.

The nucleotide sequence of mitochondrial DNA thus changes rapidly. Molecular biologists and anthropologists have joined forces to use mitochondrial DNA sequence variation to trace lineages within human populations. Their studies have pinpointed Africa as the location where the human species originated. Since mitochondrial DNA is inherited only through the egg (sperm contribute no mitochondria to the embryo), recent research has revealed that in all likelihood each member of the human race (including you and I!) can be traced back to a single ancestral mother.

In order to keep the rapidly increasing volume of molecular biology details in proper perspective, *Essentials of Molecular Biology* will endeavor to highlight the various concepts that unify the many sub disciplines of the field.

Summary

Serious and diligent attention to learning molecular biology from this textbook offers several rewards to the beginning student. First, the use of the “layering approach” to constructing a knowledge base, which is used in this textbook (see the Preface), provides a learning tool that, once mastered, can be applied to other disciplines, ranging from poetry to physics. Second, studying molecular biology provides an excellent opportunity for enhancing one’s analytical thought processes (so-called “critical thinking skills”). From time to time, experimental data are encountered. Learning how to interpret it usually involves analyzing entries in tables or graphs, and thereby improves problem-solving skills. Third, by learning the concepts and details of molecular biology, the beginning biology student is provided a gateway to virtually all other disciplines in biology, ranging from cell biology to genetics to population biology. Finally, should you, the beginning student, develop an interest in molecular biology, you might consider choosing a career in a field related to this discipline.

Model questions

1.What is molecular biology and add a note on it?
2.What are the different Model organisams used to study the molecular Biology?

Reference Books

1.Cell and Molecular Biology By Lodish
2.The Cell by Bruce Albert
3. Molecular Biology By David Frifeilder

Sudhakar

LESSON 1.3.2
GENE DISCOVERY

Objective:

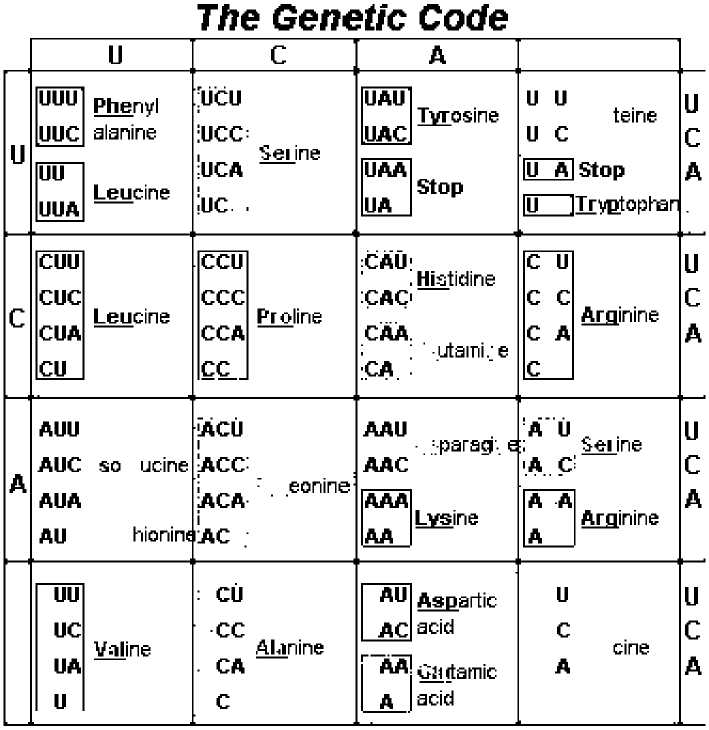
- To know about methods of gene discovery and their evaluation.
- To have a glance towords the protocol of discovery

1.3.2.1 Introduction

With the advent of whole-genome sequencing projects, there is considerable use for computer programs that scan genomic DNA sequences to find genes, particularly those that encode proteins. Once a new genomic sequence has been obtained, the most likely protein-encoding regions are identified and the predicted proteins are then subjected to a database similarity search. The genomic DNA sequence is then annotated with information on the exon–intron structure and location of each predicted gene along with any functional information based on the database searches. This procedure is summarized in the gene prediction flowchart.

The simplest method of finding DNA sequences that encode proteins is to search for open reading frames, or ORFs. An ORF is a length of DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid. There are six possible reading frames in every sequence, three starting at positions 1, 2, and 3 and going in the 5_ to 3_ direction of a given sequence, and another three starting at positions 1, 2, and 3 and going in the 5_ to 3_ direction of the complementary sequence. In prokaryotic genomes, DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated directly into proteins without significant modification. The longest ORFs running from the first available Met codon on the mRNA to the next stop codon in the same reading frame generally provide a good, but not assured prediction of the protein-encoding regions. A reading frame of a genomic sequence that does not encode a protein will have short ORFs due to the presence of many in-frame stop codons. These predictions have to take into account the observation in *E. coli* and its phages of the presence of multiple genes on mRNA and sometimes of overlapping genes in which two different proteins may be encoded in different reading frames of the same mRNA, either on the same or complementary DNA strands. In eukaryotes, prediction of protein-encoding genes is a more difficult task. In eukaryotic organisms, transcription of protein-encoding regions initiated at specific promoter sequences is followed by removal of noncoding sequence (introns) from premRNA by a splicing mechanism, leaving the protein-encoding exons. Once the introns have been removed and certain other modifications to the mature RNA have been made, the resulting mature mRNA can be translated in the 5_ to 3_ direction, usually from the first start codon to the first stop codon. As a result of the presence of intron sequences in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted by the presence of introns that usually generate stop codons. Three types of posttranscriptional events influence the translation of mRNA into protein and the accuracy of gene prediction. First, the genetic code of a given genome may vary from the universal code. For the most part, the universal genetic code, shown in Table, is used.

Table *The universal or standard genetic code*



The Genetic Code

	U	C	A		
U	UUU Phenyl UUC alanine UU UUA Leucine	UCU UCC Serine UCA UC	UAU Tyrosine UAC UAA Stop UA	U U teine U C U A Stop U Tryptophan	U C A
C	CUU CUC Leucine CUA CU	CCU CCC Proline CCA CC	CAU Histidine CAC CAA utami e CA	C U C C Arginine C A C	U C A
A	AUU AUC so ucine AUA AU hionine	ACU ACC eonine ACA AC	AAU paragi e AAC AAA Lysine AA	A U Serine A C A A Arginine A	U C A
	UU UC Valine UA U	CU CC Alanine CA C	AU Aspartic AC acid AA Glutamic A acid	U C cine A	U C A

Shown are each codon and the three-letter and one-letter codes for each encoded amino acid. ATG is the usual START codon and the three TER codons cause translational termination. Second, one tissue may splice a given mRNA differently from another, thus creating two similar but also partially different mRNAs encoding two related but partially different proteins. Understanding the molecular interactions between RNA and the RNA binding proteins that perform these modifications is an area of active investigation. Availability of this information will assist in the prediction of such variations. Third, mRNAs may be edited, changing the sequence of the mRNA and, as a result, of the encoded protein. Such changes also depend on interaction of RNA with RNA-binding proteins.

1.3.2.2 METHODS OF GENE PREDICTION

The major methods of gene prediction are as follows:

1. Laboratory-based approaches
2. Feature-based approach
3. Homology-based approach

Laboratory-Based Approaches to Gene Prediction

This is the traditional way to find a gene was to do it in the laboratory. Experimental procedures for

locating genes in new DNA are basically of three types:

1. Identification via hybridization to mRNA or cDNA.
2. Identification of the 5'-end and intron-exon junctions of the gene.
3. Exon trapping

Identification via hybridization to 'mRNA or cDNA

Northern blots

Northerns are the same as Southerns except that mRNA is run out on the Gel. Thus, transcripts resulting from expression of a gene can be detected and isolated to any given new DNA sequence by using a labeled probe of the same sequence as this new DNA sequence. This methodology can also be used to distinguish exons from introns by appropriate probe construction, although a more complete experimental approach is to sequence the mRNA via the cognate cDNA and compare the sequence directly with the genomic DNA sequence.

Zoo blots

Zoo blots are simply Southern blots of a labeled probe from the new DNA sequence against genomic DNA R.fragments from different organisms (the Zoo). The point is to determine if DNA sequences that are highly similar, and hence possibly homologous, to the new DNA sequence are present in one or more of these other organisms. An observed hybridization signal argues strongly that:

. The DNA probe comes from an intragenic region

. Both organisms encode homologous proteins that probably execute similar functions.

Zoo blots thus provide both gene location information as well as predictive gene function information.

Identification of the 5'-end and Intron-Exon Junctions of the Gene

S1 *Nuclease mapping and Primer Extension* (Fig. 12.2)

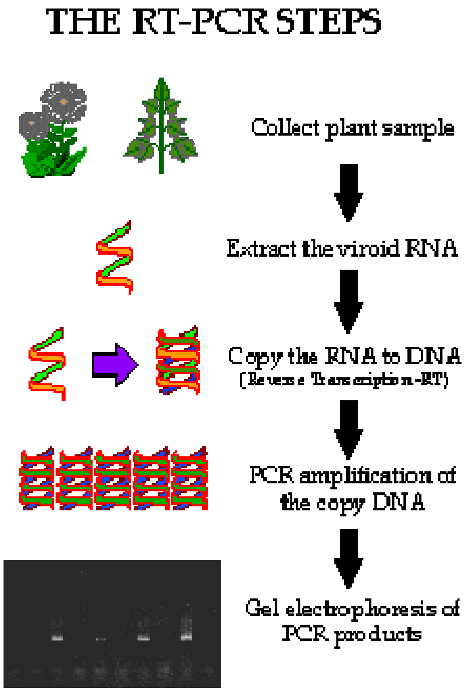
In S 1 nuclease mapping, a DNA probe labeled at its 5'-end and which overlaps the gene 5'-end or the 5'-end of an exon is hybridized to the gene DNA. S 1 nuclease, which is specific for either ssDNA or RNA, is used to digest the single-stranded DNA. The resulting 5'-labeled DNA probe is then "sized" via a Southern blot. Its size pinpoints the 5'-end of the gene or exon.

In primer extension, a DNA probe lab led at its 5'-end and which is contained within the gene is hybridized to mRNA from the gene. The probr DNA is used as a primer for a Reverse Transcriptase which will extend this primer, using the mRNA as Template, synthesizing DNA to the end of the mRNA. The resulting 5'-labled DNA probe, identical to the one produced in the S 1 nuclease mapping approach, is then "sized" via a Southern blot, to pinpoint the 5-end of the gene.

Exon Trapping

This method is used most often to isolate exons from new DNA, rather than simply to identify exon intron boundaries. In exon trapping, an R.fragment from a new DNA sequence is cloned into a cognate R.site in an intron of a cloned Gene; Cloning Vehicles have been constructed that make this relatively easy to do. This chimeric DNA is introduced into an appropriate eukaryotic host, usually *Yeast,* and the cloned gene is expressed. During processing of the initial transcript, introns are spliced out, leaving only the exon from the cloned R.fragment behind. DNA from this mRNA is obtained via R T - PCR, and its sequence determined. Comparison of this sequence, containing only the exon from the cloned R.fragment, with the sequence of the R.fragment itself shows where the intron-exon boundaries are located.

RT-PCR is a variation on PCR in which the first polymerization step of the PCR reaction is executed by a reverse transcriptase, thereby converting the mRNA into a cDNA. This cDNA is then amplified further using a thermostable DNA polymerase such as Taq polymerase. A problem with many laboratory methods is that relatively few genes tend to dominate the population of expressed sequences, and hence one discovers duplicates instead of new genes


THE RT-PCR STEPS
Collect plant sample
Extract the viroid RNA
Copy the RNA to DNA
(Reverse Transcription-RT)
PCR amplification of the copy DNA
Gel electrophoresis of PCR products

Feature-Based Approaches to Gene Prediction

Web-based gene recognition systems such as *Grail, GeneID,* and *GeneParser* work by searching for various ad hoc features of genes, and then identifying regions which score high enough. Typical features _clude codon bias, donor / acceptor sites, and coding frame length.

Since stop codons should occur every 20 codons or so, long *open reading frames* or ORFs without stop co dons are strongly suggestive of genes. The key to the analysis of an unknown DNA sequence is the identification of ORFs. ORF has the presence of a long series of codons in a DNA sequence without the series being interrupted by a termination codon. An ORF signal is enhanced even further by the presence of sequence patterns for starting and stopping transcription before and after the ORF. Dynamic programming can be used to identify the highest scoring regions. The best gene recognition systems tend to be species-specific, trained on examples of known genes in the given organism.

Scanning DNA for ORFs : The transcription initiation site is always an ATG codon and it is always about 30 base pairs downstream from a TAATAA sequence. This is enough infohnation to specify a pattern for the GCG program FINDP A TTERNS . It may be even easier to just produce a map of ORFs in all 6 reading frames and look for a long one. Simple software that maps an ORF starting at every A TG and stops it at every stop codon is available in a wide variety of forms.

GCG provides the FRAMES program. The MAP program can also be used to identify open reading frames. GeneWorks, MacVector, and Sequencher all handle this function quite elegantly. Introns can often be identified as breaks in ORFs and with moderate reliability by the occurrence of consensus splice signal sequences. However the only way to truly prove the existence of an intron is experimen. tally by comparing RNA (cDNA) to genomic sequences.

ORFs are easy to find with computers, however there are two major problems:

(i) *Small Proteins:* Even in prokaryotes, with no exons, what "cutoff' should be used for a mini. mum sized protein? In practice, a cutoff of 100 amino acids is often used. However, in so doing, some true small proteins containing fewer than 100 amino acids are not annotated and some ORFs containing more than 100 putative amino acids are annotated even though they in fact do not encode a protein.

(ii) *SmallExons:* Exons smaller than about 30 nucleotides cannot be reliably predicted by normal computational methods. However such exons do exist. Missing a small exon can result in pre. diction of a protein sequence that has an internal "frame shift", (i.e) the protein coding frame has shifted. Such a shift changes all the amino acids after the frame shift position, resulting in major errors in prediction of the protein sequence.

Three tests of ORFs have been devised to verify that a predicted ORF is in fact likely to encode a protein. These are described below:

1. This is based on an unusual type of sequence variation that is found in ORFs - every third base tends to be the same one much more often than by chance alone. This property is due to non. random use of codons in ORFs and is true for any ORF, independent of the species. The pro. gram TESTCODE (from GCG) provides a plot ofthe non-randomness of every third base in the sequence.
2. This is based on the analysis to determine whether the codons in the ORF correspond to those used in other genes of the same organism. For this test, information on codon use for an organ. ism is necessary, averaged over all genes.
3. The ORF may be translated into an amino acid sequence and the resulting sequence then compared to the databases of existing sequences. If one or more sequences of significant similarity are found, there will be much more confidence in the predicted ORFs.

Homology-Based Approaches to Gene Prediction

Searching for a known homolog is the most widely understood means of identifying new protein. coding genes. Such searches depend only on evolutionary relatedness, and so are widely applicable. A major advantage offmding homologous product is that some of the biology of the gene may be already elucidated. Usually databases are searched for ACRs (Ancient Conserved Region) and ESTs (Expressed Sequence Tags).

Hence evidence for genes can consist of matches to

. Known proteins

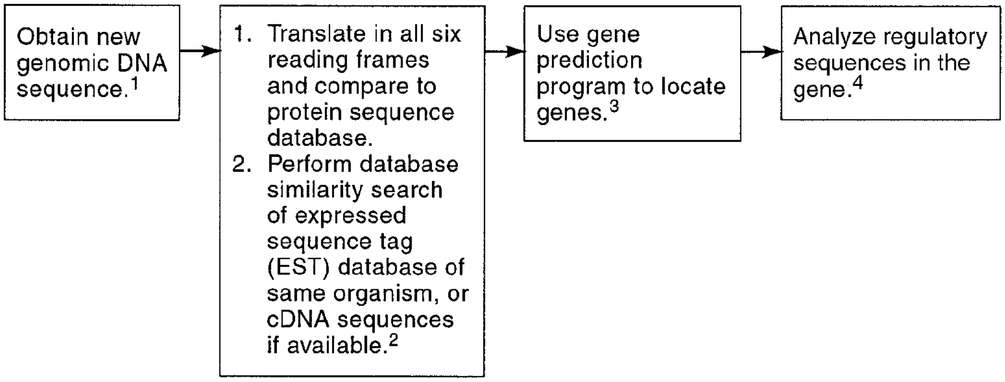
. Protein motifs (e.g. zinc finger, ATP and GTP-binding motifs, etc.)

. ESTs and ACRs

Homology-based gene prediction systems such as *Procrustes* scan databases find similarities to previously identified coding regions. Such homology-based approaches can only identify previously; known genes, of course, but the fraction of known genes is growing rapidly. A different homology \based approach to identify totally unknown genes is to compare two whole genomes and look for conserved regions, on the theory that sequence is only conserved if it is important.

,*Finding coding regions by similarity searching* :An approach to this problem is to translate the sequence in all six reading frames (3 forward and three reverse) and do a similarity search against the protein databanks. There is a variant of the BLAST program (BLASTX) that automatically translates a DNA query sequence and performs a similarity search against protein databanks. If a protein sequence matches, get its DNA sequence and align it with your unknown sequence. The start and stop codons should line up nicely. If the query sequence is genomic, then the introns should also be obvious.

1.3.2.3 Gene Prediction Methods Using Bioinformatics


Obtain new genomic DNA sequence.1
1. Translate in all six reading frames and compare to protein sequence database.
2. Perform database similarity search of expressed sequence tag (EST) database of same organism, or cDNA sequences if available.2
Use gene prediction program to locate genes.3
Analyze regulatory sequences in the gene.4

1.

The purpose of gene prediction is to identify regions of genomic DNA that encode proteins, although searches for RNA-encoding genes are also performed). The genomic DNA sequence may be that of an insert of genomic DNA in a bacterial artificial chromosome (BAC) or similar vector or that of an assembled chromosome or chromosomal fragment. Genome sequencing centers often search through newly acquired sequences with gene prediction programs and then annotate the sequence database entry with this information. This annotation includes gene location, gene structure (positions of predicted exons/introns and regulatory sites), and any matches of the translated exons with the protein sequence databases. The amino acid sequence of the predicted gene may also be entered in the protein sequence databases. Because the standards for identification are not uniform, and because gene predictions can be incorrect, it is a good idea to reconfirm any gene prediction of interest, perform alignments of the predicted sequence with matching database sequences to confirm statistical and biological significance, and confirm the predicted gene sequence by cDNA sequencing. If EST sequences are available in a sufficient coverage of the genome, these are also useful for confirmation of predicted gene sequences. The final goal of the gene annotation procedure for an organism is to produce a genome database that includes a rich supply of biological information on the function of each gene. This information will come from laboratory experimentation and manual entry of relevant published data into the genome database.

2. For genes of prokaryotic organisms, step 1 identifies open reading frames (ORFs, a series of amino acid-specifying codons) that encode a protein similar to one found in another organism. ORFs without a similar gene in another organism may also be found, as described in the text. Genes of eukaryotic organisms often have intron and exon sequences in the genomic DNA sequence. Step 1 provides the approximate locations of exons that encode a protein similar to one in another organism. Eukaryotic genomes may also have ORFs that do not match a database sequence, and these ORFs may or may not encode a protein. In the Genome Annotation Assessment Project (GASP) of the *Drosophila* genome, one study showed that combining gene prediction methods with homology searches generally provides a reliable annotation method. Step 2 is an additional type of database similarity search that identifies protein-encoding ORFs. Because cDNA sequences and partial cDNA sequences correspond to exons, genomic ORFs that can be aligned to these expressed gene sequences include exon sequences. This analysis can be enhanced by using databases of indexed genes in which overlapping ESTs have been identified . EST_GENOME is a program for aligning EST and cDNA sequences to genome sequences. Collections of EST sequences for an organism are often only partial collections; thus, failure to find a matching EST is not a sufficient criterion for rejecting an ORF by this test. Searching the EST collections of related organisms, e.g., another mammal or plant, may be helpful in identifying such missing EST sequences. An additional type of gene analysis is to use an already-identified

ORF as a query sequence in a database search against the entire proteome (all of the predicted proteins) of an organism to find families of paralogous genes.

3. There are a large number of gene prediction programs available. They all have in common to varying degrees the ability to differentiate between gene sequences characteristic of exons, introns, splicing sites, and other regulatory sites in expressed genes from other non-gene sequences that lack these patterns. Because these gene sequences as well as gene structure (the number and sizes of exons and introns) vary from one organism to another, a program trained on one organism, e.g., the bacterium *E. coli* or the worm *Caenorhabditis elegans*, is not generally useful for another organism, e.g., another bacterial species or the fruit fly *D. melanogaster*. Reliability tests of gene prediction programs have shown that the available methods for predicting known gene structure are, in general, error-prone. Referring to Web sites with this information or performing one's own reliability check is recommended. Some "reliability checks" should be eyed with suspicion because they are based on a comparison of new predictions with previous gene annotations. When gene predictions are made using gene-sized rather than large-sized, multigene sequence genomic DNA fragments, the predictions are generally more reliable.

4. In prokaryotes, the predicted genes may have conserved sequence patterns such as those for promoter recognition by RNA polymerases and transcription factors, for ribosomal binding to mRNA, or for termination of transcription, as found in the model prokaryote *E. coli*. Similarly, in eukaryotes, the region at the 5_ end of the gene may also have characteristic sequence patterns such as a high density and periodicity of putative transcription-factor-binding sites and sequence patterns characteristic of RNA polymerase II promoters. These types of analyses are enhanced by searching for similar sequence patterns in genes that are regulated by the same set of environmental conditions or that are expressed in the same tissue. Regulatory predictions are enhanced when information about conserved oligomers found in the promoters of co-regulated genes is available, as described in the text.

1.3.2.4 Gene Prediction In Microbial Genomes

Predicting protein-encoding genes is generally easier in prokaryotic than eukaryotic organisms because prokaryotes generally lack introns and because several quite highly conserved sequence patterns are found in the promoter region and around the start sites of transcription and translation, at least in the *E. coli* model of prokaryotes. When a set of different patterns characteristic of a gene are found in the same order and with the same spacing in an unknown sequence, the prediction is more reliable than if only one pattern is found, and this type of information can be obtained in *E. coli*.

1.3.2.5 Gene Prediction in Eukaryotes

A simple method for discovering protein-encoding genes within a eukaryotic genomic sequence is to perform a sequence database search by translating the sequence in all possible reading frames and comparing the sequence to a protein sequence database using the BLASTX or FASTX programs. Alternatively, if a genomic sequence is to be scanned for a gene encoding a particular protein, the protein can be compared to a nucleic acid sequence database that includes genomic sequences and is translated in all six possible reading frames by the TBLASTN or TFASTX/TFASTY programs. For proteins that are highly conserved, these methods can give a very good, albeit approximate, indication of the gene structure. If the proteins are not highly conserved, or if the exon structure of a gene is unusual, these methods may not work. Additional information as to the locations of genes in genomic DNA sequences may be found by using cDNA sequences of expressed genes.

1.3.2.6 Gene Prediction from a genomic DNA sequence

In this section we use several gene prediction programs on a particular genomic DNA sequence. For each of these programs we obtain a prediction of a candidate gene and we will analyze the differences between predictions and the annotation of the real gene. The programs we are going to use are geneid, genscan and fgenesh, which are available through a web interface. In these, and in many other tools in the web, we access a form where we can *paste*, or *submit*, the sequence we want to analyze, and then we press a button in the form that starts the computing process in some computer where the program runs. Once this process is finished, we get a new page in our browser with the results, which in this case should be a predicted gene.

A genomic DNA sequence

We are going to work with the sequence HS307871, which is stored in FASTA format. This sequence contains one gene, annotated in the following EMBL and NCBI records. Try to identify in these records the different pieces of information related to the annotation of the gene. geneid

In order to use geneid follow these steps:

1. Connect to the geneid server.

2. Paste the DNA sequence.

3. Select organism (*human*).

4. Run geneid with different (output) parameters:

Searching signals: Select acceptors, donors, start and stop codons. For each type of signal, try to find the real ones.

Searching exons: Select All exons and try to find the real ones.

Finding genes: You do not need to select any option (default behavior).

5. Compare the prediction with the real annotation. By taking a look to the graphical representation of the predicted sites and exons. By inspection of the output and the EMBL/NCBI record. By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.

6. Improve the prediction from some confirmed evidence. Below the text box where we pasted the DNA sequence, we can find a text box where we can paste *evidences*, which should consist of one or more exons (in GFF format) that are, e.g., experimentally confirmed. In this case we are going to paste as evidence the first exon which has not been predicted. Select and copy the GFF line corresponding to this exon contained in this file. Paste the line into the *evidences* text box, and run again geneid on the sequence. Compare the result with the real annotation. What has changed from the previous prediction?

genscan

In order to use genscan follow these steps:

1. Connect to the genscan server.
2. Paste the DNA sequence.
3. Select organism (*vertebrate*).
4. Compare the prediction with the real annotation.

By inspection of the output and the EMBL/NCBI record. By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.

fgenesh

In order to use fgenesh follow these steps:

1. Connect to the fgenesh server.
2. Paste the DNA sequence.
3. Select organism (*human*).
4. Compare the prediction with the real annotation.

By inspection of the output and the EMBL/NCBI record. By taking a look to the graphical representation of both, the output and the EMBL/NCBI annotation in this link.

Current annotations in the genomic DNA sequence

We can see the annotation of the gene together with the three predicted genes by geneid, genscan and fgenesh by following this link. Go to the page where we saw the NCBI record, click on the link *CDS*, and, next to the *Display* button, unroll the menu box and select the display option *FASTA*. Now press the button *Display*, and we will obtain the protein-coding DNA sequence of this gene in FASTA format. Select the entire sequence (first line is not necessary) and go to the UCSC genome BLAT search by following this link. In the big text box, paste the coding sequence we just copied, and press the *Submit* button on the top-right corner of this page. The result is a single match, where we find two links, *browser* and *details*. Visit first the *details* link and try to understand the the information provided there. Then go backwards and visit the *browser* link where we will see where this gene is located within the Human genome, as well as other annotated information as EST spliced alignments, etc.

In this section we will run several *ab initio* gene prediction programs on a particular genomic DNA sequence and we will compare the results against predicted genes from a gene finding program that uses genomic homology. For each of these programs we will obtain a prediction of a candidate gene and we will analyze the differences between predictions and the annotation of the real gene both in human and mouse. The programs we are going to use are geneid, genscan and fgenesh, which have been used in the previous practical exercise. blast will be used to compare human and mouse sequences. Finally, sgp2 (syntenic gene prediction tool) will predict genes taking into account the homology found between these two species.

A genomic DNA sequence

We are going to work with this Human sequence, which is stored in FASTA format. We also provide the homologous region in the mouse genome in this Mouse sequence.

Ab initio gene finding

In the first approach, we will use all the *ab initio* tools from the Gene Prediction section and compare the result of the three programs. You could open a simple word processor and paste the results of each gene-finding program in order to compare the coordinates of the predicted exons.

In order to use geneid follow these steps:

1. Connect to the geneid server.
2. Paste the DNA sequence.
3. Select organism (*human*)
4. Finding genes: You do not need to select any option (default behavior).

In order to use genscan follow these steps:

1. Connect to the genscan server.
2. Paste the DNA sequence.
3. Select organism (*vertebrate*)
4. Run gene predictions.

In order to use fgenesh follow these steps:

1. Connect to the fgenesh server.
2. Paste the DNA sequence.
3. Select organism (*human*)
4. Run gene prediction.

Using comparative gene finding tools

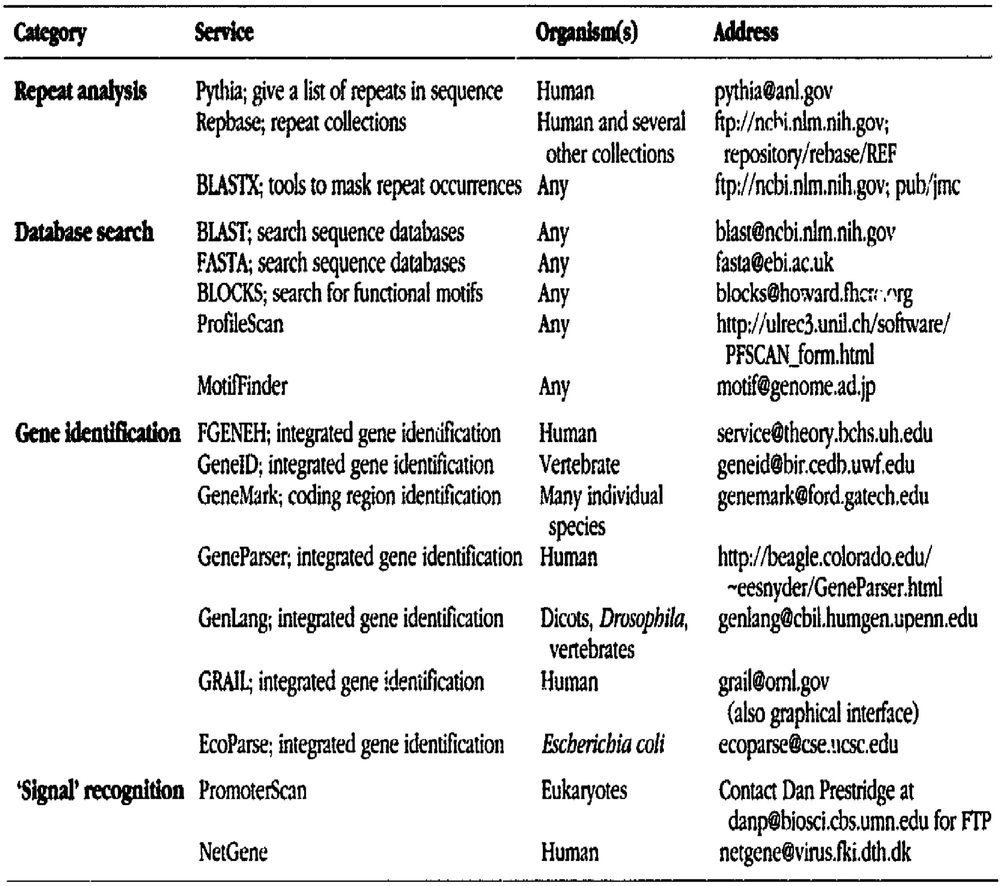
In this section we will use sgp2 to make the predictions using the conservation pattern between human and mouse.

In order to use blastn follow these steps:

1. Connect to the sgp2 server by following this link.
2. Paste the Human sequence in the "Sequence 1".
3. Paste the Mouse sequence in the "Sequence 2".
4. Select *Homo sapiens* vs *Mus musculus* parameters.
5. Select Prediction in both sequences.
6. Select geneid output format

Here you can find the human predictions, the mouse predictions and the human and mouse predictions with the tblastx similarity regions. There are other program that uses genomic comparison to improve gene prediction: twinscan and slam.

TABLE 1. Internet tools for gene discovery in DNA sequence data[a]



Category	Service	Organism(s)	Address
Repeat analysis	Pythia; give a list of repeats in sequence	Human	pythia@anl.gov
	Repbase; repeat collections	Human and several other collections	ftp://ncbi.nlm.nih.gov; repository/rebase/REF
	BLASTX; tools to mask repeat occurrences	Any	ftp://ncbi.nlm.nih.gov; pub/jmc
Database search	BLAST; search sequence databases	Any	blast@ncbi.nlm.nih.gov
	FASTA; search sequence databases	Any	fasta@ebi.ac.uk
	BLOCKS; search for functional motifs	Any	blocks@howard.fhcrc.org
	ProfileScan	Any	http://ulrec3.unil.ch/software/ PFSCAN_form.html
	MotifFinder	Any	motif@genome.ad.jp
Gene identification	FGENEH; integrated gene identification	Human	service@theory.bchs.uh.edu
	GeneID; integrated gene identification	Vertebrate	geneid@bir.cedb.uwf.edu
	GeneMark; coding region identification	Many individual species	genemark@ford.gatech.edu
	GeneParser; integrated gene identification	Human	http://beagle.colorado.edu/ ~eesnyder/GeneParser.html
	GenLang; integrated gene identification	Dicots, *Drosophila*, vertebrates	genlang@cbil.humgen.upenn.edu
	GRAIL; integrated gene identification	Human	grail@ornl.gov (also graphical interface)
	EcoParse; integrated gene identification	*Escherichia coli*	ecoparse@cse.ucsc.edu
'Signal' recognition	PromoterScan	Eukaryotes	Contact Dan Prestridge at danp@biosci.cbs.umn.edu for FTP
	NetGene	Human	netgene@virus.fki.dth.dk

Summary:

The simplest method of finding DNA sequences that encode proteins is to search for open reading frames, or ORFs. An ORF is a length of DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid. A simple method for discovering protein-encoding genes within a eukaryotic genomic sequence is to perform a sequence database search by translating the sequence in all possible reading frames and comparing the sequence to a protein sequence database using the BLASTX or FASTX programs.

Model Questions:

1. Briefly explain the methods of gene discovery?
2. Outline the steps involved in Gene Discovery?

References:

1. Bioinformatics, Concepts, Skills, and Applications by S.C.Rastogi & NAmita MEndiratta.
2. Bioinformatics – A practical guide to the Analysis of Genes and Proteins – Andreas D. Baxevanis and B F Francis Quelette.
3. Bioinformatics - Sequence and Genome Analysis- David W. Mount.
4. A review on finding genes by computer by JAMES W. FICKEI'C

B.M.REDDY **M.Tech. (HBTI, Kanpur)**

Lesson 1.3.3

GENETIC CODE

Objective :

How the 20 proteins will be coded by total 4 nucleotide bases, this concept is called the genetic code. Cracking of genetic code & properties of genetic code were discussed.

1.3.3.1 Introduction

Three major advances set the stage for our present knowledge of protein bio-synthesis.

1. In the early 1950s Paul Zamecnik and his colleagues designed a set of experiments to investigate where in the cell, proteins are synthesized.

They injected radioactive aminoacids into rats and, at different time intervals after injection, the liver were removed, homogenized and fractionated by centrifugation. The sub cellular fractions were then examined for the presence of radioactive protein.

When hours or days were allowed to elapse after injection of the labeled aminoacids, all the subcellular fractions contained labeled proteins. However, when only minutes had elapsed, labeled protein was found only in a fraction containing small ribonucleo protein particles

These particles visible in animal tissues by electron microscopy, were therefore identified as the site of protein synthesis from aminoacids, and later were named as ribosomes.

2. The second key advance was made by Mahlon Hogland and Zamecnik, when they found that aminoacids were activated when incubated with ATP and the Cytosolic fraction of liver cells. The aminoacids became attached to a heatstable RNA later called tRNA forming aminoayl – tRNA.

3. The third major advance occurred with the rise of a question, how genetic information encoded in the four-letter sequence codes for 20 aminoacids of proteins.

Crick reasoned that a small nucleic acid could serve the role as an adaptor, one part of adapter molecule binding a specific aminoacid and another part recognizing the nucleotide sequence encoding that aminoacid in mRNA. This idea was soon verified. The tRNA adapter translates the nucleotide sequence of an mRNA into the aminoacid sequence of a polypeptide.

These three developments soon lead to recognition of the major stages of protein syntheis and ultimately to the elucidation of the genetic code that specifies each aminoacid.

1.3.3.2 Genetic Code

The sequence of nucleotides in the mRNA molecule is read consecutively in groups of three. Since RNA is a linear polymer of 4 different bases, there are 4^3=64 possible combinations. Each group of three consecutive nucleotides in RNA is called a CODON , and each specifies.

The mRNA codons cannot directly recognize aminoacids, Rather they specifically bind molecules of tRNA that each carry a corresponding aminoacid. Each tRNA contains a trinucleotide sequence anticodon, which is complementary to an mRNA codon specifying the tRNAS's aminoacid.

In 1961, Marshall Nirenberg and Himrich Matthaei reported an observation, that solved the genetic code. They used i*nvitro* translation systems. Basically such systems generally include the following:

- Ribosomes, mRNA, and Mg++ - the polysome component
- Aminoacids, activating enzymes, 20 different tRNAs, ATP and mg++ to provide activated aminoacids.
- Protein factors (initiation, elongation and termination factors) – auxillary requirements.
- Phosphoenol pyruvate and pyruvate kinase – ATP generation system.
- GTP, Thiol compounds – auxillary requirements.

They incubated the synthetic polyribonucleotides polyuridylate(poly-U) with an E.coli extract, GTP and a mixture of 20 aminoacids in 20 different tubes. In each tube a different aminoacid is radiolabelled. A radio active polypeptide was formed in only one of the 20 tubes, that containing radioactive phenylalanine. Nirenberg and Matthari therefore concluded that the triplet UUU codes for phenyl alanine. The same approach revealed that synthetic poly © codes for proline and poly(A) codes for poly Lysine.

The synthetic polynucleotides used in such experiments were made by the action of polynucleotide phosphonylase. If it is presented with a mixture of 5 parts of ADP and I part of CDP, it will make a polymer in which about 5/6th, if the residues are adenylates and 1/6th cytidylates. Such a random polymer is likely to have many triplets of the sequence AAA, lesser numbers of AAc, CAA, relatively few ACC, CCA, CAC and very few CCC triplets.

AAC	:	Asn
AAC	:	Gln
AC	:	His
AAA	:	Lys
AAC, CCC	:	Pro
CCA	:	Thr

With the use of different artificial mRNAs made by polynucleotide phosphorylase, from different starting mixtures of ADP, GDP, UDP and CDP, the base compositions of the triplets were soon identified. However, these experiments could not reveal the sequence of the bases in each coding triplet.

1.3.3.3 Triplet binding assay

In 1964, Nirenberg and Philip Leder found that isolated E.Coli ribosomes will bound a specific aminoacyl tRNA, if the corresponding synthetic polynucleotide messenger is present.

For example, ribosomes incubated with poly(u) and phenylalanyl –tRNA ^{phe} binds both RNAs, but if the ribosomes are incubated with poly (U) and some other aminoacyl tRNA, the aminoacyl-tRNA is not bound because it doesnot recognise the UUU triplets in poly(U)

Trinucleotide	C^{14} – labeled aminoacyl –tRNA bound to ribosomes		
	Phe-tRNA^{phe}	Lys-tRNA^{lys}	Pro-tRNA^{pro}
UUU	4.6	0	0
AAA	0	7.7	0
CCC	0	0	3.1

Each number represents the factor by which the amount of bound ^{14}C-increased when the indicated trinucleotide was present, relative to a control is which no trinucleotide was added.

Even trinucleotides could promote specific binding of appropriate tRNAs, allowing the use of chemically synthesized oligonucleotides. Researchers identified, aminoacyl –tRNAs bound to 50 of the 64 possible triplet codons. For some codons, either no aminocyl-tRNAs or more than one would bind.

Chemical Methods

H.gobind Khorana, developed chemical methods to synthesize polyribonucleotides with defined, repeating sequences of 2-4 bases. The polypeptide produced by these mRNAs had one or few animoacids in repeating patterns. These patterns, when combines with information from the random polymers used by Nirenberg and colleagues permitted unambiguous codon assignments.

Example

The copolymer$(AC)_n$, has alternating ACA and CAC codons: ACA CAC ACA CAC. The polypeptide synthesized from this messenger contained equal amounts of Threonine and Histidine. Given that Histidine codon has One A & Two C, CAC must code for his and ACA for threonine.

An RNA with three bases in a repeating pattern should give three different types of polypeptide, each derived from a different reading frame and containing a single type of aminoacid.

Example : $(GUA)_n$

GUA GUA GUA GUA

An RNA with 4 bases in a repeating pattern should yield a single type of polypeptide with a repeating pattern of four aminoacids. Consolidation of the results from all such experiments permitted the assignment of 61 of the 64 codons. The other three were identified as termination codons, because they disrupted aminoacid coding patterns, when they occurred in a synthetic RNA polymer.

Example: $(GUA)_n$

Reading frame1:5^1 ....GUA AGU AAG UAA GUA AGU AA ....3^1

Reading frame2:5^1.........GUAA GUA AGU AAG UAA GUA A.......3^1

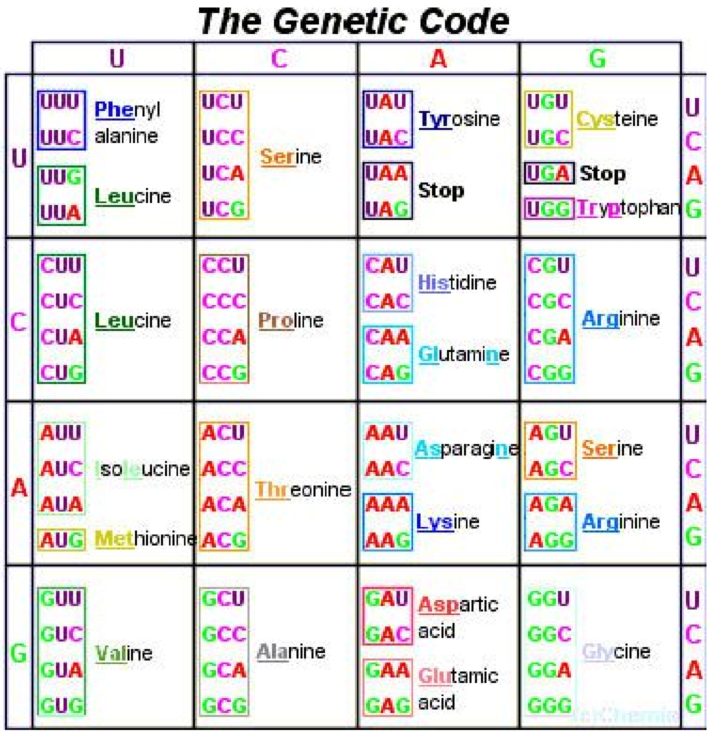
Reading frame3:5^1..........GU AAG UAA GUA AGU AAG UAA.....3^1

Dipeptides and tripeptides are synthesized depending on where the ribosome initially binds. Termination codons encountered every fourth codon in all three reading frames.

Meanings for all the triplet codons were established by 1966 and have been verified in many different ways.

1.3.3.4 START and STOP codons

Codons are the key to translation of genetic information, allowing the synthesis of specific proteins. Several codons serve special functions. The initiation codon, AUG signals the beginning of a polypeptide in all cells, in addition to coding for Methionine residues in internal positions of polypeptides. Three of the 64 possible codons do not code for protein.



The Genetic Code

	U	C	A	G	
U	UUU Phenyl alanine	UCU Serine	UAU Tyrosine	UGU Cysteine	U
	UUC	UCC	UAC	UGC	C
	UUG Leucine	UCA	UAA Stop	UGA Stop	A
	UUA	UCG	UAG	UGG Tryptophan	G
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Glutamine	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lysine	AGA Arginine	A
	AUG Methionine	ACG	AAG	AGG	G
G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glutamic acid	GGA	A
	GUG	GCG	GAG	GGG	G

Table: Dictionary of aminoacid code words

Any known aminoacids. These termination codons (stop/non sense codons) normally signal the end of polypeptide synthesis.

In a random sequence of nucleotides, one in every 20 codons in each reading frame, or on average a termination codon. In general, a reading frame without a termination codon among 50 or more codons is called an open reading frame (ORF). Long ORFs usually correspond for a typical protein with a molecular weight of 60,000 would require an ORF with 500/more codons.

1.3.3.5 Codon Degeneracy

The code is highly degenerate. Three aminoacids-Arginine, Leucine and serine are each specified by six codons and most of the rest are specified by either 3, or 2 codons. Only Methionine and Tryptophan, two of the least common aminoacids in proteins, are represented by a single codon. Codons that specify the same aminoacids are termed synonyms

When an amino acid has multiple codons. The difference between the codons usually lies in the third base (at the 3'end) for example, Alanine is coded by the triplets GCU,GCC, GCA & GCG

-XYC and XYU always specify the same aminoacid

-XYG and XYA do so in all but two cases

AUA Ile	UAA	
AUG Met	UAG	stop codons
UGG Trp	UGA	

- codons with a second position pyrimidine encode mostly hydrophobic aminoacids.
- codons with second position purines encode mostly polar aminoacids.

As a consequence of the genetic codes' degeneracy, many point mutations at a 3rd codon position are phenotypically silent, i.e. the mutated codon specifies the same aminoacid as the wild type.

1.3.3.6 Wobble hypothesis

In protein synthesis, the proper tRNA is selected only through codon-anticodon interactions; the aminoacyl groups does not participate in this process. Many tRNAs bind to 2nd or 3rd of the codons specifying their cognate aminoacids.

For example:

- yeast tRNA^{phe}, which has the anticodon mGAA, recognizes the codon UUU and UUC

anticodon : 3` – A-A-Gm-5`   3`-A-A-Gm-5`

codon: 5`-U-U-C-3`   5`-U U-U-3`

-yeast tRNA ^{Ala}, which has the anticodon IGC recognizes the codons GCU, GCC, GCA

anticodon: 3`-C-G-3-5`   3`-D G I – 5`   3`-C G I – 5`

codon : 5`-G-C-U-3`   5`-G C C –3`   5`-G C A-3`

The 3rd base of the codons form weak H-bonds with Inosinate residue at the 1st position of anticodon. Examination of such codon-anticodon pairings lead crick to conclude that the 3rd base of most codons pair rather loosely with the corresponding base of its anticodon. The third bases of such codons were called `wobble' crick proposed a set of relationships called the wobble hypothesis

1. The first two bases of codon is mRNA always form strong Watson-crick basepairs with the corresponding bases of the anticodon and confer most of the coding specificity.

2. The first base of some anticodons (in 5^1 -3^1 direction) determines the number of codons read by a given tRNA.

5^1 -anticodon base	3^1 -codon base
C	G
A	U
U	A or G
G	U or C
I	U, C or A

3. When an aminoacid is specified by several different codons, those codons that differ in either of the first two bases require different t RNAs.

4. A minimum of 32 tRNAs are required to translate all 61 codons

The wobble base of the codon contributes to specificity, but because if pairs only loosely with its corresponding base in the anticodon, it permits rapid dissociation of the tRNA from its codon during protein synthesis. It all three form strong Watson-crick base pairs. t RNAS would dissociate too slowly and severly limit the rate of protein synthesis. This codon-anticodon interactions optimize both accuracy and speed.

1.3.3.7 Universality

For many years it was thought that the sandard genetic code was universal. This assumption was based on the observation that one kind of organism (ex:E.coli) can accurately translate the genes from quite different organisms (ex:humans). DNA sequencing studies in 1981 revealed that the genetic codes of certain mitochondria are variants of the standard genetic code.

For example, in mammalian mitochondria AUA as well as the standard AUG, is a methionene or initiation codon.

*UGA specifies Trp rather than stop

*AGA and AGG are `stop' rather than Arg

*more recent studies, however revealed that in ciliated protozoa, the codons UAA and UGA specify G/N rather than stop.

Normalcode Assigned :	UGA Stop	AUA Ile	AG^AG Arg	CUN Leu	CCG Arg
Vertebrates :	Trp	Met	Stop	+	+
Drosophila :	Trp	Met	Ser	+	+
S.cerevisiae :	Trp	Met	+	Thr	+
Filamentous fungi :	Trp	+	+	+	+
Trypanosomes :	Trp	+	+	+	+
Higher plants :	+	+	+	+	Trp

+ - the codon has same meaning as in 1 normal code

** in E.coli UGA some times codes for selenocysteine, in addition to termination

At any rate, the standard genetic code is widely used, but not universal. The limited scope of code variants strengthens the principle that all life on this planet evolved on the basis of a single (very slightly flexible) genetic code.

Summary

The sequence of nucleotides in the mRNA molecule is read consecutively in groups of three nucleotides known as codons .The mRNA codons cannot directly recognize aminoacids but Each tRNA contains a trinucleotide sequence anticodon, which is complementary to an mRNA codon specifying the tRNS's aminoacid.In 1961, Marshall nirenberg and Himrich Matthaei reported an observation, that solved the genetic code..They decipher the genetic code by using homopolymers(poly-U,polyC,poly-A),tripletbindingassay,and polymers.the genetic code is universal and degenerate.In somecodons even though the third base is varid is recognized by same t-RNA which is known as wobble hypothesis

Model Questions

1) Explain the properties of genetic code

Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork.

Dr.N.Srinivasa Reddy,

Lesson 1.3.4

DNA AS HERIDITARY MOLECULE AND ITS STRUCTURE

Objective

This chapter explains the different experiments that proves DNA as hereditary material and its structure is also clearly explained.

1.3.4.1 Introduction

In 1865, Mendel showed that genes transmitted genetic information. The classical genetics of early twentieth century showed that the genetic material must perform three essential functions.

1. Replication – genotypic function
2. Gene expression – phenotypic function
3. Mutation – evolutionary function

Other early genetic studies established a precise correlation between the patterns of transmission of genes and the behaviour of chromosomes during sexual reproduction, providing strong evidence that genes are usually located on chromosomes.

Chromosomes are composed of two types of large organic molecules (macromolecules) called proteins and nucleicacids. The nucleic acids are of two types : deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA). During the 1940s and early 1950s, the results of elegant experiments clearly established that the genetic information is stored in nucleic acids, not in proteins. In most organisms, the genetic information is encoded in the structure of DNA. However, in many small viruses, the genetic information is encoded in RNA.

1.3.4.2 Proof that genetic information is stored in DNA

Several lines of indirect evidences suggested that DNA harbors the genetic information of living organisms.

For example:

- Most of the cell's DNA is located in the chromosomes, whereas RNA and proteins are also abundant in cytoplasm.

- A precise correlation exists between the amount of DNA per cell and the number of sets of chromosomes per cell.

- Most somatic cells of diploid organisms contain twice the amount of DNA as the haploid germ cells (gametes) of the same species.

- The molecular composition of the DNA is the same (with rare exceptions) in all the cells of an organism, where as the composition of RNA and proteins is highly variable from one cell type to another.

- DNA is more stable than RNA or proteins.

Transformation

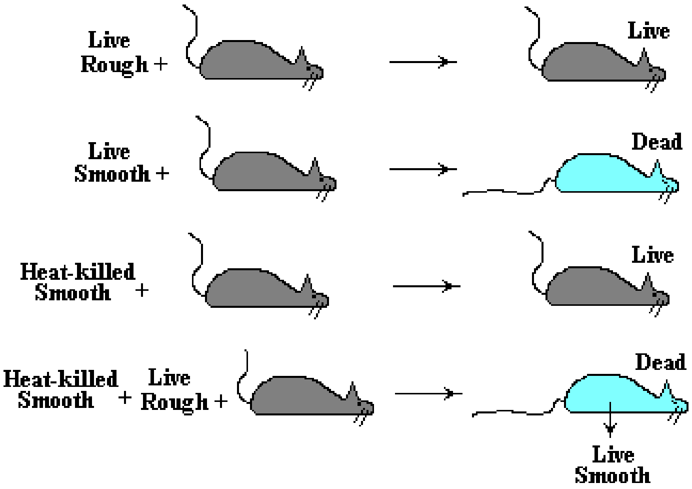
It involves the uptake of naked DNA molecules from one bacterium (the donor) by other bacterium (the recipient). It was discovered by Frederick Griffith in 1928 which laid foundation for the identification of DNA as the genetic material in the bacterium Streptococcus pneumoniae.

1.3.4.3 Griffith's experiment

The wild-type organism is a spherical cell surrounded by mucous coat called capsule. They form large, glistening, smooth (s) colonies. These cells are virulent, capable of causing lethal infections upon injection into mice.

A certain mutant strain of S.pneumoniae has lost the ability to form a capsule. As a result, it forms small, rough ® colonies and is avirulent.

Fig.



When virulent strains were injected into mice, they died. The mice was live when both avirulent and heat killed virulent strains were injected independently. However, the mice was died when heat killed virulent strain+avirulent strain is injected Griffith called this conversion of avirulent strains to virulent strains as transformation. This transformation was not transient, the ability to make capsule and therefore to kill host cells, once conferred upon the avirulent bacteria, was passed to their descendants as a heritable trait. In otherwords, the gene for virulence, missing in avirulent cells, was somehow restored during transformation. It means that the transforming substance in the heat killed bacteria was probably the gene for virulence itself.

DNA as the transforming principle

In 1944, Avery Mc Leod and Madyn Mc Carty showed that the transforming principle was DNA.

- First, they removed the protein from the extract with organic solvents and found that it still transformed.

- Trypsin, chymotrypsin which destroy protein had no effect on transformation. Neither did Ribonuclease which destroy RNA.

- On the otherhand they found that the enzyme Dnase which breaksdown DNA, destroyed the transforming ability of the virulent cell extract.

Finally, direct physico-chemical analysis showed the purified transforming substance to be DNA.

Ultracentrifugation

The material with transforming activity sedimented rapidly suggesting a very high molecular weight, characteristic of DNA.

Electrophoresis

Transforming activity had a relatively high mobility, also characteristic of DNA.

UV-absorption spectrophotometry

Its absorption spectrum matched that of DNA, i.e, maximum absorption at 260nm.

Elementary chemical analysis

This yielded an average Nitrogen / Phosphorous ratio of 1.67, equal to that of DNA which is rich in both elements.

Genetic transformation has been demonstrated in several other genera including; Haemophilus, Bacillus

Salmonella

Streptococcus

Rhizobium

Neisseria and in

Higher organisms

Drosophila

Ephestia insects

Bombyx

Mice and humans cells cultured invitro. Even in these species, all cells in a given population are not capable of active uptake of DNA.

Only competent cells, which possess a so called competence factor, are capable of serving as recipients in transformation. Competence of bacteria is not a permanent feature but occurs only at certain times in life cycle. Competence is commonly observed towards the ends of the `log' phase of growth just before the stationary phase.

There are two theories to explain development of competence.

1. Structure of cellwall is critical. It permits uptake of DNA only during the restricted competence phase and its permeability to macromolecules may change with growth conditions.

2. Competence results from the synthesis of specific receptor sites on the surface of the cell. This view is supported by the fact that synthesis of new proteins is necessary for development of competence. Inhibition of proteins or RNA synthesis inhibits the transformation.

Stages in transformation

1. DNA comes into contact with the bacterial cell surface as a result of random collision. The binding becomes irreversible after a very short period (5-6 seconds).

2. Permanently bound DNA penetrates the bacterium. Double stranded DNA is converted into single stranded DNA by the action of exonuclease. Penetrating DNA must have a minimum length of about 750 base pairs.

3. SsDNA is stabilized by a competence-specific protein. SsDNA migrates from the periphery of the cell to the chromosome DNA.

4. The homologous portion synapses with the recipient chromosome. The unsynapsed DNA is cut by means of nuclease action.

5. The transformation heteroduplex undergoes replication to form transformation homoduplexes. One of these is a normal duplex, while the other is transformed duplex. The clone produced from the transformed duplex is the transformed duplex. The normal duplex will give rise to a non-transformed duplex.

Proof that DNA is the Genetic Material in T_2 Bacteriophage

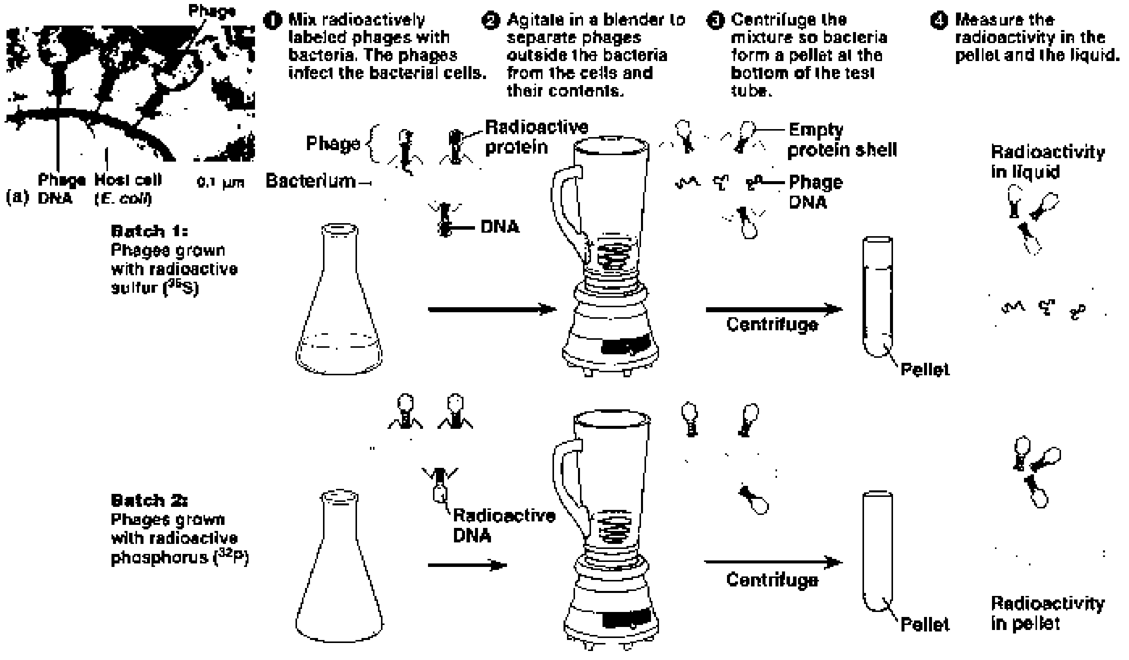
Finally in 1952, A.D. Hershey and Martha chase performed an experiment to prove that DNA was the genetic material. Their experiment involved a bacteriophage, called T_2 that infects the bacterium E.Coli. During infection, the phage genes enter the host cell and direct the synthesis of new phage particles.

Their experiment depended on radioactive labels on the DNA and protein, a different label for each. ^{32}p for DNA, ^{35}s for protein, DNA is rich in phosphorous while phage protein has none, and protein contains sulfur but DNA does not.

Hershey and chase allowed the labeled phages to attach by their tails to bacteria and inject their genes into their hosts. Then they removed the empty phage coats by homogenizing in a blender.

Since the phage genes must enter the cell, they reasoned that the type of label found in the infected cell would indicate the nature of the genes. From the above experiment the conclusion was that the genes are made of DNA.

Fig.


Phage
Phage DNA
Host cell (E. coli)
0.1 µm
(a)
❶ Mix radioactively labeled phages with bacteria. The phages infect the bacterial cells.
❷ Agitate in a blender to separate phages outside the bacteria from the cells and their contents.
❸ Centrifuge the mixture so bacteria form a pellet at the bottom of the test tube.
❹ Measure the radioactivity in the pellet and the liquid.
Phage
Radioactive protein
Bacterium
DNA
Empty protein shell
Phage DNA
Radioactivity in liquid
Batch 1: Phages grown with radioactive sulfur (^{35}S)
Centrifuge
Pellet
Batch 2: Phages grown with radioactive phosphorus (^{32}P)
Radioactive DNA
Centrifuge
Pellet
Radioactivity in pellet

(b) The experiment showed that T2 proteins remain outside the host cell during infection, while T2 DNA enters the cell.

1.3.4.4 Deoxyribonucleic acid and its chemical composition

Friedrich Miescher, in 1869 had isolated a previously unidentified macromolecular substance, to which he gave the name nuclein. Nuclein was later renamed nucleic acid. Development of DNA-specific staining techniques by

Feulgen and Rossenbeck in 1924 enabled Feulgen to demonstrate in 1937 that most of the DNA content of a cell is located in the nucleus.

Nucleic acids are macromolecules present in all living cells, either in Free State or in combination with proteins. Nucleic acids are polymers consisting of units called nucleotides. They are hence called polynucleotides.

Nucleotides

These are the compounds consitiuted by purine or pyrimidine bases, deoxyribose sugars and phosphoricacid.

Importance

1. Purine nucleotides act as the high energy sources: ATP, GTP.
2. Serve as monomeric precurosors of RNA & DNA.
3. Play an important role in carbohydrate, fat, protein metabolism.
4. They also serve as chemical signals Ex: cAMP, cGMP.
5. Function as components of coenzymes FAD, NAD^+ etc. and an important methyl donor, SAM.
6. Act as high energy intermediates such as UDP-glu & UDP-gal in carbohydrate metabolism and CDP-acyl glycerol in lipid synthesis.

Nitrogenous bases

The bases are derivatives of two parent compounds: purines & pyrimidines. These are weakly basic.

Pyrimidine bases

Pyrimidine bases found in Nucleic acids are mainly three.

- Cytosine – found in both DNA and RNA
- Thymine – found in only DNA
- Uracil - found in only RNA

All the pyrimidine bases can exist in lactam form and lactim form. If the group is –NH-CO- it is called Lactam (keto) type, while the same if isomerizes to

–N=C-OH, it is called lactim (enol) type. At the physiological pH, the lactam forms are predominant.

Cytosine

- chemically 2-oxy-4-amino pyrimidine
- is found in all nucleic acids except DNA of certain viruses.

Thymine

- Chemically, if is 2,4-dioxy – 5- methyl pyrimidine
- Also called as 5-methyl uracil
- Occurs only in DNA, however, minor amounts have recently been found in tRNA.

Uracil

- chemically it is 2,4-dioxy pyrimidine
- is confined to RNA only, not found in DNA

Purines

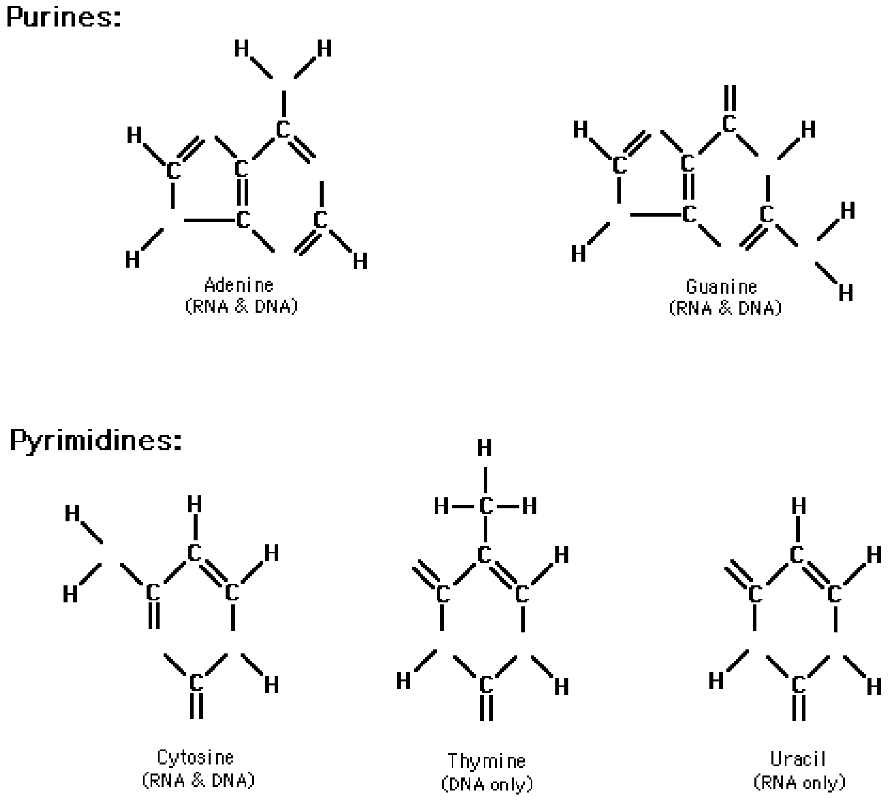
- Purine ring is more complex than pyrimidine
- It can be considered as the product of fusion of a pyrimidine ring with an imidazole ring.
- Adenine & Guanine are the two principal Purines
- Found in both DNA & RNA.

Adenine

- chemically it is 6-amino Purine

Guanine

- chemically it is 2-amino- 6-oxy purine



The Nitrogen Bases Occurring in Nucleic Acids:

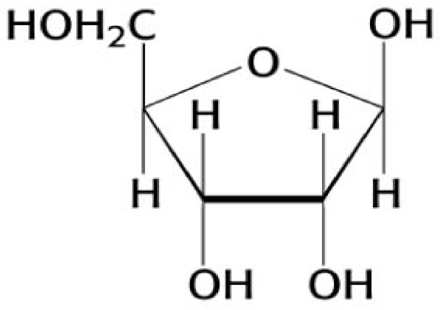
Minor bases in DNA

- 5-methyl cytosine occur in plants
- N^6-methyl adenine in bacterial DNA
- 5-hydroxy methyl cytosine in bacteria infected with certain bacteriophages.

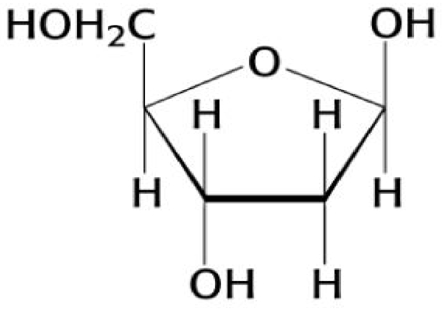
Sugar

2-deoxy β-D-ribose. The sugar is in its furanose form in nucleic acids. An important property of the pentose is their capacity to form esters with phosphoric acid. The –OH group of the pentose, especially those at C_3 & C_5 are involved forming a 3^1, 5^1 – phosphodiester bond.

Fig.



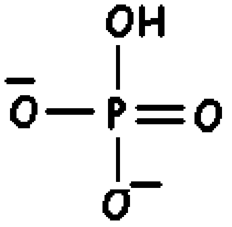
D-Ribose



2-Deoxy-D-ribose

Phosphoric acid

The molecular formula of phosphoric acid is H_3PO_4. It contains 3 monovalent –OH groups and a divalent oxygen atom, all linked to the pentavalent phosphorous atom.



Phosphoric acid

Nucleotides are the phosphoric acid esters of nucleosides. These occur either in the free form or as subunits in Nucleic acids.

Deoxyribonucleotides

Deoxy adenylic acid

Deoxy cytidylic acid

Deoxy thymidylic acid

Deoxy guanylic acid

Nucleosides are composed of a Purine or pyrimidine base and a deoxyribose sugar.

The base is joined covalently, at N-1 of pyrimidine and N-9 of purines in an N-glycosyl linkage to the 1 carbon of the pentose.

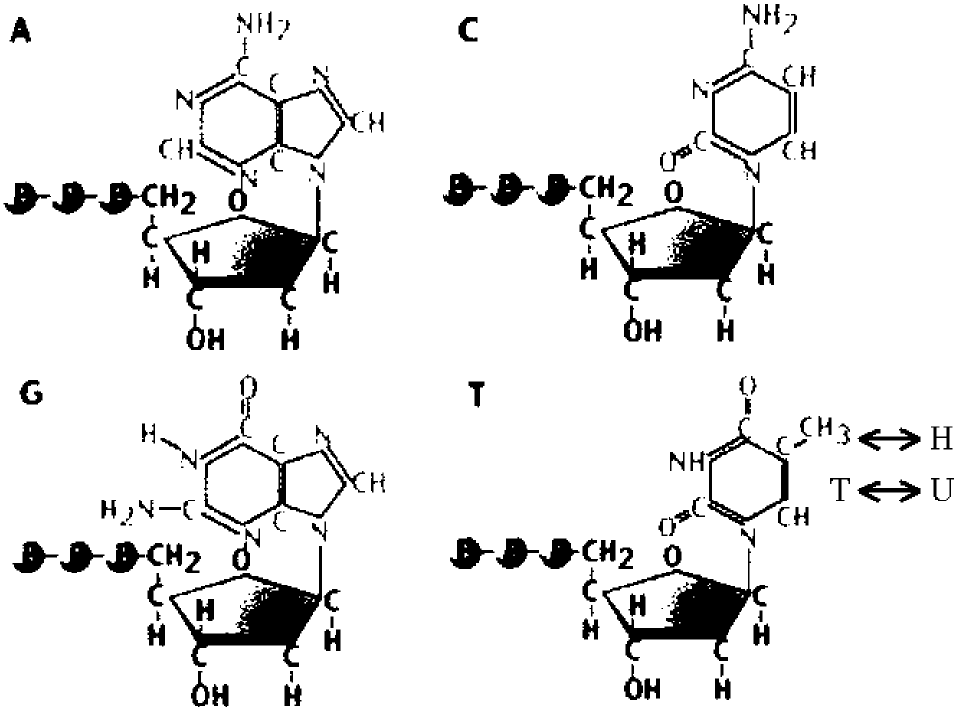
The anti form is necessary for the proper positioning of the complementary purine and pyrimidine bases in the dsDNA.

Deoxyribonucleosides : Deoxy adenosine

Deoxy Guanosine

Deoxy Cytidine

Deoxy Thymidine



Synthetic derivatives

Synthetic nucleobases, nucleosides, nucleotides are widely used in the medical science & clinical medicine. Changes in heterocyclic ring structure and sugar moiety, induces toxic effects when incorporated into cells and inhibits the activities of enzymes.

- 6-thioguanine
- 6-mercaptopurine
- 4- hydroxy pyrazole pyrimidine

- also called as allopurinol
- inhibitor of xanthine oxidase.

- cytarabine (arabinosyl cytosine) – used in the chemotherapy of cancer & viral infections
- Vidarabine (arabinosyl adenine)
- Azathioprine – useful in organ transplantation

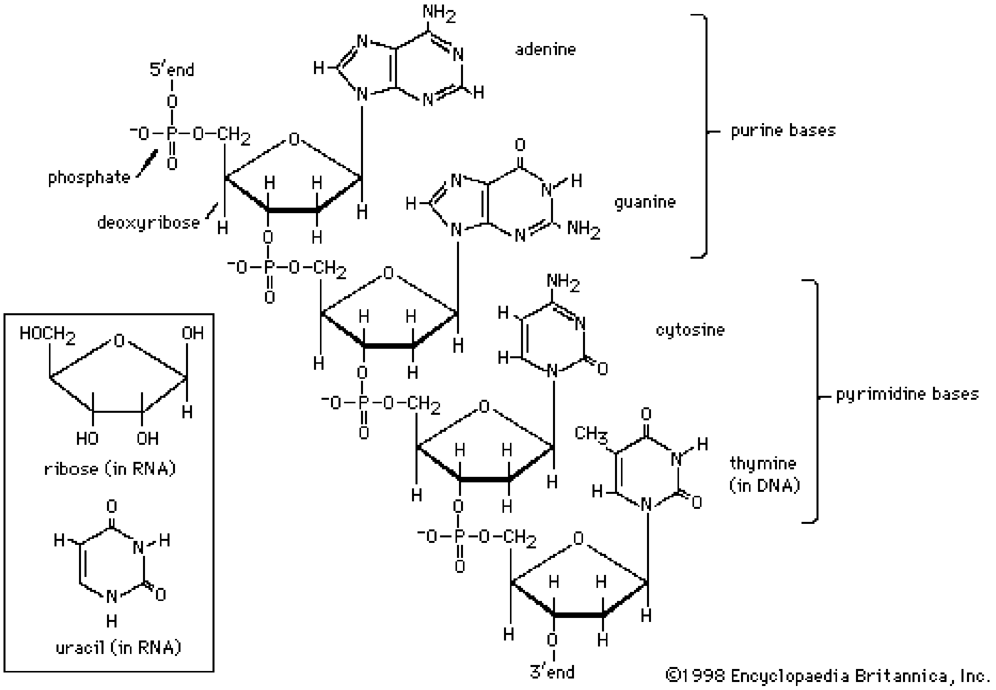
Nucleic acid

A nucleic acid is a polymer of a nucleotide monomer and can be considered as a polynucleotide.

The successive nucleotides in DNA are covalently linked through phosphate group bridges, specifically the 5^1 -OH group of one nucleotide unit is joined to the 3^1 -OH group of the next nucleotide by a phosphodiester bond. Thus the back bone of nucleic acids consist of alternating phosphate and pentose residues, and the characteristic bases may be regarded as side groups joined to the backbone at regular intervals.

Each linear nucleicacid has a specific polarity and distinct 5^1 and 3^1 ends.

Fig.



The back bone of phosphate and sugar is hydrophilic, where as the bases is hydrophobic.

1.3.4.5 Structure of DNA

DNA structure contains hierarchical levels of complexity.

1. Primary structure → covalent structure of nucleotides forming a linear chain.

2. Secondary structure → any regular, stable structure taken up by some or all of the nucleotides.

3. Tertiary structure → The complex folding of large chromosomes with in the bacterial nucleoid & eukaryotic chromatin.

A most important clue to the structure of DNA came from the work of Erwin chargraff and his colleagues in the late 1940s. They concluded that:

- the base composition of DNA generally varies from one species to another.

- DNA specimens isolated from different tissues of the same species have the same base composition.

- The base composition of DNA in a given species does not change with the organism's age, nutritional state or changing environment.

- In all DNAs, regardless, of species, the no. of A residues is equal to the no. of T residues and the no. of G is equal to the no. of C residues, i.e., the no. of purines = the no. of pyrimidines. (A+G = T+C). This is sometimes referred as chargraff's rule and is the key for establishing 3-dimensional structure of DNA.

Secondary structure / double helical structure

Evolution

- W.T. Astbury was the first person to propose the 3-D-structure of DNA. By his X-ray crystallography study on DNA, he concluded that, because DNA has a high density, its polynucleotide was a stack of flat nucleotides, each of which was oriented perpendicularly to the long axis of the molecule & was placed 3-4 A^O=Angstroms apart from each other.

- Continued crystallography studies by Wilkins & R. Franklin confirmed Astbury's 3-4 A^o internucleotide distance and suggested a helical configuration for DNA molecule. They also suggest that, the helix is folded into many turns and each turn causes a vertical rise of 34 A^o.

- An analytical study also suggests that the polynucleotide chains were held together by H-bonding between the base residues.

Watson and Crick model of DNA

- In 1953, Watson & Crick postulated a three-dimensional model of DNA structure from the available data.

- It consists of two helical DNA chains coiled around the same axis to form a right handed helix.

- They hydrophilic backbones of alternating deoxyribose & negatively charged phosphates are on the outside of the double helix, facing the surrounding water.

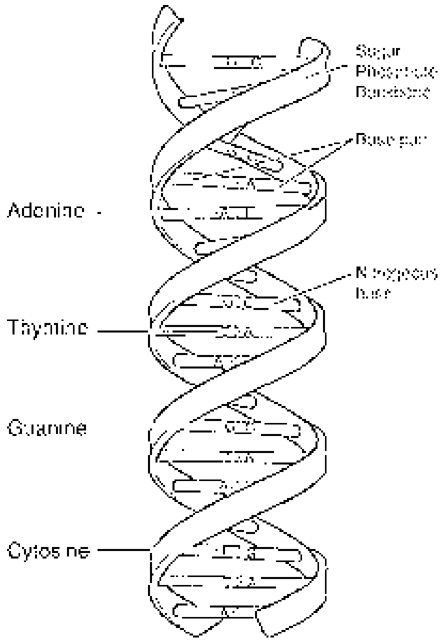
- The purine & pyrimidine bases of both strands are stacked inside the double helix with their hydrophobic & nearly planar ring structure very close together and as perpendicular to the long axis.

- The spatial relationship between these two strands creates a major groove and a minor groove between the two strands.

- The diameter of the helix is 20 A^o, the bases are 3.4 A^o apart along the helix axis. Each turn of the helix contains 10 nucleotide residues. Therefore the helical structure repeats at intervals of 34 A^o.

- The two chains are held together by H-bonds between pairs of bases. Adenine always pairs with thymine by two Hydrogen bonds and guanine always pairs with cytosine by three H-bonds.

Fig.



- The two chains or strands of the helix are antiparallel, Their 5^1x3^1-phosphodiester bonds run in opposite directions.

- The two strands are complementary to each other. Where ever adenine appears in one strand, thymine is found in the other; Similarly wherever guanine is found in one chain, cytosine is found in the other.

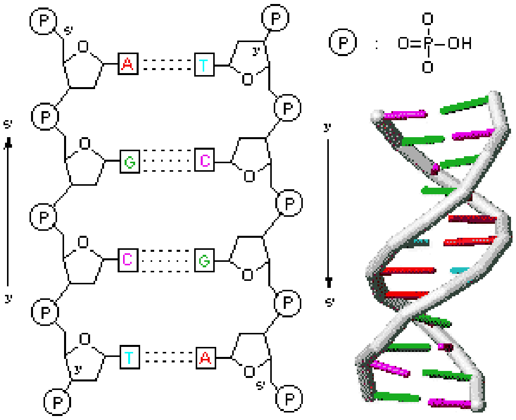
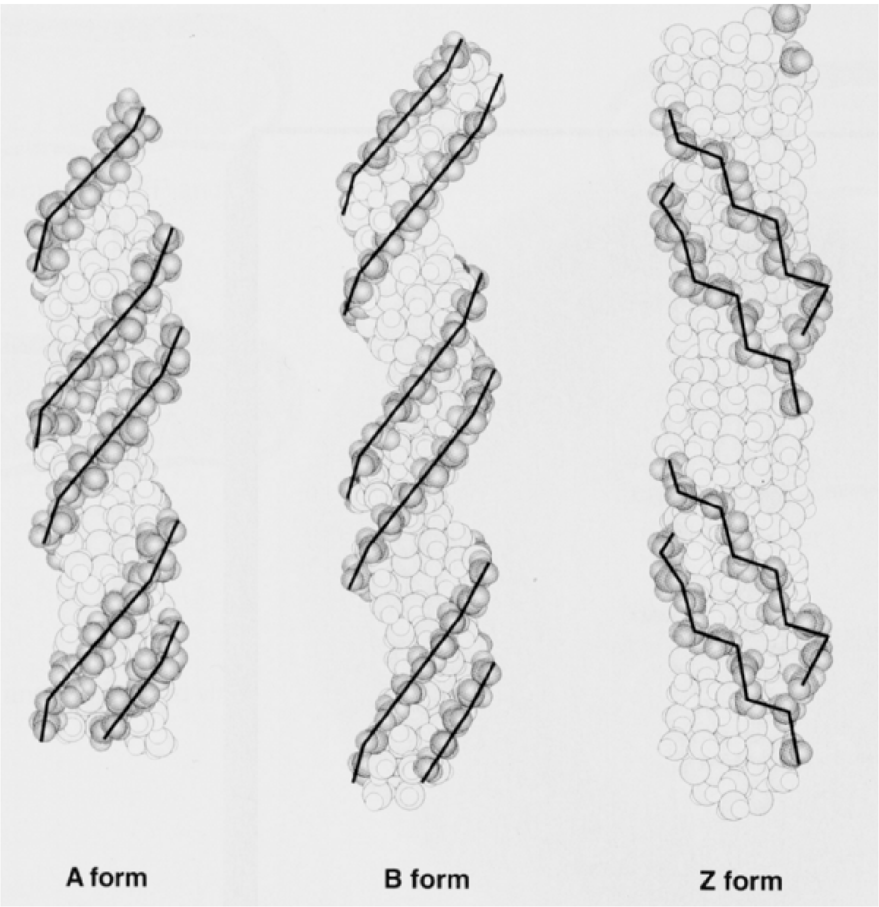


Fig.



Different structural forms of DNA

DNA is remarkably flexible molecule. The Watson-Crick structure is also referred as B-form; which is the more stable, right handed DNA molecule. Many significant deviations from the Watson crick DNA structure are found in cellular DNA, and some (or) all of these may play an important role in DNA metabolism.

A-form

- it is a right handed helix
- it is favoured in many solutions that are relatively devoid of water.
- Number of base pairs per helical turn is 11.
- Rise per basepair is 2.3 A^o
- Is shorter and have a greater diameter
- The reagents used to promote crystallization of DNA tend to dehydrate it, and this leads to a tendancy for many DNAs to crystallize in the A-form.

Z-form

- left-handed double helix
- 12-bp per helical turn
- rise per basepair is 3.8 A^o
- DNA back bone takes zig-zag appearance.
- Sequences in which pyrimidine alternating with purine will give Z-forms.
- There is evidence for short stretches of Z-DNA both in prokaryotes & eukaryotes.
- These Z-DNA tracts may play an undefined role in the regulation of gene expression or genetic recombination.

H-DNA

- triple helical DNA
- unusual structure
- occur in polypyrimidine / polypurine tracts

- pairing and interwinding of three strands.
- They form spontaneously only within long sequences containing only pyrimidines or only purines in one strand.
- Two of the 3 strands in the H-DNA triple helix contain pyrimidines and the third contains purines.
- Found within regions involved in the regulation of expression of a number genes in eukaryotes.

Unusual structures of DNA

These are formed during initiation of DNA metabolism (Replication, transcription, translation) and / or in regulation of gene expression.

1.3.4.7 Denaturation & Renaturation

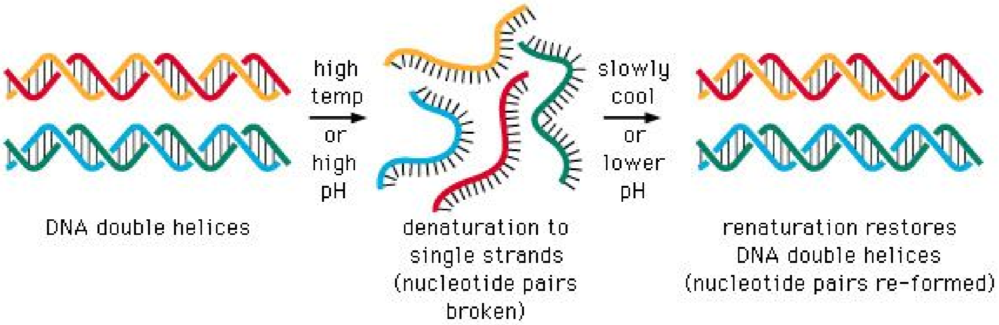
Solutions of carefully isolated, native DNA are highly viscous at pH 7.0 and room temperature (20-25^{o}C). When such solution is subjected to extreme temperature or pH its viscosity decreases sharply. This is due to a process called denaturation.

Denaturation of helical DNA involves disruption of the H-bonds between the paired bases and the hydrophobic interactions between the stacked bases. As a result, the double helix unwinds to form two single strands, completely separate from each other along the entire length or part of the length (partial denaturation) of the molecule. No covalent bonds in DNA are broken.

When the temperature or pH is returned to normal biological range, the unwound segments of the two strands spontaneously rewind or anneal to yield the intact duplex. This is called renaturation. If the two strands are completely denatured, the process of renaturation occurs in two steps.

1. The step is slow, because two strands must find each other by random collision and form short segment of complementary double helix.

Fig.


high
temp
or
high
pH
slowly
cool
or
lower
pH
DNA double helices
denaturation to single strands (nucleotide pairs broken)
renaturation restores DNA double helices (nucleotide pairs re-formed)

2. The second step is much faster, the two strands `Zipper' themselves together to form the double helix.

Careful, controlled denaturation cause, denaturation of A-T rich region, while the other length of DNA is double helix. This partial denaturation is required for initiation of DNA replication and transcription.

T_M

The transition of double helical DNA to single stranded DNA can be accomplished by heating a solution of DNA or by adding acid or alkali to ionize its bases. The unwinding of double helix is called melting, because it occurs abruptly at a certain temperature.

Melting temperature is the temperature at which half of the helical structure is lost. This transition indicates that the DNA is a highly cooperative structure, stabilized by the stacking of bases as well as by basepairing.

The T_M of a DNA molecule depends markedly on its base composition. DNA molecules rich in G-C bps have a higher T_M than those having an abundance of A-T basepairs. The T_M of DNA from many species varies linearly with G-C content, rising from 77^oc-100^oc as the amount of G-C pairs increases from 20-75%.

Hybridization

The double helical DNA contains two complementary strands. If the separated strands from two different species are reannealed to form a double helix, there occurs a hybrid DNA. This process is called hybridization.

Ex: one strand from human DNA & one strand from mouse DNA can form a duplex (hybrid) DNA.

- two DNAs from different species are completely denatured by heating.

- When mixed and slowly cooled, complementary DNA strands of each species will associate & anneal to form duplexes.

- If the homology is high, greater no. of hybrids are formed.

  - It is the basis for essential techniques in molecular genetics
  - Can be used to detect specific RNA
  - To detect sequence homology among different species
  - To detect evolutionary heritage.

Hybrid formation can be measured by different procedures such as chromatography or isopycnic centrifugation. Usually one of the DNAs is labeled with a radioisotope to simplify the measurement.

1.3.4.8 Tertiary structure – suppercoiling

The DNA double helix can undergo coiling about its own axes to produce a supercoiled tertiary structure. Thus DNA can exist in forms other than a linear molecule. In bacterial, viral replicative forms, plasmids, mitochondrial, and chloroplast DNA, the ends of the DNA molecules are covalently joined to form a closed, circular duplex molecule. In the much larger eukaryote chromosomes, supercoiling arises when the DNA coils around histones.

The terms supercoiling, superhelicity and supertwisting are employed for the twisting of DNA duplex upon itself. This property of DNA is an integral feature of all chromosomes, whether circular or linear. It has been shown to be essential for the stages of replication, transcription and recombination.

Discovery of supercoiling in the polyoma virus

In 1965, vinograd and his associates discovered that the genetic material of the polyoma virus was easily renatured after denaturation by heat. They also found that the individual strands of the DNA duplex did not separate from each other. It was therefore concluded that the two strands were covalently closed and intertwined with each other. Sedementation studies of the polyoma DNA showed the existence of 3 components.

- a linear or broken open form of the genome.
- A twisted loop or supercoiled form
- A relaxed loop containing at least one nick or break in one of the strands.

If the two ends of the linear DNA are joined, a covalent relaxed circle is formed. If the supercoiled DNA undergoes a nick in one strand, a relaxed circle is formed. If it undergoes nicks in both strands, the linear form results.

Negative and positive superhelices

The structure of the double helix which is thermodynamically favored contains one complete turn per 10 basepairs. For each turn of the helix, the strands cross twice. This structure is found in linear DNA. The circular structure consists of fewer turns of the helix. This double helix characterized by a deficit of turns is called a negative superhelix. The deficit in turns can accommodate by breaking of the H-bonds and the opening of the double helix over a small region. Another way of accommodating the strain is by the formation of the tertiary structure with supercoils. DNA from natural sources is negatively supercoiled. Usually there is one negative twist in the DNA double helix per 15 turns of the helix. In a different form of the superhelix structure, an excess of helical turns is present. This type of helical structure is called positive superhelix. In this, strain can be accommodated only through the formation of supercoils.

Linking number : α or L

The linking number, also known as the linkage number, is a topological parameter which characterizes closed circular dsDNA. It specifies the number of times the two complementary strands of DNA duplex twist around each other in the DNA circle. The linkage number can change only by breaking and resealing covalent bonds in DNA, as in the case of DNA topoisomerase treatment. For relaxed B-form DNA, the linking no is the no. of basepairs in the molecule divided by 10 conventionally, the linkage number is counted so that it is positive for each crossover in a right-handed helix.

Enzymatic activity altering DNA supercoiling

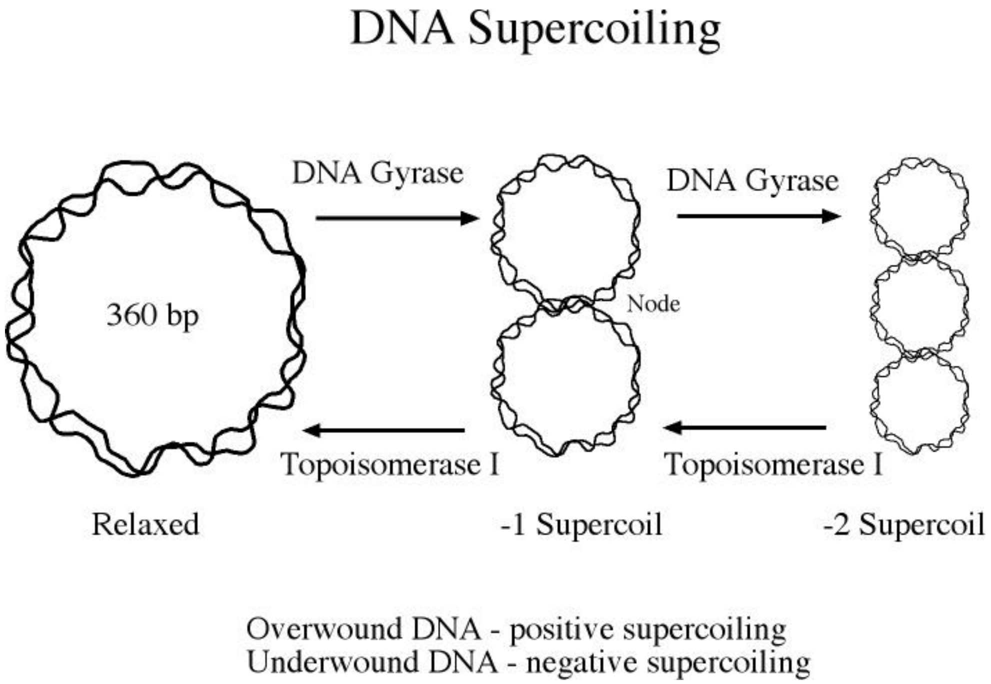
Topoisomerases

These are nicking-closing enzymes whose functions depend on supertwising of DNA. They catalyze the breaking (nicking) and rejoining (closing) of phosphodiester bonds. This alters the topology of DNA without affecting its primary structure.

DNA topoisomerases have been isolated from viruses and from bacterial, plant, & animal cells. The E.coli omega (w) protein was the first topoisomerase discovered, and has been renamed as Eco DNA topoisomerase I. Topoisomerases convert or isomerise one topological version of DNA into another by changing its linkage no. (the no. of times two DNA chains twist around each other). Topoisomerase action is implicated in replication, segregation of replicas, transcription, recombination and nucleosome assembly.

There are two classes of topoisomerases, type I (topo I) and type II (topo II, gyrase) with counterbalancing action.

Supercoiled dsDNA relaxed DNA



Topoisomerase I

- The E.coli type I topoisomerase is a monomer (100 Kda), encoded by the top A gene. It breaks and reseals one strand of DNA, changing the linkage no. in steps of 1. The enzymes binds to duplex DNA and unwinds the double helix locally. It then nicks one strand, and the free phosphate on the DNA becomes covalently attached to a tyrosine residue in the enzyme. Free rotation of the helix is prevented by the cut ends of the DNA remaining bound to the enzyme. The other strand is passed through the break, and the complex rotates, relieving a supercoil. The enzyme now ligates the cut ends. The linkage no. is increased by one. The enzyme becomes separated from the DNA, which undergoes renaturation. The reaction does not require energy. The end result is DNA with one less negative supercoil.

Topoisomerase II (gyrase)

E. coli type II topoisomerase is an A_2B_2 tetramer (mol. Wt. 400 Kda), encoded by gyrA or gyrB gene. Each polypeptide has a molecular weight of 105 Kda. The eukaryote (Hela) type II isomerase is a dimer (mol. wt 309 Kda) each subunit of which has a molecular weight of 172 Kda.

Type II isomerase break and reseal both strands of DNA, changing the linking number in steps of two. The enzyme can cut a ds DNA molecule, pass another duplex through the cut, and reseal the cut. This activity requires ATP. The effect of enzyme action is to change a positive supercoil into a negative supercoil.

Summary

Chromosomes are composed of two types of large organic molecules (macromolecules) called proteins and nucleicacids. In 1940,s there are several experiments were conducted by different groups of scientists that proves the genetic information is stored in DNA not in proteins.the DNA contains deoxyribose sugar,phosphate and nitrogenbases such as adenine ,Guanine ,thymine and Cytosine.The structure of DNA was proposed by Watson and crick i.e double helical model.the DNA helix can undergo coiling about its own axes to produce a supercoiled tertiary structure

Model Questions

1)write in detail about s the components of nucleic acids

2)Prov the DNA as genetic material

Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork

Dr.N.Srinivasa Reddy

Lesson 1.4.1

REPLICATION

Objective

1.4.1.1 Introduction

1.4.1.2 Replication models

1) Semi-conservative model

2) Conservative model

3) Dispersive model

4) Meselson & Stahl experiment

5) Replication Origin

6) Replication fork

7) Replication Direction

1.4.1.3 Prokaryotic Replication

1) Enzymology of Replication

2) Replication of *E.Coli* Chromosome

1) Initiation

2) Elongation

3). Termination

1.4.1.4 Replication Mechanism of Bacteriophage M 13

1) Replication Mechanism of 17 A

1.4.1.5 Fidelity of Replication

1.4.1.6 Eukaryotic Replication

1) Enzymology of Eukaryotic Replication

2) Mechanism of Replication

3) Eukaryotic Chromosome Origin

4) Mitochondrial DNA Replication

5) Termination

OBJECTIVE

During the synthetic phase of cell division the content of DNA increases by a process known as replication. In this chapter the process of replication, types and inhibitors were clearly explained. Different DNA repair mechanisms were also discussed.

1.4.1.1 INTRODUCTION

The ability to reproduce is one of the most fundamental properties of all living organisms. This duplication is observed at various levels:

Organisms duplicate by sexual / asexual methods

Cells duplicate by cellular division

Genetic material duplicates by Replication.

The capacity of duplication is thought to be the first critical properties to have appeared on the path toward evolution.

1.4.1.2 Replication Models

The formulation of DNA structure by WATSON & CRICK in 1953 accompanied the proposal for its self duplication. (Watson and Crick envisioned that gradual separation of the helix by the successive breakage of H-bonds is possible and as the two strands are complementary and follow strict base pairing rules: each strand contains information for the synthesis of other. Thus once strands are separated, each can act as template to direct the assembly if nucleotides).

1. Semi-conservative Model

Watson and Crick proposed several predictions concerning the behavior of DNA and, most important is the physical separation of the two strands.

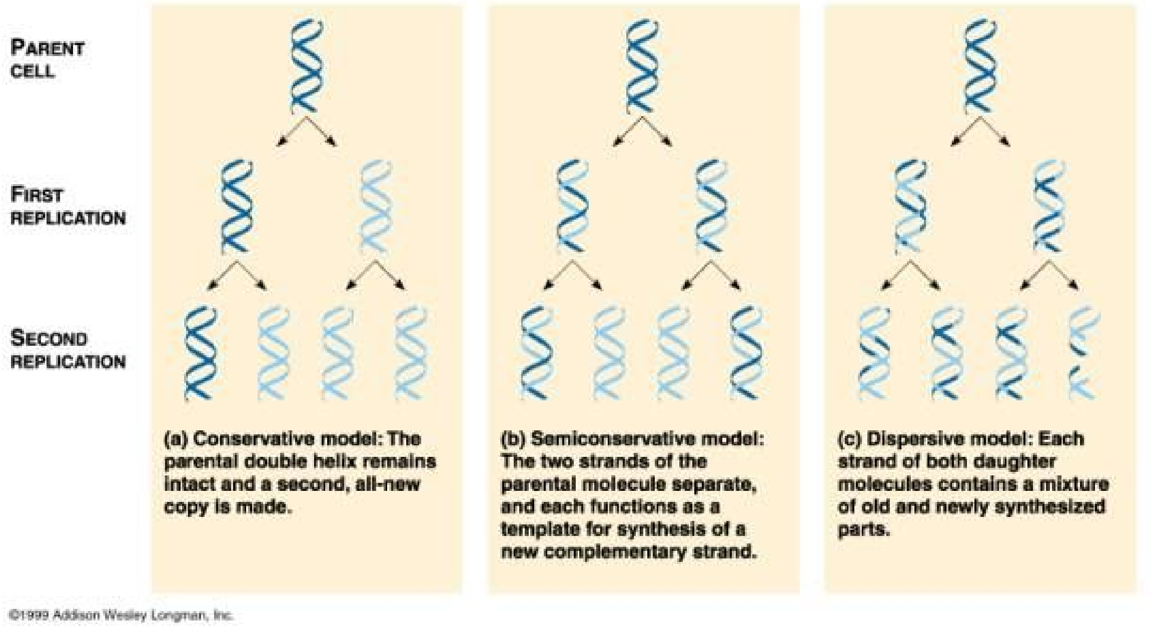
According to Watson and Crick, each of the daughter duplexes should be composed by one parent strand and one new strand.

2. Conservative Model

The two original strands would remain together and two newly synthesized strands remain together. One daughter cell contains fully conserved duplex, while the other cell contains newly synthesized duplex.

3. Dispersive Model

The integrity of the each of parental strands would become disrupted, as a result, the daughter cell contains duplexes in which each strand is composite of old and new DNA i.e., neither the strands nor the duplex itself is conserved.



(a) Conservative model: The parental double helix remains intact and a second, all-new copy is made.

(b) Semiconservative model: The two strands of the parental molecule separate, and each functions as a template for synthesis of a new complementary strand.

(c) Dispersive model: Each strand of both daughter molecules contains a mixture of old and newly synthesized parts.

©1999 Addison Wesley Longman, Inc.

To decide the possibilities among the three models, it is necessary to distinguish newly synthesized DNA from the original DNA that served as template. This was first accomplished in studies on bacteria in 1958, by Matthew Meselson, Franklin Stahl of California Institute of Technology.

4. Meselson and Stahl's Experiment

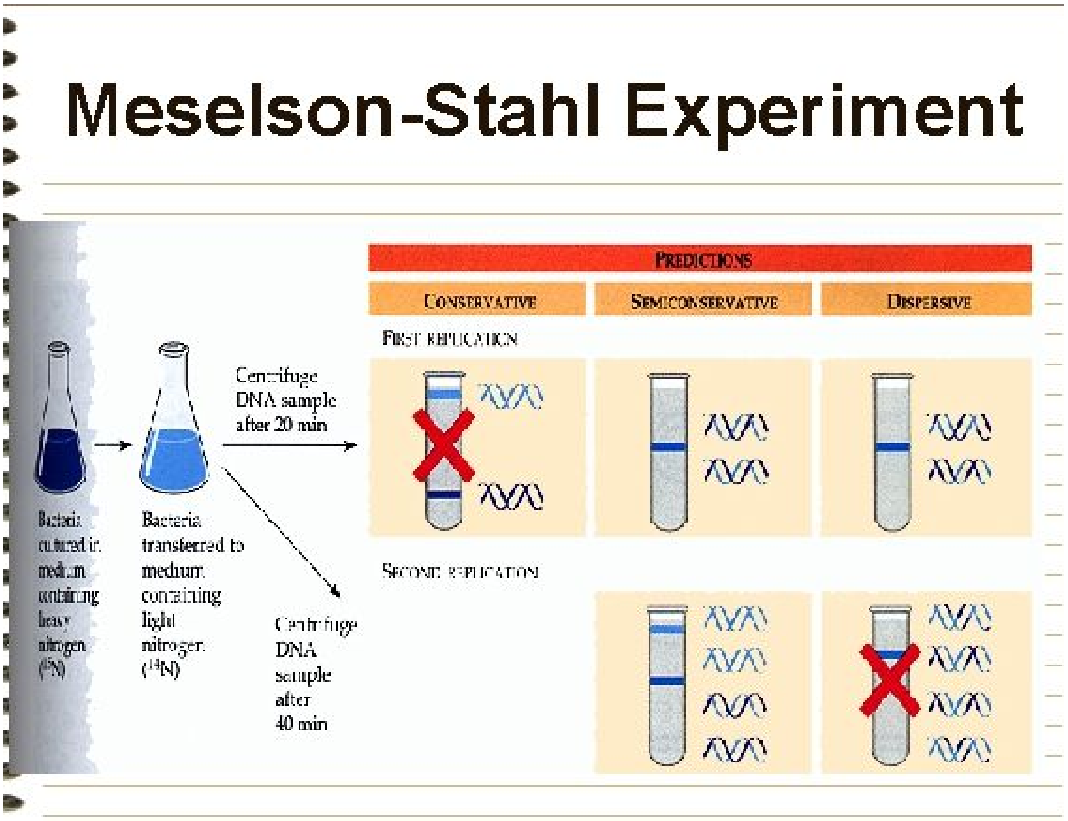
-Meselson and stahl grew bacteria (E.Coli) for many generations in media containing N^{15} – NH_4Cl as sole source of nitrogen until the entire cellular DNA contained the isotope.

- The cells were then transferred to a medium containing the normal `light' isotope N^{14} and samples were removed periodically from the cultures.

- The DNA in each sample was analyzed by density-gradient equilibrium centrifugation which can separate heavy-heavy, (N^{15}-N^{15}) , light-light (N^{14}-N^{14}), heavy-light (N^{15}-N^{14}) duplexes into distinct bands.

- The results showed that replication occurs by semi-conservative model.

  DNA extracted & centrifuged to equilibrium in Cscl Density gradient



Meselson-Stahl Experiment

PREDICTIONS
CONSERVATIVE
SEMICONSERVATIVE
DISPERSIVE
FIRST REPLICATION
SECOND REPLICATION
Bacteria cultured in medium containing heavy nitrogen (^{15}N)
Bacteria transferred to medium containing light nitrogen (^{14}N)
Centrifuge DNA sample after 20 min
Centrifuge DNA sample after 40 min

Semi-conservative replication of chromosomes can also be visualized through an examination of chromosomes that are allowed two rounds of replication in a medium containing Bromodeoxy uridine. These duplicated chromosomes are then stained with a fluorescent dye and Giemsa stain to produce harlequin chromosomes. The newly synthesized DNA stains differently than the parental DNA.

5. Replication Origin

Replication begins at a specific site on the chromosomes called ORIGIN. A replication origin is analogous to a promoter of transcription in that – both regulatory sequences acts as biding sites for sequence specific DNA-binding proteins that initiate DNA or RNA synthesis at a specific site along the template.

John Cairns, in the early 1960s, developed an autoradiographic technique, to visualize the semi-conservative model. In this process, bacteria are lysed very gently and their chromosomes spread out on a surface without any further manipulation. If the bacteria had been growing on the H^3 thymidine prior to the preparation of the chromosomes, then the light microscopic autoradiogrphy could reveal the outline picture of the labeled DNA.

Autoradiographs of this type confirmed circular nature of the bacterial chromosome and revealed the overall process of replication. In many cases the outlines of the silver grains formed a structure resembled θ - which is expected to be caught during replication of a circular duplex chromosome.

Each θ structure is composed of three distinct lengths of DNA – two newly synthesizing daughter strands and one unreplicated portion of the parent duplex.

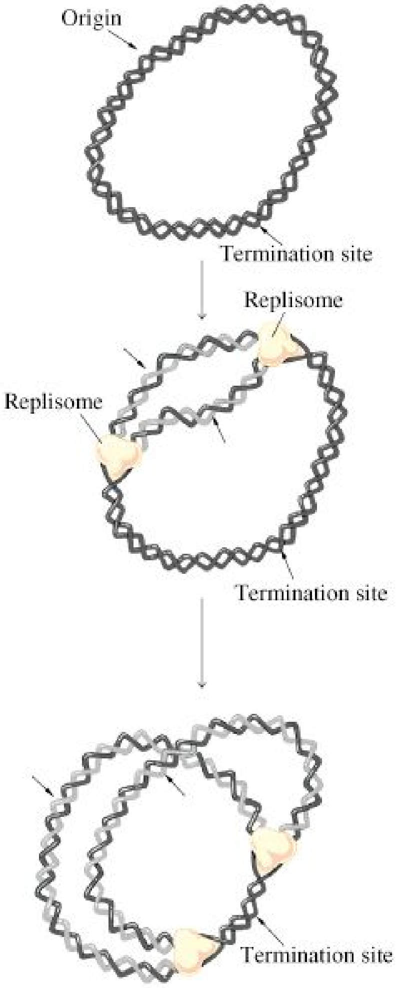
6. Replication Fork

The point at which the pair of replicated segments come together is called the replication fork. Each replication fork corresponds to a site where the parental double helix is undergoing strand separation, and the nucleotides are being added into the new strands. The replication fork marks the advance of DNA synthesis within a replication intermediate a duplex chromosome.

Common intermediates include:

θ Structures

Resulting in replication in a specific region of a circular Chromosome.



D-loop Structures

Resulting from the synthesis of one daughter strand in a circular

Or linear chromosome with synthesis initially in one direction on one template strand.

Bubble or Y Structures

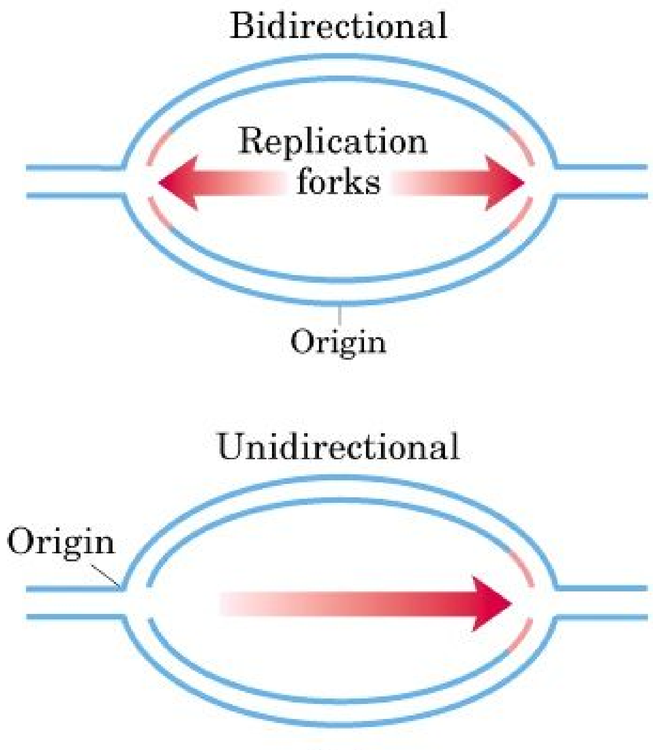
Resulting from initiation of a linear chromosome.

Lariat forms

Resulting from covalent extension of a parental strand in a circular chromosome.

7. Replication Direction

DNA replication can be unidirectional or bi-directional, depending upon whether the replication from the point of origin proceeds only in one direction or proceeds in both the directions. In unidirectional replication, one of the two replication forks will be stationary and the other moves with replication. In bi-directional replication, none of the two replication forks will be stationary and both will be moving.



(b)

Autoradiographic experiment

If cultured mammalian cells undergoing replication are exposed first to high concentration and then to low concentration of 3_H thymidine, the resulting DNA will be heavily labeled (hot) near replication origins (ori) and lightly labeled farther away. When such labeled DNA is dried on a microscopic slide as long fibers and their exposure to radiation – sensitive enulsion, autoradiographs should be produced corresponding to hot-warm regions. Such autoradiographs from cultured mammalian cells show the replication to be a bi-directional.

1.4.1.3 Prokaryotic Replication

1. Enzymology of Replication

DNA Polymerase I

- It was the first polymerase recognized.It was discovered by Arthur Kornberg.It was the most extensively studied DNA polymerase.A cell contains 400 polymerase I molecules.The enzyme is a single chain of 928 aminoacid residues. Molecular weight is 103 Kda. It is a globular protein, with a diameter nearly of 65A^{o}. The enzyme is readily cleaved by proteases too. A small 35Kda C-terminal fragment with $5^{1} \rightarrow 3^{1}$ exonuclease activity. A large 68Kda C-terminal fragment with polymerase and 3^{1}->5^{1} exonuclease activity. This fragment is called as klenow fragment. The active site of the enzyme contains binding sites for the template chain, growing chain (the primer), dNTP's, and divalent metal ion. General reaction of :

Binding of Mg^{2+} - dNTP, in the dNTP binding pocket of polymerase I and nucleophilic attack by the 3'-OH group of the growing chain or primer on the α phosphorous of the incoming dNTP.

- The enzyme is inactive on complete ds DNA, ss DNA which are covalently linked and nicks with 5'-OH and 3'-P, but are active at nicks with 3'-OH and 5'-P. The enzyme is able to bind at nicks with 5'-OH & 3'-P but is unable to polymerize.

- 3^{1}->5^{1} exonuclease activity: In polymer chains with correctly paired 3'-OH ends and in the absence of dNTPs required for polymerization, the primer terminus will be exposed to repeated nuclease actions, which can be inhibited by a correct dNTP. This activity is called proof reading.

- 5^{1} ->3^{1} exonuclease activity :

  - Cleaves a diester bond only at a base paired region.
  - Can excise oligonucleotides upto 10 residues long from the 5' end.
  - It has endonucleolytic activity, that it cleaves diester bonds as distant as 8 from the 5' end or even further.
  - Or the excision enzyme and polymerizing enzyme are present to gather, the gap formed by excision is filled immediately without giving a chance for endonucleases.
  - Removes RNA primer from 5' and of a DNA chain.

- Polymerase I turnover is 600 nucleotides / min/enzyme at 37^{o}c.

DNA Polymerase II

The identification of mutants lacking polymerase I by Paula De Lucia and John Cairns in 1969 lead to the discovery of polymerase II and polymerase III enzymes. Pol II was isolated by T.Kornberg and Gefter in 1970. The presence of Pol II & Pol III was masked by the high level of activity of Pol I. Molecular Weight: 90Kda. Unlike Pol I unable to use a template-primer that is simply nicked. Lacks the 5'->3' exonuclease activity. Strains lacking the gene for this enzyme show no replicational or growth defects.

DNA Polymerase III

Isolated by T.Kornberg & Gefter. It is the most complex of the DNA polymerases. pol III holoenzyme is the principal replicative enzyme of E.Coli. Only 10 molecules / cell are present. It has a collective mass of 900KD, with 20 or more polypeptides. It dissociates readily upon dilution into a variety of subassemblies. The core pol III, while effective in filling gaps in DNA, incapable of rapidly replicating ss viral circles that are several thousand residues. However, a novel activity was observed in E.Coli extracts which was highly efficient in replicating long stretches of template. This proved upon purification to be pol III complexed with many auxillary subunits that clamp the core to the template and endow it with high processivity. The β - subunit is the most important processivity factor for subassemblies of pol III holoenzyme. Although, it is not demonstrably a DNA- binding protein, β can become tightly attached to a template – primer to form a reinitiation complex.

Pol III holoenzyme lacks5^1->3^1exonuclease activity.

Primase

- One of the basic rules of replication is that a DNA pol cannot start a chain and must rely on a priming device. So, every initiation event requires a primer. The opposite polarities of the two strands of DNA duplex and the exceusively 5^1->3^1 polymerization by DNA pol necessitates a semi-discontinuous mechanism of replication. Two different modes of priming operates for continuous (leading) and discontinuous (lagging) strands. The continuous strand needs to be primed only ones, usually at or near a chromosomal origin. The discontinuous strand must be primed repeatedly to generate the short nascent strand (1000-2000 nt long). The enzyme is a single polypeptide of 60Kda. About 50-100 copies / cell. *E. coli* primase acts alone, most commonly it teams up with the multifunctional DNA B protein in the synthesis of primers to start DNA chains. Primase – DNA B complex on a template requires additional prepriming proteins Pri A, Pri B, Pri C. The resulting mobile protein complexes track progressively along the DNA template and are called primosomes.

Pri A

Pri B + DNA B → Preprimosome Primosome

Pri C

Helicase

- It couples the energy of NTP hydrolysis for melting the hydrogen bonds and dissipating the other forces that hold together the strand of the DNA duplex. The helicase activity of DNA B protein is central to replication in E.Coli. DNA B, a hexamer of 50Kda subunits, activates priming by primase and melts the DNA strands at replication fork. As a fuel for DNA melting, ATP is preferred. *E.coli* SSB & primase together stimulate the helicase approximately 6 folds. Helicase binds templates very poorely and for proper binding requires a large number of proteins. This is overcome by the origin – initiator proteins. At origin, the DNA A and DNA C proteins efficiently load DNA B protein at a ratio of 2-4 molecules per template. The DNA B and DNA C proteins fro a 6:6 complex that activates DNA B protein for entry into the Ori C-DNA A protein comples. The distance between polymerase and helicase on DNA appears small, indicating that melting is what limits the rate of DNA synthesis.

SSBs

Single stranded DNA binding proteins. They impart regular structure to DNA single strands, a structure required for the action of a variety of enzyme of replication. Had strong preference for DNA than RNA and for ssDNA than for dsDNA. Binds tightly and cooperatively. A cell contains 300 copies. The protein is a tetramer of 18.9 Kda subunits. A polynucleotide, at least 4-6 residues long is required for binding. This protein contributes to opening and unwinding of a duplex at an origin of replication in a super coiled DNA. They direct priming of DNA synthesis to specific origins by covering single strand regions. Sustains unwinding of duplex DNA by helicase actions at replication forks and loci of repairs. Inhibits both exo and endonucleolytic activities of nucleases.

Ligase

By 1967, models for recombination, repair & replication predicted an enzyme, which could reseal the breaks of DNA backbone. DNA ligases – phosphodiester bond in DNA polymers. Less effective in joining ssDNA. More active with DNA than with RNA. It is a 75 Kda polypeptide.300 molecules per cell, a value close to that for DNA pol I, inview of their closely related functions in filling gaps and resealing segments of DNA. Ligases from *E. coli* & *B. subtilis* use NAD^+ as coenzyme and as the energy source for the synthesis of the phosphodiester bond. In eukaryotes coenzyme is ATP. Ligases utilize the group transfer potential of the phospho-anhydride bonds of NAD^+ or ATP to form a phosphodiester bond between nucleic acid chains. The reaction occurs in 3 steps.

- The formation of an enzyme – nucleotide intermediate by transfer of the adenylyl group of NAD / ATP to the $E\text{-}NH_2$ of a Lysine residue in the enzyme. Product is NMN or Ppi.
- Adenylyl activation of the 5'-P terminus of DNA by transfer of the adenylyl group from the enzyme.
- Phosphodiester bond formation by attack of the 3'-OH terminus of the DNA on the activated 5'-P group with release of AMP.

- The joining reaction is inhibited by the addition of coenzymes NAD^+ or ATP. The enzyme cannot link ss nicks and blunt ends. It joins only nicks with 3^1-OH and 5^1-P of DNA and 5^1-P of DNA to 3^1-OH of RNA, but not the reverse.

Topoisomerases

Topoisomerases catalyse the interconversion of topological isomers (topoisomers) of a DNA molecule. By introducing a transient break in the phosphodiester backbone, through formation of a covalent protein – DNA intermediate, the enzyme allows the DNA strands to pass through one another there by changing the topological state of the DNA molecule. Other DNA characteristics are not affected. The linking number may be changed by the introduction or relaxation of super helical turns.

Topoisomerase I

Is a single polypeptide of about 100 Kda. The relaxing activity of topoisomerase I has several features. Negative supercoil turns are removed but no other change in the DNA occurs. Relaxation of the DNA occurs gradually. A covalent protein – DNA intermediate conserves energy of the nicked phosphodiester bond for concerted resealing. The enzyme reacts with ss circles to introduce topological knots. It is also involved in the formation of covalently closed duplex circles from complementary single stranded ones. It also catenates and decatenates duplex DNA circles, provided that at least one of the molecule contain a nick or gap in one strand.

Topoisomerase II or DNA Gyrase

Gyrase converts relaxed, closed circular duplex DNA to the negatively superhelical form. ATP is required and is hydrolysed in the reaction. It is a tetramer (dimer of dimers) A_2B_2 and has a molecular weight of 400 Kda.

The enzyme catalyses.

- negative supercoiling of circular DNA.
- Relaxation of negatively supercoiled DNA in the absence of ATP.
- Double strand breakage of DNA.
- ATP hydrolysis to ADP + Pi in the presence of duplex DNA.
- Positive wraping of DNA around Gyrase.
- Catenation and decatenation of duplex DNA.

Topoisomerase II is necessary to maintain a negatively supercoiled template for the assembly of the initiation complex and opening of the duplex at the origin, to provide swivel for the progress of replication fork and to decatenate and superc oil the daughter molecules during the terminal stages.

2) Replication of E. Coli Chromosome

The synthesis of a DNA molecule can be divided into 3 stages.

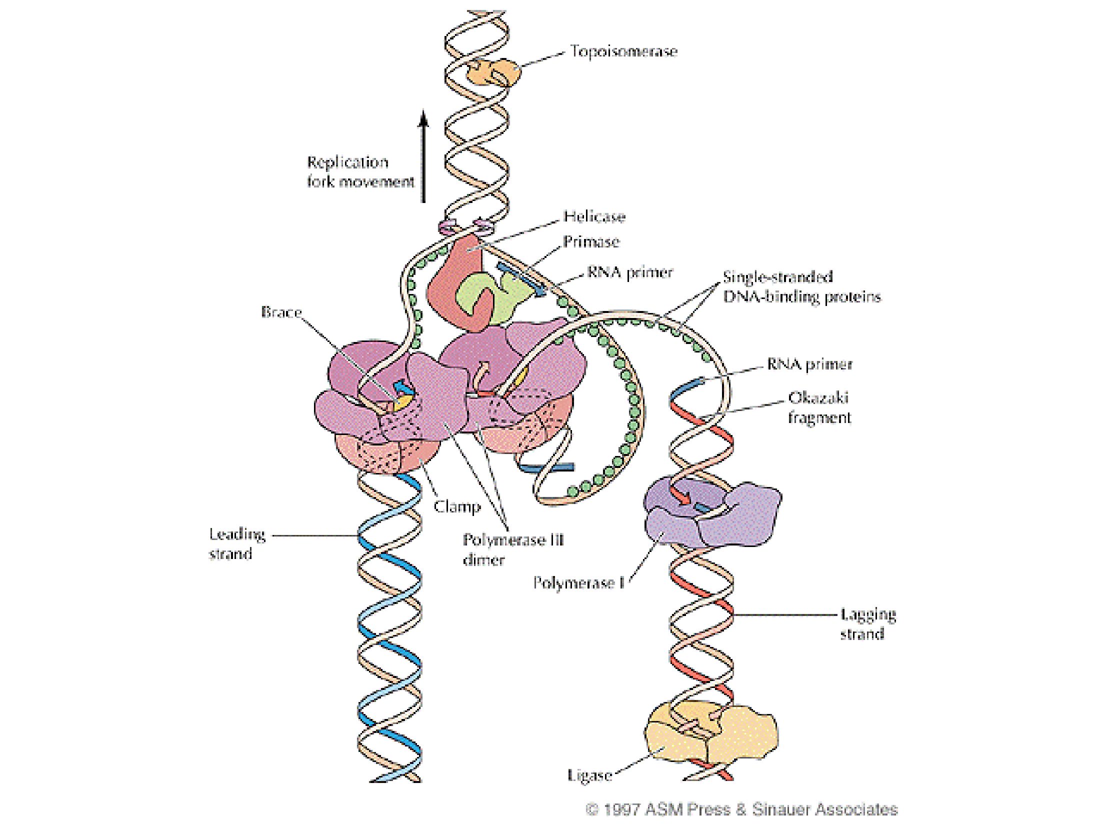
1. Initiation
2. Elongation
3. Termination.

1. Initiation

E.coli replication origin, ori C consists of ---- 245 base pairs, many of which are highly conserved among bacteria. The key sequence includes 2 series of short repeats.

- 3 repeats of a 13 bp sequence.
- 4 repeats of a 9 bp sequence.

The key enzyme in the initiation process is the DNA A protein. DNA A protein binds to the four 9 mers in oric, forming an initial complex that contains 10-20 protein subunits. Although DNA A can bind to E. Coli Ori DNA in the relaxed state, it can initiate replication only in the negatively super coiled DNA. This specificity is that, DNA molecules with negative super coils are tightly wound and are easily melted, thus providing single strand templates. Binding of DNA A protein to the Ori c 9-mers facilitates the initial strand separation or melting of duplex DNA, which occurs at the Ori C 13-mers. This process requires ATP and yields, the so called `open complex'. The DNA B protein then binds to this region and unwinds the DNA bi directionally and requires the presence of DNA C proteins and creates two replication forks. Multiple molecules of SSBP bind cooperatively to ssDNA and stabilizes the separated DNA strands and prevents renaturation. Gyrase relieves the topological stress created by the Dna B helicase action.



© 1997 ASM Press & Sinauer Associates

Regulation

DNA replication is precisely regulated so that it occurs only ones in each cell cycle. The DNA A protein hydrolyses its tightly bound ATP slowly (about 1 hr) to form an inactive DNA A –ADP complex. Reactivating this complex (replacing ADP with ATP) is facilitated by interaction with acidic phospholipids in the bacterial plasma membrane.

Initiation at inappropriate time is prevented by the presence of the inactive DNA A – ADP complex, by the binding of a protein called Ici A (Inhibitor of Chromosomal Initiation) to the 13-mer repeats.

Elongation

The elongation phase of replication consists of two seemingly similar operations that are mechanistically quiet different i.e, leading & lagging strand synthesis.

Several enzymes at the replication fork are important for the synthesis of both strands.

- DNA helicases unwind the parental duplex.
- DNA topoisomerases relieve the topological stress induced by helicase.
- SSBs stabilize the separated strands.
- In other respects, synthesis of DNA is two strands is sharply different.

Continuous / leading strand synthesis

- Begins with synthesis of short RNA primers (10-60nt) at the replication ori. Deoxy NTPs are then added to this primer by DNA pol III holoenzyme. Onces began leading strand synthesis proceeds continuously keeping pace with replication fork.

Discontinuous / Lagging strand synthesis

Synthesis must be accomplished in short fragments, called okazaki fragments after their discoverer Reiji Okazaki. Synthesis occurs in direction opposite to the fork movement. Each fragment must have its own RNA primer, synthesized by primase and its positioning must be controlled. The regulatory apparatus for lagging strand synthesis is called primosome. Primosome moves along the lagging strand template in 5^1->3^1 direction. Primosome at intervals compels primase to synthesize a short RNA primer to which DNA is then added by the DNA pol III holoenzyme. When the new okazaki fragment reaches the old primers, then DNA pol I removes that primer by its 5^1->3^1 exonuclease activity and is replaced with dNTPs by the same pol V. The nick left was sealed by Ligase. In *E. coli* synthesis of the leading and lagging strand synthesis may be actually coupled. This can be accomplished by looping of the lagging strand, so that synthesis can be carried concurrently on both strands by a single pol III dimeric enzyme. As the okazaki fragment synthesis proceeds, the loop between pol III and replication fork increases.

Discovery of okazaki fragments

The most recently synthesized or nascent / DNA molecular in preparations of T_4 phage are short pieces. They sediment at about 85-105s in a gradient of alkaline sucrose, representing chain lengths of 1000-2000 nt.

To capture, T_4 replication pieces before they join the main body of growing chains, it is necessary to

- Reduce the replication rate by lowering the growth temperature (Ex: to 8^{o}c).

Then use a brief pulse of a DNA precursor of high radio-activity that enters DNA rapidly and directly. Quench the pulse efficiently. Under these conditions, much of the precursor label is captured as small pieces. After a prolonged pulse or after subsequent exposure to large concentration of unlabeled precursors (chase), the radioactive precursor is found exclusively in high molecular weight DNA. Mutant *E. coli* deficient in ligase or in pol I or in both accumulate large amounts of *E. coli* replication fragments. It seems likely that, in cell systems, initiations are limited to one of the two strands at the fork and that a “semi-discontinuous” model best accounts for kinetic observation of the DNA. **Termination**

- The *E. coli* replication terminus is a large region flanked by six nearly identical, non-palindromic --- 23 bp terminator sites.

➔ Ter E, Ter D, Ter A → counter clockwise

➔ Ter F, Ter C & Ter B → clockwise

A replication fork traveling counter clockwise passes through Ter E, D, A. A clockwise traveling replication fork transits Ter E, D, A but halts at Ter C or failing that Ter B or Ter F. Thus they act as one way valves that allow replication forks to enter the terminal region but not to leave it. The arrest of replication fork motion at Ter sites requires the action of Ter protein (monomer of 309 aminoacid residues).

(Terminator utilization substance)

Thus protein specifically bids to a Ter site, where it prevents strand displacement by helicase, and there by arresting the replication fork. However, this termination system is not essential when the replication terminus is deleated, replication simply stops. Through the collision of opposing replication forks. The last few helical turns in the parental DNA could be removed by changing the topology of the nearby completed replicated regions, leaving the two daughter helices linked together as catenazes. Replication then could be completed before or after decatenation to yield two separated complete daughter helices. This decatenation a catalysed by DNA gyrase and topoisomerase IV. Topo IV is responsible for separating newly replicated molecules in vivo. DNA gyrase cannot fully substitute for Topo IV, although if decatenates in vitro.

1.4.1.4. Replication Mechanism of Bacteriophage M_{13}

Bacteriophage M_{13} carries a 6408 nt single straded circular DNA, known as Viral or (+) strand. When it infects a cell (bacterium), this strand directs the synthesis of its complementary or (-) strand and thus forms the circular duplex, replicative form (RF). The RF may be nicked (RF II) or supercoiled (RF II). It gives a good paradigm for leading strand synthesis in duplex DNA. As the M13 (+) strand enters the E. Coli cell, it becomes coated with SSBP except at a palindromic 57-nt segment. That forms a hairpin. RNA polymerase commences the synthesis of primer, 6nt before the start of the hairpin and extends the RNA 20-30 residues to form a segment of RNA-DNA hybrid duplex. The DNA that is displaced from the hairpin becomes coated with SSB so that when RNA polymerase reaches it, primer synthesis stops. Pol III holoenzyme extends the RNA primer around the circle to form the (-) strand. The primer is removed by Pol I there by forming RF II, which is converted to RF I by the sequential actions of DNA ligase and DNA gyrase.

1) Replication mechanism of bacteriophage ϕ x 174

- Bacteriophage ϕ x 174 carries a small (5386 nt) single stranded circular DNA. The conversion of its DNA to replicative form is a much more complex process than that for M_{13} phage DNA. ϕ x 174 replication requires the participation of a nearly 600 Kda protein assembly called primosome.

a) (-) strand synthesis

It acts as a good model for lagging strand synthesis. The (+) strand is coated with SSBP except for a 44-nt hairpin near position 2300. A 70nt sequence containing this hairpin known as pas (primosome assembly site), is then recognized and bound by the Pri A, Pri B and Pri C proteins. DNA B and DNA C proteins in the form of a DNA B_6. DNA C_6 complex add to the DNA with the help of DNA T protein. This process requires an ATP molecule. DNA C is then released yielding the preprimosome. The preprimosome, inturn, binds primase yielding primosome. The primosome moves in 2^1 -> 3^1 direction along the (+) strand by Pri A and DNA B catalysed ATP hydrolysis. This motion, displaces SSBs in its path. At randomly selected sites, the primosome reverses its migration while pimase synthesizes an RNA primer.

The initiation of primer synthesis requires the participation of Dna B protein, which through concomitant ATP hydrolysis is thought to alter DNA template conformation in a manner required by primase. Pol III holoenzyme extends the primers to form okazaki fragments. Pol I excises the primers and replaces them by DNA. The fragments are then joined by DNA ligase and supercoiled by DNA gyrase to form the ϕ x 174 RFI.

b) (+) strand synthesis

The ϕ x 174 (+) strand is synthesized on an RF-I template by a variation of rolling circle or σ - replication mode called the looped rolling circle mode. (+) strand synthesis begins with primosome-aided binding of the phage-encoded enzyme gene A protein (60 Kda) to its ---- 30-bp recognition site. There gene A protein specifically cleaves the phosphodiester bond preceeding (+) strand nucleotide 4306, by forming a covalent bond between a Tyr residue and the DNA's 5^1 -phosphoryl group. Rec protein subsequently attaches to the (-) strand at the gene A protein and, with the help of the primosome still associated with the (+) strand, commences unwinding the duplex DNA from the (+) stand's 5^1 end. The displaced (+) strand is coated with SSBP, which prevents it from reannealing to the (-) strand. Pol III holoenzyme extends the (+) strand from its free 3^1 -OH group. The extension process generates a looped rolling circle structure in which the 5^1 end of the old (+) strand remains linked to the gene A protein at the replication fork. When it has come full circle around the (-) strand, the gene A protein again makes a specific cut at the replication origin so as to form a covalent linkage with the new (+) strand's 5^1 end. Simultaneously, the newly formed 3^1 -OH group of the old, looped out (+) strand attacks its 5^1 -P attachment to the gene A protein, there by liberating a covalently closed (+) strand. The replication fork continues its progress about the duplex circle, producing new (+) strands in a manner reminiscent of linked sausages being pulled off a reel.

In the intermediate stages of a ϕ x 174 infection each newly synthesized (+) strand directs the synthesis of the (-) strand to form RF –I as described in 1.7.4.1. In the latter stages of infection, however, the newly formed (+) strands are packaged into phage particles.

1.4.1.5 Fidelity of Replication

Preservation of the genome of a species is entrusted to a replication process with an error frequency, in E. Coli cells, of 10^{-10} or less per bp. These rare replication errors are, principal source of the so-called spontaneous mutation. High replication accuracy comes from 5 mechanisms.

1. Maintenance of balenced levels of dNTPs

These are provided by fine regulation of key biosynthesis steps. Aberrantly high levels of a dNTP favor its misincorporation. Low levels of a dNTP invites the incorporation of miscorrect base. Ex. Incorporation of Uracil inplace of dTTP is minimized by the action of UTPase and maintenance of dTTP levels.

2. Watson-crick basepairing

Of a dNTP to the template acts as a checkpoint. The error frequency at their step is estimated to be 10^{-3} to 10^{-4} / bp.

3. Induced fits of polymerase and DNA

The polymerase's active site adapts to the size and shape of a correct base pair and conformational features of the DNA template are adjusted by bending and base stacking. At this sage the error frequency is reduced to 10^{-5} – 10^{-6} per bp.

4. Proof reading

The 3^{1-} 5^{1} exonuclease functions of pol I and pol III detect and eliminate the occassional errors made by their polymerase functions.

5. Mismatch correction / Repair

A remarkable battery of enzyme systems, contained in all cells, function to repair residual errors in the newly synthesized DNA, as well as any damage that it may occur after its synthesis through chemical and / or physical insults.

1.4.1.6 Eukaryotic Replication

1) Enzymology

Polymerase α

It is essential for the replication.The enzyme contains – a catalytic core subunit 170Kda.Two associated primase subunits 50 & 60 Kda. An additional polypeptide of uncertain function.

An essential role for pol α in chromosomal replication is based on many lines of evidence.

- Mouse cells with temperature sensitive mutant pol α fail to replicate DNA at the restrictive temperature.
- Antibodies against pol α inhibit DNA replication when introduced into permeabilised nuclei or cultured mammalian cells.
- The levels of Pol α are high in rapidly growing mammalian cells in culture and in proliferating lymphocytes.
- Pol α mutants are not diminished in repair of DNA damage. Pol α is not responsible for DNA repair.

- it has no 3^{1} -5^{1} exonuclease activity even then, the fidelity is high, one in 10^{6} nts.
- It is involved in discontinuous replication.

Polymerase δ

- it has unique intrinsic 3^1 -5^1 exonuclease activity.
- Pol δ requires PCNA (proliferating cell nuclear antigen), which is a processivity factor and needed for initiation of replication.
- Involved in continuous replication.

Polymerase ε

- It has similarities to pol δ
- It is distinguished from pol δ, by its high processivity independent of PCNA.

Polymerase β

- It is the smallest of eukaryotic polymerases
- It shows no correlation with chromosome replication.
- In regenerating rat liver and hepatomas pol α levels increase enoromously while those of pol β remain the same.
- Levels of the pol β transcript respond to doses of DNA-damaging agents.
- Pol β fills the single nucleotide gap in-short patch repair.

Polymerase γ

- It is the replicative polymerase of mitochondrial DNA.
- Has potent 3^1 -5^1 exonuclease activity.
- Low error rate, nearly one per 10^6 nts.
- Although pol γ represents only about 2% as much activity as α, and 10-25% as much as β.

Primase

- Primase activity is commonly a component of DNA pol α, a nuclear enzyme responsible for chromosomal replication.
- Primer length is 12-14 nts.
- The relatively low processivity of pol α, in addition to its associated primase makes it possible candidate for the synthesis of the discontinuous strand.

- Purified pol α contains

  - 180 Kda core
  - 50 & 60 Kda primase
  - 70 Kda no known function

The two primase units are active in pol-primase complex and also when they are separated. The two subunits function as a unit and not been separated from each other under primase activity conditions are maintained. Low the primer is moved from the active site on the small primase subunits to the DNA pol site on the 180 Kda polypeptide is not known.

Helicase

Proteins from many sources have the capacity to unwind DNA duplexes. For lack of adequate genetic foundation and purified systems for replication, recombination and repair, the functions of most of these helicases are still unknown. The exceptions are certain viral systems in which the invitro replication is advanced and genetic analysis is available.

SV40 large T antigen

The only viral protein required for SV40 replication. The large T – antigen binds specifically to the origin sequence, unwinds it during initiation and has NTP-dependent DNA & RNA helicase activities. By genetic analysis, T-antigen is essential for replication.

Translocates in 3^1 -5^1 direction and depends NTP hydrolysis with ATP preferred. The rate of duplex melting is only 75-100bp per min compared to that of 300-800 bp per second achieved by the prokaryotic replicative helicases.

SSBPs

Discoveries of such SSBPs have been limited, perhaps due to the abundance and prominence of histones in the distinctive nucleosomal organization of eukaryotic genomes. The most convincing examples of SSBPs are those disclosed by viral infections.RF-A (human SSBP; also called RP-A & protein A) It is made up of three tightly associated polypeptides of 70-76 Kda, 32-34 Kda, and 11-14 Kda. The purified protein binds tightly to singlestranded DNA and is required for the helicase action of the virus –encoded T-antigen in opening the SV 40 origin

Ligases

Coenzyme is ATP

- Product is Ppi

Topoisomerases

Topo I :

Type I topoisomerase

Eukaryotic topo I is a monomeric protein of --- 95 Kda. Although the reactions catalysed are generally the same as those of E. Coli topo I; +vely and –vely supercoiled DNAs are relaxed equally well. Relaxation is independent of ATP. DNA – covalent protein intermediate is formed with a specific tyrosine residue via a 3'-phosphate group. Rather than the 5'-p used by the E. Coli type I enzyme.

Topo II :

Type II Topoisomerase

Homodimers of 150-180 Kda subunits.Relaxes negative and +ve supercoils at equal rates. Doesnot introduce negative supercoils like E.Coli Gyrase. Requires ATP hydrolysis. U-terminal region is related to the B subunit of E.Coli Gyrase and it is the ATPase domain.The central portion is similar in sequence to A subunit which is involved in the nicking and closing activity. Reactions catalysed by eukaryotic topo II and E.Coli Gyrase are different.

2) Mechanism of Replication

In Drosophila, the DNA molecules are replicated in bi-directional fashion. Instead of beginning from a single point of origin, replication of eukaryotic DNA begins at multiple points of origins. This forms a number of eyes, as the parental strands of DNA are separated and replicated until they meet.

Multiple origins are necessary because of the great length of eukaryotic DNA and the relatively slow movement of replication fork, i.e, about 2600 bp per min compared to 16,000 bp per min in E.Coli.

If there were only a single replicating fork in eukaryotes, complete replication of the nuclear DNA would require atleast two weeks. However, a Drosophila egg cell completes replication in about 3 min and is believed to employ upto 6000 replication forks simultaneously.

Eukaryotic chromosome origin

Finding of specific initiation sequence has been difficult, however, due to the complexity of eukaryotic genomes and the limitations of assays to detect the initiation sites. The clearest evidence for specific replication origin is found in yeast.

Yeast ARS

Autonomous Replication Sequences isolated from chromosomes and naturally occurring plasmids in yeast. About 400 ARS elements are estimated to be present in the yeast genome. ARS elements are mode of two functional domains.

- Domain A, always present, consists of a II-bp sequence that has the consensus sequence (A/T) TTTAT (A/G) TTT (A/T). Mutations that alter this core sequence destroy ARS function. Analogous to the binding site of an initiator protein.
- Domain B is AT rich and extends 50-100 bp to the 3^1 side of the core. At rich character is thought to provide a region of DNA melting.

It is thought that initiation of replication is triggered by the binding of another component to the multiprotein ORC (origin recognition complex) that is already residing at the origin. This binding event is thought to stimulate the local unwinding of the DNA at the adjacent sequences.

The origin ones utilized becomes inactivated, by the formation of new histones and methylation of newly synthesized DNA by methyl transferases.

3) Mitochondrial DNA Replication

Mitochondrial DNA is replicated by a process in which leading strand synthesis preceeds lagging strand synthesis. The leading strand therefore displaces the lagging strand template to form a displacement or D loop.

During replication the D loop is extended. When it has reached a point ---- 2/3 of the way around the chromosome, the lagging strand origin is exposed and its synthesis proceeds in the opposite direction around the chromosome. Logging strand synthesis is therefore only --- 1/3 complete when leading strand synthesis terminates.

4) Termination

No polymerase is known that can extent a chain from the 5' end and thus fill the gap created at the end after the RNA primer has been excised on the lagging strand. How, then, are the DNA sequences at the ends of eukaryotic chromosomes, the telomeres, are replicated:

Telomeric DNA has an unusual sequence. It consists of upto 1000 or more tandem repeats of a simple, species – specific depedent, G-rich sequence concluding the 3' ending strand of each chromosomal terminus.

For example, the ciliated protozoan Tetrahymena has the repeating telomeric sequence TTGGGG, where as in humans it is TTAGGG. More over, this strand ends with a 12 to 16-bp overhang.

The synthesis of telomeric sequences follow the instructions of a 3'-AACCCCAAC (ribonucleotide) sequence contained within an essential RNA (about 160 nt) in the enzyme Telomerase.

The proposed mechanism entails hybridization of a long protruding 3' end of the DNA chain to the RNA sequence to permit elongation by TTG, further by GGGTTG by translocation back on this RNA template. Thus, growth of the 3' end of the telomere is by 6nt lengths of GGGTTG. So, a 3' end terminating at any nucleotide within the TTTGGGG sequence can be elongated to yield the perfect tandem repeats.

Synthesis of the $(AACCCC)_n$complementary chain to form the duplex telomere may be primed by a primase. The problem of filling gaps at the 5' end created by removal of the RNA primer is solved by the enoromous redundancy of the telomeric tandem repeats.

1.4.1.7 Inhibitors of DNA Replication

Inhibitors of DNA replication continue to serve as prime drugs for suppressing proliferative viral, bacterial and autoimmune diseases.To provide the laboratory investigator with the means to analyse biochemical pathways Invivo & Invitro.Inhibitors are especially attractive for studies of eukaryotic system in which mutant selection and genetic analysis are different.

Inhibitors of Topoisomerases

DNA replication at many stages, as well as other DNA transactions (Recombination, transcription) depends on the topological state of the DNA. Type I and II topoisomerases, found in all cells relax –ve supercoils, except E.Coli type II topoisomerase, which is unique in inducing –ve supercoils.

Inhibitors of topoisomerases have proved to be outstanding antibacterial drugs & highly promising as antitumor agents.

Coumarins

- Novobiocin, coumermycin A & chlorobioan.
- Related streptomyces – derives antibiotics containing comermycin & sugar moieties.
- Inhibits bacterial Gyrase (B subunit), eukaryotic topoisomerase II, vaccinia type I topoisomerase.

Alkaloids

Camptothecin – inhibits eukaryotic topo I.

Quinolones

Nalidixic acid, oxolinic acid, Norfloxacin – inhibits bacterial Gyrase (A subunit).

- Some replication systems are inhibited because RNAP provides the primer for DNA replication.
- Blocks the formation of first two phosphodiester bonds.

Inhibitors of DNA polymerases

- a tetracyclic diterpenoid antibiotic, inhibits replicative eukaryotic DNA polymerases.
- α and δ pols of yeast and animal, viral encoded pols, α-like pols of plants can be inhibited by aphidicolin.
- It doesnot affect the β and α - pols.

Phosphnoacetic acid and phosphonoformic acid

- Despite their simple structures, are effective antiviral drugs.
- They selectively inhibits the DNA pol encoded by herpes simplex, Vaccinia viruses.
- The viral pols are generally over 100 times more sensitive than the host cell enzymes.

Inhibitors of postreplicational modifications

Except in some phage systems, which synthesize alternative precursors, modifications of DNA and RNA are made after the chain has assembled. As the functional significance of these modifications become clearer, there is greater interest in the development of agents that affect these processes.

For example: The methylation of certain cytosine residues in mammalian DNA is likely to have an important influence on gene activation and cell differentiation.

The methyl donor is invariably 5-Adenosyl Methionine compounds that affect – methyl transferases, hydrolytic enzymes that remove inhibitory products, enzymes of Ado Met regeneration can have major sequences.

Example : 5-azacytidine → inhibits the methyl transferase –7-Deaza – S-adenosyl homocysteine → potent inhibitor of Ado met-dependent methylases and decreases the level of methylation.

1.4.1.8 DNA Repair

DNA is the only molecule which, when altered or damaged is repaired by the cell. A bacterial gene has a 50% chance of remaining unaltered even after having been duplicated 100 million times. This remarkable stability of DNA in all cells is in considerable measure due to variety of devices, for preserving its integrity and repairing any lesion it may sustain.

The repair systems are extra ordinarily diverse and effective. Perhaps, 100 loci in E.Coli are involved in DNA repair and related functions. The molecular mechanism for repair can be divided into –

- Those that reverse the damage

  i.e., photoreactivation, dealliylation

- Those that excise and replace the damaged unit by replication, recombination or the mis-match repair pathways.

When repair fails, continuity of the genome may be preserved through error-prone replication, in which bypass of the lesion permits replication to proceed.

1) Direct reversal of the damage

Photoreactivation by photolyases

The energy of visible light is used by photolyases to break the cyclobutyl pyrimidine dimer rings, restoring the bases to their monomeric form. Photoreactivation, as an alternate of excision, repairs any UV-induced cyclobutyl dimers.

All photolyases contain two chromophores.

1. $FADH_2$

2. Pterin)folate coenzyme in yeast and E.Coli)

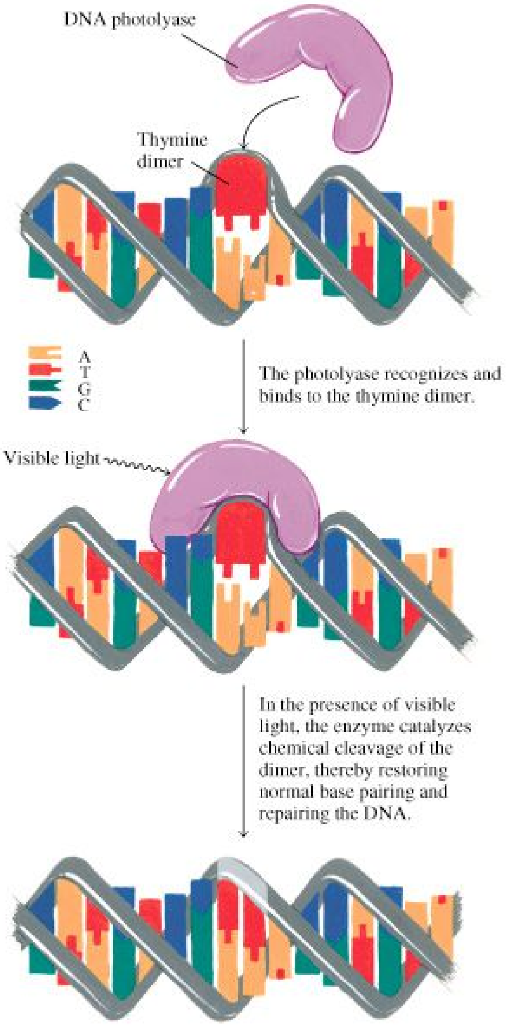
Or

Deaza flavin (other classes)

- In the first stage of photoreactivation, the enzyme recognizes and binds specifically to the dimer in the dark.

- When the lesion absorbs light (of a wavelength characteristic to the chromophore), the energy is used by the stable enzyme-DNA complex to convert cyclobutyl dimer to pyrimidine monomers.

- The enzyme then dissociates from the DNA.



The photolyase recognizes and binds to the thymine dimer.

In the presence of visible light, the enzyme catalyzes chemical cleavage of the dimer, thereby restoring normal base pairing and repairing the DNA.

Dealkylation

Another example of direct reversal of damage is the transfer of a methyl group from the precarcinogenic O^6-methyl guanine to a cysteire residue of an O^6-methyl guanine – DNA methyl transferase (Mtase).

The enzyme also removes an alkyl group from a phosphotriester by alkylation of another cysteine residue.

* Mtases are present in E.Coli, Yeast, and Mammalian cells.

2) Excision Repair

Repair of great variety of damaged and modified bases is achieved by excisions that remove

- the damaged base, creating an AP site
- a fragment containing an AP site
- the nucleotide lesion and neighboring region of DNA or
- an inserted crosslink.

The gap generated by excision is filled by DNA pol and then covalently joined by a ligase.

1. Base exicision

- Removes the lesion, there by creating an AP site.
- It is carried out by one of several N-glycosylases that recognize a de aminated or altered base or a helical deformation caused by the lesion and then hydrolyzes the bond linking the base to the sugar.

2. Excision of an AP region

This is achieved by either of two ways.

- By one pathway, a class II endonuclease makes the initial incision next to the AP site.

  After which the fragment is removed by an exonuclease.

- By the second route, a class I AP endonuclease, which possesses both N-glycosylase and AP endonuclease domains, uses the latter for an incision of -O-P-O bond to the 3'-side of AP site.

- removal of the AP gragment is then carriedout by a class II endonuclease (incises the backbone to the 5' side of the AP site).

AP II endonuclease → 3'→5' exonuclease activity

AP I endonuclease → the –O-P-O-bond is broken on either side of the AP site to initiate the removal of the abasic deoxyribose for its eventual replacement by the proper nucleotide.

DNA Glycosylase → hydrolyses N-glycosyl bond that links the base to the

Deoxyribose of the DNA backbone.

3. Oligonucleotide excision

Includes the region containing a UV dimer, bulky adduct, or interstrand cross link. Their excision is performed by an excinuclease, such as Uvr ABC complex of E. Coli.

- Uvr ABC, cut's out a 12-13 nt fragment from one strand.
- The resulting gap is filled by pol I and is sealed by ligase.
- The hydrolytic incisions are made at the 8th –o-p-o bond 5' to the lesion and 4th or 3rd bond 3' to it.
- Cross links are produced by many carcinogenic and chemotherapeutic agents.
- Excision requires repair of both strands.
- In E. Coli and mammalian cells, cross links are repaired by a combination of nucleotide exceision and recombinational repair, which requires Rec A recombinase and the assistance of Rec BCD nuclease, helicase.

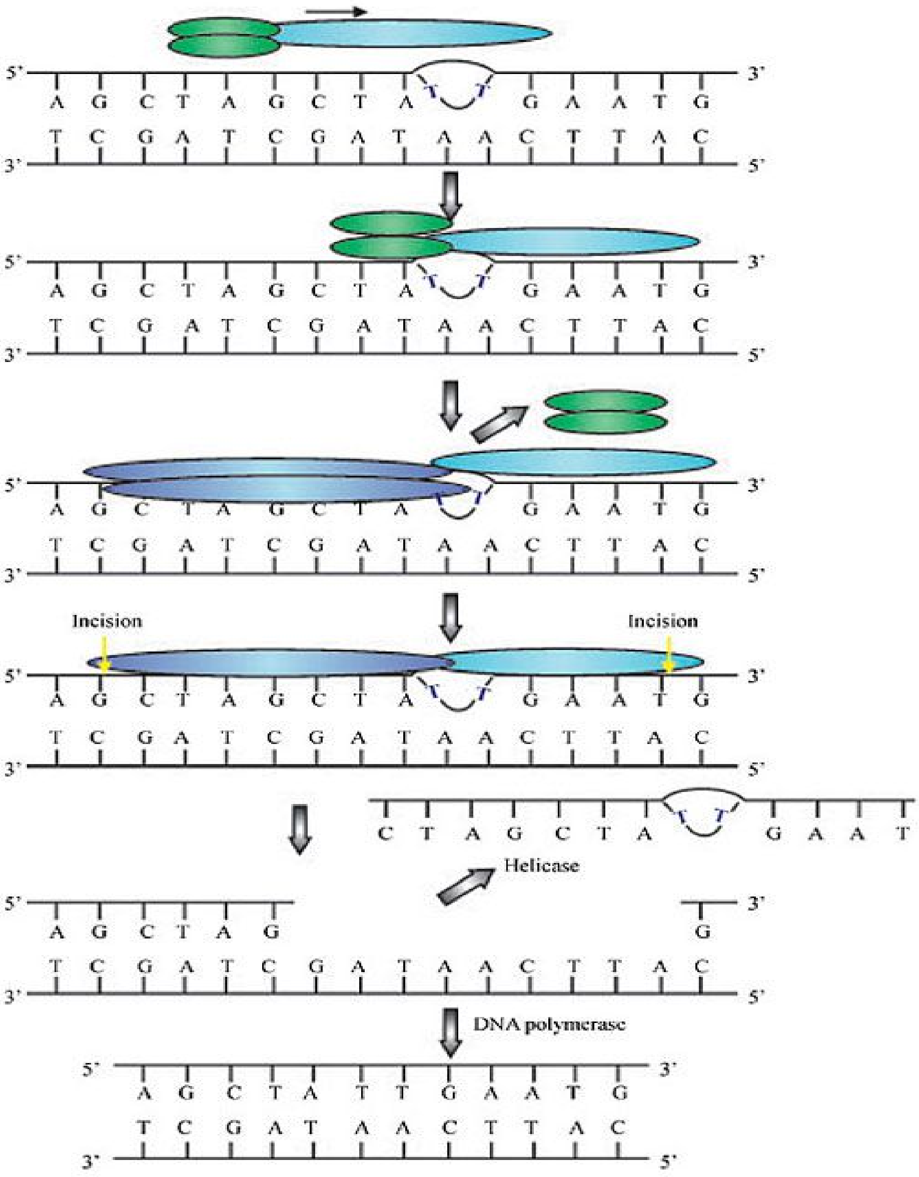


Figure 2. Schematic representation of nucleotide excision repair. The UvrAB heterodimer scans the DNA searching for large distortions in the helix such as the ones caused by pyrimidine dimers. Once a damaged site is found, UvrA proteins (dark green) dissociate, and a stable UvrB-DNA (light green) complex is formed. UvrC (blue) associates to bound UvrB and enables UvrB protein to nick the DNA at the fourth nucleotide 3' to the site of damage. Following the 3' incision, UvrC protein catalyzes nicking of the DNA at the seventh nucleotide, 5' to the damage. The potential oligonucleotide fragment that is generated is removed by a helicase. The remaining gap is filled up by polymerase synthesis and repair is completed by ligase.

4. Excision – Repair Patches

- Is produced by pol I in *E. coli* and is --- 20 nt long.

- These short-patches are longer in pol A mutants deficient in 5'→3' exonuclease function. These long patches may be several thousand nucleotides long.

- In mammalian cells. The short patch is only 3-4 nt long

  The long patch is about 35 nt.

- Short patches are produced by the damage of DNA from ionizing radiation and alkylation.

- Long patches are produced from DNA distorsion caused by UV and strand linkages.

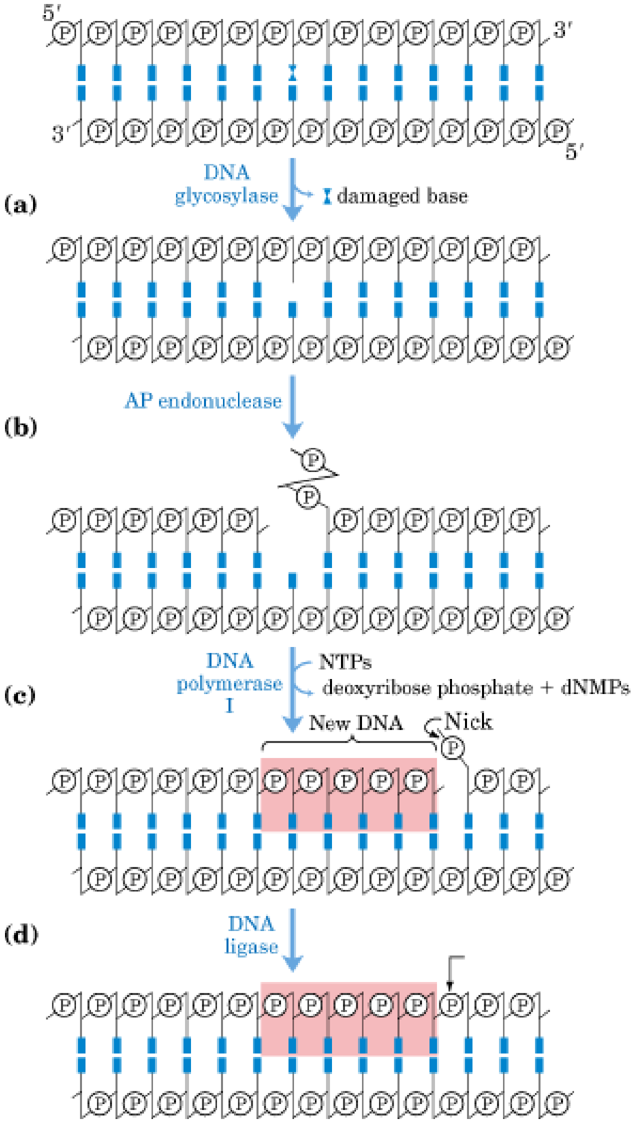
5. Translesion replication and recombinational repair

When a DNA polymerase encounters a pyrimidine dimer or certain other lesion in the template, either one of two things happens.

- The polymerase, as part of the SOS response, fills that spot by non template – directed (error-prone) replication across the lesion, employing the Uma C and Uma D proteins and pol III holoenzyme.

- The polymerase stops and then resumes 1000 nt or so down stream; the discontinuity or post replication gap may be filled with a complementary strand from the sister-duplex by Rec A mediated.

By either pathway, the lesion remains and must be removed subsequently by one or another of the direct-repair or excision-repair systems.



6. Mis Match Repairs

This repair system corrects mismatches within recombination intermediates. Ex: Thymine of a GT bp produced by spontaneous deamination of 5-methylcytosine.

Methyl-directed mis-match repair

- The newly synthesized DNA strand of a duplex can be identified because its GATC sites have not been methylated.

- Any of the possible base-base mispairs be discovered by this repair system, then that section of the unmethylated strand containing the mispaired base is removed.

- Mis Match repair depends on seven proteins.

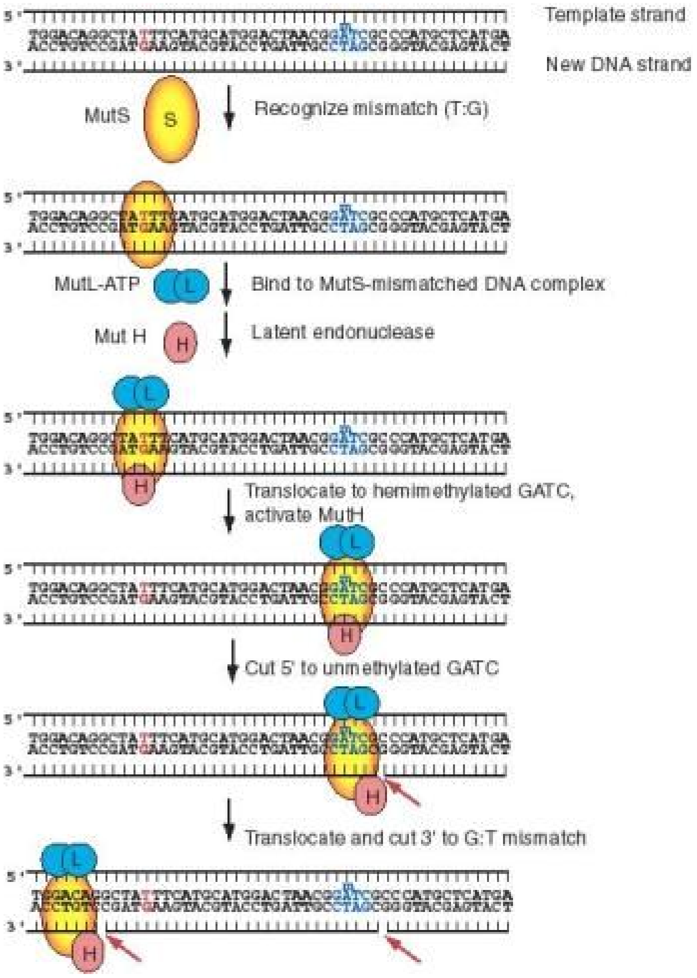
  Mut S, Mut L, Mut H, Mut U (helicase II, Uvr Dhelicase)

  Exonuclease I, SSBP and pol IV holo enzyme.

- The sequence of events is as follows.

  - Mut S forms a complex with the heteroduplex and is then joined by Mut L, Mut H.
  - The S.L.H complex is then translocated along the DNA, in either direction, for several thousand bps of necessary until it encounters an unmethylated GATC sequence in one strand of the duplx.
  - The endonuclease function of Mut H in the presence of ATP, incises that strand at the GATC sequence.
  - Excision of the incised strand, from the break upto and including the mismatch, depends on exonuclease I, helicase II and SSBP.
  - Concerted replacement of the DNA by pol III holoenzyme synthesis.

Mut S can recognize the slight helical distortion of an incorrect bp and the DNA can be tracked for great distances to identify which of the strand is newly synthesized in order to remove and replace the entire section with high efficiency.


5'
TGGACAGGCTATTTCATGCATGGACTAACGGATCGCCCATGCTCATGA
ACCTGTCCGATGAAGTACGTACCTGATTGCCTAGCGGGTACGAGTACT
3'
Template strand
New DNA strand
MutS
S
Recognize mismatch (T:G)
MutL-ATP
L
Bind to MutS-mismatched DNA complex
Mut H
H
Latent endonuclease
Translocate to hemimethylated GATC,
activate MutH
Cut 5' to unmethylated GATC
Translocate and cut 3' to G:T mismatch

Summary

The duplication of DNA is known as replication. There are different models were proposed for DNA replication. They are Semi conservative replication in which the daughter duplexes should be composed by one parent strand and one new strand Conservative Model in which the two original strands would remain together and two newly synthesized strands remain together. And Dispersive model in which the daughter cell contains duplexes in which each strand is composite of old and new DNA i.e., neither the strands nor the duplex itself is conserved. In these three models the semiconservative model is the correct model proved by Messelson and Sthal experiment.the DNA polymerase is the enzyme that synthesizes the DNA.IN addition to DNA pol number of proteins like ligase, topoisomerase, ssb, s helicase etc are involved in replication. There are different models of DNA replication like rolling circle replication. Looped circle mechanism etc. There different agents that damage the DNA by causing mutations that should be repaired otherwise they affects the normal cellular functions sometimes they causes cell death. The cell contains different repair mechanisms to repair the DNA damage. They are Methyl-directed mis-match repair recombinational repair Photo reactivation by photolyases, sos repair.

Model questions

1)Compare the process of replication of projkaryotes with eukaryotes

2)explain the different repair mechanismss

Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork.

Dr.N.Srinivasa Reddy

Lesson 1.4.2

TRANSCRIPTION

Objective

The process of synthesizing RNA from DNA is known as transcription .In this chapter we have explained the process of transcription and the inhibitors of transcription

1.4.2.1 Introduction

The information in DNA, encoded in the sequence of the four bases, is used to direct the assembly of 20 aminoacids in the correct sequence so as to produce the protein for which a given gene is responsible. A gene does not participate directly in the protein synthesis; in eukaryotes the DNA is enclosed inside the nuclear membrane while the protein-synthesizing machinery is outside in the cytoplasm and the two never meet.

The protein synthesis is done by sending out copies of its coded information to the cytoplasm. (In E.Coli, the copy is in immediate contact with the cytoplasm). Since, the information is in a sequence of bases, the copy must also be a nucleic acid, but this time it is RNA and not DNA, called mRNA or messenger RNA.

- mRNA is single stranded, not a duplex, i.e., mRNA is a copy of only one of the two strands of the DNA of a gene.

- Its 4 bases are A, C, G and U. There is no T.

The flow of information in gene expression is DNA Transcription mRNA Translation protein.

In copying DNA to RNA there is transcription of the information. Hence mRNA production is called gene transcription or simply TRANSCRIPTION and the DNA is said to be transcribed. The RNA molecules produced are called transcripts.

The DNA strand that acts as the template for mRNA synthesis is called the template and the other one is called the non-template strand. But the other terms coding and non-coding, sense & non-sense strands are used more commonly.

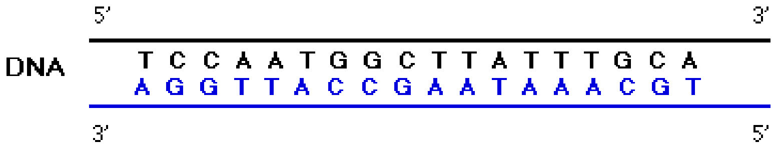
5^1 CGATGCAT 3^1 Non-template (coding / sense) strand DNA

3^1 GCTACGTA 5^1template (non-coding / non-sense) strand

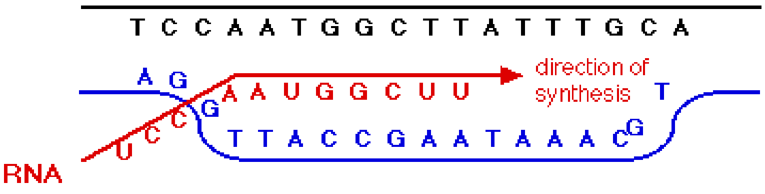
5^1 CGAUGCAU 3^1 mRNA strand.

Fig: Relationship of transcribed mRNA to template and non-template strands of DNA terminologies.

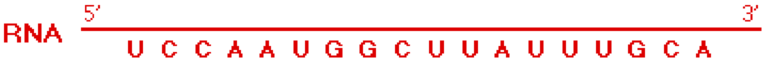
Transcription of RNA from DNA


5'
3'
DNA
T C C A A T G G C T T A T T T G C A
A G G T T A C C G A A T A A A C G T
3'
5'

- The bottom strand of the DNA molecule above is the template for RNA synthesis.
- RNA polymerase makes a copy of the DNA sequence but substitutes uridine (U) in place of thymine (T).


T C C A A T G G C T T A T T T G C A
A G
direction of synthesis
T
G A A U G G C U U
U C C C G
T T A C C G A A T A A A C G
RNA

- The bottom strand of the DNA duplex is used as the template to synthesize RNA. However, the sequence of bases in the RNA is the same as in the top strand of the DNA, with U in place of T



The base sequence of mRNA is same that of the non-template strand, this has the information for the sequence of aminoacids in protein. Hence this non-template strand is called coding or sense strand or (+1) strand in viruses. The template strand is called the non-coding or non-sense strand or (-) strand inn viruses.

1.4.2.2 RNA Polymerase (RNAP)

The major enzyme of transcription is RNA polymerase, specifically DNA-dependent RNA polymerase, simply called RNAP.

Similarities:

RNA & DNA pols have basically identical catalytic properties in the growth of polynucleotide chains. RNAP require all 4 dNTPs for complementary base-pairing with a DNA template, just as DNA pols require all 4 dNTPs. The newly synthesized chain propagates in 5'→ 3' direction.

Distinctions

RNA POL	DNA POLS
1. It can both start and stop a new chain when copying a duplex.	- Cannot start chain and generally rely on RNPs for primers.
2. Terminates chain at the end of a gene or an operon.	- Normally copy a template until it is exhausted. Specific replication termination signals have been identified; but they exert their effect by impeding helicase action on a duplex, rather than by affecting a DNA pol directly.
	- Inert on duplex DNAs, require nicked or frayed regions with auxillary proteins to melt the helical duplex.
3. The true template for RNAP is a duplex that undergoes localized and transient melting during transcription.	- Had ultrahigh fidelity in copying any template.
4. Fidelity in copying is uncertain.	

1.4.2.3 RNA Polymerase

- E. coli has a single DNA-derected RNA pol that synthesizes all types of RNA.

- It is the most extensively studied and is representative of the enzymes isolated from other bacterial genera-salmonella, Bacillus, serratia, proteus, Aerobacter etc..

- It is a large and complex enzyme containing totally six subunits with a composite mass of 448 Kda.

- The core enzyme contains 5 subunits, $\alpha_2\beta\beta^1$ w.

- Association of a σ subunit with the core constitutes the holoenzyme.

- all the RNAPs of bacterial genera are complex in organization and similar in properties except for those of phages T_7 and N_4.

- The structure of the enzyme contains a channel similar to the active site cleft on DNA pol I. This channel is about 25 A^o in diameter and 55 A^o in length.

- The structural similarity between polymerase I & RNA Polymerase is supported by a small amount of aminoacid sequence homology between pol I and the β-subunit of RNA P.

The σ subunit

- present only about 1/3rd the abundance of the core.

- Required only during initiation of transcription at a specific site.

- **σ** dissociates from the RNA P-template complex shortly thereafter and cycles to a new core.

- Both E.Coli and B.subtilis have multiple forms of σ**,** each responsible for recognizing a particular class of promoters.

- The predominant σ is 70Kda (σ 70) in E.coli and 43 Kda (σ 43) in β. Subtilis.

β' subunit

- antibiotics that bind RNA P and inhibit the action have provided important insig hts into the structure and functions.

- Rifampicin, for example, which binds firmly to the β subunit, completely blocks productive initiation of RNA chains by the enzyme in vivo and vitro.

- The polymerase – rifampicin complex apparently fails in performing the translocation step that follows the formation of the 1st or 2nd phosphodiester bond.

- Cells gain resistance to rifampicin by changing its β-subunit that fails to bind the drug.

- In addition to functioning in rifampicin binding, in transcription termination and interacting with the σ subunit, β also binds rNTPs and possesses an atom of Zn^{++}.

Theβ' subunit

- Compared toβ, little is known about the functions of other core subunits.

- Theβ' polypeptide also has a tightly bound Zn^{++}, is involved in DNA binding and is the site of action of polyanionic inhibitors such as heparin.

The α subunit

- No specific function is ascribed to α subunit.

The ω subunit

- it can be dissociated from the intact enzyme (holoenzyme) without apparent loss of activity and is not required in the reconstitution of an active enzyme.

- Its presence is required, however for transcription to be inhibited by Guanosine tetra phosphate (PPGPP) in vitro.

- PPGPP is thought to be a signaling molecule involved in the stringent control.

- Thus the `w' subunit is implicated in the regulation of transcription rather than directly in RNA synthesis.

Reconstitution of Holoenzyme

- Separate polypeptides, individually inert, become active when reassembled in the following order:

- $\alpha + \alpha \rightarrow \alpha_2 \rightarrow \alpha_2\beta \rightarrow \alpha_2\beta\beta' \rightarrow \alpha_2\beta\beta'\sigma$

- The requirement for all the subunits in reconstituting enzyme activity indicates that each is an essential component of RNA P.

- Assembly of the enzyme is required for subunit stability: overproduced normal polypeptides are also rapidly degraded in vivo for lack of subunits with which to assemble.

General reaction

The fundamental chemistry of RNA synthesis is has much in common with DNA synthesis. The overall reaction is

$$\underset{\text{RNA}}{(NMP)_n} + NTP \longrightarrow \underset{\text{Lengthened RNA}}{(NMP)_{n+1}} + Ppi$$

RNA P both initiates and terminates the chain. Formation of phosphodiester bond for chain growth is thus only one stage in a complicated sequence of reactions – template binding, site selection, initiation, elongation and termination.

1.4.2.4 Promoters

RNA synthesis is normally initiated only at specific sites on the DNA template called `promoters' that are recognized by the corresponding σ factor.

The existence of promoters was first recognized through mutations, which enhance or diminish the transcription rates of certain genes. Promoters lie on the upstream side of the RNA's starting nucleotide (+1), towards 5' end of the transcribing gene.

Analysis and comparison of sequences in many different bacterial promoters have revealed similarities in two short sequences located about to & 35 bp away from transcription start site and are represented as –10 & -35 consensus sequences. For most promoter in E.coli and related bacteria, the consensus sequence for the –10 region (also called pribnow box) is 5' TATAAT 3' and for –35 region is 5' TTGACA 3'.

Trp	TTGACA	N_{17}	TATAAT	N8	A
lac	TTTACA	N_{17}	TATGTT	N_8	A
rec A	TTGATA	N_{16}	TATAAT	N_7	A
ara BAD	CTGACG	N18	TACTGT	N_6	A

a) Promoter recognition

The σ subunit, as part of the holoenzyme recognizes both the –35 and –10 sequences in the promoter and probably contributes to the melting of the duplex to form open complex.

The multiple σ forms that are normally present in bacteria, those induced during spore formation in β subtilis and those appearing during infection by certain phages have provided insights into their structure and functions. At least 17σ factors have been identified, along with the promoter sequences which they recognize.

The major σ forms from E.coli (σ 70) and β subtilis (σ43) recognize the same sequences and direct the initiation of most genes.

The minor σ factors, commonly around 30Kda redirect the core polymerase to different classes of promoters, such as those for the heat shock genes in E.coli and the sporulation genes in Bacillus subtilis.

b) Chain initiation

- The σ subunit allows holoenzyme to move rapidly along a DNA strand in search of the σ subunit's corresponding promoter.
- The holoenzyme then migrates to the –35 region, forming the closed complex.
- The DNA is then unwound for about 17 base pairs beginning at the –10 region, exposing the template strand at the initiation site.
- The RNAP binds more tightly to this unwound region, forming an open complex.
- The RNA synthesis begins.
- The binding of RNAP to promoters is facilitated by the super coiling of the DNA.
- The initiating reaction of transcription is the coupling of two nucleoside triphosphate in the reaction:

  PPPA + PPPN ⟶ PPPApN + Ppi

- bacterial RNAs therefore have 5'-triphosphate groups as was demonstrated by the incorporation of radioactive label into RNA when it was synthesized with (- P) ATP.
- Only the 5' terminus of the RNA can retain the label because the internal phosphodiester groups of RNA are derived from the α-phosphates of NTPs.

c) Chain elongation

- the σ subunit is required only to ensure the specific recognition of the promoter by the RNAP.
- Once a few phosphodiester bonds re formed the σ subunit dissociates, leaving the care polymerase to complete the syntheses of the RNA molecule.
- The σ is recycled to initiate another round of transcription.
- The chain lengthens in 5'→ 3' direction (50nt/see), as was detected by using the antibiotic cordycepin.

- The topological problems caused by transcription are relieved through the action of topoisomerases.

- Moving of RNAP, produces +ve supercoils ahead of and –ve supercoils behind the point at which transcription is occurring.

d) Chain termination

- RNA synthesis proceeds until the RNAP encounters a sequence that triggers its dissociation.

- In E.coli there are atleast 2 classes of such fermination signals or terminators.

- One class relies on a protein factor called `p' (rho) and the other is p-independent.

- The P-independent class has two distinguishing features. The first is a region that is transcribed into self-complementary sequences, resulting in the formation of a hairpin structure, 15-20 nt before the end of the RNA.

- The 2nd is, a run of adenylates in the template strand that are transcribed into uridylates at the end of the RNA.

- It is thought the formation of hairpin disrupts part of the RNA-DNA hybrid in the transcription complex. The remaining hybrid duplex (oligo u-dA) contains a particularly unstable combination of bases and the entire complex simply dissociates.

ρ-dependent

- RNAP needs no help to terminate transcription at a hairpin followed by several u residues. At other sites, however, termination requires the participation of an additional factor.

- Some RNA molecular synthesized in vitro by RNAP alone are longer than those synthesized in vivo. The missing factor, a protein, was isolated and named `rho' `p'.

- The p factor is active as hexamer and specifically binds ssRNA, a stretch of 72nt, 12 per subunit. It also has an associated RNA-dependent NTPase activity (specially ATP)

- The ATPase activity of ρ enables it more unidirectional along nascent RNA toward the transcription bubble. It then breaks stops the transcription process.

Evidences for P function

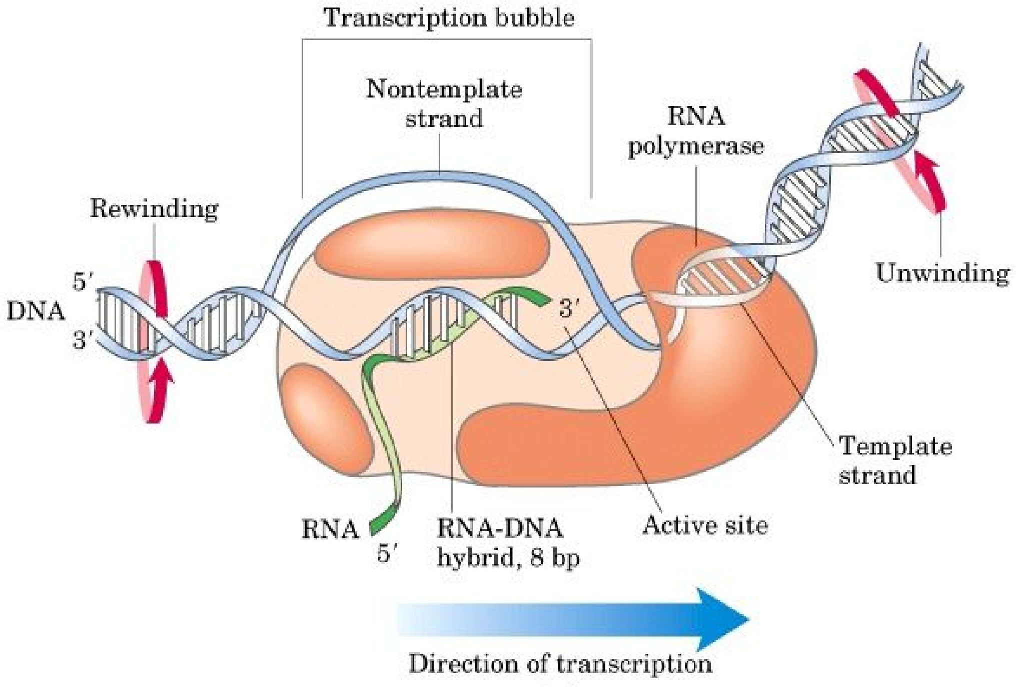
Additional information about the action of ρ was obtained by adding this termination factor to an incubation mixture at various timer after the initiation of RNA synthesis. RNAs with sedementation coefficients of 13S, 17S and 23S were obtained, when P was added a few seconds, 2 min, and 10 min respectively after initiation.

It is evident that the template contains atleast three termination sites that respond to ρ (yielding 10S, 13S & 17S) and one termination site that doesnot require --- (23s RNA)

However, ρ detects additional termination signals that are not recognized by RNAP alone.

1.4.2.5 Eukaryotic RNAP

In contrast to pro karyotes, eukaryotes have three different RNAPs, each transcribing a particular class of genes. The three RNAPs are designated as RNA pol I, RNA pol II, and RNA pol III, were first resolved by RoEDER and RUTTER in 1969, as three distinct proteins electing at different salt concentrations during ion-exchange chromatography.



(a)

1) RNA pol I

- Present in nucleus
- Synthesizes only one type of RNA, called pre-rRNA, which consists the precursors for the 18s, 5,8s, 28s rRNAS.
- Very insensitive to α-amanitin (upto 10μg/ml)

2) RNA pol II

- Present in nucleoplasm
- Transcribes all protein-coding genes, ie, it functions in the production of mRNAs.
- It also produces four small nuclear RNAs that participates in RNA splicing.
- Very sensitive to α-amanitin (0.1 μg/ml)

3) RNA Pol III

- present in nucleoplasm and cytoplasm
- transcribes the genes encoding tRNAs, 5S rRNA and also small, stable RNAs.

Ex : U_6 → involved in RNA splicing

7S RNA of SRP (signal recognition particle) involved in transport of protein into the Emdoplasmic reticulum.

- moderately sensitive to α-amanitin (1-10 μg/ml)

Each of the eukaryotic RNA polymerase is more complex than E.coli RNA polymerase. All 3 contain two large subunits and 12-15 smaller, some of which are present in two or all three polymerases.

The best characterized eukaryotic RNAPs are from the yeast S.cerevisiae.

- The largest subunit (160-220Kda) and 2nd largest subunit (128-150Kda) of each eukaryotic RNAPs are related to the E.coli β' & β subunits respectively.
- Both yeast RNAP I and III contain α subunits (19 7 40Kda), that have sequence homology with E.Coli α-subunit.

- Yeast RNAP II contains 2 copies of a different subunit (44 Kda) that exhibit most distant similarity with E.Coli α-subunit.

- In addition to their core subunits related to the E.coli polymerase subunits, all 3 yeast RNAPs contain 5 small common subunits (10-27 Kda).

- In addition, each RNAP has 4-7 enzyme specific subunits that are not present in the other two RNAPs.

The functions of these multiple polymerase subunits are not understood, but gene-knockout experiments in yeast indicate that most of them are essential for cell viability.

4) Promoters

- As prokaryotic genes, eukaryotic genes also contain promoters to the 5'side of start site.

- The start site itself has a recognizable short sequence between nucleotides –3 and +5, called the initator region.

- About 80% genes have a TATA box centered at about ---- 25 to 35 basepairs upstream of the start site. This highly conserved sequence is also called Goldberg – Hogness box

- Further upstream, within about 100-200 or so basepairs of the startsite, are elements or boxes – short specific DNA sequences that are recognized by specific proteins called. Transcriptional factors. Three common upstream elements are the CAAT box, the GC box and an 8 bp octamer box. These are located at sites within –100 to –200 region in different genes.

- These control elements, together with the TATA box and/or initiator often are referred to as the promoter of the gene they regulate. However, we prefer to reserve the term promoter for the TATA box or initiator sequences and the term promoter-proximal elements for control regions lying within 100 – 200 bps upstream of the start site.

- 20% of genes, including many house keeping genes, lack TATA boxes, some lack other elements, some have multiple copies of an element.

5) Enhancers

The activities of many promoters in higher eukaryotes are greatly increased by sequences called `enhancers'.

- They are relatively large elements, often including several 100 bps and sometimes contains repeated sequences that are independently functional.
- They can act over considerable distances, upto several thousand bps.
- They can be upstream, downstream or even in the middle of a transcribed gene.
- The enhances element and the gene must always be present on the same DNA molecule.
- A particular, the Ig enhances functions in β-lymphocytes but not else where.
- Enhancers act only in cells that contain cognate stimulatory proteins.

Ex: The steroids interact with soluble receptors. These hormone – Receptor complexes bind to the glucocorticoid enhancers and then lead to stimulation of transcription of a distinctive set of genes.

- These control elements, together with the TATA box and/or initiator often are referred to as the promoter of the gene they regulate. However, we prefer to reserve the term promoter for the TATA box or initiator sequences and the term promoter-proximal elements for control regions lying within 100-200 bps up stream of the start site.
- 20% of genes, including many house keeping genes, lack TATA boxes, some lack other elements, some have multiple copies of an element.

Enhancers

The activities of many promoters in higher eukaryotes are greatly increased by sequences called `enhancers’

- They are relatively large elements, often including several 100 bps and sometimes contain repeated sequences that are independently functional.
- They can act over considerable distances, upto several thousand bps.
- They can be upstream, downstream or even in the middle of a transcribed gene.
- The enhancer element and the gene must always be present on the same DNA molecule.
- A particular enhance is effective only in certain cells, for example, the Ig enhancer functions in β-lymphocytes but not elsewhere.
- Enhancers act only in cells that contain cognate stimulatory proteins.

Ex: The steroids interact with soluble receptors. These hormone- Receptor complexes bind to the glucocorticoid enhancers and then lead to stimulation of transcription of a distinctive set of genes.

Enhances are not the only DNA elements that can act at a distance to control transcription of a gene. There are certain elements which prohibit transcription, called Silencers. It is clear that silencers depend on silencer proteins, that bind to the silencer DNA and some how cause repression of surrounding genes. The mechanism of this repression is still unknown.

1.4.2.6 Transcriptional factors

In prokaryotes, RNAP recognizes the correct binding site on a promoter and binds directly to the DNA helped in somecases by for example, CAP. This does not occur in eukaryotes. All the genes in chromatin are basically in `shutdown' or repression state because, a nucleosome blocks the initiating region of each gene promoter. A gene cannot be transcribed until the nucleosome is displaced, thus allowing the basal initial complex formation. This formation involves several proteins, collectively called as Transcriptional factors.

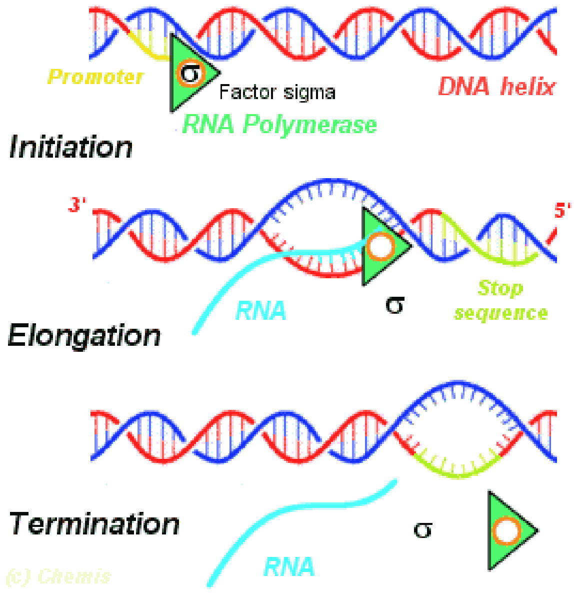
1.4.2.7 RNA synthesis and termination

Transcription by eukaryotic RNAP leads to synthesis of primary transcripts which, are longer than the mature, functional RNAs consequently, the primary transcripts require and undergo one kind of processing or the other before they finish up as mature RNAs.

Electron microscopic pictures of transcription of rRNA genes show that definite start and termination sites exist for RNAP I. Termination occur a few 100 bases past the 3' end of the 28S RNA gene.

Transcription of 5S RNA genes by eukaryotic RNAP III involves precise termination to produce 5S RNA straight away. Transcription terminates in a run of U s.

RNAP II termination is closely linked to processing. However, one general feature is that termination occurs anywhere between a few 100 to a few 1000 bases downstream of the 3' end of the coding sequence. It has been suggested that loss of processivity of pol II rather than a precise signal may determine termination.



1.4.2.8 Inhibitors of transcription

A variety of antibiotics and other inhibitors affect transcription at different stages.

Actinomycin- D

- it is a streptomyces antibiotic
- acts as antitumor agent but is highly toxic.
- It is a planar, tricyclic compound with a phenoxazone ring containing two –COOH groups, each of which carries a cyclic peptide chain.
- Binds to ds DNA at GC rich regions.
- Inhibits replication as well as transcription.
- At low concentrations (10^{-6} M), specifically inhibits DNA – dependent RNA polymerases.

Rifampicin, Rifamycin

- inhibitors of bacterial RNA polymerase.
- Have long aliphatic side chains attached at both ends to planar aromatic ring systems.
- Rifampicin specifically inhibits the β subunit.
- Elongation is not affected.

α - amanitin

- an extremely toxic octapeptide from mushrooms
- Inhibits eukaryotic RNAPs.
- Has one residue each of L-hydroxy proline, L-Aspargine, L-hydroxy isoleucine and 2 glycine residues linked to a central tryptophan, itself attached to an oxidized cysteine moiety.
- Obtained from poisonous mushroom Amanita phalloides.
- Forms a tight 1:1 complex with RNAP II and a looser with RNAP III so as to specifically block their elongation steps.
- RNAP I, as well as mitochondrial, chloroplast and prokaryotic RNAPs are insensitive.

Cordycepin

- naturally produced by cordyceps militaris
- inhibits transcription by getting incorporated at 3' end of the growing transcript.
- Results in chain termination, since chain elongation is not possible due to the absence of free 3'-OH group.

Summary

The synthesis of RNA from DNA is known as Transcription carried out by the enzyme RNA polymerase.In case of prokaryotes a single RNA polymerase synthesizes all the RNA's. But in the case of Eukaryotes three RNA polymerases are present RNA polymerase 1 synthesizes 5.8s, 18s, 28s rRNA's .RNA polymerase synthesizes mRNA, snRNA's. RNA polymerase III synthesizes 5s rRNA,tRNA,U_6s_nRNA. RNA synthesis is normally initiated only at specific sites on the DNA template called `promoters'.the eukaryotes contain specific DNA sequences that increase the rate of transcription known as Enhancers.The prokaryotic RNA polymerase doesn't require any proteins for the transcription ,it can initiate ,elongate and terminate the RNA synthesis. But eukaryotic RNA polymerases need proteins known as transcriptional factors.The transcription in prokaryotes is inhibited by Rifampicin, Rifamycin,Actinomycin wher as eukaryotic RNA polymerase are inhibited by amanitin,cordycepin

Model questions

1) Explain the process of transcription

2) Write a note on inhibitors of transcription

Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork

Dr.N.Srinivasa Reddy

Lesson 1.4.3

PROTEIN SYNTHESIS

Objective

This chapter gives an idea about protein synthesis from m-RNA with the help of t-RNA, ribosomes, and inhibitors of translation

1.4.3.1 Introduction

Proteins are the end products of most information pathways A typical cell requires thousands of different proteins at any given moment. These must be synthesized in response to the cell’s current needs, transported to their appropriate cellular locations, and degraded when no longer needed.

-It is the most complex biosynthetic process

-Overall, almost 300 different macromolecules cooperate to synthesize polypeptides

- To different ribosomal proteins
- 20 or more enzymes to activate the amino acid precursors
- a dozen or more auxiliary enzymes

- other protein factors for the initiation, elongation, termination of polypeptide synthesis
- 100 additional enzymes for the final processing of different proteins.

If accounts for up to 90% of the chemical energy used by a cell for all biosynthetic reactions. The 20,000 ribosomes, 100,000 related protein factors and enzymes and 200,000 t-RNAs in a typical bacterial cell can account for more than 35% of the cell's dry weight. Despite the great complexity proteins are made at exceedingly high rates.

Example: a polypeptide of 100 residues is synthesized in E.Coli cell in about 5 sec.

1.4.3.2 Transfer of RNA or t-RNA

In 1955, crick proposed "adaptor hypothesis", according to which translation occurs through the mediation of adaptor molecules. Each adaptor was postulated to carry a specific enzymatically appended aminoacids and to recognize the corresponding codon.

Crick suggested that these adaptors contain RNA because codon recognition would then occurs by complementary basepairing. At about this time, Zamecnik & Hogland discovered that, in the course of protein synthesis, c^{14}-labeled aminoacids became transiently bound to a low molecular weight fraction of RNA. Further investigations indicated that these are tRNAs which were first called as sRNAs. (Soluble RNA's) . tRNAs are relatively small and consists of a single stand of RNA folded into 3-dimensional structure. In bacteria & in cyosol of eukaryotes, tRNAs have between 73 and 93nt, corresponding to molecular weight 24-31Kda. There is at least one kind of tRNA for each aminoacid, for some aminoacids there are two or more tRNAs.

Structure

Many t-RNAs have been isolated in homogenous form. In 1965, Robert W.Holley & his collegues reported the first known base sequence of yeast Alanine tRNA ($tRNA^{Ala}$)

- This is the very first nucleic acid to be sequenced.

- It has 76nt residues, 10 of which have modified bases.

Fig: The base sequence of yeast $tRNA^{Ala}$.

- Since then, the base sequences of many other tRNAs from various

Species have been solved and revealed many common features.

- 8 or more of the nucleotide residues of all tRNAs have unusual

modified bases, many of which are methylated derivatives of the bases.

- Most tRNAs have a guanylate residue at 5^1end and all have the

trinucleotide sequence CCA(3^1) at the 3^1end.

- All the tRNAs have a common secondary structures, called cloverleaf structure and have the following common features.

- a 5^1-terminal phosphate group
- a 7bp stem including the 5^1 terminal nucleotide and that may contain non-watson-crick basepairs such as G.U. This is known as acceptor arm or aminoacid stem because the aminoacid residue carried by the tRNA is appended to its 3^1-terminal OH.
- A 3- or 4- basepair stem ending in a loop that frequently contains the modified base DHU (dihydrouracil). The stem and loop are therefore collectively termed the D arm.
- A 5 basepair stem ending in a loop that usually contains the anticodon, the triplet of bases complementary to the codon. This is called anticodon arm.
- A 5 basepair stem ending in a loop that usually contains the sequence TψC, and called TψC arm.
- The longer tRNAs have a short variable extra arm. It has 3-21 nucleotides and may have a stem consisting of upto 7 basepairs.

Tertiary Structure

- In 1974, Alexander Rich in collaboration with, Sung Hou Kin eludicated

the X-ray crystal structure of Yeast $tRNA^{Phe}$.

The molecule assumes an L-shaped conformation in which one leg of L is formed by the acceptor and T-stems folded into a continuous double helical like structure and the other leg is similarly composed of the D and anticodon stems. Each leg of the L is ~60 A° long and the antocodon and aminoacid acceptor sites are at opposite ends of the molelcules, some 76 A^0 apart. The narrow 20-25 A^0 width of native tRNA molecule is essential to its biological function.

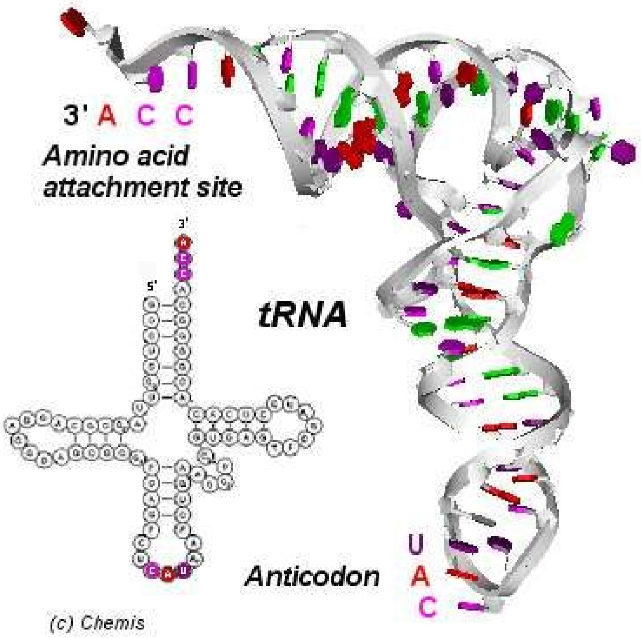


Fig: tertiary structure of tRNA molecule.

tRNAs complex tertiary structure is maintained by H-bonding as well as stocking interactions. The structure also contains 9 basepairing interactions that cross link its tertiary structure. Remarkably all but one of these tertiary ineractions are non-watson-crick interactions, and they are all located near the corner of the L structure. The structure is also stabilized by several unusual H-bonds between bases and either phosphate groups on 2`-OH group of ribose residues.

Aminoacyl –tRNA synthetases

Accurate translation requires the covalent attachment of the correct aminoacid to a tRNA, which is catalysed by aminoacid specific enzymes known as aminoacyl-tRNA synthetases (aa RSs). These enzymes append an aminoacid to the 3`-terminal ribose residues of corresponding tRNA to form an aminoacyl-tRNA. This unfavourable process is driven by the hydrolysis of ATP in two sequential reactions that are catalyzed by a single enzyme.

1. The aminoacid is first "activated" by reaction with ATP to form an aminoacyl – adenylate

$$\begin{array}{lcl} \text{R-CH-COOH+ATP} & \rightarrow & \text{R-CH-C-(AMP) + Ppi} \\ \quad\;\; | & & \quad\;\; | \\ \quad NH_2 & & \quad NH_2 \end{array}$$

This intermediate may be isolated although if normally remains lightly bound to the enzyme.

2. This anhydride then reacts with tRNA to form the aminoacyl-tRNA

Aminoacyl – AMP +tRNA → aminoacyl – tRNA + AMP

Some Aminoacyl – tRNA synthetases exclusively append an aminoacid to the terminal 2^1OH group of their corresponding tRNAs, and other do so at the 3^1-OH group. This selectivity was established with the use of chemically modified tRNAs that lack either the 2^1or 3^1OH group of their 3^1-terminal ribose residues.

The overall aminoacylation reaction is:

Aminoacid +tRNA +ATP→aminoacyl+tRNA+AMP+Ppi

These reaction steps are radily reversible because the free energies of hydrolysis of the bonds formed in both aminoacyl – AMP and aminoacyl – tRNA are comparable to that of ATP hydrolysis. The overall reaction is driven to completion by the pyrophosphatase catalysed hydrolysis of the pyrophosphate generated in the first reaction step.

As elucidated by paul Berg, tRNA is the acyl acceptor is aminoacid activation, where as CoA performs this function in fattyacid activation.

There are two unrelated families of aminoacyl – tRNA synthetases, termed class I and class II aminoacyl RNA synthetases. Each have the same 10 member in all organisms. Detailed sequence and structural comparison of these two families is given by Diro Moras.

Many class I aaRSs require anticodon recognition to aminoacylate their cognate tRNAs. Incontrast, several class II enzymes do not interact with their bound tRNAs anticodon. All class I enzymes aminoacylate their bound tRNAs 3`-terminal 2^1-OH group. Where as class II enzymes charge the 3^1-OH group.

Prokaryotic aaRSs occur as individual proteins. In many higher eukaryotes aa RSs associate to form a multienzyme complex. The advantages of this system are unknown.

Non-sense Suppression

Non-sense mutations are usually lethal when they prematurely terminate the synthesis of an essential protein. Restoring the gene to its normal function requires a 2nd mutation that either converts the termination codon to a codon specifying an aminoacid or alternatively suppresses the effects of the termination codon. The 2nd class of restorative mutations are called non-sense suppresors – generally involve mutations in tRNA genes that produce altered tRNAs that can recognizes the termination codon and insert an aminoacid at that position. Most suppressor tRNAs are produced by single base substitutions in the anticodons of minor tRNA species.

Non sense suppression doesnot completely disrupt information transfer in the cell. This is because, usually there are several copies of the genes for some tRNAs in any cell; some of these duplicate genes are weakly expressed and account for only a minor part of cellular pool of a particular tRNA. Supressor mutations usually involve these minor tRNA species, leaving the major tRNA to read its codon normally.

For example

There are three identical genes for $tRNA^{Tyr}$ in E.coli, each producing a tRNA with the anticodon 5^1GUA 3^1. One of these is expressed at relatively high levels-major $tRNA^{Tyr}$; the other two are transcribed in only small amounts. A change in the anticodon from 5^1GUA to 5^1 CUA, produces a minor $tRNA^{Tyr}$ species that will insert tyrosine at UAG stop codons. This insertion of Tyr at UAG is inefficient, but can permit production of enough useful full-length protein from a gene with a non-sense mutation to allow the cell to live.

Some E.coli Nonsesnse suppressors

	Codn suppressed	aminoacid iserted
SU1	UAG	Ser
SU3	UAG	Tyr
SU6	UAA	Leu
UGA1	UGA	Trp
UGA2	UGA	Trp

1.4.3.3 Ribosomes

-1st observed by albert claude (1930s) and referred them to as Microsomes – by darkfield microscopy.

-in mid 1950s, George palade observed them in cells by electron microscopy – ribosomes.

Contain 2/3rd of RNAS & 1/3rd of protein.

Bacterial ribosomes

-There are 15,000 ribosomes or more in E.Coli cell and make up 1/4th of the cell's dry weight.

-Contain 65% RNA and 35% protein

-Have a diameter of about 18nm and sedementation coeffieicnt of 70s

-Can be disassociated into two unequal subunits as given by lalatson

-The larger one having a S of 50s and

-The smaller 30s subunit

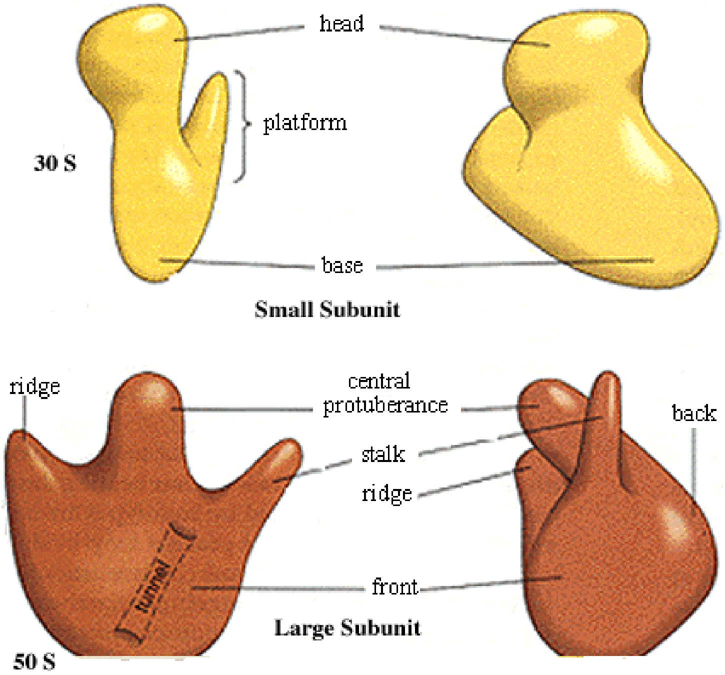
-The 50s submit contains one molecule of 5s rRNA, one molecule of 23S rRNA and 34 proteins.

-The 30s submit contains one molecule of 16s rRNa and 21 proteins.

-The proteins are designated by numbers. Those in the large 50s subunit are numbered L1-L34 and those in smaller subunit S1 to S21.

-Each of the 55 proteins in E.coli ribosome is believed to play a role in the synthesis of polypeptides, either as enzymes or as a structural component in the overall process.

-The two oddly shaped subunits fit together in such a way that a cleft is formed through which the mRNA, passes as the ribosome moves along it during the translation.



Prokaryotic Ribosomes (70 S)

Bacterial ribosome

Eukaryotic ribosomes

-larger and more complex than bacterial ribosomes

-have adiameter of about 23nm and as of about 80s

-they also have subunits which vary in size between species but on average are 60s and 40s

-the small subunit contains 18s rRNA and ~ 33 proteins.

-The large subunit contains 5s rRNA, 5.8s rRNA, 28s rRNA and 49 proteins

-in contrast, the ribosomes of mitochondria and chloroplasts are some what smaller and simpler than bacterial ribosomes.

Structure and functionalities of ribosome subunits

-The 3`end of the rRNA participate in mRNA binding – located on the small subunit's platform

-The anticodon binding sites occur in the small subunit's cleft region

-Large subunit's stalk participate in the ribosome's various GTP^{ase} reactions.

-The peptiyl transferase function (p) occupies the valley, between the large subunit`s other two protrubarences

Thus the large subunit appears to be mainly involved in mediating biochemical tasks such as catalyzing the reactions of polypeptide elongation, where as the small subunit is the major ancor in ribosomal recognition process, such as mRNA and tRNA binding processes.

1.4.3.4 Polypeptide synthesis

<u>Polypeptide synthesis begins at the Amino-terminal end</u>

The direction of ribosomal polypeptide synthesis was established in 1961, by Howard Dintzis. Reticulocytes, that were actively synthesizing hemoglobin were incubated withradioactive leucine. (Leu occurs frequently along both the α-and β-globin chains) samples of completed β-chains were isolated from the reticulocytes at various times after addition of radioactive leu. The distribution of radioactivity along the β-chain was determined, with the expectation that it would be concentrated at the end that was synthesized last.

Isolated after 60min of incubtion, rearly all the Leucine residues were radioactive in β-chains. However, in completed globin chains that were isolated only a few min after *leu was added, * Leu were concentrated at the c-terminal end.

Fig. Proof of polypeptide chains grow in N→C terminal

From these observations it was concluded that polypeptide chains are begun at the N-terminal end and are elongated by sequential addition of residues to C-terminal end. This applies to all proteins in all cells.

Ribosomes read mRNA in 5`→ 3` direction

The direction that the ribosome reads mRNAs was determined through the use of a cell-free protein synthethesis system in which the mRNA was poly(A) with the 3-terminal cytosine.

5` A-A-A----------------------A-A-A-C-8`

Such a system synthesizes poly(Lys) that has C-terminal Aspargine.

H_3N^+-Lys-Lys-Lys……………………Asn – coo-.

This together with the knowledge that AAA and AAC code for Lys and Asn and the polarity of polypeptide synthesis, indicates that the ribosome reads mRNA in the 5`→ 3` direction. Since, mRNA is synthesized in the 5`→ 3` direction, this accounts for the observation that, in prokaryotes, ri9bosomes initiate translation on nascent mRNA.

1.4.3.5 Prokaryotic Translation

The complex process of translating mRNA into protein can be divided into three stages

- initiation
- Elongation
- Termination

Initiation

The initiation of polypeptide synthesis in bacteria requires

- the 30s ribosomal subunit which contain 16s rRNA
- the mRNA coding for the polypeptide to be synthesized
- initiating fMet - tRNA^{tmet}
- a set of three proteins called Ifs
- GTP
- 50^{s} ribosomal subunit

A specific aminoacid initiates protein synthesis

Although there is only one codon for Met (AUG), there are two tRNAs for Met in all organisms. One tRNA is used exclusively when AUG represents the initiation codon for protein synthesis. The 2nd is used when Met is added at an internal position in a polypeptide.

In bacteria, the two tRNAs specific for Met are designated as tRNA^{Met} and tRNA$^{iMet.}$ The starting aminoacid at N-terminal end is N-formyl Methionine. It enters the ribosome as N-formyl Methionyl-tRNA^{fMet}, which is formed in two successive reactions.

1. Methionine is attached to tRNA^{tMet} by the Met=tRNA synthetase. There is only one of these enzymes in E.Coli and aminoacylates both tRNA^{Met} and tRNA^{tMet} .

$$\text{Met} + \text{tRNA}^{tMet} + \text{ATP} \rightarrow \text{Met-tRNA}^{tMet} + \text{AMP} + \text{ppi}$$

2.A formyl group is transferred to the $-NH_2$ group of the Met residue from N^{10} – formyl tetrahydrofolate by a trans-formylase enzyme.

N^{10}-formyl FM_4+Met-$tRNA^{tMet}$→fMet-$TRNA^{fMet}$+FM_4

This transformylase is more slective, it is specific for Met residues attached to $tRNA^{tMet}$. Presumably recognizing some unique structural feature of that tRNA.

Initiation Complex

To begin the assembly of a bacterial translation complex, sequential interactions occur between specific proteins initiation factors and the small 30s ribosomal subunit. The resulting preintiation complex, together with fMet-$tRNA^{fMet}$ then binds the mRNA at specific sequence near the initiating AUG and thus forms the 30s initiation complex. The formation of initiation complex takes place in three steps.

1. The 30s ribosomal submit binds IF-3 which prevents the 30s and 50s subunits from combining prematurely.

-binding of mRNA to the 30s subunit then takes place in such a way that the initiation codon binds to precise location on 30s Ribosomal subunit.

-the initiating AUG is guided to the correct position on the 30s ribosomal subunit by an initiating signal called the shine-dalgarno sequence in the mRNA,centered 8-13 bases from the 5`end of AUG.

-The sequence generally contains 4-9 purine residues and basepairs with the complementary pyrimidine rich sequence at the 3`end of 16s rRNA of 30s ribosomal subunit.

-This mRNA –rRNA interaction fixes the mRNA so that the AUG is correctly positioned for initiation of translation.

-AUG is positioned in the P or peptidyl site, which is the only site to which fMet-$tRNA^{fMet}$ can bind.

2. Complex consisting of the 30s subunit, IF-3, mRNA now forms a still larger complex by binding IF-2, which already is bound to GTP and the initiating fMet-$tRNA^{fMet}$. The anticodon of this tRNA pairs correctly with initiation codon in this step.

3. This large complex combines with the 50s ribosomal subunit. Simultaneously, the GTP molecule bound to IF-2 is hydrolyzed to $GDP^{+}Pi$. IF-3 and IF-2 also depart from the ribosome.

- The correct binding of the fMet – tRNA^{tMet} to the P site in the complete 70s initiation complex is assured by Codon-anticodon interactions

- Binding interactions between the P site and f Met-tRNA$^{fMet.}$

Elongation

The stepwise addition of aminoacids to the polypeptide chain requires – the initiation complex

-the next aminoacyl-tRNA sepecified by the next codon in mRNA

-cytosolic protein factors – Elongation factors

-GTP

Three steps takes place in the addition of each aminoacid residue,and this cycle is repeated as many times as there are residues to be added.

1.The next aminoacyl-tRNA is first bound to a complex of EF- Tu containing a molecule of bound GTP.

-The resulting aa-tRNA –EF-Tu.GTP complex is then bound to the a site of the 70s initiation complex.

-The GTP is hydrolyzed, an EF-Tu. GDP complex is released from the 70s ribosome and an EF-Tu. GTP complex is regenerated.

2. – a new peptide bond is formed between the aminoacids bound their tRNAs to the A & P sites on the ribosome.

-This accurs by the transfer of the initiating N-formyl methionyl group from its tRNA to the – NH_2 group of the second aminoacid now in the A-site.

-NH_2 group of the aminoacid in A site, acts as a nucleophile, displacing the tRNA in the P site to form the peptide bond. This reaction produces a dipeptidyl – tRNA in the A site and the now deacylated tRNA remains bound to the P site.

-The enzyme activity that catalyzed peptide bond formation has been referred to as peptidyl transferase and was widely assumed to be intrinsic for 1 or more proteins in the 50s ribosomal subunit. In 1992, Harry Noller and colleagues discovered that this activity was catalyzed not by a protein, but by the 235 rRNA, another critical biological function for ribozymes.

3. In the final step of elongation, called translocation, the ribosome moves by the distance of one codon toward the 3`end of the mRNA.

-because the dipeptidyl – tRNA is still attached to the 2nd codon of mRNA, the movement of the ribosome shifts the dipeptidyl – tRNA from the A site to P site and the deacylated tRNA is released from the P site back into the cytosol.

-the 3rd codon of the mRNA is now in the A site and the 2nd codon in Psite.

-This shift of the ribosome along the mRNA requires EF-G(translocase) and the enrgy provided by the hydrolysis of another molecule of GTP.

-the ribosome, with its attached dipeptidyl – tRNA and mRNA is now ready for another elongation cycle to attach the 3rd aminoacid residue.

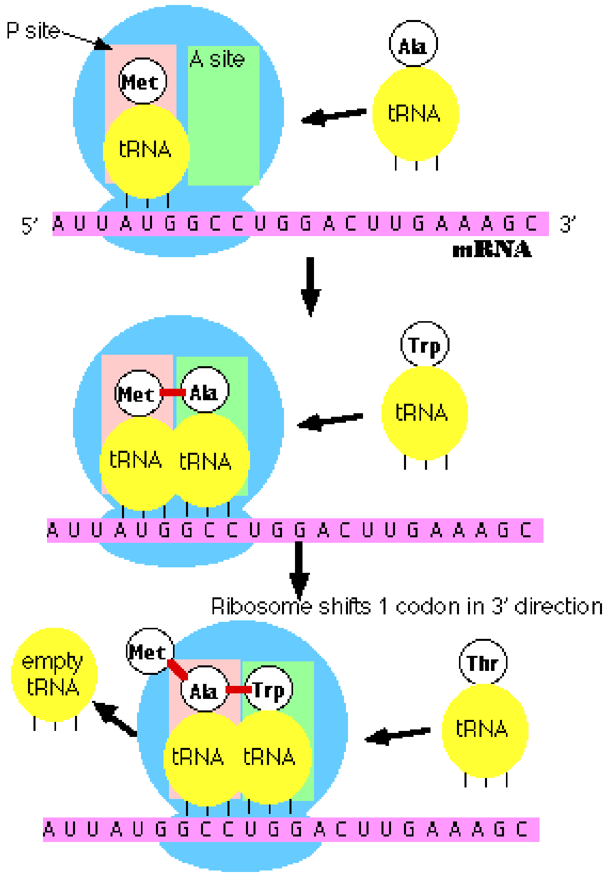
-the ribosome moves from codon to codon along the mRNA toward the 3`end, adding one aminoacid residue at a time to the growing chain.

Termination

Elongation continues until the ribosome adds the last aminoacid, completing the polypeptide coded by the mRNA. Termination, is signaled by one of the three termination codons in the mRNA (UAA, UAG,UGA), immediately following the last aminoacid codon.

In bacteria, once a termination codon occupies the ribosomal A site three termination factors/releasing factors, RF_1, RF_2, and RF_3 contribute to the hydrolysis of the terminal peptidyl – tRNA bond release of the free poly peptide and the last tRNA, now unchanged from P site the dissociation of the 70s ribosome into 30s & 50s subunits, ready to start a new cycle of polypeptide synthesis.

RF 1 recognizes termination codons UAA & UAG and RF2 recognises termination codons UAA & UGA. Either RF1 & RF2 bind at termination codon and induces peptidyl transferase to a water molecule rather than to another aminoacid. The specific function of RF3 has not yet been firmly established.


P site
A site
Met
tRNA
Ala
tRNA
5'
A U U A U G G C C U G G A C U U G A A A G C
3'
mRNA
Trp
Met
Ala
tRNA
tRNA
tRNA
A U U A U G G C C U G G A C U U G A A A G C
Ribosome shifts 1 codon in 3' direction
empty tRNA
Met
Ala
Trp
Thr
tRNA
tRNA
tRNA
A U U A U G G C C U G G A C U U G A A A G C

1.4.3.6 Eukaryotic translation

Initiation

Two eukaryotic factors eIF$_3$ (a large multimeric protein with about 8 subunits) and eIF$_6$ keep the ribosomal subunits apart. Eukaryotic preinitiation complex will be formed from an active ternary complex of eIF$_2$ bound to a GTP molecule & Met-tRNAi^{Met} associated with a small (40s) ribosomal subunit complexed withtwo other factors eIF$_3$ and eIF 1A.

Cells can regulate protein synthesis by phosphorylating a ser residue on eIF$_2$ bound to GDP, this complex is then unable to bind Met- tRNA^{Met}, thus inhibiting protein synthesis.

Most eukaryotic mRNAs have a single start site near the 5` capped end of the mRNA. A cap binding protein, eIF$_4$ recognises the 5`-cap structure present on all eukaryotic mRNA. After recognition any secondary structure at the 5`end is removed by an associated helicase activity.

The bound preinitiation complex then probably slides along the mRNA, most often stopping at the first AUG, However, selection of the initiation AUG is facilitated by specific surrounding sequences called KOZAK sequences.

mRNA 5` - ACCAUGG---------

The A preceeding the AUG seems to be most important nucleotide affecting initiation efficiency.

Scanning of the mRNA by preinitiation complex yields 40s initiation complex, in which Met-tRNA^{Met} is correctly positioned at the translation start site. Once,40s initiation complex is formed with its correctly positioned Met-tRNA^{Met} at the start codon, the 60s ribosomal submit binds completing the formation of 80s initiation complex.

Elongation

-same as in prokaryotes

-Elongation factors only differ

EF1 &-GTP ~ EF – Tu = GTP

EF2 – GTP ~ EF – G-GTP

Termination

- same as in prokaryotes

- Contains a single releasing factor CRF 1

Proof Reading

The GTPase activity of EF-Tu makes an important contribution to the rate and fidelity of the overall biosynthetic process. The EF-Tu. GTP complex exists for a few milliseconds and the EF-Tu. GDP complex also exist for the same period, before it dissociates. Both of these intervals provide an opportunity for the codon – anticodon interactions to be verified(i.e. proof red). Incorrect, aa-tRNAs normally dissociate during one of these periods.

If the GTP analog, GTPs us used in place of GTP, hydrolysis is slowed, improving the fidelity but reducing the rate of protein synthesis.

This proof reading mechanism establishes only that the proper codon-anticodon pairing has taken place. The identity of aminoacids attached to tRNAs is not at all checked on the ribosome. This was demonstrated experimentally by two research groups led by Fritz Lipmann and Seymour Benzer in 1962.

They isolated enzymatically formed cys-tRNA^{cys} and then chemically converted it into Ala-tRNA^{cys}. This hybrid aa-tRNA, whichcarries Ala but contains the anticodon for cys, was then incubaked with a cell free system capable of protein synthesis. The newly synthesized polypeptide was found to contain Ala residues in positions that should have been occupied by cys. So, on ribosomes only the codon-anticodon interactions will be verified and the fidelity of this proteinsynthesis process depends on the central role of aminoacyl t-RNA synthetases.

Proof reading by aminoacyl tRNA synthetases

The potential for any enzymes to descriminte between two different substates is limited by the available binding energy that can be derived from Enzyme-substate interactions.

-The first filler is the initial aminoacid binding and activation to aminoacid – AMP

-The second filler is the separate active site, which catalyses the deacylation of incorrect aminoacid – AMPs

-in addition to proof reading after formation of the aminoacid – AMP, most aminoacid tRNA synthetases are also capable of hydrolyzing the ester linkage between aminoacid and tRNA in aa-tRNA. This hydrolysis is greatly accelerated for incorrectly charges tRNAs – 3rd filter.

-in a few aminoacyl –tRNA synthetases that activate aminoacids, that have no close structural relatives little or no proof-reading occurs, in these cases the active site can sufficiently discriminte between the proper substrate aminoacid and the incorrect aminoacids.

1.4.3.7 Inhibitors of Protein Synthesis

Protein synthesis is the primary target of a wide variety of naturally accruing antibiotics and toxins. This is presumably a consequence of translational machinery's enoromous complexity which makes it vulnerable to disruption in many ways. Antibiotics have become vulnerable tools in the study of protein synthesis, nearly every step in protein synthesis can be specifically inhibited by one antibiotic or another.

PUROMYCIN

-made by the mold streptomyces alboniger

-has structural similarity to the 3`end of aa-tRNA

-it binds to the A site and participates in all elongation steps including peptide bond formation, producing a peptidyl puromycin.

-However, puromycin will not bind to the P site.

-it dissociates from the ribosome shortly after it is linked to the – coo terminus of the peptide, prematurely terminating synthesis of the polypeptide.

Streptomycin

- Medically important number of a family of antibiotics known as aminoglycosides, that inhibit prokaryotic ribosome's in a variety of ways.

-a low concentrations, it induces the ribosome to characteristically misread the mRNA – one pyrimidine may be mistaken for the other in 1st and 2nd codon positions and either pyrimidine may be mistaken for A in 1st codon position.

-At higher concentrations, however streptomycin prevents proper chain initiation and there by causes cell death.

-certain streptomycin – resistant mutants(str^R) have ribosomes with an altered protein S12 compared with streptomycin sensitive bacteria.

Chloroamphenicol

-inhibts the peptidyl transferase activity on the large subunit of prokaryotic ribosomes.

-protein L 16 is necessary for chloramophenicol binding

-chloramphenicol`s binding site must lie near the A site since it competes for binding with the 3`end of aa-tRNAs and puromycin but not with peptidyl-tRNAs.

Tetracyclins

-Tetracyclin and its derivatives are broad spectrum antibiotics

-binds to the small subunit of prokaryotic ribosome

-inhibts aa-tRNA binding

-Tetrocyclin resistant bacterialstrains have become quite common. Resistence is conferred by a decrease in bacterial cell membrane permeability to the drug rather than any alteration of ribosomal components.

1.4.3.8 Post-translational Modifications

Newly synthesized polypeptides in the membrane and lumen of the ER undergo 5 principal modifications before they reach their final destinations.

*formation of disulfide bonds

*proper folding

*addition and processing of carbohydrates

*specific proteolytic cleavages

*assembly into multimeric proteins

Only properly folded and assembled proteins are transported from the rough ER to the Golgi complex and ultimately to the cell surface or other final destination. Unfolded, misfolded, or partly folded proteins are retained in the rough ER or are retrieved from the cis-Golgi network and returned to the ER. Misfolded proteins and unassembled subunits of multimeric proteins often move from ER lumen back through the translocon into the cytosol where they are degraded.

1. Formation of disulfide bonds

Intermolecular & intramolecular disulfide bonds help stablize the 3^0 and 4^0 structure of many proteins. These covalent bonds formed by the oxidative linkage of $^{-SH}$ groups, on two cysteine residues in the same or different polypeptide chains. In eukaryotic cells, disulfide bonds are formed in the lumen of the rough ER but not in cytosol. Thus disulfide bonds are found only in secretory proteins and in the exoplasmic domains of membrane proteins synthesis on the rough ER.

Because of higher amounts of reduced glutathione (GSH) cytosolic proteins synthesized on free ribosomes lack –s-s-bonds and depend on other interactions to stabilize their structures. (GSH=GSSG ratio 50:1 in cytosol)

In proteins that contain more than one –s-s-bond, the proper pairing of cys residues is essential for normal structure and activity. Disulfide bonds sometimes are formed sequentially while a polypeptide is still growing on the ribosome.

For example: During synthesis of the Ig light chain, which contains two –s-s-bonds, the first and second cysteins closest to the N-terminus from a-s-bond before the 3rd cys has even been added to the nascent chain, automatically ensuring the correct pairing of cys. Similarly, the 3rd pairs with the 4th to create the second –s-s-bond.

The –s-s-bonds in some proteins however, do not link cys that occur sequentially in the aminoacid sequence.

For example: Proinsulin has three –s-s-bonds that link cystine 1-4,2-6, and 3-5. In this case the 1st-s-s-bonds that form spontaneously by oxidation of –SH groups may have to undergo rearrangements.

In cells, the rearrangement of –s-s-is accelerated by the enzyme protein –s-s-isomerase(PDI) which is found in abundance in the Er of secretory tissues in such organs as the liver & pancreas. In catalyzing rearrangement, PDI forms a disulfide bonded substate – enzyme intermediate. –s-s-bonds generally form in a specific order, first stabilizing small domains of a polypeptide, then stabilizing the interactions of more distant segments.

2. Proper folding

The ER contains several proteins that accelerate the folding of newly synthesized proteins with in the ER lumen. Protein disulfide isomerase(PDI) is one such folding catalyst; the chaperone HSP 70 is another. Like cytosolic HSP 70, this ER chaperone transiently binds to proteins and prevents them from misfolding or forming aggregates, there by enhancing their ability to fold into proper conformation. Two other ER proteins, the homologous lectins – calnexin and calreticulin bind to certain carbohydrates attached to newly made proteins and aid in protein folding.

Other important protein folding catalysts are peptidyl-prolyl isomerase, a family of enzymes that accelerate the rotation about peptidyl-prolyl bonds in unfolded segments of a polypeptide.

For example: In drosophila an ER peptidyl –prolyl isomerase called Nina A is required for the folding of opsin, the membrane protein that absorbs light and triggers the visual response.

3.Glycosylation

Most plasma membrane and secretory proteins contain one or more carbohydrate chains. The addition and subsequent processing of carbohydrates is the principal chemical modification to most proteins some glycosylation reaction occur in the lumen of ER, others in lumina of Cis-medial – or trans golgi. Thus the presence of certain carbohydrates provide useful markers for their movement from the ER and through the Golgi cisternae.

The oligosaceharide bound to either Asn or Ser/Thr are referred as N-linked and 0-linked oligo saccharides respectively. O-linked oligosaccharides are generally short, often containing only 1 to 4 sugar residues. Typical N-linked digosaccharide always contain mannose as well as N-acetyl glucosamine and usually have several branches each terminating with a negatively charged sialic and residues.

NANA→ N-acetyl Neuraminic acid or sialic acid

Gal→ Galactose

Glc→ Glucose

GalNAc→ N-acetyl galactosamine

Glc Nac→ N-acetyl glucosamine

Fuc→ fucose

Man→ Mannose

O-linked sugars are added one at a time, and each sugar transfer is catalyzed by a different glycosyl transferees enzyme. In contrast, biosynthesis of N-linked oligosaccharide begins with the addition of a large preformed oligosaccharide containing 14 sugar residues; subsequently certain sugars are removed and others are added, one at a time in a defined order with each reaction catalyzed by a different enzyme.

4. Assembly into multimeric proteins

Many important secretory & membrane proteins are built of two or more polypeptides. In all cases, these multimeric proteins are assembled in the ER. One important example is provided by the Igs which contain two heavy and two light chains, all linked by –s-s-bonds.

Haemagglutinin(HA) –the trimeric protein that forms the spikes protruding from the surface of the influenza virus particle.

-each spike is formed within the ER of an infected host cell from three copies of a precursor protein termed HAO, which has a single membrane spanning &-helix.

-In the Golgi complex, each of the 3 HA_0 proteins is cleaved to form two polypeptides, HA_1 and $HA_2$2; thus each spike in the virus particle contains 3 copies of HA_1 and HA_2.

-The timer is stabilized by interactions between the exoplasmic domains of the constituent polypeptides as well as by interactions between the 3 cytosolic and membrane spanning domains.

5. Proteolytic cleavage

Mutant misfolded secretary & Membrane proteins, as well as, the unassembled subunits of multimeric proteins, often are degraded within an hour or two after this synthesis in the RER. For many years researchers thought that the proteolytic enzymes in the ER catalyzed degradation of misfolded or unassembled polypeptides, but such proteases were never found. Recent studies have shown that misfolded membrane and secretory proteins are transported from ER lumen "backwards" through translocon into the cytosol where they are degraded by the ubiquitin-mediated proteolytic pathway.

Summary

Proteins must be synthesized in response to the cell's current needs, transported to their appropriate cellular locations, and degraded when no longer needed protein synthesis can be divided into three stages. Initiation, Elongation, Termination.The formation of peptide bond between aminoacids is thermodynamically unfavourable and the aminoacids are unable to reconise the nucleotides on the m-RNA these two problems are solved by the activation of aminoacids carried out by aminoacyl t-RNA synthetases. The activated aminoacids bind to the ribosome m-RNA complex with the help of translation factors. After binding to the ribosomes the peptide bond is formed by the peptidyl transeferase.

Then the de acylated t RNA's are removed by translocation step .This process is repeated for number of cycles until it reaches the termination codons UAA,UAG & UGA.These terminating codons are recognized by releasing factors which changes the specificity of peptidyl transferase leads to theformation bond betweenwater and amino group of aminoacid resulting in the release of nascent polypeptide chain.the polypeptidechin undergo number of post translational modifications such as *formation of disulfide bonds *proper folding *addition and processing of carbohydrates *specific proteolytic cleavages *assembly into multimeric proteins

Model questions

1) Compare the process of translation in prokaryotes with eukaryotes

2) Write a note on post translational modifications

Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork.

Dr.N.Srinivasa Reddy

Lesson 1.4.4

MUTATIONS

Objective

Mutation is nothing but a sudden change. This chapter explain different types of mutations both at sequence of DNA level and at the level of chromosomes.

1.4.4.1 Introduction

The term mutation refers to all the heritable changes in the genome, excluding those resulting from incorporation of genetic material from other organisms. A mutation is an abrupt qualitative or quantitative change in the genetic material of an organism.

Hugo de Vries was the first to use the term `mutation' to describe phenotypic changes in oenothera lamarkiana, which were heritable.

First, the changes occur in the DNA sequence itself. These changes in the nucleotide sequence may occur in a gene. Finally different kinds of phenotypic changes can arise in an organism as a result of changes in a gene.

Changes in nucleotide sequence may result in changes in gene may lead to altered phenotypes in organisms

1.4.4.2 Mutations at the level of DNA sequence

A point mutation is the replacement of one nucleotide or few nucleotide pair(s) by another. A point mutation can be classified as

- Substitution of one nitrogen base by the other

  Transition : of purine substituted by another purine or if pyrimidine is replaced by another pyrimidine.

  Transversion : if purine is replaced by pyrimidine or vice versa.

- an insertion or deletion is the addition or removal of anything from one basepair upto quite extensive pieces of DNA.
- Inversion : excision of portion of the double helix followed by its reinsertion at the same position but in a reverse orientation.

1 2 3 4 5 6 7

-A-T-G-T-T-C-A

-7-A-C-A-A-G-T

Fig : Various types of sequence changes in the given DNA sequence.

1)Mutations at the level of Gene

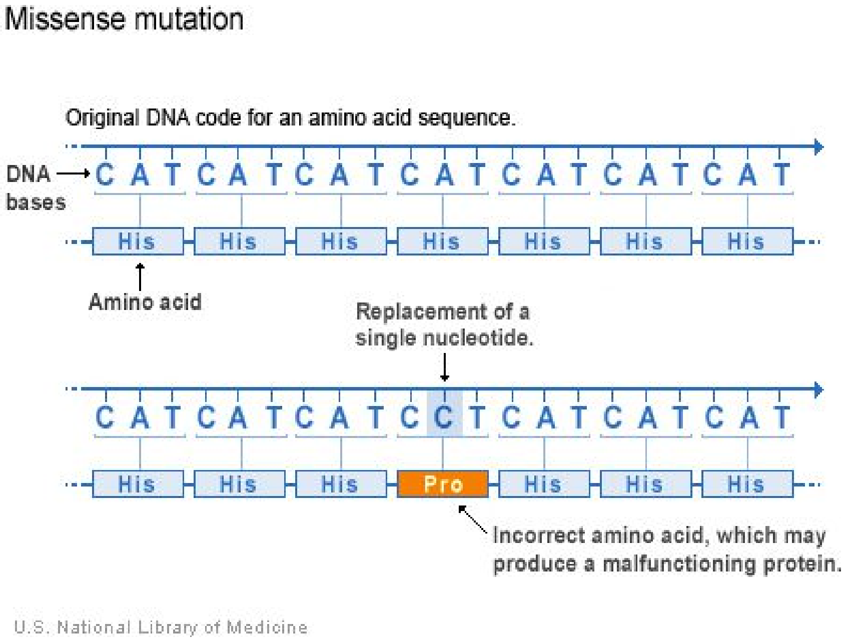
1. Silent mutation

Occur of a point change takes place at the third nucleotide position of a codon and changes the codon, but owing to the degeneracy of the genetic code, not the aminoacid sequence of the gene product and doesnot give rise to a mutant phenotype.

Missence mutation

This is also a point mutation, but in this case it does change the aminoacid. Most point changes at the first or second nucleotide positions of a codon result in missence mutation. A few third position nucleotides also.

Example :


Missense mutation
Original DNA code for an amino acid sequence.
DNA bases
C A T C A T C A T C A T C A T C A T C A T
His His His His His His His
Amino acid
Replacement of a single nucleotide.
C A T C A T C A T C C T C A T C A T C A T
His His His Pro His His His
Incorrect amino acid, which may produce a malfunctioning protein.
U.S. National Library of Medicine

Changing nucleotide 4, from G to A produces an Arg codon instead of a Gly codon.

Similarly changing nucleotide 15, from A to T specifies phe ratherthan Leu.

A missence mutation gives rise to a polypeptide with a single aminoacid change. Whether or not it causes phenotypic change depends on its precise location in the protein. Many proteins can tolerate some changes in their aminoacid sequence, although a missence mutation that alters an aminoacid essential for structure or function will inactivate the protein, that lead to a mutant phenotype.

Non-sense mutations

This is also a point mutation that changes a codon specifying an aminoacid into a termination codon. The result is a truncated gene which codes for a polypeptide that has lost a segment at its carboxy terminees. In many cases, although not always, thus segment will include aminoacids essential for the proteins activity and a mutant phenotype results.

Ex:

5’ – ATG GGA GCT CTA TTA ACC TAA – 3’

Met Gly Ala Leu Leu Thr Stop

5' – ATG GGA GCT CTA TGA ACC TAA – 3'

Met Gly Ala Leu Stop

Frameshift mutations

This is the usual consequence of an insertion or delection event because the addition or removal of any number of basepairs that is not a multiple of three causes the ribosome to read a completely new set of codon down stream the mutation. It usually produces mutant phenotypes.

ATG GGA GCT CTA TTA ACC TAA TTT GA

Met Gly Ala Leu Leu Thr stop

ATG GGG AGC TCT ATT AAC CTA ATT TGA

Met Gly Ser Ser Ile Asn Leu/ Ile Stop

Deletion

ATG GGG CTC TAT TAA CCT AAT TT GA

Met Gly Leu Tyr Stop

1.4.4.3 Mutations at the level of organism

In order to produce a mutant phenotype, the nucleotide sequence alteration must produce a mutated gene product, that is unable to fulfill its function in the cell. In many cases the cell will be unable to tolerate the loss of the function and will die. Such mutations are called lethal mutations.

Some mutations inactivate proteins that are not essential to the cell and others result in proteins with reduced or modified activities. This is true for both pro-and Eu-karyotes.

Auxotrophic mutants

This type of mutant lacks a gene product involved in synthesis of an essential metabolite such as an aminoacid. However, these mutants can be kept alive if, the metabolite is supplied as a nutrient in the culture medium.

Example : A Trp auxotroph to E.coli that lacks one of the enzyme involved in Trp biosynthesis.

Conditional – Lethal mutants

This type of mutations can survive, but only if cultured under a particular set of conditions. The most common example are temperature, sensitive mutants which are able to survive at one temperature range (<30^{o}c, say) but die if the temperature is raised above this permissive threshold.

The mutations carried by these temperature sensitive organism is often one that affects an aminoacid product. The mutated protein is able to retain its correct structure at a low temperature, but is essentially denatured and inactivated by heat.

Antibiotic resistant mutants

Antibiotics kill wild-type bacteria but have no effect on resistant mutants. Resistance to an Antibiotic can arise in several ways, but the commonest is when the biomolecule that is target for the antibiotic becomes altered.

Example :

Streptomycin interferes with protein synthesis by binding to ribosomal protein S_{12}, one of the small subunit component of E.coli ribosome.

When streptomycin is bound to the small subunit, the initiator tRNA cannot enter the p site and mRNA is not translated and the bacterium dies.

In streptomycin resistant mutant, the gene for the ribosomal protein S_{12} is mutated. This leads to an altered S_{12} protein, which is still able to fulfill its role as a ribosomal protein, but can no longer bound to streptomycin. Has no effect on the mutated bacterium.

Regulatory mutants

These have lost the ability to control expression of a gene or operon normally subject to regulation. For instance, it is possible to obtain E.coli mutants that express the genes of the lac operon even in the absence of lactose. These are called constitutive mutants.

Constitutive mutants arise through a mutation in the gene for the lac repressor, so that the repressor is either not produced or has an altered structure.

.. no longer able to bind the operator.

Reverse mutations

A point mutation can be reversed by a 2nd point mutation, an insertion event by a subsequent deletion and so on. These events are called back mutations and occur rarely.

Many mutations can also be corrected by second site reversions → a 2nd mutation that restores the original phenotype but doesnot return the DNA sequence to its precise unmutated form.

ATG GGA GCT CTA TTA ACC TAA

Met Gly Ala Leu Leu Thr stop

Missence mutation

ATG GGA GCT CTA TTT ACC TAA

Met Gly Ala Leu Phe Thr Stop

2nd site reversion

ATG GGA GCT CTA CTT ACC TAA

Met Gly Ala Leu Leu Thr Stop

Mutations can occur at any stage during development.

- if mutations occur in a germinal cell, before differentiation of gametes, it would influence several gametes and will thus affect all the individuals derived from these affected gametes.
- If mutation occurs in a gamete / zygote, a single individual will carry the mutation.
- If a mutation occurs in a cell after the zygote has undergone one or more divisions, only a part of the body will show the mutant character.

The first two are called germinal mutations and the last is called the somatic mutation.

Spontaneous and induced mutations

The background or spontaneous mutations occur suddenly in the nature. They have been reported in oenothera, maize, bacteria, bread molds, viruses, Drosophila, mice, man etc.

Besides naturally occurring mutations, the mutations can be induced artificially in the living organisms. Those agents which cause mutations artificially are called mutagens and the mutations are called as induced mutations.

Chemical Mutagens that effect replicating DNA

Base analoguer

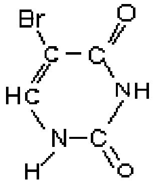
5-Bromo Uracil (5-bU) is derived from thymine by replacement of the methyl group with Bromine. It is sufficiently similar to thymine to be incorporated into a polynucleotide chain, in place of normal nucleotide.

5-bu, as tymine analog, base pairs with A and also pairs with G after a tautomeric shift.

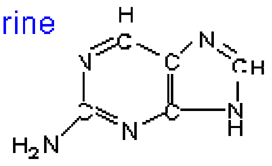
If the tautomeric shift happens during DNA replication, then one of the daugher molecules will have a 5bu-G bp instead of the original A-T bp. A further round of replication of the mutant molecule produces a G-C pair in one double helix, in which the mutation is now established and 5-bu-G/5-bu-A in the other daughter cell. This is an example how a point mutation can be brought about.

I. Base analogs

5-Bromouracil

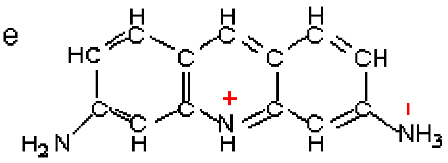


2-Aminopurine



II. Acridines

2,8,-Diamino acridine (Proflavin)



Intercalating agents

Ex: Acridine dyes (Ethidium bromide, EtBr). Et Br is a A ringed molecule whose dimensions are similar to those of a purine – pyrimidine basepair, so that the compound can intercalate into a double helix, moving adjucent basepairs slightly apart. An insertion of a single nucleotide is likely to occur at the intercalation position and cause a frameshift mutation if the position lies within a gene.

Chemical mutagens that effect non replicating DNA

These are the substrates that can alter a base that is already incorporated in DNA and there by change its Hydrogen bonding specificity. Three commonly used mutagens are

Nitrous acid(HNO_2)

Primarly converts – NH_2 groups to keto groups by oxidative deamination. Thus Cytosine, Adenine and Guanine are converted to Uracil Hypoxanthine (H) and xanthine (X) respectively. These bases can form base pairs :

U – A

H – C

X – C

Conversion of G→ X is not directly mutagenic, since both G and X pairs with C. But in some single stranded DNA phages, x undergoes an undiscovered tautomeric change and is able to pair with thymine.

Hydroxylamine (HA)(NH_2OH)

HA probably causes hydroxylation of cytosine at amino group giving rise to hydroxyl cytosine, which then subsequently pairs with Adenine. So, a G-C pair can ultimately become an A-T pair.

EMS(CH_3-CH_2-O-SO_2-CH_3)

It is an alkylating agent and is a potent mutagen extensively with eukaryotes. Many sites in DNA are alkylated by these agents.

Of prime importance is the addition of an alkyl group to the H-bonding oxygen of Guanine and Thymine. These alkylations impair the normal H-bonding of these bases and cause mispairing of Guanine and Thymine leading to the transitions.

Another phenomenon resulting from alkylation of Guanine is depurination. Depurination is not always mutagenic, since the gap left by loss of purine can be repaired by AP endonuclease.

However, sometimes the replication fork may reach the apurinic site before repair has occurred, when this happens replication stops just before the

AP site (or) if the cell contains a functional SOS system, replication restarts after a brief pause.

However, with high probability an adenine is put in the daughter strand opposite to AP site. It gives a mismatch Purine-Adenine.

Physical mutagens

Heat

It is probably the most important environmental mutagen. Its effect on DNA molecule is to cause cleave the bonds between purine bases and their sugars, resulting in apurinic sites in polynucleotides. Pyrimidines are also removed but at a much slower rate.

Upto 10,000 apurinic sites are created everyday in every human cell, equivalent to one per chromosome every min or so. Those that escapes repair can cause point or deletion mutations when DNA is replicated.

Radiation

Several types of radiation are mutagenic. UV radiation of about 260nm is absorbed by purine and pyrimidine and can cause structural damages, particularly result in the formation of cyclobutyl dimers between adjacent pyrimidine in a polynucleotide. Dimerization causes the bases to stack closer together and can give rise to deletions during DNA replication.

2.5.4.4 Detection of mutants

Induced mutations and visible mutations. Both these types could be located either on sex chromosomes or on autosomes.

The procedures used to identify and isolate the mutants are referred to as Genetic screens and they depend on whether the experimental organism is haploid or diploid. If it is diploid whether dominant or recessive.

Detection of lethal mutants

In Haploids

In haploid organims all mutations are in effect dominant, so that the mutant phenotype is exhibited immediately in the progeny of the mutagenized population. For instance, mutations that disrupt Arginine synthesis lead to cells that require Arginine for growth. Such mutations are easily detected by growing them in the presence and absence of Arginine.

In prokaryotes and haploid eukaryotes viz, yeast, essential genes can be studied through the use of conditional mutations.

Example: Temperature sensitive mutants.

L.H. Hartwell & colleagues studied a particulary important temp. sensitive screen in the yeast Saccharomyces cerevisiae in lak 1960s and early 1970s. They setout to identify genes important in regulation of the cell cycle.

Screening

- Yeast cells were grown in a large liquid culture, treated with a chemical mutagen and then subcultured into a – small aliquots.
- After a – 5 he growth period at 23°c, aliquots from each tube were separately plated onto a medium and incubated at 23°C.
- Then, the colonies developed were replica-plated onto two plates → incubated at permissive temp 23°C

  → other at non-permissive temp. 36°c
- the temp sensitive colonies grow at 23°c but not at 36°c were assessed to determine whether they were blocked at specific stages in the cell cycle.
- The cell cycle stage at which cell growth was arrested at the non-permissive temp. indicated when the protein encoded by the mutated gene was required.

In diploids

In diploid organism (ex. Drosophila), phenotypes resulting from recessive mutations can be observed only in individuals homozygous for the mutant allele.

Autosomal mutations

Mutation on chromosome 3.

This approach requires. Three sequential crosses. Many males are treated with a mutagen (EMS), producing flies carring various mutations (M_1, M_2 etc..) in their germline cells. The level of mutagen used is sufficiently to induce atleast one mutation on each chromosome. These males carry a non-lethal recessive mutation that gives rise to a visible phenotype in homozygotes, the marker in this example is rosy (ry) eye color.

1st cross: In the first cross ($P_{1)}$, mutagenized males are mass-mated to a large no.of females.

The females carry dominant visible markers (D_1 and D_2) on chromosome 3. These are non-lethal in heterozygous condition, but are lethal in homozygoles.

2nd cross: In the second cross (P_2), individual heterozygous F_1 ----- carrying mutagenised chromosome 3 are mated individually to non – mutagenized organisms.

The F_2 progeny homozygous for either dominant marker will die, those heterozygous for both markers are easily identified and excluded. The F_2 heterozygotes includes O and O that have the identical mutagenized chromosome carrying `ry' marker and one non-mutagenized chromosome carrying a single dominant visible marker.

Fig

3rd cross :

F_2 generation, heterozygous brothers and sisters are mated individually in the third cross (P_s).

The absence of flies with rasy colored eyes in F_3 progeny indicates the presence of an induced lethal mutation (M_1 for ex). Although flies homozygous for M_1 donot survive, heterozygotes carrying M_1 on one chromosome and one of the dominant marker on the other will survive.

Fig.

The mutation can be maintained in heterozygous flies.

Sex-linked lethals

Muller-5 method: This method makes use of a muller-5 Drosophila stock, which carrier two marker genes, dominant `Bar' (barred eye) and recessive `apricot' but doesn't contain a lethal gene.

In F_2 generation 50% -- are muller-5 in phenotype and remaining 50% are wild type. If a lethal mutation is induced in x-chromosome of irradiated ---, no wild type ---would appear in F_2 generation. Therefore, the absence of wild type ---- in F_2 is an indiation of an induced lethal mutation.

Fig.

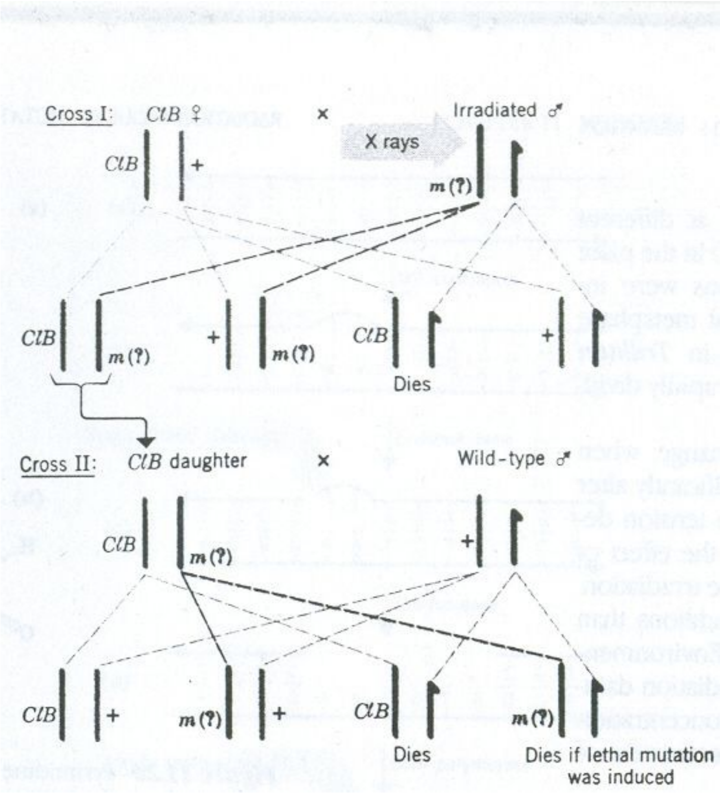


Fig: Mullers clb method for detecting sex linked lethal mutations.

Sexlinked visible mutations

For detection of sex-linked visibles, muller-5 and attached x-chromosome were used. The attached x ---- (xxy) have a special advantage when these --- are crossed to an irradiated ---, x-chromosome of irradiated --- goes either to superfemale daughter or to the sons. Since in sons there is only a single – x chromosome, any visible induced mutation will immediately express itself and can be easily scored.

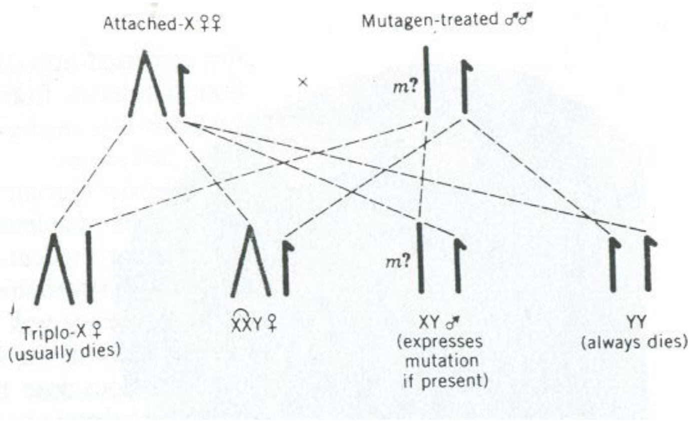


Fig: Attached-X method for detecting sex linked visual mutations

Isolation of mutants

There are a number of selective techniques to isolate mutants, when they arise in a parent organism.

Enrichment

It is possible to selectively pickout strains of microorganisms with specific capabilities. For example, we may wish to isolate a rare bacterial cell, present in the culture of a prototroph with the ability to utilize a particular, non-utilizable growth substance.

For this purpose, the bacterial cell culture is switched to a minimal medium, containing x as the only energy source. Although most bacterial cells cann't grow, a few cells utilize x and grow slowly. By repeated sub-culturing on medium having x, we could enrich the culture and exclude, large population which lack the ability to utilize x.

Filteration enrichment (fungal spores)

Concentration of auxotrophs can also be increased by filtration of a culture grown in suspension culture with minimal medium on which only protorophs grow, and yield mycelia, Auxotrophic spores cannot do so. Suring filtration, only prototrophs will be filtered out due to filamentous mycelia and non-growing auxotrophs will be in the filtrate. This filtrate can be used for culturing of auxotrophs on supplemented media and thus they can be concentrated. This method was derived by Woodward, 1954 and is known as Woodward's filtration technique.

Penicillin enrichment

Bacteria and fungi (Neurospora) in proliferative stage are sensitive to penicillin as it kills growing cells by specifically blocking the synthesis of cell wall precursors and is not lethal to non-proligerating cells. Therefore in a suspension culture with minimal medium, if penicillin is added, the prototrophs will be killed. The auxotrophs survive due to lack of growth. After killing prototrophs, the penicillin can be removed by washing the cells on a filter. If the culture is now plated on a supplemented medium, auxotrophs will grow. Multiple auxotrophs can be selected by this method.

Replica plating

It is a simple and powerful technique deviced by `Lederberg' in 1952. This tecynique is conveniently utilized for microorganisms forming colonies.

The material is first grown on complete medium. For ex: if streptomycin resistant mutants are to be isolated, material should be allowed to grow on medium lacking streptomycin, so that both mutant and wild types may grow. These colonies are imprinted on velveteen by inverting the potriplate. As a result, all colonies are imprinted and leave cells at corresponding position on velveteen. Other plates having streptomycin can then be pressed on velvet to get an impression. On this plate now, only resistant cells will form colonies. Knowing the position of mutants in the original plate, they could be isolated and multiplied.

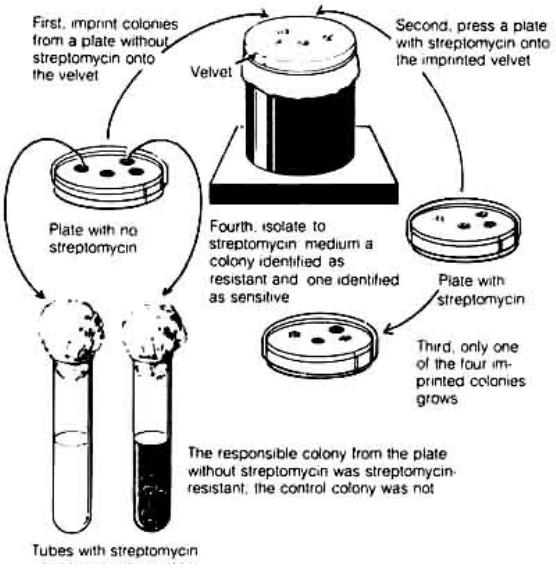

First, imprint colonies from a plate without streptomycin onto the velvet
Velvet
Second, press a plate with streptomycin onto the imprinted velvet
Plate with no streptomycin
Fourth, isolate to streptomycin medium a colony identified as resistant and one identified as sensitive
Plate with streptomycin
Third, only one of the four im-printed colonies grows
The responsible colony from the plate without streptomycin was streptomycin-resistant, the control colony was not
Tubes with streptomycin

Fig: Replica plating technique

Utility of mutants

Mutations are normally deleterious and recessive and therefore majority are of no practical value. A gustafsson estimated that < 1 in 1000 mutants produced may be useful in plant breeding. .

- Mutant microorganisms are useful in industry.
- Mutation breeding helps to generate varieties of plants with characters like resistance to drought, pests and pesticides.
- One of high yielding varieties of Rice, Reimei was developed through mutations isolated after irradiation.
- Protein and Lysine content was increased through mutations in wheat.
- Crop harvesting time is decreased in Aruna variety castor from 270 days to 140 days.
- Increased yield of penicillin was obtained from mutant penicillium fungus.
- Many of our fruit varieties resulted from spontaneous somatic mutations.
- Mutants are important tools for unraveling metabolic pathways.

1.1.4.4.5

The term mutation refers to all the heritable changes in the genome . A point mutation or gene mutation is the replacement of one nucleotide or few nucleotide pair(s) by another. A point mutation can be classified asSubstitution, ransition,Transversion or Inversion and these mutations may be natutal or induced by some chemical and physical agents.the mutant organisms can be isolated by replicaplating method. Mutations are normally deleterious and recessive and therefore majority are of no practical value However, mutants had some uses. in industry,or for getting good yielding varities etc

1.4.4..6 Model questions

1) Explain different types of mutations in detail

2) Write in detail about the detection of mutations

1.4.4.7. Reference books

Freifelder, David., Physical Biochemistry, W.H.freeman & company

Griffiths, Anthony JF. , Wessler, Susan R. , Lewontin, Richard C. , Gelbart William M., Suzuki, David T. , Miller, Jeffrey H. *An Introduction to Genetic Analysis* 8/e, W.H. Freeman

Lewin B., Genes, Oxford University Press, Newyork.

Dr.N.Srinivasa Reddy

Lesson 1.5.1
Scope of Bioinformatics

Objective

- To know the importance and scope of bioinfprmatics.
- To know other useful areas of bioinformatics and biotechnology

1.5.1.1 Introduction

One of the earliest branches of scientific study to evolve is biology. Biology as a science has living beings as the focus of study. As the facts of biology started accumulating because of diligent study vast amounts of data have been pouring which need sophisticated means of analysis rather than manual analysis. In other words for a quick analysis of data, automation seems to be highly essential. With molecular biology making making rapid strides, data on two important biomolecules – DNA and proteins have been mounting. If we have to have any meaningful understanding of these data, then naturally we must know what do they mean ? How do they function ? How do they compare with others ? For instance when we get the sequence of a DNA molecule obviously we would like to know which part of this codes for amino acids ? How does this sequence compare with other seemingly similar molecules from another source ? In the same way in proteins also – how does a new protein function ? When we know the sequence can we predict its structure ? Whether two peptide chains with a similar sequence will have a similar structure?

Naturally, all the above questions need a quick answer, which is possible only by automation. The one modern machine that has revolutionized automation of data-collection, storage, retrieval, comparison and analysis is the computer.

The necessity for a fast, accurate and logical analysis of biological data requires the use of computers and thus was born Bioinformatics. Though various definitions for Bioinformatics are available, the simplest one would be – It is an interdisciplinary field using computational techniques to analyze the biological data. A person who is specialized in this is called a Bioinformatist. We must understand however, that thought computational techniques including software development are all essential, all these are mainly focused on biological facts with computers only providing the tools. The computer tools are used to find, separate, store, compare and analyze the data for a meaningful interpretation.

Which data of biology are used in bioinformatics ? Normally however one can say any biological data that can be stored and analyzed using logic and a pattern will form the subject of bioinformatics. We can elaborate the above with a specific example. For instance a library is a storehouse of books. If a library has few volumes, it is easy to pick up a specific book, but if the library has large number of volumes, naturally then an arrangement is necessary. This arrangement should have a pattern – the arrangement may be authorwise or subjectwise etc. But if the library has diverse books where pattern recognition is difficult, arrangement is difficult. When arrangement is difficult accessibility is also difficult.

The above example would suffice to indicate that biological data are amenable to computer techniques only when they have a general similarity, pattern and inbuilt diversity.

As has been pointed out already, any biological data can form the subject of bioinformatics. In practice however data on two biomolecules only – Nucleic Acids (DNA and RNA) and proteins have been the subject matter of bioinformatics. The reasons for this are not far to seek. These two molecules – one the hereditary determinant forming a link between one generation and the next (DNA) and the other – the vital molecule deciding and executing all the functions of living organisms (proteins) are probably the most important ones in bioinformatics. Besides this, the structure of these molecules – general similarity with inbuilt diversity will be readily amenable to computer analysis based on pattern and sequence identification and comparison. Thus bioinformatics concentrates on Genomics and Proteomics. While genomics deals with the analysis of sequence of nucleotides in a particular fragment of DNA, proteomics deals with sequence analysis and structure prediction of protein molecules.

1.5.1.2 Importance of bioinformatics

Of what use are computational techniques in dealing with biological data – in particular data from molecular biology ? Is it just another branch of science of interest only to academicians ? Or has it any impact, which is far reaching on human society. We will try to find answers to the above in the following paragraphs.

With the progress of the human society being rapid, industrialization taking long leaps and at the same time human population increasing forever, unforeseen problems are cropping up and human society has to find an urgent answer to maintain a healthy and happy state. Problems like systemic diseases infectious diseases (that were once conquered) are cropping up once again, crop plants suffering a large scale destruction due to pathogens and above all the curse and challenge to human beings from the tiniest of the creatures – the AIDS virus, all these need a solution urgently. In the quest to find an answer to this – speed is the by word and that is possible only with computers.

Let us take a specific example of drug resistance in bacteria. An antibiotic that was controlling particular bacterium will no longer be so. In other words due to its association with the antibiotic the bacterium has become resistant. What is the basis of this resistance? How to study it and finally how to overcome it? Obviously the drug resistance is due to the changed molecular configuration (of DNA probably due to mutation) of the bacterium when the antibiotic or drug molecule can no longer inhibit it. There are two methods to understand and solve this problem.

In the traditional microbiological way, the drug resistant pathogen has to be cultured in the pure from and a number of other drugs are to be tried (in vitro) to find out whether any other drug can control the bacteria. This is a trial and error method and possibly time consuming also.

How to tackle the problem of drug resistance using Bioinformatics? If we know the genomic sequence of the original bacterium (drug sensitive) it can be compared with the mutated (drug resistant) bacterium. Using the computational techniques we can compare the two DNA sequences and we know exactly which segment has changed. The next step would be, if the original architecture of the drug molecule is ineffective against the new sequence (of DNA) of the resistant bacterium we can use computers to design a molecule to match the new sequence of the bacterium. In fact if we have in store in computers a large number of related sequences and the architecture of a large number of drug molecules with the flick of a button we can find out which drug molecule will suit (inhibit) which DNA sequence of the pathogenic bacterium; thus the whole process takes very little time. This method of new drug discovery can be called `in silico' experiment as against the traditional molecular biology experiment of `in vitro'. In `in silico' method, the entire experiment is conducted `in simulation' or `in virtual reality' in the computeer chips and it provides not only faster but accurate results.

The above is one example as to how Bioinformatics ca revolutionize new drug discovery to alleviate human suffering.

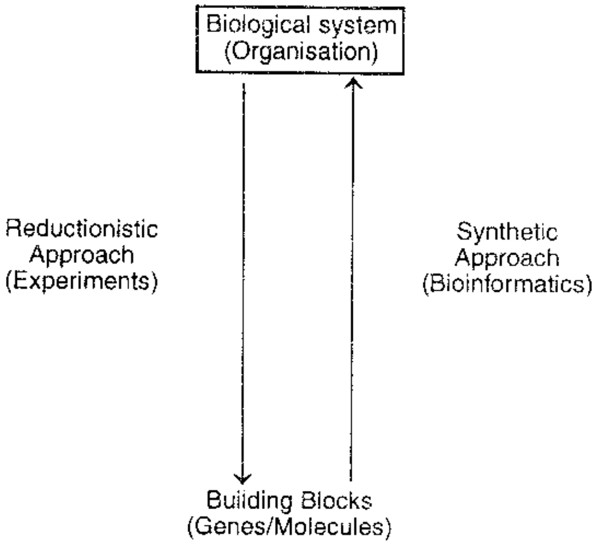


Fig: Bioinformatics approach to the understanding of the Biological systems.

1.5.1.3 Biological databases

Biological databases store data related to biology. It may be from literature, sequencing propjects, structure determination, etc. In recent years, biological databases have become a part of the biologist's everyday toolbox. With the increase in biological knowledge (sequencing, structure determination), computer-based databases have become essential for managing them. There are a large numbers of databases and among them – Genbank, EMBL, DDB) are most popular for storing nucleotide sequences; -SWISS-PROT; PIR, TrEMBL, GenPept store protein sequences; -GDB, MGB, SGB etc., store complete Genomic data; -ESTs, STSs, REPBASE etc., are some of specialized databases; -PDB stores 3D structures and there are many more databases similar to this or resulted from them. As the biological data is increasing rapidly day by day, it becomes essential to know how to access this information. Retrieval systems like SRS, Entrez, DBGET are the most commonly used by biologists. By choosing proper matrices (Blosum, PAM) and algorithms (FASTA, BLAST.) you can proceed further in searching your query sequences to obtain the best homologue sequence. Once you obtain the hit (from the database) to your query sequence you can get all the additional information for your sequence and proceed with your experimental work.

1.5.1.4 Bioinformatics and its Scope

Bioinformatics area is not a new field. Although it existed in 1970's and the word Bioinformatics was coined inn mid 1980's this area has gained its importance, hype and recognition only after the release of Human Genomic project. Massive amounts of money has been invested in genomic projects in order to obtain the complete genomic picture of a particular organism like Human, Drosophila, Yeast, Rice, *E.coli* etc. Genomic sequencing are done in order to obtain the complete genome picture of that particular species and find out the exact number of genes that are involved in coding a protein. If we take the example of Human Genome, once we identify all the genes we can then go on studying the genes causing hereditary diseases (this is done by studying the genomic data from individuals belonging to same family members) also create drugs for that particular individual (human). As of today we see a large application in the pharma or medical area in order to provide a better life for the human being and make him fight against the diseases that are invading him, and also provide him with the best nutritional food using the present technology (Bioinformatics, Biotechnology). To increase the rate of interpreting things at a faster rate we depend on the computational work the data's efficacy is proved experimentally and applied in the respective way. Some of the other areas where Bioinformatics has been applied largely are pharmacogenomics, Medical Informatics, functional genomics, comparative genomics, proteomics, Agri Informatics, etc.

In last decade, Bioinfonnatics has emerged as a new discipline. Bioinformatics uses advances in the area of computer science, information science, computer and information technology, communication technology to solve complex problems in life sciences and particularly in biotechnology.

Data capture, data warehousing and data mining have become major issues for biotechnologists and biological scientists due to sudden growth in quantitative data in biology such as complete genomes of biological species including human genome, protein sequences, protein 3-D structures, metabolic pathways databases, cell line & hybridoma information, biodiversity related information.

Advancements in information technology, particularly Internet, are being used to gather and access ever increasing information in biology and biotechnology. Functional genomics, proteomics, discovery of new drugs and vaccines, molecular diagnostic kits and phrmacogenomics are some of the areas in which bioinformatics has become an integral part of Research & Development.

The knowledge of multimedia databases, tools to carry out data analysis and modeling of molecules and biological systems on computer workstations as well as in a network environment has become essential for any student of Bioinformatics.

Bioinformatics, the multidisciplinary area, has grown so much that one divides it into molecular bioinformatics, organal bioinformatics and species bioinformatics. Issues related to biodiversity and environment, cloning of higher animals such as Dolly and Polly, tissue culture and cloning of plants have brought out that Bioinformatics is not only a support branch of science but is also a subject that directs future course of research in biotechnology and life sciences.

As geneticists, microbiologists and other researchers continue to gather huge amounts of new information about the human genome and biological molecules, there is a growing need for sophisticated, computerized approaches for compiling and analyzing that data. The process by which that is done is called bioinformatics. Every major university in the world is trying to get its share in this field.

There is a great scope for Bioinformatics in India. Companies have to work hard to gain respect and credibility. Bioinformatics hasn't and cannot create a million jobs like IT as it is only a subset of IT. The numbers will increase but in small percentages.

"I wouldn't advice everyone to jump into this field as it would only dilute the market with excess supply of professionals. On the other hand, it might be good for companies as it would give us enough people to choose from", said Ocimum's Anuradha.

Another observation was that for a Bioinformatics company which hires 100 people, about 70 percent are people with core knowledge with some understanding of bioinformatics. The number of people with bioinformatics resumes have increased rapidly but the quality of these "professionals" hasn't.

"Companies like us are always looking for good people but it takes us, on an average, 100 shortlisted resumes to finally pick one qualified person," she added

According to Rajendran, Sr. Executive - Business Development, BrainWave Bioinformatics Ltd, a lot of universities and institutes are into bioinformatics.

Almost every university in Andhra Pradesh and Karnataka offer a diploma in Bioinformatics. The prominent ones are the University of Hyderabad, Osmania University, IICT, IIIT.

Many private institutions which started during the hype have shut down. There are very good universities like the University of Pune, Madurai Kamaraj, Bose Institute and Jawaharlal Nehru University. The IIT's at Kharagpur and Delhi also have a very good biotechnology department.

A lot of IT companies like TCS and Infosys have ventured in to this area but most of them do not have very large teams. Many large companies and research institutions are hiring hundreds of bioinformatics professionals.

"Bioinformatics as a career is very lucrative and has a great future. Requirement from an individual is the ability to contribute either in life sciences or in IT when working in a team comprising of professionals from both fields. Typical qualifications would be Masters and Ph.D. The salaries are benchmarked against industry standards and would be comparable with any other industry including IT. The sector is growing at an impressive rate and companies which understand the 'real issues' of the industry will only survive in the long run. Working with such companies will result in overall development for professionals in this sector," says Sowmya Narayan of Strand Genomics.

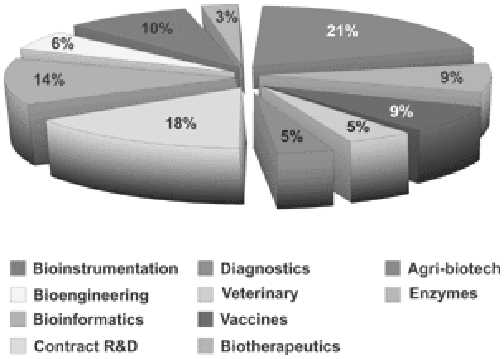
According to Dr GPS Raghava, Scientist & Co-ordinator of Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, there is a big gap between the demand and expertise available. The gap is not only in India but in the US also. Despite the hype and the presence of large number of bioinformatics training centers in India our contribution is too limited.

Is it easier to move from biology to computers or the reverse?

The answer depends on whether you are talking to a computer scientist who 'does' biology or a molecular biologist who 'does' computing. Most of what you will read in the popular press is that the importance of interdisciplinary scientists cannot be over-stressed and that the young people getting the top jobs in the next few years will be those graduating from truly interdisciplinary programs.

However, there are many types of bioinformatics jobs available, so no one background is ideal for all of them. The fact is that many of the jobs available currently involve the design and implementation of programs and systems for the storage, management and analysis of vast amounts of DNA sequence data. Such positions require in-depth programming and relational database skills which very few biologists possess and so it is largely the computational specialists who are filling these roles.

This is not to say the computer-savvy biologist doesn’t play an important role. As the bioinformatics field matures there will be a huge demand for outreach to the biological community as well as the need for individuals with the in-depth biological background necessary to sift through gigabases of genomic sequence in search of novel targets. It will be in these areas that biologists with the necessary computational skills will find their niche.



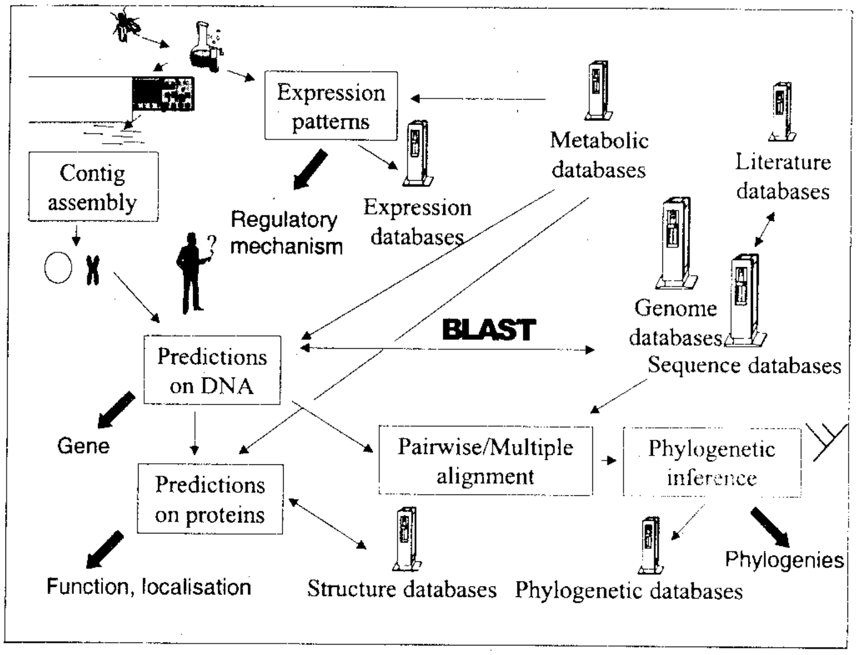
1.5.1.5 Other useful areas

Bioinformatics is today seen as primarily applied to speeding up new drug discovery. But the other area that assumes increasingly higher significance is the application of IT to the entire life sciences sector- for the same purpose it is done in other industrial sectors- improving efficiency, reducing costs, wider access, etc. For example bio-diversity data management is an area that requires application of the best database design techniques and planning for data warehousing and data-mining. Knowledge management as applied to corporations will also become relevant in the scientific context to ensure that Indian scientists get relevant and timely information related to their research to help them network and collaborate to create new intellectual property.

"There may be around 200-300 employed in this sector every year. There are a lot of private institutions getting into the foray, but then quality is indeterminate," said Rajiv Vasudevan, who is an expert both in IT and biotechnology.

Bioinformatics in India is at an early stage of development. But at 4 to 5 centers in the country, one sees mature understanding of the needs of this sector and world class development of tools and applications. These centers will ensure that India’s traditional strengths in IT are leveraged to place us on par with the developed countries.

One view of Bioinformatics



DNA microarrays

DNA microarray technology offers the first great hope for providing a systemic way to explore the genome. It permits very rapid analysis of thousands of genes for the purposes of gene discovery, sequencing, mapping and expression, and polymorphism detection. Powerful bioinformatics tools are used to manage the massive data generated by microarray technology.

Functional genomics

The term “genomics”, is the study of genes and gene activity within an organism. In humans, the genome consists of three billion nucleotides, comprising roughly about 30,000-100,000 genes on 23 pairs of chromosomes. “Functional genomics” is the study of the function of these genes and an understanding of how they affect normal physiology. Because of the rapid progress in sequencing of the human genome, In the post-genome era, when all genes are known the challenge then will be to ascribe to each gene its biological relevance and function in health and disease. Functional genomics thus is the derivative science of database driven discovery tools and biological techniques to ascribe the function of genes.

Comparative genomics

The functions of human genes and other DNA regions often are revealed by studying their parallels in nonhumans. To enable such comparisons, HGP researchers have obtained complete genomic sequences for the bacterium Escherichia coli, the yeast Saccharomyces cerevisiae, the roundworm Caenorhabditis elegans, the fly Drosophila melanogaster and the laboratory mouse. The availability of complete genome sequences generated both inside and outside the HGP is driving a major breakthrough in fundamental biology as scientists compare entire genomes to gain new insights into evolutionary, biochemical, genetic, metabolic, and physiological pathways.

Comparative genomics basically involve the comparison of proteomes (the complete protein set) of two or more organisms. In addition, it involves comparison of the gene locations, relative gene order (or synteny) and comparative regulation. It also involves an examination of such events as gene loss or gene duplications and horizontal gene transfer. Such analysis eventually aims to go beyond mere descriptions, of the absence or presence of similarities & differences, and eventually attempts to evolve models and rules that might explain such events.

Pharmacogenomics

Pharmacogenomics or pharmacogenetics is intersection of the fields of pharmacology and genetics. In other words pharmacogenomics is the study of how genetic variations affect the ways in which people respond to drugs. Pharmacogenomics involves the identification and elucidation of genetic variations that will alter the efficacy of a drug or suggest new drug targets. By understanding the components of a metabolic pathway, researchers can tailor drugs more specifically and reduce or eliminate potential side effects. The main goal of genomics research is to develop better, more specific drugs (single gene-drug) and move away from the "one drug for all" mode of pharmaceutical development of the present trend. Moreover, pharmacogenomic analysis can identify disease susceptibility genes representing potential new drug targets.

Chemoinformatics

During development of a new pharmaceutical therapy, no matter how long gene sequences, gene expression or protein activities are studied finally some chemistry has to be performed. There are literally dozens of databases containing information about chemical structures, reaction kinetics, and synthetic methods, but they are not typically comprehensive and, although they are useful, finding significant information can be cumbersome. To address these kinds of problems various groups of people have tried to coordinate the content of these sources through their own Internet search engines.

The Comprehensive Medical Chemistry database contains compounds that are biologically active and have been used to pharmacological agents. The Toxicity database allows structure-based searches of more than 250,000 toxic chemical substances, and drugs and drug-development compounds constitute 65% of the substances covered. Once a chemical has shown some biological activity and become a lead drug candidate, it is necessary to generate or purchase a series of compounds similar in structure to determine their efficacy. For those unable or unwilling to buy their compounds, reaction databases are available.

Medical informatics

Medical informatics is also known as Biomedical informatics or clinical bioinformatics. Medical informatics can be simply defined as computer applications in medical care more precisely it is defined as an emerging discipline that has been defined as the study, invention, and implementation of structures and algorithms t improve communication, understanding and management of medical information. The end objective of biomedical informatics is the coalescing of data, knowledge, and the tools necessary to apply that data and knowledge in the decision-making process, at the time and place that a decision needs to be made.

Neural networks

Neural networks are a new method of programming computers. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well. An ANN consists of many simple computational neural units connected to each other. An input is presented to some of (or all) its input units, this input vector is propagated through the whole network and finally, some kind of output is splitted out. So, essentially, they are functions. The network gets an input as an argument and gives output for that particular input. Because input and output can consist of many units or components, they are considered as vectors. However, ANN's real power is on its ability to learn, that is, the function is not constant but can be changed dynamically.

Phylogeny

Phylogeny is the study of the evolution of life forms. Phylogenetic tree, also called a cladogram or a dendrogram is a tree of several life forms and their relations. The problem in phylogenetic tree construction, is to find the tree which best describes the relationship between a set of species or genes or proteins. Throughout the evolutionary history of life many things have happened: organisms have consumed, been consumed, made babies, interacted with other organisms, and through the years, they have evolved. An important component of this evolutionary history is the passage of genetic information from parent to offspring. One species makes two, those two each make more, and so on. This process gives rise of a tree of species lineages descending and splitting. At the same time. Some of the species go extinct, thus pruning the growing tree. In the end, this process has produced many species of organisms we see around us today.

Whole cell simulation

Genome sequencing projects and further systematic functional analyses of complete gene sets are producing an unprecedented mass of molecular information for a wide range of model organisms. This provides a detailed account of the cell with which models can be build for simulating intracellular molecular processes to predict the dynamic behavior of living cells. Previous work in biochemical and genetic simulation have isolated well-characterized pathways for detailed analysis, but methods for building integrative models of the cell that incorporate gene regulation, metabolism and signaling have not been established. E-CELL System, a generic software package was developed for building such integrative models based on gene sets, and running simulations to conduct experiments in silico.

Human genome project

In 1988 the Human Genome Project (HGP) was initiated by founding the Human Genome Organisation (HUGO) in the USA. The Human Genome Project is a worldwide research effort initiated as a multi-disciplinary effort to understand the basis of human heredity. This international collaboration is being carried out at several genome centers located in different parts of the world. The focus of the Human Genome Project is the characterization of the human genome by determining the complete nucleotide sequence of our chromosomes, including the estimated 30,000 to 100,000 genes contained in human DNA. Sequencing the entire human genome will enable us to identify all human genes, investigate how the control of gene expression contributes to the development of humans and other organisms and understand the molecular basis of evolution. In addition several model organisms are also being sequenced including yeast, nematode and fruit fly.

Opponents of this project, however, fear that it will have negative consequences for future human biological and social life. Program planners have recognized that this information would raise many complex ethical, legal and social issues, including the interpretation and use of genetic information and issues of privacy. The Ethical, Legal and Social Implications (ELSI) Program was established in 1989 to address these complex issues while the scientific studies were being carried out, so that solutions to these issues could be developed in parallel to the scientific developments.

The results of this project are supposed to further the understanding of genetic diseases and shall render possible new ways of their diagnosis and therapy. New maps developed through the Human Genome Project will enable researchers to pinpoint specific genes on our chromosomes. The most detailed map will allow scientists to decipher the genetic instructions encoded in the estimated 3 billion base pairs of nucleotide bases that make up human DNA. By understanding the genetic makeup of an organism in excruciating detail, HGP hope to better understand how organisms develop from single eggs into complex multicellular beings, how food is metabolized and transformed into the constituents of the body, and how the nervous system assembles itself into a smoothly functioning ensemble. From the medical point of view, the wealth of knowledge that will come from knowing the complete DNA sequence will greatly accelerate the process of finding the causes of, and potential cures for human diseases. Analysis of this information, likely to continue throughout much of the 21st century, will revolutionize our understanding how genes control the functions of the human body. It will also help us to explain the mysteries of our evolutionary past. Like the big question “How life began on earth” whose answers rely on understanding the tiniest ones.

1.5.1.6 The Potential of Bioinformatics

The potential of bioinformatics in the identification of useful genes leading to the development of new gene products, drug discovery and drug development has led to a paradigm shift in biology and biotechnology - these fields are becoming more and more computationally intensive. The new paradigm, now emerging, is that all the genes will be known "in the sense of being resident in databases available electronically", and the starting point of biological investigation will be theoretical and a scientist will begin with a theoretical conjecture and only then turning to experiment to follow or test the hypothesis. With a much deeper understanding of the biological processes at the molecular level, the Bioinformatics scientists have developed new techniques to analyse genes on an industrial scale resulting in a new area of science known as 'Genomics'.

The shift from gene to genome biology has resulted in the development of strategies - from lab techniques to computer programmes to analyse whole batch of genes at once. Genomics is revolutionizing drug development, gene therapy, and our entire approach to health care and human medicine.

The genomics discoveries are getting translated in to practical biomedical results through Bioinformatics applications. Work on proteomics and genomics will continue using highly sophisticated software tools and data networks that can carry multimedia databases. Thus, the research will be in development of multimedia databases in various areas of life sciences and biotechnology. There will be an urgent need for development of software tools for datamining, analysis and modelling, and downstream processing. Security of data, data transfer and data compression, auto checks on data accuracy and correctness will also be major research areas of Bioinformatics. The use of Virtual Reality in drug design, metabolic pathway design, and unicellular organism design, paving the way to designand modification of multicellular organisms, will be the challenges which Bioinformatics scientists and specialists have to tackle. It has now been universally recognized that Bioinformatics is the key to the new grand data-intensive molecular biology that will take us into 21 st century.

Summary

As the facts of biology started accumulating because of diligent study vast amounts of data have been pouring which need sophisticated means of analysis rather than manual analysis. Biological databases store data related to biology. It may be from literature, sequencing projects, structure determination, etc. Bioinformatics, the multidisciplinary area, has grown so much that one divides it into molecular bioinformatics, organal bioinformatics and species bioinformatics. Bioinformatics is today seen as primarily applied to speeding up new drug discovery.

Model Questions

1. Write notes on scope and importance of bioinformatics?
2. Briefly explain the scope and other useful areas of Bioinformatics?

References

Introduction to Bioinformatics by Arthur M. Lesk.

1. Bioinformatics: wave of the future by Roby Ajith, Available online at Biospectrum home page.

B.M.REDDY **M.Tech. (HBTI, Kanpur)**

Lesson 1.5.2

CHALLENGES IN INFORMATION PROCESSING

Objective:

The objective of this lesson is to explain about the new technologies in information processing and the challenges facing in processing the biotechnology information.

1.5.2.1 Introduction

A very exciting development in current intelligent information processing is the Semantic Web and the innovative e-applications it promises to enable. This promise will not come true, however, if research limits itself to the technological aspects and challenges only. Both supply-demand sides and business-technology sides need to be investigated in an integrated fashion. This implies that we simultaneously have to address technological, social, and business considerations. Therefore, a comprehensive research strategy for the next decade of intelligent information processing must be of an integrated socio-technical nature covering different levels: (1) Definition and standardization of the baseline infrastructures, content libraries and languages that make up the Semantic Web; (2) The associated construction of generic smart web services that dynamically bridge the low-level (for the end user) infrastructures and the high-level user applications; (3) Designing and studying innovative e-services, information systems, and business processes at the domain, customer, and business level; (4) Understanding and influencing the business and market logics and critical success factors that will determine the social adoption of smart web-based innovations.

1.5.2.2 Exciting New Technologies

A very exciting development in current intelligent information processing is the Semantic Web and the innovative applications it promises to enable. The Semantic Web will provide the next generation of the World Wide Web. The current Web is a very interesting and successful, but also passive and rather unstructured storage place of information resources. This makes it increasingly difficult to quickly find the right information you need, a problem that becomes even more pressing with the scaling up of the Web. The vision of the Semantic Web is to make the Web from a passive information store into a proactive service facility for its users. This is done by equipping it with information management services, based on semantic and knowledge-based methods, that let the Web act - in the eyes of its users - as understanding the contents and meaning (rather than just the syntax) of the many information resources it contains and, moreover, as capable of knowledge processing these resources. In the words of Tim Berners-Lee: "The Semantic Web will globalise knowledge representation, just as the WWW globalised hypertext". This globalised semantic approach offers concrete research lines how to solve the problem of interoperability between systems and humans in a highly distributed but connected world.

Designing the infrastructure of the Semantic Web poses major technical and scientific challenges. This is already evident if we look at the envisaged technical architecture of the Semantic Web (see Figure 1) that somewhat resembles a delicately layered cake made from a variety of cyberspace ingredients.

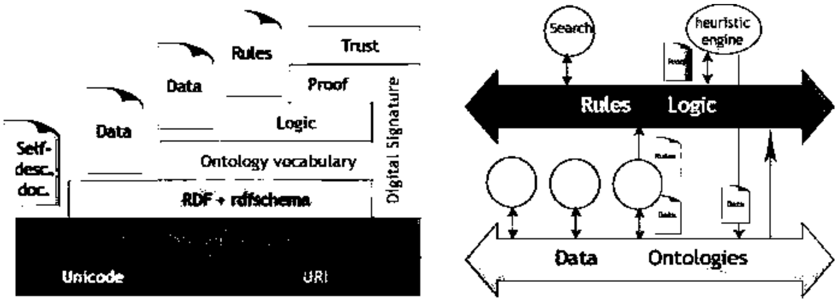


Figure 1. Ingredients and envisaged technical architecture of the Semantic Web.

Challenging and interesting as this is, it is a necessary but not yet sufficient condition to realize the full potential of the Web. For a comprehensive R&D strategy it is necessary to look at the broader picture (depicted in Figure 2) of the Semantic Web: how it is going to be useful in practical real-world applications, and how it will interact with and be beneficial to its users.

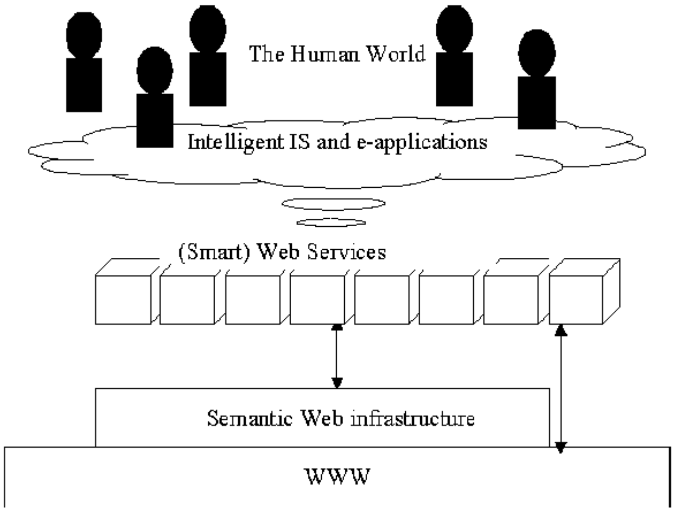

The Human World
Intelligent IS and e-applications
(Smart) Web Services
Semantic Web infrastructure
WWW

Figure 2. The broader picture: Semantic Web infrastructure, smart services, e-applications and their human-world context.

The ongoing worldwide research effort related to the Semantic Web currently shows an emphasis on those technological issues that are indicated in Figure 2 as web infrastructure and, to a lesser extent, smart web services. This is highly important research because generic semantic infrastructure (such as web ontology languages and content libraries) and associated generic smart web services (such as semantic search, semantic browsing, reasoning, knowledge processing and ontology management services – these services are highly non-trivial because they must be able to deal with the unavoidable evolutional dynamics of web-based knowledge) are a *conditio sine qua non* for the Semantic Web. Nevertheless, it is also important to look already from the start from an *outside-in* perspective. What are the new business, domain, or user/customer applications that are not yet possible today but will be tomorrow as a result of the Semantic Web? Why would businesses, markets or individuals be willing to adopt such innovations? After all, many great innovations fail or have very long lead times because of significant upfront investments. These are in many cases not just of a financial nature: in addition they require behavioural or -even more problematic- cultural changes from their adopters (whether individuals or organizations). We must recognize that the Semantic Web is such a great innovation. Consequently, there is no reason to assume that the new wave of intelligent information processing is immune to the age-old established social laws that govern innovation adoption.

1.5.2.3 Innovative E-Applications

To illustrate some of the pertinent issues I will consider a few specific examples of advanced intelligent information processing that aim the creation and introduction of innovative e-applications for end users (the third level in Figure 2). In addition to the Web becoming smarter (which is denoted by the Semantic Web effort), it will also become more universal in the sense that it will not just connect computers, but essentially any device. This is variously referred to as "ambient intelligence", "universal connectivity" or "pervasive computing". Mobile commerce applications are one step in this direction, but basically all equipment, including home appliances such as personal audio and video, telecom and home control systems, and even heaters, coolers or ventilation systems, will become part of the Web. This enables a broad spectrum of e-applications and e-services for end consumers in many different industry areas: home security, e-health, e-entertainment, e-shopping, distance learning, digital media services, and smart buildings that are able to manage themselves. All of these new imagined e-services are technically challenging, but will also require and induce different behaviours and attitudes from the end consumers as well as from the businesses delivering these e-services.

As a specific example, I take smart buildings. With several colleagues from different countries, we are researching how smart buildings can serve those who live or work in. This work has progressed to the point that actual field experiments are carried out (Figure 3), whereby the social aspects are investigated as an integrated part of the research. One of the issues studied is comfort management: how buildings can automatically provide a optimally comfortable climate with costs and energy use that are at the same time as low as possible.

Technically, smart comfort management is based on intelligent agents (so-called *HomeBot* agents) that act as software representatives of individual building users as well as of various types of equipment that play a role in the energy functionality, usage and production in a building (e.g. heaters, sun-blinds, ventilators, photovoltaic cells). These *HomeBot* agents communicate with each other over Internet and PLC media, and negotiate in order to optimise the overall energy efficiency in the building. This optimisation is based on multi-criteria agent negotiations taking place on an electronic marketplace. These take place in the form of a multi-commodity auction, where energy is being bought and sold in different time slots. They are based on the current energy needs, local sensor data, model forecasts (e.g. weather, building physics), and the going real-time power prices. The e-market outcome then determines the needed building control actions in a fully distributed and decentralised way.

The calculation model optimises the total utility, which is a trade-off between cost and comfort, over the coming 24 hours, taking into account both the customer preferences and the actual energy prices. This optimisation is redone every hour, because expected energy prices, outside temperatures, etc. may change, which results in different optimal device settings. Needed forecasts of comfort aspects in a building are based on simple thermodynamic climate models. Energy prices are in general known a certain period (typically 24 hours) in advance. The system reacts on electricity prices, trying to use as little energy as possible when prices are high. In simulations we have concentrated on two dimensions: the economic aspect and the inside climate. The economic aspect is illustrated by a scenario featuring two archetypes: Erika, a yuppie who wants to make no concessions to her comfort level whatsoever irrespective of cost; and Erik, a poor student who wants to keep comfort levels acceptable when at home, but also needs to economise as much as possible. Some typical results are presented in Figure 4. They do show that significant savings without loss of comfort are possible in smart self-managing buildings.

1.5.2.4 The Promise and Challenge of Emerging Technologies

The Internet has evolved into a global information network and has developed beyond its original purpose of sharing information into a global commercial trading system. Electronic commerce is straining existing trade regimes, protocols for the protection of intellectual property, and concepts of currency. It is creating problems and jurisdictional issues for taxation and regulation. A number of governance issues have already arisen, such as the fairness of the current system of allocating Internet domain names in an international environment.

In the future, problems related to information security will require a high degree of international cooperation to govern or resolve. These include both the use of the Internet for crime and the misuse of the network by public and private groups in ways that invade personal privacy. Some suggest that contractual relationships will replace regulation and trade protocols. What organization is capable of negotiating and implementing new rules or enforcing net-based contracts? What court of law will adjudicate international contracts agreed to on the global information infrastructure. Can these simply be folded into the World Trade Organization, or do they need a separate institution?

Recent developments in biological sciences, particularly in genetics, raise the question of international organizational and legal governance. Procedures that are judged ethically or medically objectionable in one country may become available elsewhere through market mechanisms, leading to the development of foreign sites where individuals may go to avoid regulations. Can existing organizational structures and laws be adapted to the products of biological science? Will new forms of political organization and law be needed to address these changes?

Similar questions will be raised as biological sciences and computer sciences converge into applications called *bioinformatics*. As science explores creating information technology that can be used as a human prosthetic—either worn on the body or implanted under the skin—questions about when it is appropriate to use these technologies and under what conditions will arise. Science is also exploring the use of biological materials as information processors in objects, such as "biochips." Technologists suggest that miniature biological sensors detecting chemical and biological information may soon be available that will be capable of providing instant feedback on individual or group activities and, further, of linking this information into ultrascale networked computing. How can abuses of these technologies, such as surveillance and large-scale information-gathering among the population, be anticipated and regulated or countered? This section explores the possibilities and challenges that areas of technological change posing particular challenges to global governance may offer.

1.5.2.5 Challenges for Bioinformatics Industry

There are numerous challenges for the bioinfonnatics industry.

It must be able to deal with increasingly complex data and to integrate data sources into a single system.

The amount of data is staggering. For example, rnicroarrays now allow for the generation of data points for thousands of genes instead of the single genes that were investigated in traditional experiments. This flood of data has resulted in a glut of candidates to develop into drugs. As a result, phannaceutical companies must learn portfolio management to detennine which drug candidates to pursue.

Diverse types of data must be handled simultaneously to provide a better understanding of what genes do. New analysis strategies, are required to detennine patterns in information. Once the infonnation is in an organized state, bioinformaticians can apply data mining strategies from other fields-those used by spy satellites and astronomy, where there is a need to pick up weak signals.

Another major challenge for the bioinfonnatics industry is trying to anticipate wh(lt _ill be the next llspful data SOllr_es_ analysis tools This sItuation is reflected in the wet labs, in which hands-on test tubes and spectrometers have given way to automated DNA sequencers.

.Further more, as new technologies are created, different data types must be melded into software systems. Relate one te of data to the next is not as simple content as recalling it from a database. The data has to be normalized content so e researchers can make comparisons from one technique or type of data to another.

As companies with a bioinformatics business consolidate, a related problem springs up: integrating the multiple databases that must communicate with each other. Each company has its-own sYStem of registering or numbering compounds _em that must be made compatible with the system used by the other company. The lack of standards hurts companies that are attempting to merge data and annotations.

The bioinformatics industry has to meet four chief informational challenges if it is to make sense of the avalanche of data coming out of genomics and proteomics research: (i) Annotating data better

(ii) Filtering, visualizing

(ill) Analyzing data using better algorithms

(iv) Integrating genomic and gene expression data more effectively.

These challenges share the need for data organization.

Bioinformatics is the development and use of computational and mathematical methods for acquiring, storing, analysing and interpreting biological data to solve biological questions. Today's biological research generates a huge quantity of data. This is growing exponentially with a shift in emphasis from individual biomolecules, to analysis of how they interact in complex networks which control the developmental and physiological processes of whole biological systems, and research into how this relates to human health. This transition has increased the importance of bioinformatics and raises key challenges which make it imperative that computer scientists work closely with biologists to refine existing bioinformatics tools and develop new ones. Bioinformatics can be described as the science of collecting, modelling, storing, searching, annotating and analysing biological information. It involves a range of activities from data handling, publication, to data mining and analysis. An essential part of bioinformatics is to create new algorithms for the analysis of complex and/or large data sets.

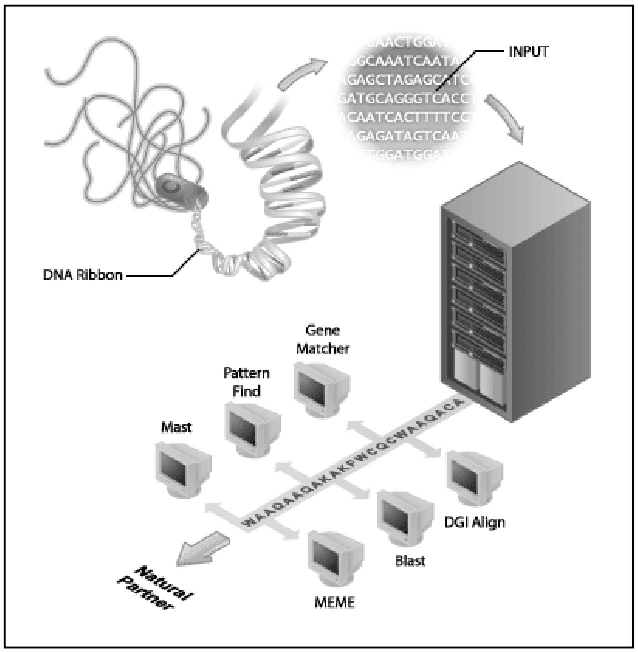


Figure 1. The Use of Computers to Process Biological Information. The wealth of genome sequencing information has required the design of software and the use of computers to process this information.

1.5.2.6 New Challenges of Bioinformatics

The use of informatics to organize, manage, and analyze genomic data (the genetic material of an organism) has become an important element of biology and medical research. A new IT discipline— *bioinformatics*—fuses computing, mathematics, and biology to meet the many computational challenges in modern molecular biology and medical research. The two major themes in bioinformatics—data management and knowledge discovery—rely on effectively adopting techniques developed in IT for biological data, with IT scientists playing an essential role.

Bioinformatics deals with the issues created by the massive amounts of new types of data obtained through novel biological experiments. Informatics has helped launch molecular biology into the genome era. In the 1990s, the Human Genome Project and other genome sequencing efforts generated large quantities of DNA sequence data. Informatics projects in algorithms, software, and databases were crucial in the automated assembly and analysis of the genomic data. The Internet also played a critical role: the World Wide Web let researchers throughout the world instantaneously share and access biological data captured in online community databases. Information technologies produced the necessary speedup for collaborative research efforts in biology, helping genome researchers complete their projects on time. Many genomes have already been completely sequenced, and genome research has migrated from raw data generation to scientific knowledge discovery. Likewise, informatics has shifted from managing and integrating sequence databases to discovering knowledge from such biological data. Informatics' role in biological research has increased and it will certainly become increasingly important in extending our future understanding of biological life.

The basic data has so far usually been sequence data (nucleotide or protein), but other types of data (microarray, functional analysis, interactions) are now rapidly coming into focus.

1.5.2.7 Technical issues

Collecting, storing, searching and using biological information entails a number of technical problems, ranging from the trivial but important (file formats, database interactions) to the sophisticated (algorithm design, data modelling, ontologies). For example:

- Algorithms for analysis: properties, implementation
- Software libraries implementing analysis or access methods
- Data modelling: optimal ways to represent heterogenous data
- Object-oriented vs. relational databases
- Database technologies, implementations
- Centralised database systems, or distributed data networks?
- Update policies, data tracking (who modifies what)

Sharing of software: languages, licenses, machine independence

Data Management

The many genome mapping and sequencing initiatives of the 1990s resulted in numerous databases. The hot topics then were managing and integrating these databases and comparing and assembling the sequences they contained.

Data integration

No single data source can provide answers to many of biologists' questions; however, information from several sources can help satisfactorily solve some of them. GDB,a relational database from the company Sybase supporting

Structured Query Language (SQL) queries, was located in Baltimore, Maryland. Entrez, which users accessed through an ASN.1 (Abstract Syntax Notation One) interface supporting simple keyword indexing, was in Bethesda, approximately 38 miles south. Kleisli, a powerful general query system developed at the University of Pennsylvania in the mid-1990s, solved this problem. Kleisli lets users view many data sources as if they reside within a federated nested relational database system. It automatically handles heterogeneity, letting users formulate queries in an SQL-like highlevel way independent of

• the data sources' geographic location,

• whether the data source is a sophisticated relational database system or a dumb flat file, and

• the access protocols to the data sources.

Kleisli's query optimizer lets users formulate queries clearly and succinctly without having to worry about whether the queries will run fast. Several additional approaches to the biological data integration problem exist today.Ensembl,SRS, and Discovery-

Link are some of the better-known examples.

• EnsEMBL (http://www.ensembl.org) provides easy access to eukaryotic genomic sequence data. It also automatically predicts genes in these data and assembles supporting annotations for its predictions. Not quite an integration technology, it's nonetheless an excellent example of successfully integrating data and tools for the highly demanding purpose of genome browsing.

• SRS (http://srs.ebi.ac.uk) is arguably the most widely used database query and navigation system in the life science community. In terms of querying power, SRS is an information retrieval system and doesn't organize or transform the retrieved results in a way that facilitates setting up an analytical pipeline.However, SRS provides easy-to-use graphical user interface access to various scientific databases.For this reason, SRS is sometimes considered more of a user interface integration tool than a true data integration tool.

• IBM's DiscoveryLink (http://www.ibm.com/discoverylink) goes a step beyond SRS as a general data integration system in that it contains an explicit data model—the relational data model. Consequently, it also offers SQL-like queries for access to biological sources, albeit in a more restrictive manner than Kleisli, which supports the nested relational data model. Recently, XML has become the de facto standard for data exchange between applications on the Web. XML is a standard for formatting documents rather than a data integration system. However, taken as a whole, the growing suite of tools based on XML can serve as a data integration system. Designed to allow for hierarchical nesting (the ability to enclose one data object within another) and flexible tag definition, XML is a powerful data model and useful data exchange format, especially suitable for the complex and evolving nature of biological data. It's therefore not surprising that the bioinformatics database community has rapidly embraced XML. Many bioinformatics resource and databases such as the Gene Ontology Consortium (GO, http://www.geneontology.org), Entrez, and the Protein Information Resource (PIR, http://pir.georgetown.edu) now offer access to data using XML. The database community's intense interest in developing query languages for semistructured data has also resulted in several powerful XML query languages such as XQL and XQuery. These new languages let users query across multiple bioinformatics data sources and transform the results into a more suitable form for subsequent biocomputing analysis steps. Research and development work on XML query optimization and XML data stores is also in progress. We can anticipate robust and stable XML-based general data integrating and warehousing systems in the near future. Consequently, XML and the growing suite of XML-based tools could soon mature into an alternative data integration system in bioinformatics comparable to Kleisli in generality and sophistication.

Data warehousing

In addition to querying data sources on the fly, biologists and biotechnology companies must create their own customized data warehouses. Several factors motivate such warehouses:

- Query execution can be more efficient, assuming data reside locally on a powerful database system.

- Query execution can be more reliable, assuming data reside locally on a high availability database system and a high availability network.

- Query execution on a local warehouse avoids unintended denial-of-service attacks on the original sources.

- Most importantly, many public sources contain errors. Some of these errors can't be corrected or detected on the fly. Hence, humans—perhaps assisted by computers—must cleanse the data, which are then warehoused to avoid repeating this task. A biological data warehouse should be efficient to query, easy to update, and should model data naturally. Biological data's complex structure makes relational database management systems such as Sybase unsuitable as a warehouse. Such DBMSs force us to fragment our data into many pieces to satisfy the third normal form requirement. Only a skilled expert can perform this normalization process correctly. The final user, however, is rarely the same expert. Thus, a user wanting to ask questions on the data might first have to figure out how the original data was fragmented in the warehouse. The fragmentation can also pose efficiency problems, as a query can cause the DBMS to perform many joins to reassemble the fragments into the original data.

Kleisli can turn a relational DBMS into a nested relational DBMS. It can use flat DBMSs such as Sybase, Oracle, and MySQL as its updateable complex object store. In fact, it can use all of these varieties of DBMSs simultaneously.This capability makes Kleisli a good system for warehousing complex biological data. XML, with its built-in expressive power and flexibility, is also a great contender for biological data warehousing. More recently, some commercial relational DBMSs such as Oracle have begun offering better support for complex objects. Hopefully, they'll soon be able to perform complex biological data warehousing more conveniently and naturally.

Knowledge Discovery

As we entered the era of post-genome knowledge discovery, scientists began asking many probing questions about the genome data. The genome projects' success depends on the ease with which they can obtain accurate and timely answers to these questions. Informatics therefore plays a more important role in upstream genomic research. Three case studies illustrate how informatics can help turn a diverse range of biological data into useful information and valuable knowledge.This can include recognizing useful gene structures from biological sequence data, deriving diagnostic knowledge from postgenome experimental data, and extracting scientific information from literature data. In all three examples, researchers used various IT techniques plus some biological knowledge to solve the problems effectively. Indeed, bioinformatics is moving beyond data management into a more involved domain that often demands in-depth biological knowledge; postgenome bioinformaticists are now required to be not just computationally sophisticated but also biologically knowledgeable.

Biological Sequence Analysis

In addition to having a draft human genome sequence, we now know many genes' approximate positions. Each gene appears to be a simple-looking linear sequence of four letter types (or *nucleotides*)—As,Cs,Gs, and Ts—along the genome. To understand how a gene works, however, discovery of the gene's underlying structures along the genetic Sequence is essential, such as its transcription start site (point at which transcription into nuclear RNA begins), transcription factor binding site, translation initiation site (point at which translation into protein sequence begins), splice points, and poly(A) signals.Many genes' precise structures are still unknown, and determining these features through traditional wet-laboratory experiments is costly and slow.

Computational analysis tools that accurately reveal some of these features will therefore be useful, if not necessary. Informatics lets us solve the TIS recognition problem using computers.Translation is the biological process of synthesizing proteins from mRNAs.The TIS is the region where the process initiates.

Scientific Literature Mining

Other than the molecular sequence databases generated by the genome projects, much of the scientific data reported in the literature have not been captured in structured databases for easy automated analysis. For instance, molecular interaction information for genes and proteins is still primarily reported in scientific journals in free-text formats. Molecular interaction information is important in postgenome research. Biomedical scientists have therefore expended much effort in creating curated online databases of proteins and their interactions, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG,www.kegg.org) and the Cell Signaling Networks Database (CSNDB, geo.nihs.go.jp/csndb). However, such hand-curated databases are laborious and unlikely to scale. Natural language processing (NLP) of biomedical literature is one alternative to manual text processing.

The system collects numerous abstracts and texts from biological research papers in scientific literature databases such as NCBI's Pies is one of the first systems capable of analyzing and extracting interaction information from English-language biology research papers.

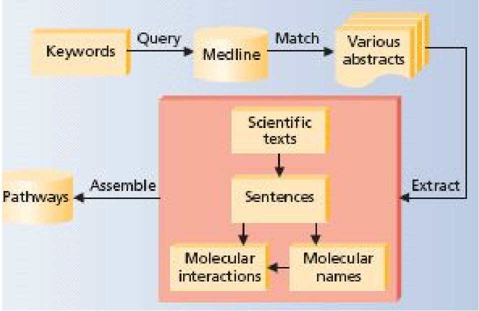


Figure : Mining literature protein-protein interactions

Pies is a rule-based system that recognizes names of proteins and molecules and their interactions. The future of molecular biology and biomedicine will greatly depend on advances in informatics.

Summary:

Bioinformatics deals with the issues created by the massive amounts of new types of data obtained through novel biological experiments. The Internet also played a critical role: the World Wide Web let researchers throughout the world instantaneously share and access biological data captured in online community databases. Information technologies produced the necessary speedup for collaborative research efforts in biology, helping genome researchers complete their projects on time. Many genomes have already been completely sequenced, and genome research has migrated from raw data generation to scientific knowledge discovery. Likewise, informatics has shifted from managing and integrating sequence databases to discovering knowledge from such biological data.

Model Questions:

1. Why information processing is a challenge for Bioinformaics industry?
2. How computers are helping to overcome the information processing challenges of bioinformatics?

References:

1. Being Smart In Information Processing, Technological And Social Challenges And Opportunities ,Hans Akkermans, Free University Amsterdam VUA, The Netherlands
2. Recent Biotechnology Development: challenges and opportunities to the consolidation of its knowledge "building blocks". Maria G. Derengowski Fonseca
3. Bioinformatics – David Mount
4. Introduction to Bioinformatics – S. Sundara Rajan and R.Balaji

B.M.REDDY **M.Tech. (HBTI, Kanpur)**

Lesson 1.5.3

Knowledge Based Data Analysis

Objective

To know the need of knowledge based analysis and how it differs from machine learning approach.

1.5.3.1 Introduction

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD).

One of the most exciting areas of modern biology is the application of data mining methods to biological databases. Many of these methods can equally well fall into the category of machine learning, the name used in the artificial intelligence community for the larger family of programs that adapt their behavior with experience.

A word of caution: anthropomorphisms have a tendency to creep into discussions of data mining and machine learning, but there is nothing magical about them. Programs are said to "learn" or be "trained," but they are always just following well-defined sets of instructions. Data mining tools are supplements, rather than substitutes, for human knowledge and intuition. No program is smart enough to take a pile of raw data and generate interesting results, much less a publication-quality article ready for submission to the journal of your choice. As we've stressed before, the creation of a meaningful question, the experimental design, and the meaningful interpretation of results are your responsibility and yours alone.

1.5.3.2 Knowledge Discovery & Data Mining in Bioinformatics

Biological databases continue to grow rapidly. This growth is reflected in increase in both size and complexity of individual databases as well as in the proliferation of new databases. A huge body of data is thus available for the extraction of highlevel information including the development of new concepts, concept interrrehttjonships and interesting patterns hidden in the databases. Knowledge Discovery in Databases (KDD) is an emerging field combining techniques from Database, Statistics and Artificial Intelligence, which is concerned with the theoretical and practical issues of extracting high level information (or knowledge) from volumes of low level data. At the core of KDD is data mining- the application of specific tools for pattern discovery and extraction. KDD process comprises several data pre-processing steps as well as data mining and knowledge interpretation steps.

Studies of biological data involve access to multiple databases using a variety of query tools. The amounts of biological data are growing faster than the capability to analyse them. KDD offers the capacity to automate complex search and data analysis tasks.

We can distinguish two types of goals of KDD systems: verification and discovery. With verification, the system is limited to verifying the users hypothesis. With discovery, the system autonomously finds new patterns. Discovery can be subdivided into prediction and description (explanation) goals. Biological sources are highly heterogeneous, geographically dispersed, constantly evolving, and often high in volume. They represent data from a highly complex domain. Numerous tools suitable for data mining in biology are available, yet the selection of an appropriate tool is a non-trivial task. The KDD process provides for the selection of the appropriate data mining methods by taking into account both domain characteristics and general KDD process requirements.

1.5.3.3 Why Do We Need KDD?

The traditional method of turning data into knowledge relies on manual analysis and interpretation. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changes in health-care data, say, on a quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-care organization; this report becomes the basis for future decision making and planning for health-care management. In a totally different type of application, planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloging such geologic objects of interest as impact craters. Be it science, marketing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products.

For these (and many other) applications, this form of manual probing of a data set is slow, expensive, and highly subjective. In fact, as data volumes grow dramatically, this type of manual data analysis is becoming completely impractical in many domains. Databases are increasing in size in two ways: (1) the number N of records or objects in the database and (2) the number d of fields or attributes to an object. Databases containing on the order of $N = 10^9$ objects are becoming increasingly common, for example, in the astronomical sciences. Similarly, the number of fields d can easily be on the order of 10^2 or even 10^3, for example, in medical diagnostic applications. Who could be expected to digest millions of records, each having tens or hundreds of fields? We believe that this job is certainly not one for humans; hence, analysis work needs to be automated, at least partially.

The need to scale up human analysis capabilities to handling the large number of bytes that we can collect is both economic and scientific. Businesses use data to gain competitive advantage, increase efficiency, and provide more valuable services to customers. Data we capture about our environment are the basic evidence we use to build theories and models of the universe we live in. Because computers have enabled humans to gather more data than we can digest, it is only natural to turn to computational techniques to help us unearth meaningful patterns and structures from the massive volumes of data. Hence, KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: data overload.

1.5.3.4 KDD Applications

In many industries, large amounts of quickly changing data may make it difficult to get a handle on important business trends. For example, by starting with general trends, the retail industry has been able to use knowledge discovery to uncover details, such as how purchase patterns relate to demographics, time'of day, and amount spent. Other industries have not caught up in their level of knowledge discovery sophistication, missing opportunities to leverage current customers.

One of the differences between business-based data warehouses and genetics-based data warehouses, however, is that business data tends to be better defined and more complete than genetics data. In fact, the scarcity of complete genetic information is the reason that knowledge discovery in biotechnology is a growing field. Molecular biologists recognize the opportunity to extract valuable knowledge from the existing genetics databases. They are now trying to create complete genomic databases through projects like the Human Genome Project and other less publicized projects.

The first goal of the projects - to build a complete database of the DNA sequence of a human, animal, or plant - has been completed at a coarser level of resolution. However, these efforts are complicated by the fact that there is no one DNA molecule for a given species. Your DNA differs from mine by a small percentage, which is why you are different from me. With hundreds of millions of base pairs in a DNA molecule, however, a small percentage adds up to a large absolute number. Therefore, your DNA may differ from mine by a million base pairs. So whose DNA do they store in the Human Genome Project database?For the first attempt, genetic scientists build a genetic map of one individual. Later, they will fill in the gaps and differences among a diversity of individuals.

Because the information is incomplete, opportunities exist in discovering new knowledge about how genes work-functional genomics-not in the database itself. The database only stores the data of what the genes are, not tIiemowledge of how they work. The human genome databastis similar to a directory to the homes of the movie stars that includes the addresses and phone numbers, but not the names. The names, like gene functions, are what make the directory and the database useful.

For that knowledge, biotech organizations are creating data warehouses that include other parameters along with the genetic database. And as complex as the genetic database is, the surrounding parameters are orders of magnitude more complicated. To collect the new knowledge about these parameters, biotech companies will build a data warehouse that scientists can use for knowledge discovery.

IT experts in the biotech industry will build a data warehouse that ideally contains DNA sequences and genetic information such as gene locations and functions, when genes are active, and genes' metabolic environment across different species. Then the IT expert will define metadata that provides information about the data in the databases that biotech researchers can exploit for knowledge discovery.

One potential solution is the use of a variable pattern match. Variable matches can range from simple wildcard searches that match any character or element at a specific location in a search string to complex routines that match elements based on the coexistence *of* other predefined elements. A complex routine might allow a search request such as "find all street comers that have traffic lights and do not allow right-on-red turns." With the appropriate data representation, a similar genetic pattern match can discover functional analogs among different genes from different species.

One pattern matching method relates to how a spelling checker works. The old spelling checkers used a method called the Soundex Algorithm that mapped sounds, such as hard k to letter combinations like k, c, ch, qu, and so forth. The checker maps the letters *of* a misspelled word to various combinations ofletters with the same sound value, matches those to a list *of* words in a dictionary, and then provides the resulting list back to the user. A similar routine for genetic pattern matching to the genetic equivalent ofletter combinations can map the various DNA sequences.

Imagine being able to use genetic information to design a genetic treatment for cancer or heart disease that is specific to you, or gaining immunity from diseases by eating a potato or banana rather than undergoing a series *of* shots. Genetic knowledge discovery will support these developments and others. With these goals in mind, knowledge-management techniques can create a general, genetic data warehouse model for medical, pharmaceutical, and agricultural knowledge discovery.

1.5.3.5 Problems in Data Mining and Machine Learning

The topics addressed by data mining are ones that statisticians and applied mathematicians have worked on for decades. Consequently, the division between statistics and data mining is blurry at best. If you do work with data mining or machine learning techniques, you will want to have more than a passing familiarity with traditional statistical techniques. If your problem can be solved by the latest data-mining algorithm or a straightforward statistical calculation, you would do well to choose the simple calculation. By the same token, please avoid the temptation to devise your own scoring method without first consulting a statistics book to see if an appropriate measure already exists. In both cases, it will be easier to debug and easier to explain your choice of a standard method over a nonstandard one to your colleagues.

Supervised and unsupervised learning

Machine learning methods can be broadly can be divided into supervised and unsupervised learning. Learning is said to be supervised when a learning algorithm is given a set of labeled examples from which to learn (the training set) and is then tested on a set of unlabeled examples (the test set). Unsupervised learning is performed when data is available, but the correct labels for each example aren't known. The objective of running the learning algorithm on the data is to find some patterns or trends that will aid in understanding the data. For example, the MEME program introduced in Chapter 8, Multiple Sequence Alignment, Trees, and Profiles, performs unsupervised learning in order to find sequence motifs in a set of unaligned sequences. It isn't known ahead of time whether each sequence contains the pattern, where the pattern is, or what the pattern looks like.

Cluster analysis is another kind of unsupervised learning that has received some attention in the analysis of microarray data. Clustering, as shown in Figure 14-5, is the procedure of classifying data such that similar items end up in the same class while dissimilar items don't, when the actual classes aren't known ahead of time. It is standard technique for working with multidimensional data. Figure 14-5 shows two panels with unadorned dots on the left and dots surrounded by cluster boundaries on the right.

1.5.3.6 A Collection of Data Mining Techniques

In this section, we describe some data mining methods commonly reported in the bioinformatics literature. The purpose of this section is to provide an executive summary of the complex tricks for data analysis. You aren't expected to be able to implement these algorithms in your programming language of choice. However, if you see any of these methods used to analyze data in a paper, you should be able to recognize the method and, if necessary, evaluate the way in which it was applied. Like any technique in experimental biology, it is important to have an understanding of the machine learning methods used in computational biology to know whether or not they have been used appropriately and correctly.

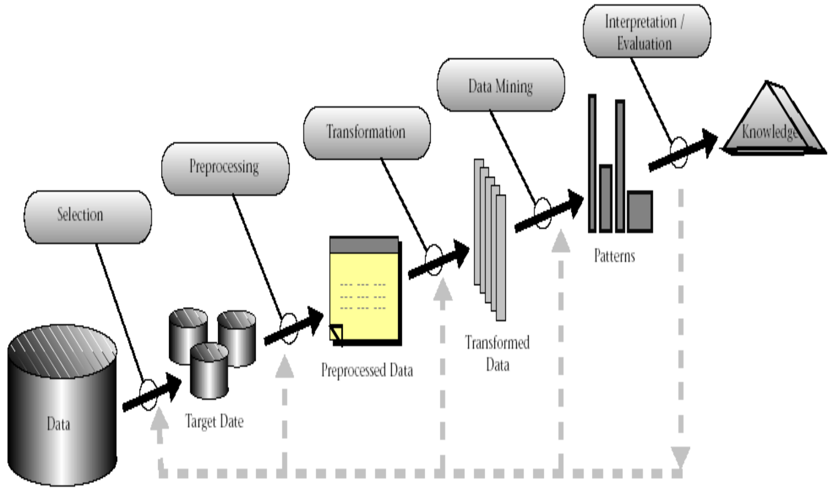


Figure 1. An Overview of the Steps That Compose the KDD Process.

Decision trees

In its simplest form, a decision tree is a list of questions with yes or no answers, hierarchically arranged, that lead to a decision. For instance, to determine whether a stretch of DNA is a gene, we might have a tree like the one shown in Figure 14-6.

A tree like this one is easy to work through, since it has a finite number of possibilities at each branch, and any path through the tree leads to a decision. The structure of the tree and the rules at each of the branches are determined from the data by a learning algorithm. Techniques for learning decision trees were described by Leo Breiman and coworkers in the early 1980s, and were later popularized in the machine learning community by J. R. Quinlan, whose freely available C4.5 decision tree software and its commercial successor, C5, are standards in the field.

One major advantage of decision trees over other machine learning techniques is that they produce models that can be interpreted by humans. This is an important feature, because a human expert can look at a set of rules learned by a decision

Tree and determine whether the learned model is plausible given real-world constraints. * In biology, tree classifiers tend to be used in pattern recognition problems, such as finding gene splice sites or identifying new occurrences of a protein family member. The MORGAN gene finder development by Steven

Salzberg and coworkers in an example of a decision tree approach to gene finding.

1.5.3.7 Neural networks

Neural networks are statistical models used in pattern recognition and classification. Originally development in the 1940s as a mathematical model of memory, neural networks are sometimes also called connectionist models because of their representation as nodes (which are usually variables) connected by weighted functions. Figure 14-7 shows the process by which a neural networks is constructed. Please note, though, that there is nothing particularly “neural” about these models, nor are there actually physical nodes and connections involved. The idea beyond neural networks is that, by working in concert, these simple processing elements can perform more complex computations.

A neural network is composed of a set of nodes that are connected in a defined topology, where each node has input and output connections to other nodes. In general, a neural network will receive an input pattern (for example, an amino acid

* The canonical decision-tree urban legend comes from an application of trees by a long-distance telephone company that wanted to learn about churn, the process of losing to other customers to other long-distance companies. They discovered that an abnormally large number of their customers over the age of 70 were subject to churn. A human recognized something the program did not: humans can die of old age. So, being able to interpret your results can be useful.

1.5.3.8 Data Mining and KDD

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term *data mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. The phrase *knowledge discovery in databases* was coined at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991) to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the AI and machine-learning fields.

In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data. The distinction between the KDD process and the data-mining step (within the process) is a central point of this article. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data. Blind application of data-mining methods (rightly criticized as data dredging in the statistical literature) can be a dangerous activity, easily leading to the discovery of meaningless and invalid patterns.

Summary

One of the most exciting areas of modern biology is the application of data mining methods to biological databases. We can distinguish two types of goals of KDD systems: verification and discovery. With verification, the system is limited to verifying the users hypothesis. With discovery, the system autonomously finds new patterns. The topics addressed by data mining are ones that statisticians and applied mathematicians have worked on for decades. In its simplest form, a decision tree is a list of questions with yes or no answers, hierarchically arranged, that lead to a decision.

Model Questions

1. Briefly explain the Knowledge based data analysis and its applications in bioinformatics
2. Write a short note on datamining techniques?

References

1. From Data Mining to Knowledge Discovery in Databases by Usama Fayyad
2. Introduction to Bioinformatics by Arthur M. Lesk.

B.M.REDDY **M.Tech. (HBTI, Kanpur)**

Lesson – 1.5.4

APPLICATIONS OF DRUG DISCOVERY

1.5.4.1 INTRODUCTION

The drug industry is one of the major players involved in the development of the field of bioinformatics. Many pharmaceutical companies have internal teams conducting bioinformatics research. The objective of these teams is commercial. The main purpose is to beat the competition to solutions of a problem that may give their company that crucial edge in producing the next major drug.

Most of the drugs are small molecules that are designed to bind, interact, and modulate the activity of specific biological receptors. Receptors are proteins that bind and interact with other molecules to perform the numerous functions required for the maintenance of life. Receptors include an immense array of cell-surface receptors (hormone receptors, cell-signaling receptors, neurotransmitter receptors, etc.), enzymes, and other functional proteins. Due to genetic abnormalities, physiologic stressors, or some combination thereof, the function of specific receptors and enzymes may become altered to the point that our well being is diminished. These alterations may manifest as minor physical symptoms, as in the case of a running nose due to allergies, or as life threatening and debilitating events, such as sepsis or depression. The role of drugs is to correct the functioning of these receptors.

1.5.4.2 Drug Discovery-Historical Perspective

Drug research, began as an extension of chemistry about a century ago, By 1870, many of the basic chemical principles had been elucidated-Avogadro's atomic hypothesis, a periodic table of elements and the theory of acids and bases.

Penicillin was discovered in 1929 by Alexander Fleming and a large number of antibiotics were studied between 1877 and 1939. In 1938, E. Chain, Howard Florey et al selected penicillin for further study and made it one of the most effective antibiotics. After the discovery of penicillin and subsequently of other antibiotics, many drug companies established departments of microbiology and fermentation units, which added to their technological scope.

Biochemistry has influenced drug research in numerous ways. The most important concepts introduced by biochemistry have been of enzymes and receptors, which have been empirically found to be useful drug targets.

The description and characterization of carboanhydrase in 1933 was followed by the discovery that sulphanilamide, the active metabolite of the sulphonamide (sulpha drug) Prontosil, inhibited this enzyme and that this effect led to an increase in natriuresis and the excretion of water.

Sulphanilamide gave impetus to develop batter carboanhydrase, inhibitors such as acetazolamide. This led to development of more effective diuretics such as hydrochloro-thiazide and furosemide. There are structural genealogies that link sulphanilamide with more advanced sulphonamides like sulphathiazole, with sulphonylureas like tolbutamide, used in the treatment of diabetes mellitus type II, and with diuretics that are used to treat edema and essential hypertension.

J.N. Langley and Luchsinger described the mutual antagonism between the actions of agents on biological responses and further developed Ehlrich's theory. This was due to the formation of complexes with a "receptive substance" where the two antagonistic agents elicit distinct effects. The concept of antagonism has been an important part of drug discovery programs. As an example, a physiological antagonist can be used to elicit a biological response in certain types of tissues. The addition of a range of compound that is possibly antagonistic in the effects of the physiological antagonist can then be added to the test system. If a compound is antagonistic, there is the potential to develop that compound for clinical use. This type of drug screen has been used for the development of new pharmacological agents.

R.P. Ahlquist proposed the existence of two types adrenergic receptors. The pharmacological characterization of receptors in various organs provided the basis for a large number of very diverse drugs: β-blockers; β-antagonist; benzondiazepines, which enhance the effects of γ-aminobutyric acid and chloride flux by way of the benzo-diazepine receptor; and monoclonal antibodies, which block receptors of growth or differentiation factors on tumor cells.

1.5.4.3 Types of Drugs

The top 10 therapy classes accounted for 31% of the total audited world market in 2002 (Table 1). Antiulcerants were the largest class, followed by cholesterol and triglyceride reduces. The systemic antihistamine class of drugs dropped out of the top 10 in 2002, to be replaced by the erythroietins, which debuted at number seven. Products such as Epogen, Aranesp and Erypo/procrit dominate this class.

The maximum selling drug in 2000 was prilosec, which earned $4.102 billion in sales. Prilosec is used to treat stomach ulcers and gastro-esophageal acid reflux. Prilosec targets the proton pump, which is located in the acid producing cells lining the stomach wall. Proton pump is responsible for the production of stomach acid. Due to genetic reasons, such as deficient enzymes that regulate acid secretion, or physiologic causes, such as stress or heavy food, excess acid may be produced. This leads to ulceration of the stomach lining or acid reflux and heart burn.

1.5.4.4 IMPORTANT PARAMETERS IN DRUG DISCOVERY

The development of any potential drug begins with years of scientific study to determine the biochemistry behind a medical problem for which pharmaceutical intervention is possible. The result is the determination of specific receptor targets that must be modulated to alter their activity in some way. Once these targets have been identified, the goal is then to find compounds that will interact with the receptors in some fashion. At this initial stage of drug development, it does not matter what effect the compounds have on the targets. We simply wish to find anything that binds to the receptor in any fashion.

The modern day drug discovery pipeline is outlined in Figure 14.4. The first step is to determine an assay for the receptor. An assay is a chemical or biological test that turns positive when a suitable binding agent interacts with the receptor. Usually, this test is some form of colorimetric assay, in which an indicator turns a specific color when complementary ligands are present. This essay is then used in *mass screening* , a technique whereby hundreds of thousands of compounds can be tested in a matter of days to weeks. A pharmaceutical company will first screen their enter corporate database of known compounds. The reason is that if a successful match is found, the database compound is usually very well characterized. Furthermore, synthetic methods will be known for this compound, and patent protection is often present. This enables the company to rapidly prototype a candidate ligand whose chemistry is well known and within the intellectual property of the company.

Drug Discovery Pipeline

The process of drug development has evolved into an extremely complex procedure. According to estimates, the average drug takes 12 years and $270 million from initial discovery to public usage in the USA. For every drug that is deemed marketable by the FDA, thousands of others are considered either unsafe or ineffective clinically. Beginning with preclinical research, new chemical entities (NCEs) are discovered in laboratories and tested in animals for safely and biological activity. If a compound is thought to be safe and effective as a chemical agent, a pharmaceutical company then submits an investigational new drug application (NDA) to the FDA. Once approved for clinical studies, a three-phase process begins where safety and efficacy are continually assessed with increased scrutiny and an increasing patient population. Approximately 70% of drugs entering clinical trails complete Phase I, 33% complete Phase II, and 27% complete Phase III. After Phase III is completed a company then submits an NDA to the FDA. Those drugs that are approved for marketing comprise an extremely small percentage of new chemical entities (NCEs) that are tested. From thousands of prospective molecules, only a harmful of drugs undergo clinical studies, and even fewer receive market approval.

Process of Drug Discovery

In the context of drug discovery and design, there are five major areas of interest. The five areas are given below.

- Target Identification
- Target Validation
- Lead Identification
- Lead Optimization
- Preclinical Pharmacology and Toxicology.

Target identification

Conventional drug discovery process focuses on a known pathological phenomenon and then develops a therapy to combat it. The process of therapy is chemistry-based, i.e. you need to produce compounds for screening. The approaches of identifying targets include protein expression, protein biochemistry, structure function studies, study of biochemical pathways, etc.

There are now several other methods to identify specific molecular targets like high throughput sequencing analysis, positional cloning, generation of cDNA libraries with ESTs and database mining by sequence homology. These high-throughput technologies and the knowledge of the human genome map generate a very large number of prospective targets. The problem is now of plenty. It is important to determine whether these novel targets are actually relevant to the physiology of the diseases.

Target validation

As there are a plethora of new potential therapeutic drug targets that are being discovered, selection and validation of novel molecular targets has become important. It needs to be confirmed that the targets identified will affect an appropriate biological response.

Chemogenomics and chemical genetics are likely to provide more small molecule drug candidates. Chemogenomics has been defined as the discovery and description of all possible drugs to all possible drug targets. Chemical genetics involves the use of defined chemical probes to understand some specific features of biology and can be viewed as a subset of chemogenomics.

Targeted gene disruption (TGD) is a term that refers to several different methods of target validation. TGD relates to the production of knockout or transgenic animals to study the effect of removing a particular gene coding for the putative molecular target.

This approach assumes that since the system under study is an intact higher organism, it may correlate well with a similar disruption in the intact human. There are limitations of this assumptions as there are compensatory mechanisms in any organism that may invalidate the intended result of the knockout. Also, TGD methods are time-consuming methods.

Some of the technologies used in the lead identification are:

- Virtual screening
- Chemoinformatics
- Quantitative Structure Activity Relationship (QSAR)
- High throughput docking
- NMR-based screening
- Chemical genetics

Virtual screening

Virtual screening (VS) is part of chemoinformatics. It involves protein-structure-based compound screening or docking and chemical-similarity search based on small molecules. VS technologies are used in high throughput docking, homology searching and pharmacophore searchers of 3-D databases.

Some important features to consider when developing a VS system are: knowledge about the compounds that you may screen against your receptor, knowledge about the receptor structure and receptors-ligand interactions in general and standard knowledge about drugs and drug characteristics.

It is highly desirable to design a focused virtual library that contains synthesizable and drug-like compounds. There is also the concept of multilevel compatibility (MLCC) scoring, which is used to measure drug-like characteristics. MLCC is a systematic comparison of the local environments within a compound and those within existing commercial drugs was applied to focus test sets: top selling drugs, compounds under biological scrutiny prior to preclinical testing, anticancer drugs, and compounds known have poor drug-like character.

Chemoinformatics

Chemoinformatics combines elements of biology and chemistry with mathematics, statistics and computer sciences. Analysis in chemoinformatics focuses on several types of large datasets available such as macromolecular structures, 3-D chemical database and compound libraries. Combinatorial chemistry and HTS depend on chemoinformatics include new molecular descriptors and pharmacophore mappings techniques, statistical tools and novel visualization methods.

Pharmacophore mapping

The pharmacophore search is an approach to identify lead compounds against a desired target. A pharmacophore is the specific 3-D arrangement of functional groups within a molecular framework that are necessary to bind to a macromolecule and/or an enzyme active site. The identification of a pharmacophore is an important step in understanding the interaction between a receptor and a ligand.

If you are able to establish a pharmacophore, you have the opportunity to search databases and identify novel compounds that fit the pharmocophore model. The search algorithms are sophisticated and can be effectively used to identify and optimize leads, focus combinatorial libraries and assist in virtual HTS. Filtering can be used to partition large library into trial sets of pharmacophores.

Quantitative structure activity relationship (QSAR)

QSAR analysis refers to methods that relate structural features of molecules to biological activity in quantitative terms. In most cases, QSAR analysis tries to establish linear relationships between selected structural features in a series of related molecules and their known level of activity. Models derived from training sets can be applied to predict molecules with higher potency. QSAR is a valuable timesaving alternative to labour intensive approach.

QSAR analysis has evolved to translate pharmacophore information into QSAR models. These QSAR models can be used as virtual HTS activity profiling of a library. There have been attempts to generate a general concept of descriptor pharmacophore, which uses variable selection QSAR as a subset of molecular descriptors that give the most statistically significant structure-activity correlation. These methods include partial least squares and K-nearest neighbors.

Hence, similarity searchers using descriptor pharmacophores is done to mine chemical databases or virtual libraries to discover compounds with a desired biological activity.

High-throughput docking

Docking refers to the ability to position a ligand in the active or a designed site of a protein and calculate specific binding affinities. Ligand-protein docking has evolved so that docking single or multiple small molecules to a receptor site is now routinely used to identify ligands. Optimal docking procedures need to be fast and accurate. They should also be able to generate reliable ligand geometries, score the ligand conformation correctly, and estimate the binding energy.

Docking algorithms can be used to find ligands and binding conformations at a receptor site close to experimentally determined structures. Docking algorithms are also used to identify multiple proteins to which a small molecule can bind. This approach can be used to predict either unknown and secondary therapeutic target proteins or side effects and toxicity of particular drugs.

The evaluation of scoring functions is the key to computational structure-based drug design. There are several approaches to improve their reliability and accuracy. The three major families of scoring functions are: force-field-based, knowledge-based, and empirical. The Ligand-Protein DataBase (LPDB) (http://Ipdb.scripps.edu/) has data on protein complexes with both high-resolution structure and known experimental binding affinity.

Nuclear magnetic resonance (NMR)-based screening

NMR spectroscopy, used for compound identification and conformation analysis, has long been used in drug discovery and design. NMR is used to determine the 3-Ddisposition of potential drug candidates (small organic molecules) and to reveal the tertiary structures of the biomacromolecules (proteins and DNA) that interact or are inhibited by the drug entity. NMR-based methods have been used to screen-molecule binding to proteins without any prior information on the function of the target.

NMR can be used to screen very weak binders as the global changes in the NMR events that are perturbed when small molecule binds to a macromolecule can be detected by observing either the ligand or the receptor. Changes in properties like molecular diffusion can be detected when binding events occur even if the binding constants are in the millimolar range. Since small and large molecules have very different diffusion or relaxation properties, the use of special experiments allow the determination of these differences between the free ligand and the protein-ligand complex (bound form).

There are various modifications of NMR-based techniques like diffusion ordered spectroscopy, saturation transfer differences, NOE pumping and SAR by NMR. The SAR by NMR technique was the technique that proved that detection of chemical shift changes in 2-D HSQC 15N-1H spectra could guide the design of small molecule with enhanced binding to the target protein. This method has been extended through the use of CryoProbe technology to screen up to 200,000 compounds per month in mixtures of 100 entities per experiment.

Another modified NMR technique is the SHAPES strategy. This strategy employs a limited but diverse library of fragments from known drugs or compounds with drug-like properties along with those from protein binding molecules that are screened for binding a target (target that are generally too large to analyze structurally by NMR). Weak binding “shapes” are screened by 1-D line broadening or 2-D transferred NOE measurements and used as lead scaffolds in library design and high throughput screening. WaterLogsy technique, which uses the bulk water on a protein surface as the magnetization to be transferred to the binding ligand has also been employed to screen SHAPES libraries.

Chemical genetics

Chemical genetics is the ‘study of gene-product function in a cellular or organismal context using exogenous ligands’. The method involves perturbing biological systems with small molecules. This approach is used to identify compounds that may “induce a specific cellular state”. The main objective is to discover compounds that may act as “knockouts”, i.e. inactive a gene product (protein) akin to using mutant mouse models, and be able to study the kinetic effects of the particular gene inactivation within the organism.

The chemical genetics approach has worked with selected systems to identify small molecule modulators of cellular function. For example, it has been used in the discovery of the cell cycle-arresting agent monastrol, an agent that halts cells in mitosis with monopolar spindles. It was demonstrated that this molecule inhibits the motility of the kinesin motor protein Eg5, a protein necessary for spindle bipolarity.

Lead optimization

Various techniques that are used in the lead discovery phase also play key roles in the optimization of the newly found lead. Once a lead compound is established in the identification process, you need to optimize the desirable traits of the lead.

You can use structure-activity relationship (SAR) and quantitative SAR (QSAR) for exploring various options. Computer-aided drug design (CADD) or structure-based drug design (SBDD) is now extensively used for drug candidate optimization.

You need to have a broad knowledge for de nova drug design. There are many tools for characterization of binding sites: Calculation of charge distribution, lipophilicity or pka of side-chain functionalities and identification of H-bond donors and acceptors. In addition, docking programs are used in conjunction with large 3-D databases of small molecule structures and the scoring algorithms that attempts to predict the binding affinity of designed ligands. To be considered for further development, lead structures should be amenable for chemistry optimization and have good ADME properties.

Structure-based drug design (SBDD)

Structure-based in a very powerful approach in drug design and is most effective when the 3-D structure of an existing inhibitor complex with its target is known. This technique has played a major role in designing a number of drug candidates that have progressed to clinical trails. A requirement for this approach is an understanding of the principles of molecular recognition in protein-ligand complexes. SBDD is an iterative approach.

Predicting drug-like properties

The phrase “drug-like” is defined as those compounds that have sufficiently acceptable ADME and toxicity properties to survive through the completion of Phase I clinical trails. To build a drug-like database, it is essential to apply a variety of filters to remove useless compounds such as those that contain reactive groups and exhibits false positive in a majority of assays. Future filtering will depend on the type of target and project, where for example bioavailability, pharmacokinetics (PK), or CNS penetration may dictate the requirements for the target "drug-like" molecules.

Pre-clinical pharmacology and toxicology

Prior to clinical trails in human, each new chemical entity has to be tested in animals and in many cases, several species. Data concerning toxicity, PK and metabolism in necessary to determine the feasibility and safety of drug in human. In some cases testing may include xenograft models (animals with transplanted human cancers or other tissues are called xenograft models) and a complete toxicology profile should be clearly established at this stage. A careful study of ADME/T characteristics at this phase of design is extremely important since the majority of drug candidates fail clinical trails due to ADME/T deficiency. Clearly, the benefits of enhancing the ADME/T properties of molecules through computational design in the discovery phase and actual validation of these properties in several species of animals in the pre-clinical phase are enormous.

1.5.4.5 AREAS INFLUENCING DRUG DISCOVERY

There are several approaches to discover new drugs. They range from molecular biology to combinatorial chemistry. Following are some of the approaches.

The Influence of Molecular biology on Drug Discovery

Molecular biology has exerted a profound influence on drug discovery. It applies the concept of genetic information to biochemical and chemical pathways. The protein drugs, largely recombinant proteins and monoclonal antibodies are referred to as "biotech" drugs. In 1998, biotech products, accounted for one-fourth of the drugs introduced worldwide.

Monoclonal antibodies are a specialized form of recombinant protein. They are produced in three different ways:

1. They can be generated as mouse antibiotics that are later "humanized" by recombination with human antibody genes.

2. Human antibodies can be directly raised in nude mice grafted with human immune cells.

3. Antibodies can also be made by phage display techniques. Large libraries of human antibody genes in phages allow the production and subsequent optimization of a wide array of antibodies.

Antibodies may be more attractive from a therapeutic point of view than recombinant cytokines or chemokines because they can be targeted to very specific structures with high precision. Cytokines are a family of growth factors. Secreted primarily from leukocytes, cytokines stimulate the humoral and cellular immune responses, as well as the activation of phagocytic cells. Cytokines that are secreted from lymphocytes are termed lymphokines, whereas those secreted by monocytes or macrophages are termed monokines. Various cells of the body produce a large family of cytokines. Many of the lymphokines are also known as interleukins (ILs), since they are not only secreted by leukocytes but also able to affect the cellular responses of leukocytes. IL-8 is an interleukin that belongs to the family of proteins that exert chemoattractant activity to leukocytes and fibroblasts. This family of proteins is termed the chemokines. IL-8 is produced by monocytes, neutrophils, and NK cells and is chemoattractant for neutrophils, basophils and T-cells. In addition, IL-8 activities neutrophils to degranulate.

Molecular biology can also help us to understand disease processes at the molecular (genetic) level and to determine the optimal molecular targets for drug intervention. Current drug therapy is based on less than 500 molecular targets.

High-throughout Screening (HTS)

Another method of drug discovery is based on cell-based assays, and automated high-throughout screening (HTS). In this method, large numbers of hypothetical targets are incorporated into in vitro or cell-based assays and exposed to large numbers of compounds variations on a greater number of themes in high-throughout configurations.

This experimental design can help in identifying many substances, which can modify the targets in question. Many such “hits”-compounds that elicit a positive response in a particular assay- would then give rise to more leads, i.e. compounds that continue to show the initial positive response in more complex models (cells, animals) in a dose-dependent manner. Eventually, the number of compounds also would increase.

Combinatorial chemistry and HTS

Most recent attempts towards the design of combinatorial libraries have been driven by the intent to generate a high degree of structural diversity within a library. It is, however by no means certain to what extent molecular diversity as viewed by chemists and as calculated by structural descriptors resembles diversity as “seen” by a biological target molecule.

It has been shown that a protein can bind a set of structurally diverse molecules with very similar affinities in the nanomolar, whereas a number of analogs closely related to one of the good binders display only weak affinities. The design and sampling of compound libraries should be guided not only by structural descriptors, but also by biological activity descriptors. Screening all compounds in a library against a set of functionally dissimilar protein and determining the binding affinity of each compound for all proteins can achieve this. The set of binding affinities for a given compound is termed its affinity fingerprint. TRAP (Target Related Affinity Profiling) is an operational method of molecular diversity in which a compound is described by a unique affinity fingerprint that is determined by the compound's ability to bind a select panel of proteins. The similarity of affinity fingerprints has been shown to correlate with the biological activities of drug-like substances.

Pharmacogenomics and Pharmacogenetics

Pharmacogenetics is defined as the study of the hereditary basis for differences in a population. It is useful in understanding the response to a drug by an individual. The DNA sequences of human genomes are not identical and vary from individual to individual. Pharmacogenetics is hence based on the identification of genetic variations (polymorphisms) that alter drug concentrations and responses. Same medication dose gives varying responses in different individuals because of genetic and protein differences. The related and slightly broader term is pharmacogenomics-which is used to describe the application of genomic technology in drug development.

The first demonstrated pharmacogenetic trait was chemical insensitivity to phenylthiourea (PTU). Individuals with PTU insensitivity could not taste the chemical. It was the first chemical insensitivity shown to be heritable.

The Cytochrome P450 (CYP) family of enzymes is involved in metabolism of several drugs. For example, the CYP2D6 enzyme is involved in the metabolism many common drugs. For example, Ritonivar, is a protease inhibitor used in the treatment to HIV that is metabolized by the CYP2D6 enzyme. Other examples include codeine and other opiate derivatives, DXM, an ingredient found in many cough medications and Prozac. The CYP2D6 enzyme affects beta-blockers used to treat hypertension, and drugs often prescribed to control heart problems.

Some of the variants of CYP2D6 are slow metabolizers while others are very fast. Slow metabolizers may be exposed to the active product for longer than ultra-rapid metabolizers, or have greatly diminished exposure to the active metabolite. Codeine does not work in about 7 percent of Caucasians because their bodies lack the enzyme and fails to break down the drug. CYP2D6 is lacking because some individuals carry a gene variation that prevents their bodies from making it and hence some patients get no pain relief from codeine.

Pharmacogenetic technology

Pharmacogenetics makes extensive use of the automated tools for gene and protein sequencing already discussed in earlier chapters. The determination of the nucleotide and amino acid sequence of genes and proteins is used for analysis of genetic differences at an individual level. There are two main strategies used in screening for polymorphism in an individual: phenotyping and genotyping.

Phenotyping helps determination of the presence and activity of a particular metabolic enzyme in a tissue biopsy (known as functional phenotyping). Metabolic phenotyping is used to measure the level of metabolites in a person in post-administration of a drug. Phenotyping is invasive and potentially dangerous because of the administration of drugs and their resulting side effects.

Genotyping helps determination of the specific genetic code of an individual. Genotyping is non-invasive and can be done using a tissue sample. However, the results may be ambiguous and open to several equally possible interpretations. There are various technologies being used for genotyping.

SUMMARY

This chapter has demonstrated that the scope of bioinformatics is vast and provides an opportunity for development of new drugs. From the historical perspective, it is known that many drugs in the past were either experimentally development or accidentally discovered. However, by using the power of informatics, a wealth of genomic data is being generated that can be used to identify and select suitable targets for new drugs, i.e. tailored drugs to both individual diseases and individual patients. Binding of small molecules or ligands to a receptor is an important first step in the drug discovery process. A number of screening methods are available, but screening in combination with NMR techniques has often proved fruitful, an approach recently developed during the last two years. By combining bioinformatics methods in target selection, chemo informatics methods in the selection of screening candidates, and sophisticated screening methods, there is a good prospect that we may have tailored drugs for individual diseases.

Model Questions

1. Drug discovery process is fundamentally an iterative process. Comment.
2. What high-throughput screening methods are employed in screening drugs?
3. How can you use pharmacogenetics to determine genetics differences among individuals?
4. Explain the steps that a new drug has to pass through before reaching the clinical trial stage.

References :

1.Pharmacognacy by kokata

2. Injectable Drug Development: Techniques to Reduce Pain and Irritationedited by Pramod Gupta, Gayle Brazeau - Medical – 1999

3.Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery by Jurgen Bajorath - Medical - 2004 - 544 pages.

B.M.REDDY M.Tech. (HBTI, Kanpur)