MOLECULAR BIOLOGY AND APPLIED GENETICS

FOR

Medical Laboratory Technology Students

Upgraded

Lecture Note Series

Mohammed Awole Adem

Jimma University

MOLECULAR BIOLOGY AND APPLIED GENETICS

For

Medical Laboratory Technician Students

Lecture Note Series

Mohammed Awole Adem

Upgraded - 2006

In collaboration with
The Carter Center (EPHTI) and The Federal Democratic Republic of Ethiopia Ministry of Education and Ministry of Health

Jimma University

PREFACE

The problem faced today in the learning and teaching of Applied Genetics and Molecular Biology for laboratory technologists in universities, colleges andhealth institutions primarily from the unavailability of textbooks that focus on the needs of Ethiopian students.

This lecture note has been prepared with the primary aim of alleviating the problems encountered in the teaching of Medical Applied Genetics and Molecular Biology course and in minimizing discrepancies prevailing among the different teaching and training health institutions. It can also be used in teaching any introductory course on medical Applied Genetics and Molecular Biology and as a reference material.

This lecture note is specifically designed for medical laboratory technologists, and includes only those areas of molecular cell biology and Applied Genetics relevant to degree-level understanding of modern laboratory technology. Since genetics is prerequisite course to molecular biology, the lecture note starts with Genetics

followed by Molecular Biology. It provides students with molecular background to enable them to understand and critically analyze recent advances in laboratory sciences.

Finally, it contains a glossary, which summarizes important terminologies used in the text. Each chapter begins by specific learning objectives and at the end of each chapter review questions are also included.

We welcoming the reviewers and users input regarding this edition so that future editions will be better.

ACKNOWLEDGEMENTS

I would like to acknowledge The Carter Center for its initiative, financial, material and logistic supports for the preparation of this teaching material. We are indebted to The Jimma University that support directly or indirectly for the visibility of this lecture note preparation.

I extend our appreciation to the reviewers of the manuscript during intra-workshop, Namely, Ato Tsehayneh Kelemu , Biochemistry Department, School of Medicine, and Ato Yared Alemu, School of Medical Laboratory Technology, Jimma University.We greatly appreciate them for their attitude, concern and dedication.

I also acknowledge all reviewers of the manuscript during inter-institutional workshop and those who participated as national reviewers.

Last but not least I would like to acknowledge tyhose who helped me directly or indirectly.

TABLE OF CONTENTS

List of Figures

INTRODUCTION

Molecular genetics, or molecular biology, is the study of the biochemical mechanisms of inheritance. It is the study of the biochemical nature of the genetic material and its control of phenotype. It is the study of the connection between genotype and phenotype. The connection is a chemical one.

Control of phenotype is one of the two roles of DNA (transcription). You have already been exposed to the concept of the Central Dogma of Molecular Biology, i.e. that the connection between genotype and phenotype is DNA (genotype) to RNA to enzyme to cell chemistry to phenotype.

James Watson and Francis Crick received the 1953 Nobel Prize for their discovery of the structure of the DNA molecule. This is the second most important discovery in the history of biology, ranking just behind that of Charles Darwin. This discovery marked the beginning of an intense study of molecular biology, one that dominates modern biology and that will continue to do so into the foreseeable future. .

The essential characteristic of Molecular Genetics is that gene products are studied through the genes that encode them. This contrasts with a biochemical approach, in which the gene products themselves are purified and their activities studied in vitro.

Genetics tells that a gene product has a role in the process that are studying in vivo, but it doesn't necessarily tell how direct that role is. Biochemistry, by contrast, tells what a factor can do in vitro, but it doesn't necessarily mean that it does it in vivo.

The genetic and biochemical approaches tell you different things:

Genetics → has a role, but not how direct

Biochemistry → tells what a protein can do in vitro, but not whether it really does it in vivo

These approaches therefore tell different things**.** Both are needed and are equally valuable. When one can combine these approaches to figure out what a

gene/protein does, the resulting conclusions are much stronger than if one only use one of these strategies.

DEVELOPMENT OF GENETICS AND MOLECULAR BIOLOGY

1866- Genetics start to get attention when Mendel Experimented with green peas and publish his finding

1910- Morgan revealed that the units of heredity are contained with chromosome,

1944- It is confirmed through studies on the bacteria that it was DNA that carried the genetic information.

1953-Franklin and Wilkins study DNA by X-ray crystallography which subsequently lead to unrevealing the double helical structure of DNA by Watson and Crick

1960s- Smith demonstrate that the DNA can be cleaved by restriction enzymes

1966 -Gene transcription become reality

1975- Southern blot was invented

1977- DNA sequencing methodology discovered

1981-Genetic diagnosis of sickle cell disease was first shown to be feasible by kan and Chang

1985- PCR develop by Mullis an Co-workers

2001-Draft of Human genome sequence was revealed

CHAPTER ONE
THE CELL

Specific learning objectives

- ⇒ Identify an eukaryotic and prokaryotic cell
- ⇒ Describe chemical composition of the cell membrane
- ⇒ List the structure found in a membrane
- ⇒ Describe the role of each component found in cell membrane

1.0. Eukaryotic and Prokaryotic Cell

- ► Cells in our world come in two basic types, prokaryotic and eukaryotic. "Karyose" comes from a Greek word which means "kernel," as in a kernel of grain. In biology, one use this word root to refer to the nucleus of a cell. "Pro" means "before," and "eu" means "true," or "good."
- ► So "Prokaryotic" means "before a nucleus," and "eukaryotic" means "possessing a true nucleus."

► Prokaryotic cells have no nuclei, while eukaryotic cells do have true nuclei. This is far from the only difference between these two cell types, however. Here's a simple visual comparison between a prokaryotic cell and eukaryotic cell:

Fig. 1. Prokaryotic cell

► This particular eukaryotic cell happens to be an animal cell, but the cells of plants, fungi and protists are also eukaryotic.

Fig. 2. Eukaryotic cell

- ► Despite their apparent differences, these two cell types have a lot in common. They perform most of the same kinds of functions, and in the same ways. These include:
  - Both are enclosed by plasma membranes, filled with cytoplasm, and loaded with small structures called ribosomes.
  - Both have DNA which carries the archived instructions for operating the cell.
  - And the similarities go far beyond the visible--physiologically they are very similar in many ways. For example, the DNA in the two cell types is precisely the same kind of DNA, and the genetic code for a prokaryotic cell is exactly the same genetic code used in eukaryotic cells.
- ► Some things which seem to be differences aren't. For example, the prokaryotic cell has a cell wall, and this animal cell does not. However, many kinds of eukaryotic cells do have cell walls.
- ► Despite all of these similarities, the differences are also clear. It's pretty obvious from these two little pictures that there are two general categories of difference between these two cell types: size and

complexity. Eukaryotic cells are much larger and much more complex than prokaryotic cells. These two observations are not unrelated to each other.

If we take a closer look at the comparison of these cells, we see the following differences:

1. Eukaryotic cells have a true nucleus, bound by a double membrane. Prokaryotic cells have no nucleus.
2. Eukaryotic DNA is complexed with proteins called "histones," and is organized into chromosomes; prokaryotic DNA is "naked," meaning that it has no histones associated with it, and it is not formed into chromosomes. A eukaryotic cell contains a number of chromosomes; a prokaryotic cell contains only one circular DNA molecule and a varied assortment of much smaller circlets of DNA called "plasmids." The smaller, simpler prokaryotic cell requires far fewer genes to operate than the eukaryotic cell.
3. Both cell types have many, many ribosomes, but the ribosomes of the eukaryotic cells are larger and more complex than those of the prokaryotic cell. A eukaryotic ribosome is composed of five kinds of

rRNA and about eighty kinds of proteins. Prokaryotic ribosomes are composed of only three kinds of rRNA and about fifty kinds of protein.

4. The cytoplasm of eukaryotic cells is filled with a large, complex collection of organelles, many of them enclosed in their own membranes; the prokaryotic cell contains no membrane-bound organelles which are independent of the plasma membrane.
5. One structure not shown in our prokaryotic cell is called a mesosome. Not all prokaryotic cells have these. The mesosome is an elaboration of the plasma membrane--a sort of rosette of ruffled membrane intruding into the cell.

1.1. Function of the cell

- ► Cell serves as the structural building block to form tissues and organ
- ► Each cell is functionally independent- it can live on its own under the right conditions:

- it can define its boundaries and protect itself from external changes causing internal changes
- it can use sugars to derive energy for different processes which keep it alive
- it contains all the information required for replicating itself and interacting with other cells in order to produce a multicellular organisms
- It is even possible to reproduce the entire plant from almost any single cell of the plant

► Cell wall

- protects and supports cell
- made from carbohydrates- cellulose and pectin- polysaccharides
- strong but leaky- lets water and chemicals pass through-analogous to a cardboard box

► Cell membrane

- membrane is made up from lipids - made from fatty acids water-repelling nature of fatty acids makes the diglycerides form a

- sheet or film which keeps water from moving past sheet (think of a film of oil on water)
- membrane is analogous to a balloon- the spherical sheet wraps around the cell and prevents water from the outside from mixing with water on the inside
- membrane is not strong, but is water-tight- lets things happen inside the cell that are different than what is happening outside the cell and so defines its boundaries. Certain gatekeeping proteins in the cell membrane will let things in and out.

► Cytosol - watery inside of cell composed of salts, proteins which act as enzyme

► Microtubules and microfilaments - cables made out of protein which stretch around the cell

- provide structure to the cell, like cables and posts on a suspension bridge
- provide a structure for moving cell components around the cell -sort of like a moving conveyer belt.

► Organelles - sub-compartments within the cell which provide different functions. Each organelle is surrounded by a membrane that makes it separate from the cytosol. These include nucleus, mitochondrion, vacuole, ribosome, endoplasmic reticulum, and golgi apparatus. (Refer any biology text book for detail)

1.2. The chemical components of cell membranes

The components cell membrane includes:

- Lipid -- cholesterol, phospholipid and sphingolipid
- Proteins
- Carbohydrate -- as glycoprotein

Differences in composition among membranes (*e.g.* myelin *vs.* inner mitochondrial membrane)

- Illustrate the variability of membrane structure.
- This is due to the differences in function. Example: Mitochondrial inner membrane has

high amounts of functional electron transport system proteins.

► Plasma membrane, with fewer functions (mainly ion transport), has less protein.

- Membranes with similar function (*i.e.* from the same organelle) are similar across species lines, but membranes with different function (*i.e.* from different organelles) may differ strikingly within a species.

► Carbohydrates of membranes are present attached to protein or lipid as glycoprotein or glycolipid.

1. Typical sugars in glycoproteins and glycolipids include glucose, galactose, mannose, fucose and the N-acetylated sugars like N-acetylglucosamine, N-acetylgalactosamine and N-acetylneuraminic acid (sialic acid).
2. Membrane sugars seem to be involved in identification and recognition.

1.3. Membrane structure

The amphipathic properties of the phosphoglycerides and sphingolipids are due to their structures.

1. The hydrophilic head bears electric charges contributed by the phosphate and by some of the bases.
    - These charges are responsible for the hydrophilicity.
    - Note that no lipid bears a positive charge. They are all negative or neutral. Thus membranes are negatively charged.

2. The long hydrocarbon chains of the acyl groups are hydrophobic, and tend to exclude water.
3. Phospholipids in an aqueous medium spontaneously aggregate into orderly arrays.
    - Micelles: orderly arrays of molecular dimensions. Note the hydrophilic heads oriented outward, and the hydrophobic acyl groups oriented inward. Micelles are important in lipid digestion; in the intestine they assist the body in assimilating lipids.

- Lipid bilayers can also form.
- Liposomes are structures related to micelles, but they are bilayers, with an internal compartment. Thus there are three regions associated with liposomes: -The exterior, the membrane itself and the inside.
- Liposomes can be made with specific substances dissolved in the interior compartment. These may serve as modes of delivery of these substances.

4. The properties of phospholipids determine the kinds of movement they can undergo in a bilayer.
   - Modes of movement that maintain the hydrophilic head in contact with the aqueous surroundings and the acyl groups in the interior are permitted.
   - Transverse movement from side to side of the bilayer (flip-flop) is relatively slow, and is not considered to occur significantly.

Review Questions

1. Compare and contrast eukaryotic and prokaryotic cell.
2. What are the chemical compositions of cell membrane?
3. Which chemical composition is found in high proportion?
4. What are the roles of membrane proteins?
5. What are the functions of a cell?

CHAPTER TWO

THE CELL CYCLE

Specific learning objectives

At the end of this Chapter students are expected to

- ⇒ Describe the components of cell cycle
- ⇒ List steps of cell cycle
- ⇒ Outline the steps of mitosis and meiosis
- ⇒ Distinguish the difference between mitosis and meiosis

2.0. Introduction

- o A eukaryotic cell cannot divide into two, the two into four, etc. unless two processes alternate:
  - doubling of its genome (DNA) in S phase (synthesis phase) of the cell cycle;
  - halving of that genome during mitosis (M phase).
- o The period between M and S is called G_1; that between S and M is G_2.

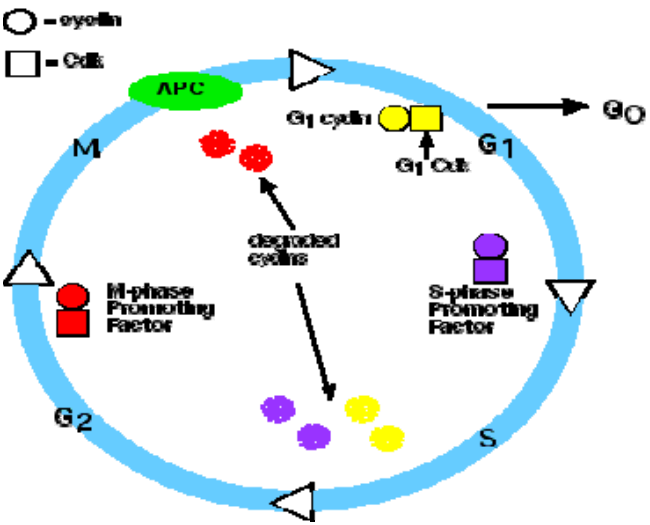


Fig. 3. The Cell Cycle

► So, the cell cycle consists of:

- **G_1** = growth and preparation of the chromosomes for replication
- **S** = synthesis of DNA (and centrosomes)
- **M** = mitosis

o When a cell is in any phase of the cell cycle other than mitosis, it is often said to be in interphase.

2.1. Control of the Cell Cycle

▪ The passage of a cell through the cell cycle is controlled by proteins in the cytoplasm. Among the main players in animal cells are:

- **Cyclins**
  - a G_1 cyclin (cyclin D)
  - S-phase cyclins (cyclins E and A)
  - mitotic cyclins (cyclins B and A)

  Their levels in the cell rise and fall with the stages of the cell cycle.

- Cyclin-dependent kinases (Cdks)

- a G_1 Cdk (Cdk4)
- an S-phase Cdk ((Cdk2)
- an M-phase Cdk (Cdk1)

Their levels in the cell remain fairly stable, but each must bind the appropriate cyclin (whose levels fluctuate) in order to be activated. They add phosphate groups to a variety of protein substrates that control processes in the cell cycle.

- The anaphase-promoting complex (APC). (The APC is also called the cyclosome, and the complex is ofen designated as the APC/C.) The APC/C
  - triggers the events leading to destruction of the cohesins thus allowing the sister chromatids to separate;
  - degrades the mitotic cyclin B.

2.2. Steps in the cycle

- A rising level of G_1-cyclins bind to their Cdks and signal the cell to prepare the chromosomes for replication.

- A rising level of S-phase promoting factor (SPF) — which includes cyclin A bound to Cdk2 — enters the nucleus and prepares the cell to duplicate its DNA (and its centrosomes).
- As DNA replication continues, cyclin E is destroyed, and the level of mitotic cyclins begins to rise (in G_2).
- M-phase promoting factor (the complex of mitotic cyclins with the M-phase Cdk) initiates
    - assembly of the mitotic spindle
    - breakdown of the nuclear envelope
    - condensation of the chromosomes

- These events take the cell to metaphase of mitosis.
- At this point, the M-phase promoting factor activates the anaphase-promoting complex (APC/C) which
    - allows the sister chromatids at the metaphase plate to separate and move to the poles (= anaphase), completing mitosis;
    - destroys cyclin B. It does this by attaching it to the protein ubiquitin which targets it for destruction by proteasomes.
    - turns on synthesis of G_1 cyclin for the next turn of the cycle;

- degrades geminin, a protein that has kept the freshly-synthesized DNA in S phase from being re-replicated before mitosis.

► This is only one mechanism by which the cell ensures that every portion of its genome is copied once — and only once — during S phase.

2.3. Meiosis and the Cell Cycle

► The special behavior of the chromosomes in meiosis I requires some special controls. Nonetheless, passage through the cell cycle in meiosis I (as well as meiosis II, which is essentially a mitotic division) uses many of the same players, e.g., MPF and APC. (In fact, MPF is also called maturation-promoting factor for its role in meiosis I and II of developing oocytes.

2.4. Quality Control of the Cell Cycle

► The cell has several systems for interrupting the cell cycle if something goes wrong.

- A check on completion of S phase. The cell seems to monitor the presence of the Okazaki fragments on the lagging strand during DNA replication. The cell is not permitted to proceed in the cell cycle until these have disappeared.
- DNA damage checkpoints. These sense DNA damage
  - before the cell enters S phase (a G_1 checkpoint);
  - during S phase, and
  - after DNA replication (a G_2 checkpoint).
- spindle checkpoints. Some of these that have been discovered
  - detect any failure of spindle fibers to attach to kinetochores and arrest the cell in metaphase (M checkpoint);
  - detect improper alignment of the spindle itself and block cytokinesis;
  - trigger apoptosis if the damage is irreparable.

- All the checkpoints examined require the services of a complex of proteins. Mutations in the genes encoding some of these have been associated with cancer; that is, they are oncogenes.
- This should not be surprising since checkpoint failures allow the cell to continue dividing despite damage to its integrity.

2.5. Regulation of the Cell Cycle

► Different types of cells divide at different rates. Skin cells divide frequently, whereas liver cells divide only in response to injury and nerve, muscle, and other specialized cells do not divide in mature humans.

1. The cell cycle control system consists of a molecular clock and a set of checkpoints that ensure that appropriate conditions have been met before the cycle advances.
2. For instance, cells must be in contact with adjacent cells before proper division can occur. Also, cells must reach a certain size and volume before they can properly divide. All of the DNA

must be properly replicated before the cell divides.

3. Checkpoints are present in the G_1, G_2, and M phases of the cell cycle. The G_1 checkpoint is the most critical one for many cells.
4. If the proper signals are not received, the cell may stay in a stage known as G_0; or the nondividing state.
5. Protein Kinases are enzymes that help synchronize the cell cycle events. Protein Kinases catalyze the transfer of a phosphate group from ATP to a target protein.
6. Phosphorylation induces a conformational change that either activates or inactivates a target protein.
7. Changes in these target proteins affect the progression through the cell cycle.
8. Cyclical changes in kinase activity, in turn, are controlled by proteins called Cyclins.
9. Protein kinases that regulate cell cycles are active only when attached to a particular Cyclin molecule.

10. Cyclin concentrations, in turn, vary throughout the cell cycle (they are highest as the cells prepare to divide). By the end of cytokinesis, cyclins are present in much smaller concentrations. The cyclins are broken down as the cells progress through the M-phase of cell division.
11. Cyclins bind with protein kinases early in the cell cycle and produce Mitosis Promoting Factor (MPF). MPF promotes chromosome condensation and nuclear membrane absorption.
12. Later in the cell cycle, MPF activates proteolytic enzymes (these enzymes break down proteins) which destroy the cyclin.
13. Thus, new Cyclin proteins must be produced during interphase, until appropriate levels build up and promote cell division.

► Certain Chemicals called Growth Factors have been isolated and are known to promote cell division as they bind to receptors of the plasma membrane. Platelet Derived Growth Factor is an example of one type of chemical signal. It may help cells to divide to heal wounds.

- If cells are too crowded, they will not divide under ordinary circumstances. Sufficient quantities of nutrients and growth factors may be lacking. Also, most cells must be adhered to an extracellular matrix in order to divide.
- Membrane proteins and cytoskeletal elements provide signals which indicate that proper anchorages exist.

2.6. Mitosis

- Mitosis is the process of separating the duplicates of each of the cell's chromosomes. It is usually followed by division of the cell.

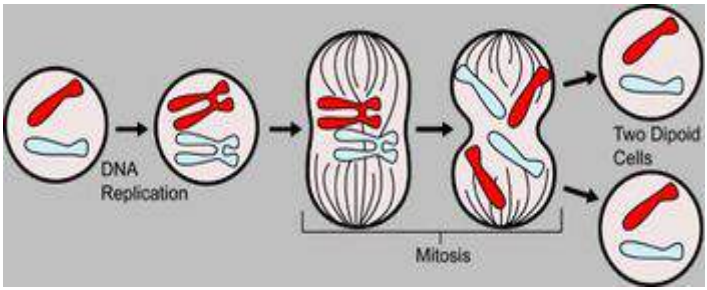


Fig. 4. Overview of Major events in Mitosis

► However, there are cases (cleavage in the insect embryo is an example) where the chromosomes undergo the mitotic process without division of the cell. Thus, a special term, cytokinesis, for the separation of a cell into two.

► When a eukaryotic cell divides into two, each daughter or progeny cell must receive
  - a complete set of genes (for diploid cells, this means 2 complete genomes, 2n)
  - a pair of centrioles (in animal cells)
  - some mitochondria and, in plant cells, chloroplasts as well
  - some ribosomes, a portion of the endoplasmic reticulum, and perhaps other organelles

o There are so many mitochondria and ribosomes in the cell that each daughter cell is usually assured of getting some. But ensuring that each daughter cell gets two (if diploid) of every gene in the cell requires the greatest precision.

1. Duplicate each chromosome during the S phase of the cell cycle.

2. This produces dyads, each made up of 2 identical sister chromatids. These are held together by a ring of proteins called cohesins.
3. Condense the chromosomes into a compact form. This requires ATP and a protein complex called condensin.
4. Separate the sister chromatids and
5. distribute these equally between the two daughter cells.

- Steps 3 - 5 are accomplished by mitosis. It distributes one of each duplicated chromosome (as well as one centriole) to each daughter cell. It is convenient to consider mitosis in 5 phases.
- When a cell is not engaged in mitosis (which is most of the time), it is said to be in interphase.
- These phases are as follows:

2.6.1. Prophase

- The two centrosomes of the cell, each with its pair of centrioles, move to opposite "poles" of the cell.
- The mitotic spindle forms. This is an array of spindle fibers, each containing ~20 microtubules.

- Microtubules are synthesized from tubulin monomers in the cytoplasm and grow out from each centrosome.
- The chromosomes become shorter and more compact**.**

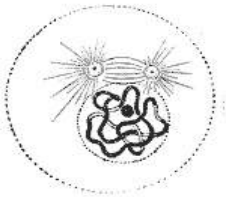


Fig. 5. Prophase: The two round objects above the nucleus are the centrosomes.

Note the condensed chromatin.

2.6.2. Prometaphase

- The nuclear envelope disintegrates because of the dissolution of the lamins that stabilize its inner membrane.
- A protein structure, the kinetochore, appears at the centromere of each chromatid.
- With the breakdown of the nuclear envelope, spindle fibers attach to the kinetochores as well as to the arms of the chromosomes.
- For each dyad, one of the kinetochores is attached to one pole, the second (or sister) chromatid to the opposite pole. Failure of a kinetochore to become attached to a spindle fiber interrupts the process.

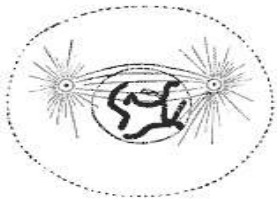


Fig. 6. Prometaphase: The nuclear membrane has degraded, and microtubules have invaded the nuclear space. These microtubules can attach to kinetochores or they can interact with opposing microtubules.

2.6.3. Metaphase

At metaphase all the dyads have reached an equilibrium position midway between the poles called the metaphase plate. The chromosomes are at their most compact at this time.

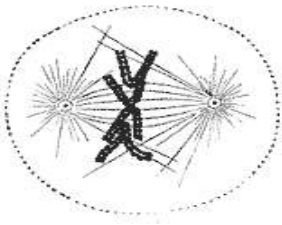


Fig.7. Metaphase: The chromosomes have aligned at the metaphase plate.

2.6.4. Anaphase

The sister kinetochores suddenly separate and each moves to its respective pole dragging its attached chromatid (chromosome) behind it. Separation of the sister chromatids depends on the breakdown of the cohesins that have been holding them together. It works like this.

- Cohesin breakdown is caused by a protease called separase (also known as separin).

- Separase is kept inactive until late metaphase by an inhibitory chaperone called securin.
- Anaphase begins when the anaphase promoting complex (APC) destroys securin (by tagging it for deposit in a proteasome) thus ending its inhibition of separase and allowing
- separase to break down the cohesins.

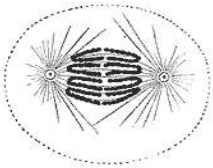


Fig. 8.Early anaphase: Kinetochore microtubules shorten.

2.6.5. Telophase

A nuclear envelope reforms around each cluster of chromosomes and these returns to their more extended form.

In animal cells, a belt of actin filaments forms around the perimeter of the cell, midway between the poles. As the belt tightens, the cell is pinched into two daughter cells.

In plant cells, a membrane-bounded cell plate forms where the metaphase plate had been. The cell plate, which is synthesized by the Golgi apparatus, supplies the plasma membrane that will separate the two daughter cells. Synthesis of a new cell wall between the daughter cells also occurs at the cell plate.

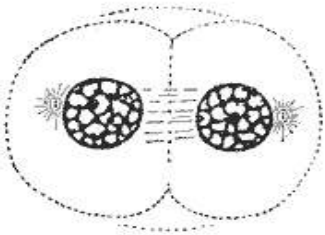


Fig.9. Telophase: The pinching is known as the *cleavage furrow*. Note the decondensing chromosomes.

2.7. Meiosis

- Meiosis is the type of cell division by which germ cells (eggs and sperm) are produced. Meiosis involves a reduction in the amount of genetic material.
- Meiosis comprises two successive nuclear divisions with only one round of DNA replication. Four stages can be described for each nuclear division:

2.7.1. Meiosis I

Prophase of meiosis I (prophase I) is a more elaborate process than prophase of mitosis (and usually takes much longer).

- **Prophase 1:** Each chromosome dupicates and remains closely associated. These are called sister chromatids. Crossing-over can occur during the latter part of this stage.
- **Metaphase 1**: Homologous chromosomes align at the equatorial plate.
- **Anaphase 1**: Homologous pairs separate with sister chromatids remaining together.
- **Telophase 1**: Two daughter cells are formed with each daughter containing only one chromosome of the homologous pair.

2.7.2. Meiosis II

Chromosome behavior in meiosis II is like that of mitosis

- **Prophase 2**: DNA does not replicate.
- **Metaphase 2**: Chromosomes align at the equatorial plate.

- **Anaphase 2**: Centromeres divide and sister chromatids migrate separately to each pole.
- **Telophase 2**: Cell division is complete. Four haploid daughter cells are obtained.

One parent cell produces four daughter cells. Daughter cells have half the number of chromosomes found in the original parent cell and with crossing over, are genetically different.

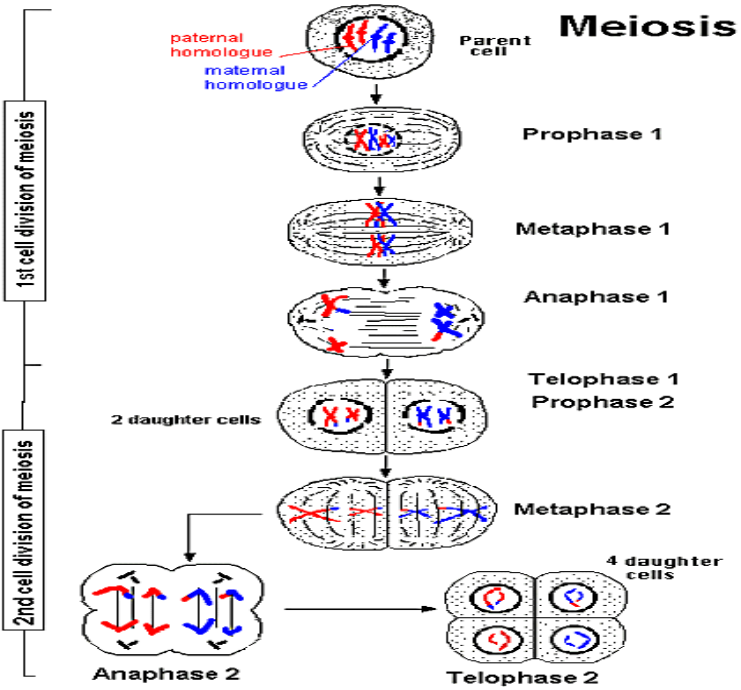


Fig. 10.Overview of steps in meiosis

Meiosis is a process of cell division in eukaryotes characterized by:

- two consecutive divisions: meiosis I and meiosis II
- no DNA synthesis (no S phase) between the two divisions

- the result: 4 cells with half the number of chromosomes of the starting cell, e.g., 2n → n. Fusion of two such cells produces a 2n zygote.

2.8. Comparison of Meiosis and Mitosis

- Chromosome behavior
    1. Mitosis: Homologous chromosomes independent
    2. Meiosis: Homologous chromosomes pair forming bivalents until anaphase I
- Chromosome number- reduction in meiosis
    1. mitosis- identical daughter cells
    2. meiosis- daughter cells haploid
- Genetic identity of progeny:
    1. Mitosis: identical daughter cells
    2. Meiosis: daughter cells have new assortment of parental chromosomes
    3. Meiosis: chromatids not identical, crossing over

2.9. Meiotic errors

- Nondisjunction- homologues don't separate in meiosis 1
    1. results in aneuploidy

2. usually embryo lethal
3. Trisomy 21, exception leading to Downs syndrome
4. Sex chromosomes
    1. Turner syndrome: monosomy X
    2. Klinefelter syndroms: XXY

- Translocation and deletion: transfer of a piece of one chromosome to another or loss of fragment of a chromosome.

2.10. Mitosis, Meiosis, and Ploidy

- Mitosis can proceed independent of ploidy of cell, homologous chromosomes behave independently
- Meiosis can only proceed if the nucleus contains an even number of chromosomes (diploid, tetraploid).
- *Ploidy* Haploid and diploid are terms referring to the number of sets of chromosomes in a cell. Ploidy is a term referring to the number of sets of chromosomes.
- Haploid organisms/cells have only one set of chromosomes, abbreviated as n. Organisms with

more than two sets of chromosomes are termed polyploid.

2.11. Meiosis and Genetic Recombination

While genes determine most of our physical characteristics, the exact combination of genes we inherit, and thus our physical traits, is in part due to a process our chromosomes undergo, known as genetic recombination.

Genetic recombination happens during meiosis, a special type of cell division that occurs during formation of sperm and egg cells and gives them the correct number of chromosomes. Since a sperm and egg unite during fertilization, each must have only half the number of chromosomes other body cells have. Otherwise, the fertilized cell would have too many.

Inside the cells that produce sperm and eggs, chromosomes become paired. While they are pressed together, the chromosomes may break, and each may swap a portion of its genetic material for the matching portion from its mate. This form of recombination is

called crossing-over. When the chromosomes glue themselves back together and separate, each has picked up new genetic material from the other. The constellation of physical characteristics it determines is now different than before crossing-over.

Tracking the movement of genes during crossing-over helps geneticists determine roughly how far apart two genes are on a chromosome. Since there are more chances for a break to occur between two genes that lie far apart, it is more likely that one gene will stay on the original chromosome, while the other crosses over. So, genes that lie far apart are likely to end up on two different chromosomes. On the other hand, genes that lie very close together are less likely to be separated by a break and crossing-over.

Genes that tend to stay together during recombination are said to be linked. Sometimes, one gene in a linked pair serves as a "marker" that can be used by geneticists to infer the presence of the other (often, a disease-causing gene).

After the chromosomes separate, they are parceled out into individual sex cells. Each chromosome moves independently of all the others - a phenomenon called independent assortment. So, for example, the copy of chromosome 1 that an egg cell receives in no way influences which of the two possible copies of chromosome 5 it gets.

Assortment takes place for each of the 23 pairs of human chromosomes. So, any single human egg receives one of two possible chromosomes 23 times, and the total number of different possible chromosome combinations is over 8 million (2 raised to the 23rd power). And that's just for the eggs. The same random assortment goes on as each sperm cell is made. Thus, when a sperm fertilizes an egg, the resulting zygote contains a combination of genes arranged in an order that has never occurred before and will never occur again. Meiosis not only preserves the genome size of sexually reproducing eukaryotes but also provides three mechanisms to diversify the genomes of the offspring.

2.12. Meiosis and Sexual Reproduction

Meiosis: Sexual reproduction occurs only in eukaryotes. During the formation of gametes, the number of chromosomes is reduced by half, and returned to the full amount when the two gametes fuse during fertilization.

Meiosis is a special type of nuclear division which segregates one copy of each homologous chromosome into each new "gamete". Mitosis maintains the cell's original ploidy level (for example, one diploid 2n cell producing two diploid 2n cells; one haploid n cell producing two haploid n cells; etc.). Meiosis, on the other hand, reduces the number of sets of chromosomes by half, so that when gametic recombination (fertilization) occurs the ploidy of the parents will be reestablished.

Most cells in the human body are produced by mitosis. These are the somatic (or vegetative) line cells. Cells that become gametes are referred to as germ line cells. The vast majority of cell divisions in the human body are mitotic, with meiosis being restricted to the gonads.

Review Questions

1. What are the basic differences between mitosis and Meiosis?
2. List the basic steps of mitosis
3. Outline the steps of meiosis
4. What are miototic errors?
5. Discuss meiosis and genetic recombination
6. What are the roles of meiosis in human life?

CHAPTER THREE
MACROMOLECULES

Specific Learning Objectives

At the end of this chapter, student are expected to

- ⇒ Describe the chemistry of biological macromolecules
- ⇒ Describe the features of each major type of macromolecule and their representative monomers
- ⇒ Be able to recognize functional groups of macromolecules
- ⇒ Explain the structures of macromolecules
- ⇒ Describe agents of denaturation

3.0. Introduction

There are three major types of biological macromolecules in mammalian systems.

1. Carbohydrates
2. Nucleic acids

3. Proteins

Their monomer units are:

1. Monosaccharide: for carbohydrate
2. Nucleotide: for nucleic acids
3. Amino acid: for proteins

3.1. Carbohydrate

Monosaccharides polymerize to form polysaccharides. Glucose is a typical monosaccharide. It has two important types of functional group:

1) A carbonyl group (aldehydes in glucose, some other sugars have a ketone group instead,
2) Hydroxyl groups on the other carbons.

Glucose exists mostly in ring structures. 5-OH adds across the carbonyl oxygen double bond. This is a so-called internal hemi-acetal. The ring can close in either of two ways, giving rise to anomeric forms, -OH down (the alpha-form) and -OH up (the beta-form)

The anomeric carbon (the carbon to which this -OH is attached) differs significantly from the other carbons. Free anomeric carbons have the chemical reactivity of carbonyl carbons because they spend part of their time in the open chain form. They can reduce alkaline solutions of cupric salts.

Sugars with free anomeric carbons are therefore called reducing sugars. The rest of the carbohydrate consists of ordinary carbons and ordinary -OH groups. The point is, a monosaccharide can therefore be thought of as having polarity, with one end consisting of the anomeric carbon, and the other end consisting of the rest of the molecule.

Monosaccharide can polymerize by elimination of the elements of water between the anomeric hydroxyl and a hydroxyl of another sugar. This is called a glycosidic bond.

If two anomeric hydroxyl groups react (head to head condensation) the product has no reducing end (no free anomeric carbon). This is the case with sucrose. If the anomeric hydroxyl reacts with a non-anomeric hydroxyl

of another sugar, the product has ends with different properties.

- A reducing end (with a free anomeric carbon).
- A non-reducing end.

This is the case with maltose. Since most monosaccharide has more than one hydroxyl, branches are possible, and are common. Branches result in a more compact molecule. If the branch ends are the reactive sites, more branches provide more reactive sites per molecule.

3.2. Nucleic acids

Nucleotides consist of three parts. These are:

1. Phosphate
2. Monosaccharide
    - Ribose (in ribonucleotides)
    - Deoxyribose, which lacks a 2' -OH (in deoxyribonucleotides), and
3. A base

The bases are categorized in two groups:

Purine	**Pyrimidine**
Adenine	Cytosine
Guanine	Uracil (in Ribonucleotides)
	or
	Thymine (in Deoxyribonucleotides)

Nucleotides polymerize to form nucleic acids. Nucleotides polymerize by eliminating the elements of water to form esters between the 5'-phosphate and the 3' -OH of another nucleotide. A 3'->5' phosphodiester bond is thereby formed. The product has ends with different properties:

- An end with a free 5' group (likely with phosphate attached); this is called the 5' end.
- An end with a free 3' group; this is called the 3' end.

The conventions for writing sequences of nucleotides in nucleic acids are as follows:

- Bases are abbreviated by their initials: A, C, G and U or T.

- U is normally found only in RNA, and T is normally found only in DNA. So the presence of U versusT distinguishes between RNA and DNA in a written sequence.
- Sequences are written with the 5' end to the left and the 3' end to the right unless specifically designated otherwise.
- Phosphate groups are usually not shown unless the writer wants to draw attention to them.

The following representations are all equivalent.

Uracil Adenine Cytosine Guanine

| | | |

A) P-ribose-P-ribose-P-ribose-P-ribose-OH

5' 3' 5' 3' 5' 3' 5' 3'

B) pUpApCpG

C) UACG

D) 3' GCAU 5'

N.B. The last sequence is written in reverse order, but the ends are appropriately designated.

3.3. Protein

Amino acids contain a carboxylic acid (-COOH) group and an amino (-NH2) group. The amino groups are usually attached to the carbons which are alpha to the carboxyl carbons, so they are called alpha-amino acids.

The naturally occurring amino acids are optically active, as they have four different groups attached to one carbon, (Glycine is an exception, having two hydrogen's) and have the L-configuration.

The R-groups of the amino acids provide a basis for classifying amino acids. There are many ways of classifying amino acids, but one very useful way is on the basis of how well or poorly the R-group interacts with water:

1. The hydrophobic R-groups which can be aliphatic (such as the methyl group of alanine) or aromatic (such as the phenyl group of phenylalanine).
2. The hydrophilic R-groups which can contain neutral polar (such as the -OH of serine) or

ionizable (such as the -COOH of aspartate) functional groups.

3.3.1. Polymerization of amino acids

Amino acids polymerize to form polypeptides or proteins. Amino acids polymerize by eliminating the elements of water to form an amide between the amino and carboxyl groups. The amide link thereby formed between amino acids is called a peptide bond. The product has ends with different properties.

- An end with a free amino group; this is called the amino terminal or N-terminal.
- An end with a free carboxyl group; this is called the carboxyl terminal or C-terminal.

3.3.2. Conventions for writing sequences of amino acids

- Abbreviations for the amino acids are usually used; most of the three letter abbreviations are self-evident, such as gly for glycine, asp for aspartate, etc.

- There is also a one-letter abbreviation system; it is becoming more common. Many of the one-letter abbreviations are straightforward, for example:

  G = glycine

  L = leucine

  H = histidine

- Others require a little imagination to justify:

  F = phenylalanine ("ph" sounds like "F").

  Y = tyrosine (T was used for threonine, so it was settled by the second letter in the name).

  D = aspartate (D is the fourth letter in the alphabet, and aspartate has four carbons).

- Still others are rather difficult to justify:

  W = tryptophan (The bottom half of the two aromatic rings look sort of like a "W").

  K = lysine (if you can think of a good one for this, let us know!)

- Sequences are written with the N-terminal to the left and the C-terminal to the right.

- Although R-groups of some amino acids contain amino and carboxyl groups, branched polypeptides or proteins do not occur.
- The sequence of monomer units in a macromolecule is called the primary structure of that macromolecule. Each specific macromolecule has a unique primary structure.

3.4. Helix

- A helical structure consists of repeating units that lie on the wall of a cylinder such that the structure is super-imposable upon itself if moved along the cylinder axis.
- A helix looks like a spiral or a screw. A zig-zag is a degenerate helix.
- Helices can be right-handed or left handed. The difference between the two is that:
    - Right-handed helices or screws advance (move away) if turned clockwise. Examples: standard screw, bolt, jar lid.
    - Left-handed helices or screws advance (move away) if turned counterclockwise. Example: some automobile lug nuts.

- Helical organization is an example of secondary structure. These helical conformations of macromolecules persist in solution only if they are stabilized.

3.4.1. Helices in carbohydrates

- Carbohydrates with long sequences of alpha (1 -> 4) links have a weak tendency to form helices. Starch (amylose) exemplifies this structure.
- The starch helix is not very stable in the absence of other interactions (iodine, which forms a purple complex with starch, stabilized the starch helix), and it commonly adopts a random coil conformation in solution.

In contrast, beta (1 -> 4) sequences favor linear structures. Cellulose exemplifies this structure.

- Cellulose is a degenerate helix consisting of glucose units in alternating orientation stabilized by intrachain hydrogen bonds. Cellulose chains lying side by side can form sheets stabilized by interchain hydrogen bonds.

3.4.2. Helices in nucleic acids

- Single chains of nucleic acids tend to from helices stabilized by base stacking. The purine and pyrimidine bases of the nucleic acids are aromatic rings. These rings tend to stack like pancakes, but slightly offset so as to follow the helix.
- The stacks of bases are in turn stabilized by hydrophobic interactions and by van der Waals forces between the pi-clouds of electrons above and below the aromatic rings.
- In these helices the bases are oriented inward, toward the helix axis, and the sugar phosphates are oriented outward, away from the helix axis.

Two lengths of nucleic acid chain can form a double helix stabilized by

- Base stacking
- Hydrogen bonds.

- Purines and pyrimidines can form specifically hydrogen-bonded base pairs.

- Guanine and cytosine can form a base pair that measures 1.08 nm across, and that contains three hydrogen bonds.
- Adenine and thymine (or Uracil) can form a base pair that measures 1.08 nm across, and that contains two hydrogen bonds.
- Base pairs of this size fit perfectly into a double helix. This is the so-called Watson-Crick base-pairing pattern.
- Double helices rich in GC pairs are more stable than those rich in AT (or AU) pairs because GC pairs have more hydrogen bonds. Specific AT (or AU) and GC base pairing can occur only if the lengths of nucleic acid in the double helix consist of complementary sequences of bases.
  - A must always be opposite T (or U).
  - G must always be opposite C.

Here is a sample of two complementary sequences:

5'...ATCCGAGTG.. 3'.
3' ...AGGCTCAC... .5'

- Most DNA and some sequences of RNA have this complementarity's, and form the double helix. It is important to note, though, that the complementary sequences forming a double helix have opposite polarity. The two chains run in opposite directions:

  5' ...ATCCGAGTG... 3'
  3' ...TAGGCTCAC... 5'

- This is described as an anti-parallel arrangement. This arrangement allows the two chains to fit together better than if they ran in the same direction (parallel arrangement). The Consequences of complementarities include:
  - In any double helical structure the amount of A equals the amount of T (or U), and the amount of G equals the amount of C
  - Because DNA is usually double stranded, while RNA is not, in DNA A=T and G=C, while in RNA A does not equal U and G does not equal C.

Three major types of double helix occur in nucleic acids. These three structures are strikingly and obviously different in appearance.

1) DNA usually exists in the form of a B-helix. Its characteristics:
   - Right-handed and has 10 nucleotide residues per turn.
   - The plane of the bases is nearly perpendicular to the helix axis.
   - There is a prominent major groove and minor groove.
   - The B-helix may be stabilized by bound water that fits perfectly into the minor groove.

2) Double-stranded RNA and DNA-RNA hybrids (also DNA in low humidity) exist in the form of an A-helix. Its characteristics:
   - Right-handed and has 11 nucleotide residues per turn.
   - The plane of the bases is tilted relative to the helix axis.
   - The minor groove is larger than in B-DNA.

RNA is incompatible with a B-helix because the 2' -OH of RNA would be sterically hindered. (There is no 2' -OH in DNA.) This is a stabilizing factor.

3) DNA segments consisting of alternating pairs of purine and pyrimidine $(PuPy)_n$ can form a Z-helix. Its characteristics:
   - Left-handed (this surprised the discoverers) and has 12 residues (6 PuPy dimers) per turn.
   - Only one groove.
   - The phosphate groups lie on a zig-zag line, which gives rise to the name, Z-DNA.

The geometry of the grooves is important in allowing or preventing access to the bases. The surface topography of the helix forms attachment sites for various enzymes sensitive to the differences among the helix types.

3.4.3. Helices in proteins

Properties of the peptide bond dominate the structures of proteins. Properties of the peptide bond include:

1) The peptide bond has partial double character. Partial double character is conferred by the electronegative carbonyl oxygen, which draws the unshared electron pair from the amide hydrogen. As a result of having double bond character the peptide bond is:

- Planar
- Not free to rotate
- More stable in the *trans* configuration than in the *cis*

These characteristics restrict the three-dimensional shapes of proteins because they must be accommodated by any stable structure.

2) The peptide bond is that the atoms of the peptide bond can form hydrogen bonds.

Stabilizing factors include:

1. All possible hydrogen bonds between peptide C=O and N-H groups in the backbone are formed. The hydrogen bonds are all intrachain, between different parts of the same chain. A although a single hydrogen bond is weak, cooperation of many hydrogen bonds can be strongly stabilizing.
2. Alpha-helices must have a minimum length to be stable (so there will be enough hydrogen bonds).

3. All peptide bonds are trans and planar. So, if the amino acids R-groups do not repel one another helix formation is favored.
4. The net electric charge should be zero or low (charges of the same sign repel).
5. Adjacent R-groups should be small, to avoids steric repulsion.

Destabilizing factors include:

1. R-groups that repel one another favor extended conformations instead of the helix. Examples include large net electric charge and adjacent bulky R-groups.
2. Proline is incompatible with the alpha-helix. The ring formed by the R-group restricts rotation of a bond that would otherwise be free to rotate.
3. The restricted rotation prevents the polypeptide chain from coiling into an alpha-helix. Occurrence of proline necessarily terminates or kinks alpha-helical regions in proteins.

The next level of macromolecular organization is Tertiary structure.

3.5. Tertiary structure

Tertiary structure is the three dimensional arrangement of helical and non-helical regions of macromolecules. Nucleic acids and proteins are large molecules with complicated three-dimensional structures.

These structures are formed from simpler elements, suitably arranged. Although structural details vary from macromolecule to macromolecule, a few general patterns describe the overall organization of most macromolecules.

3.5.1. Tertiary structure of DNA

Many naturally occurring DNA molecules are circular double helices. Most circular double-stranded DNA is partly unwound before the ends are sealed to make the circle.

- Partial unwinding is called negative superhelicity.

- Overwinding before sealing would be called positive superhelicity.

Superhelicity introduces strain into the molecule. The strain of superhelicity can be relieved by forming a super coil. The identical phenomenon occurs in retractable telephone headset cords when they get twisted. The twisted circular DNA is said to be super coiled. The supercoil is more compact. It is poised to be unwound, a necessary step in DNA and RNA synthesis.

3.5.2. Tertiary structure of RNA

Most RNA is single stranded, but contains regions of self-complementarity.

This is exemplified by yeast tRNA. There are four regions in which the strand is complementary to another sequence within itself. These regions are anti-parallel, fulfilling the conditions for stable double helix formation. X-ray crystallography shows that the three dimensional structure of tRNA contains the expected double helical regions.

Large RNA molecules have extensive regions of self-complementarity, and are presumed to form complex three-dimensional structures spontaneously.

3.5.3. Tertiary structure in Proteins

The formation of compact, globular structures is governed by the constituent amino acid residues. Folding of a polypeptide chain is strongly influenced by the solubility of the amino acid R-groups in water. Hydrophobic R-groups, as in leucine and phenylalanine, normally orient inwardly, away from water or polar solutes. Polar or ionized R-groups, as in glutamine or arginine, orient outwardly to contact the aqueous environment.

The rules of solubility and the tendency for secondary structure formation determine how the chain spontaneously folds into its final structure.

Forces stabilizing protein tertiary structure

1. Hydrophobic interactions -- the tendency of nonpolar groups to cluster together to exclude water.

2. Hydrogen bonding, as part of any secondary structure, as well as other hydrogen bonds.
3. Ionic interactions -- attraction between unlike electric charges of ionized R-groups.
4. Disulfide bridges between cysteinyl residues. The R-group of cysteine is $-CH_2-SH$. -SH (sulfhydryl) groups can oxidize spontaneously to form disulfides (-S-S-).

$$R-CH_2-SH + R'-CH_2-SH + O_2 = R-CH_2-S-S-CH2-R' + H_2O_2$$

N.B. *Under reducing conditions a disulfide bridge can be cleaved to regenerate the -SH groups.*

The disulfide bridge is a covalent bond. It strongly links regions of the polypeptide chain that could be distant in the primary sequence. It forms after tertiary folding has occurred, so it stabilizes, but does not determine tertiary structure.

Globular proteins are typically organized into one or more compact patterns called domains. This concept of domains is important. In general it refers to a region of a protein. But it turns out that in looking at protein after

protein, certain structural themes repeat themselves, often, but not always in proteins that have similar biological functions. This phenomenon of repeating structures is consistent with the notion that the proteins are genetically related, and that they arose from one another or from a common ancestor.

The four-helix bundle domain is a common pattern in globular proteins. Helices lying side by side can interact favorably if the properties of the contact points are complementary.

Hydrophobic amino acids (like leucine) at the contact points and oppositely charged amino acids along the edges will favor interaction. If the helix axes are inclined slightly (18 degrees), the R-groups will interdigitate perfectly along 6 turns of the helix.

Sets of four helices yield stable structures with symmetrical, equivalent interactions. Interestingly, four-helix bundles diverge at one end, providing a cavity in which ions may bind.

All beta structures comprise domains in many globular proteins. Beta-pleated sheets fold back on themselves to form barrel-like structures. Part of the immunoglobulin molecule exemplifies this. The interiors of beta-barrels serve in some proteins as binding sites for hydrophobic molecules such as retinol, a vitamin A derivative. What keeps these proteins from forming infinitely large beta-sheets is not clear.

3.6. Macromolecular Interactions

Macromolecules interact with each other and with small molecules. All these interactions reflect complementarity between the interacting species. Sometimes the complementarity is general, as in the association of hydrophobic groups, but more often an exact fit of size, shape and chemical affinity is involved.

Quaternary structure refers to proteins formed by association of polypeptide subunits. Individual globular polypeptide subunits may associate to form biologically active oligomers.

Quaternary structure in proteins is the most intricate degree of organization considered to be a single molecule. Higher levels of organization are multimolecular complexes.

3.7. Denaturation

Denaturation is the loss of a protein's or DNA's three dimensional structure. The "normal" three dimensional structure is called the native state. Denaturing agents disrupt stabilizing factors

Destruction of a macromolecule's three-dimensional structure requires disruption of the forces responsible for its stability. The ability of agents to accomplish this disruption -- denaturation -- can be predicted on the basis of what is known about macromolecular stabilizing forces.

Denatured macromolecules will usually renature spontaneously (under suitable conditions), showing that the macromolecule itself contains the information needed to establish its own three-dimensional structure.

Denaturation is physiological -- structures ought not to be too stable.

1. Double stranded DNA must come apart to replicate and for RNA synthesis.
2. Proteins must be degraded under certain circumstances.
    - To terminate their biological action (e.g., enzymes).
    - To release amino acids (e.g., for gluconeogenesis in starvation).

Loss of native structure must involve disruption of factors responsible for its stabilization. These factors are:

1. Hydrogen bonding
2. Hydrophobic interaction
3. Electrostatic interaction
4. Disulfide bridging (in proteins)

3.7.1. Agents that disrupt hydrogen bonding

Heat -- thermal agitation (vibration, etc.) -- will denature proteins or nucleic acids. Heat denaturation of DNA is called melting because the transition from native to

denatured state occurs over a narrow temperature range. As the purine and pyrimidine bases become unstacked during denaturation they absorb light of 260 nanometers wavelength more strongly. The abnormally low absorption in the stacked state is called the hypochromic effect.

3.7.2. Agents that disrupt hydrophobic interaction

Organic solvents: such as acetone or ethanol -- dissolve nonpolar groups.

Detergents: dissolve nonpolar groups.

Cold: increases solubility of non-polar groups in water. When a hydrophobic group contacts water, the water dipoles must solvate it by forming an orderly array around it.

The significance of cold denaturation is that cold is not a stabilizing factor for all proteins. Cold denaturation is important in proteins that are highly dependent on hydrophobic interaction to maintain their native structure.

3.7.3. Agents that disrupt electrostatic interaction.

pH extremes -- Most macromolecules are electrically charged. Ionizable groups of the macromolecule contribute to its net charge. Bound ions also contribute to its net charge. Electric charges of the same sign repel one another. If the net charge of a macromolecule is zero or near zero, electrostatic repulsion will be minimized. The substance will be minimally soluble, because intermolecular repulsion will be minimal. A compact three-dimensional structure will be favored, because repulsion between parts of the same molecule will be minimal.

The pH at which the net charge of a molecule is zero is called the isoelectric pH (or isoelectric point).

pH extremes result in large net charges on most macromolecules. Most macromolecules contain many weakly acidic groups.

- At low pH all the acidic groups will be in the associated state (with a zero or positive charge). So the net charge on the protein will be positive.
- At high pH all the acidic groups will be dissociated (with a zero or negative charge). So the net charge on the protein will be negative. Intramolecular electrostatic repulsion from a large net charge will favor an extended conformation rather than a compact one.

3.7.4. Agents that disrupt disulfide bridges

Some proteins are stabilized by numerous disulfide bridges; cleaving them renders these proteins more susceptible to denaturation by other forces. Agents with free sulfhydryl groups will reduce (and thereby cleave) disulfide bridges. It destabilizes some proteins.

e.g., 2 $HO\text{-}CH_2\text{-}CH_2\text{-}SH + R_1\text{-}S\text{-}S\text{-}R_2 = R_1\text{-}SH + HS\text{-}R_2 + HO\text{-}CH_2\text{-}CH_2\text{-}S\text{-}S\text{-}CH_2\text{-}CH_2\text{-}OH$

3.8. Renaturation

Renaturation is the regeneration of the native structure of a protein or nucleic acid. Renaturation requires removal of the denaturing conditions and restoration of conditions favorable to the native structure. This includes

- Solubilization of the substance if it is not already in solution.
- Adjustment of the temperature.
- Removal of denaturing agents by dialysis or similar means.
- In proteins, re-formation of any disulfide bridges.

Usually considerable skill and art are required to accomplish renaturation. The fact that renaturation is feasible demonstrates that the information necessary for forming the correct three-dimensional structure of a protein or nucleic acid is encoded in its primary structure, the sequence of monomer units.

Molecular chaperones are intracellular proteins which guide the folding of proteins, preventing incorrect molecular interactions. They do NOT appear as components of the final structures. Chaperones are widespread, and chaperone defects are believed to be the etiology of some diseases. Medical applications of chaperones may be expected to include things such as

- repair of defective human chaperones and
- inhibition of those needed by pathogenic organisms.

Review Questions

1. What are the major types of biological macromolecules?

2. How monomers join to form polymer in each category of macromolecule? Tertiary structure, and Quaternary structure?

3. What is the difference between the primary structure of a protein and the higher order structures (secondary, tertiary and quaternary)

4. Outline macromolecular interaction with different substance.

5. What is denaturation? List agents of denaturation.

6. Why is carbon central on biological molecules?

7. What are the elements which make up macromolecules?

8. How are lipids different that the other classes of macromolecules that we have discussed?

9. What is a peptide bond?

10. How many different amino acids are there?

11. What are some important functions for proteins in our cells?

12. What are the three components of a nucleotide?

13. Which nitrogenous bases are found in DNA?

CHAPTER FOUR

GENETICS

Specific learning objectives

At the end of this chapter, students are expected to

- ⇒ describe basics of genetics
- ⇒ describe terms used in genetics
- ⇒ explain Mendel's 1^{st} ,2^{nd} and 3rd law
- ⇒ describe exception to Mendel's law

4.1. Mendelian Genetics

A number of hypotheses were suggested to explain heredity, but Gregor Mendel, was the only one who got it more or less right. His early adult life was spent in relative obscurity doing basic genetics research and teaching high school mathematics, physics, and Greek in Brno (now in the Czech Republic).

While Mendel's research was with plants, the basic underlying principles of heredity that he discovered also apply to humans and other animals because the mechanisms of heredity are essentially the same for all complex life forms. But Mendelian inheritance not common in organelle gene

Through the selective growing of common pea plants (*Pisum sativum*) over many generations, Mendel discovered that certain traits show up in offspring plants without any blending of parent characteristics. This concept is reveled during the reappearance of the recessive phenotype in the F2 generation where allele remains particulate during transmission and are neither displaced nor blended in the hybrid to generate the phenotype.

Flower color in snapdragons is an example of this pattern. Cross a true-breeding red strain with a true-breeding white strain and the F1 are all pink (heterozygotes). Self-fertilize the F1 and you get an F2 ratio of 1 red: 2 pink: 1 white. This would not happen if true blending had occurred (blending cannot explain traits such as red or white skipping a generation and

pink flowers crossed with pink flowers should produce only pink flowers).

Mendel picked common garden pea plants for the focus of his research because they can be grown easily in large numbers and their reproduction can be manipulated. Pea plants have both male and female reproductive organs. As a result, they can either self-pollinate themselves or cross-pollinate with another plant. Mendel observed seven traits that are easily recognized and apparently only occur in one of two forms:

1. flower color is purple or white
2. flower position is axil or terminal
3. stem length is long or short
4. seed shape is round or wrinkled
5. seed color is yellow or green
6. pod shape is inflated or constricted
7. pod color is yellow or green

In his experiments, Mendel was able to selectively cross-pollinate purebred plants with particular traits and observe the outcome over many generations. This was the basis for his conclusions about the nature of genetic inheritance.

In cross-pollinating plants that either produces yellow or green peas exclusively, Mendel found that the first offspring generation (f1) always has yellow peas. However, the following generation (f2) consistently has a 3:1 ratio of yellow to green (Fig1.1)

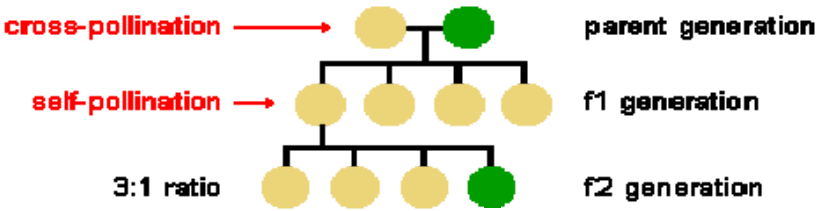


Fig. 11. Cross pollination and self pollination and their respective generation

This 3:1 ratio occurs in later generations as well (Fig.1.2). Mendel realized that this is the key to understanding the basic mechanisms of inheritance.

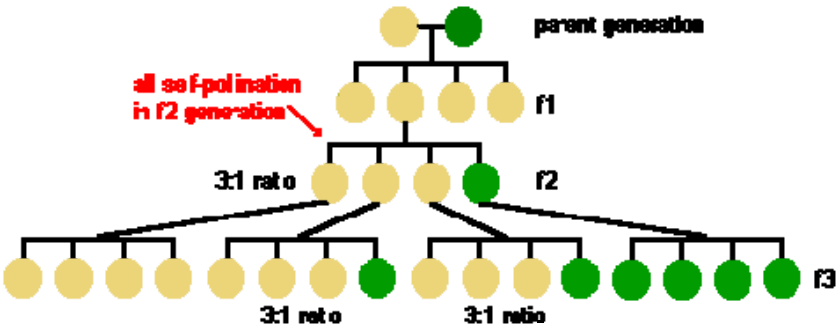


Fig.12. Self pollination of f2 generation

He came to three important conclusions from these experimental results:

- The inheritance of each trait is determined by "units" or "factors" (now called genes) that are passed on to descendents unchanged.Each individual inherits one such unit from each parent for each trait.
- A trait may not show up in an individual but can still be passed on to the next generation.

It is important to realize that in this experiment the starting parent plants were homozygous for pea color. The plants in the f1 generation were all heterozygous. It becomes clearer when one looks at the actual genetic

makeup, or genotype, of the pea plants instead of only the phenotype, or observable physical characteristics (Fig, 4.3).

Note that each of the f1 generation plants (shown below) inherited a Y allele from one parent and a G allele from the other. When the f1 plants breed, each has an equal chance of passing on either Y or G alleles to each offspring.

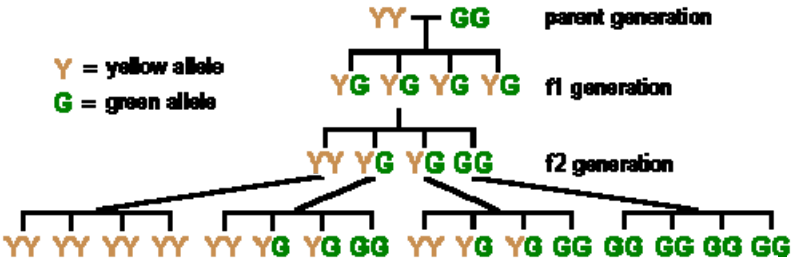


Fig.13. Genetic composition of parent generation with their f1and f2 generation

Mendel's observations from these experiments can be summarized in two principles:

- The principle of segregation
- The principle of independent assortment
- The principe of Dominance

4.2. Mendel's first law: Principle of Segregation

According to the principle of segregation, for any particular trait, the pair of alleles of each parent separate and only one allele passes from each parent on to an offspring. Which allele in a parent's pair of alleles is inherited is a matter of chance. We now know that this segregation of alleles occurs during the process of sex cell formation (i.e., meiosis).

In segregation, one allele from each parent is "chosen" *at random*, and passed in the gamete onto the offspring.

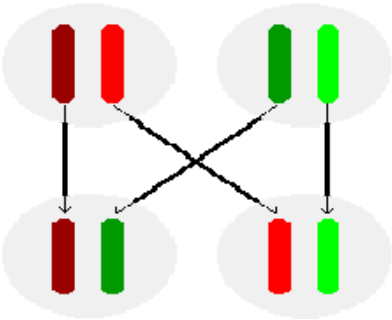


Fig. 14. Segregation of alleles in the production of sex cells

4.3. Mendel's second law: principle of independent assortment

According to the principle of independent assortment, different pairs of alleles are passed to offspring independently of each other. The result is that new combinations of genes present in neither parent are possible. For example, a pea plant's inheritance of the ability to produce purple flowers instead of white ones does not make it more likely that it would also inherit the ability to produce yellow peas in contrast to green ones.

Likewise, the principle of independent assortment explains why the human inheritance of a particular eye color does not increase or decrease the likelihood of having 6 fingers on each hand. Today, it is known that this is due to the fact that the genes for independently assorted traits are located on different chromosomes.

4.4. Mendel's third law: Principle of Dominance

With all of the seven pea plant traits that Mendel examined, one form appeared dominant over the other. This is to say, it masked the presence of the other allele. For example, when the genotype for pea color is YG (heterozygous), the phenotype is yellow. However, the dominant yellow allele does not alter the recessive green one in any way. Both alleles can be passed on to the next generation unchanged.

These two principles of inheritance, along with the understanding of unit inheritance and dominance, were the beginnings of our modern science of genetics. However, Mendel did not realize that there are exceptions to these rules.

It was not until 1900 that Mendel's work was replicated, and then rediscovered. Shortly after this, numerous exceptions to Mendel's second law were observed. These were not fully understood until Morgan.

One of the reasons that Mendel carried out his breeding experiments with pea plants is that he could observe inheritance patterns in up to two generations a year.

Geneticists today usually carry out their breeding experiments with species that reproduce much more rapidly so that the amount of time and money required is significantly reduced. Fruit flies and bacteria are commonly used for this purpose now.

4.5. Exceptions to Mendelian rules

There are many reasons why the ratios of offspring phenotypic classes may depart (or seem to depart) from a normal Mendelian ratio. For instance:

- **Lethal alleles**

  Many so called dominant mutations are in fact *semidominant*, the phenotype of the homozygote is more extreme than the phenotype of the heterozygote. For instance the gene T (Danforth's short tail) in mice. The normal allele of this gene is expressed in the embryo. T/+ mice develop a short

tail but T/T homozygotes die as early embryos. Laboratory stocks are maintained by crossing heterozygotes,

T/+ x T/+
|
|
V
T/T T/+ +/+
1 : 2 : 1 ratio at conception
0 : 2 : 1 ratio at birth

- **Incomplete or semi- dominance**

Incomplete dominance may lead to a distortion of the apparent ratios or to the creation of unexpected classes of offspring. A human example is Familial Hypercholesterolemia (FH). Here there are three phenotypes: +/+ = normal, +/- = death as young adult, -/- = death in childhood. The gene responsible codes for the liver receptor for cholesterol. The number of receptors is directly related to the number of active genes. If the number of receptors is

lowered the level of cholesterol in the blood is elevated and the risk of coronary artery disease is raised.

- **Codominance**

  If two or more alleles can each be distinguished in the phenotype in the presence of the other they are said to be codominant. An example is seen in the ABO blood group where the **A and B alleles are codominant**.

  The *ABO* gene codes for a glycosyl-transferase which modifies the H antigen on the surface of red blood cells. The A form adds N-acetylgalactosamine, the B form adds D-galactose forming the A and B antigens respectively. The O allele has a frameshift mutation in the gene and thus produces a truncated and inactive product which cannot modify H. A phenotype people have natural antibodies to B antigen in their serum and vice versa. O phenotype individuals have antibodies directed against both A and B. AB individuals have no antibodies against either A or B antigens.

ABO genotypes and phenotypes			
Genotype	Phenotype	red cell antigens	serum antibodies
AA	A	A	anti-B
AO	A	A	anti-B
BB	B	B	anti-A
BO	B	B	anti-A
AB	AB	A and B	neither
OO	O	neither	anti-A and anti-B

Silent alleles

In a multiple allele system, it is sometimes not obvious that a silent allele exists. This can give confusing results. Consider for example:

A/A x A/B (phenotype A crossed to phenotype AB)

|

|

V

A/A : A/B

1 : 1

and compare with

A/O x A/B (phenotype A crossed to phenotype AB)

|

|

V

A/A : A/O : A/B : B/O

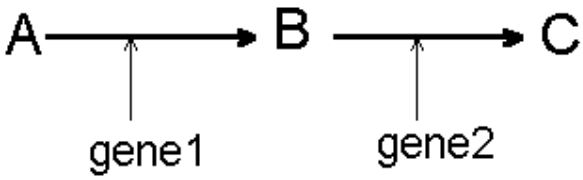
1 : 1 : 1 : 1

It would be important not to lump together these two different sorts of crosses but when there are only small numbers of offspring (which is the case in most human matings) some offspring classes may not be represented in a family and it may not be obvious which type of mating you are examining.

- **Epistasis**

This occurs where the action of one gene masks the effects of another making it impossible to tell the genotype at the second gene. The cause might be that

both genes produce enzymes which act in the same biochemical pathway.



If the product of gene1 is not present because the individual is homozygous for a mutation, then it will not be possible to tell what the genotype is at gene2. The Bombay phenotype in humans is caused by an absence of the H antigen so that the ABO phenotype will be O no matter what the ABO genotype.

- **Pleiotropy**

Mutations in one gene may have many possible effects. Problems in tracing the passage of a mutant allele through a pedigree can arise when different members of a family express a different subset of the symptoms.

Pleiotropy can occur whenever a gene product is required in more than one tissue or organ.

- **Genetic heterogeneity**

This is the term used to describe a condition which may be caused by mutations in more than one gene. Tuberous sclerosis again provides a good example of this, the identical disease is produced by mutations in either of two unrelated genes, *TSC1* on chromosome 9 or *TSC2* on chromosome 16. In such cases, presumably both genes act at different points in the same biochemical or regulatory pathway. Or perhaps one provides a ligand and one a receptor.

- **variable expressivity**

The degree to which a disease may manifest itself can be very variable and, once again, tuberous sclerosis provides a good example. Some individuals scarcely have any symptoms at all whereas others are severely affected. Sometimes very mild symptoms may be overlooked and then a person may be wrongly classified as non-affected. Clearly this could have profound implications for genetic counselling.

- **Incomplete Penetrance**

This is an extreme case of a low level of expressivity Some individuals who logically ought to show symptoms

because of their genotype do not. In such cases even the most careful clinical examination has revealed no symptoms and a person may be misclassified until suddenly he or she transmits the gene to a child who is then affected.

One benefit of gene cloning is that within any family in which a mutant gene is known to be present, when the gene is known, the mutation can be discovered and the genotype of individuals can be directly measured from their DNA. . In this way diagnosis and counselling problems caused by non-penetrance can be avoided. The degree of penetrance can be estimated. If a mutation is 20% penetrant then 20% of persons who have the mutant genotype will display the mutant phenotype, etc.

- **Anticipation**

In some diseases it can appear that the symptoms get progressively worse every generation. One such disease is the autosomal dominant condition *myotonic dystrophy*. This disease, which is characterized by a number of symptoms such as myotonia, muscular

dystrophy, cataracts, hypogonadism, frontal balding and ECG changes, is usually caused by the expansion of a trinucleotide repeat in the 3'untranslated region of a gene on chromosome 19. The severity of the disease is roughly correlated with the number of copies of the trinucleotide repeat unit.

Number of CTG repeats	phenotype
5	normal
19 - 30	"pre-mutant"
50 - 100	mildly affected
2,000 or more	severely affected
myotonic dystrophy	

The "premutant" individuals have a small expansion of the number of trinucleotide repeats which is insufficient to cause any clinical effect in itself but it allows much greater expansions to occur during the mitotic divisions which precede gametogenesis. Mildly affected individuals can again have gametes in which a second round of expansion has occurred.

Germline Mosaicism

If a new mutation occurs in one germ cell precurser out of the many non-mutant precursers, its descendent germ cells, being diluted by the many non-mutant germ cells also present, will not produce mutant offspring in the expected Medelian numbers.

- **Phenocopies**

An environmentally caused trait may mimic a genetic trait, for instance a heat shock delivered to *Drosophila* pupae may cause a variety of defects which mimic those caused by mutations in genes affecting wing or leg development. In humans, the drug thalidomide taken during pregnancy caused phenocopies of the rare genetic disease phocomelia, children were born with severe limb defects.

- **mitochondrial inheritance**

The human mitochondrion has a small circular genome of 16,569 bp which is remarkably crowded. It is inherited only through the egg, sperm mitochondria never contribute to the zygote population of mitochondria.

There are relatively few human genetic diseases caused by mitochondrial mutations but, because of their maternal transmission, they have a very distinctive pattern of inheritance.

- **Uniparental disomy**

Although it is not possible to make a viable human embryo with two complete haploid sets of chromosomes from the same sex parent it is sometimes possible that both copies of a single chromosome may be inherited from the same parent (along with no copies of the corresponding chromosome from the other parent.) Rare cases of cystic fibrosis (a common autosomal recessive disease) have occurred in which one parent was a heterozygous carrier of the disease but the second parent had two wild type alleles. The child had received two copies of the mutant chromosome 7 from the carrier parent and no chromosome 7 from the unaffected parent.

- **linkage**

When two genes are close together on the same chromosome they tend to be inherited together because of the mechanics of chromosome segregation at

meiosis. This means that they do not obey the law of independent assortment. The further apart the genes are the more opportunity there will be for a chiasma to occur between them. When they get so far apart that there is always a chiasma between them then they are inherited independently. The frequency with which the genes are separated at miosis can be measured and is the basis for the construction of genetic linkage maps.

Review Questions

1. Why Mendel uses Pea plants for his experiment?

2. What independent assortments?

3. What is principle of segregation?

4. Define the following terms:

    Co-dominance

    Allele

    Gene

    Recessive

CHAPTER FIVE
CHROMOSOME STRUCTURE AND FUNCTION

Specific learning objectives

At the end of this chapter, students are expected to

- ⇒ describe the normal and abnormal chromosome morphology
- ⇒ list the factors which affect the relative recurrence risk for a multifactorial (polygenic) trait within a family;
- ⇒ describe the various tissues which may be used to produce chromosome preparations
- ⇒ identify associated risks to carriers of structurally rearranged chromosomes. .
- ⇒ discuss chromosome abnormalities and their relevance to diagnosis, prognosis and disease progression.
- ⇒ interpret a karyotype, describe the standard notation, specify the nature of any abnormalities;

⇒ identify the major clinical features and specific genetic errors responsible for the following disorders: Down syndrome; Klinefelter syndrome; Prader-Willi syndrome; Turner syndrome; trisomy 13; trisomy 18

5.1. Chromosome Morphology

Chromosomes are complex structures located in the cell nucleus; they are composed of DNA, histone and non-histone proteins, RNA, and polysaccharides. They are basically the "packages" that contain the DNA.

Under the microscope chromosomes appear as thin, thread-like structures. They all have a short arm and long arm separated by a primary constriction called the *centromere*. The short arm is designated as *p* and the long arm as *q*. The centromere is the location of spindle attachment and is an integral part of the chromosome. It is essential for the normal movement and segregation of chromosomes during cell division.

Human metaphase chromosomes come in three basic shapes and can be categorized according to the length

of the short and long arms and also the centromere location:

- *Metacentric chromosomes* have short and long arms of roughly equal length with the centromere in the middle.
- *Submetacentric chromosomes* have short and long arms of unequal length with the centromere more towards one end.
- *Acrocentric chromosomes* have a centromere very near to one end and have very small short arms. They frequently have secondary constrictions on the short arms that connect very small pieces of DNA, called stalks and satellites, to the centromere. The stalks contain genes which code for ribosomal RNA.

5.2. Normal Chromosomes

Each species has a normal diploid number of chromosomes. Cytogenetically normal humans, for example, have 46 chromosomes (44 autosomes and two sex chromosomes). Cattle, on the other hand, have

60 chromosomes. This ratio is an important parameter for chromosome identification, and also, the ratio of lengths of the two arms allows classification of chromosomes into several basic morphologic types.

Germ cells (egg and sperm) have 23 chromosomes: one copy of each autosome plus a single sex chromosome. This is referred to as the *haploid* number. One chromosome from each autosomal pair plus one sex chromosome is inherited from each parent. Mothers can contribute only an X chromosome to their children while fathers can contribute either an X or a Y.

Cytogenetic analyses are almost always based on examination of chromosomes fixed during mitotic metaphase. During that phase of the cell cycle, DNA has been replicated and the chromatin is highly condensed. The two daughter DNAs are encased in chromosomal proteins forming sister chromatids, which are held together at their centromere.

Metaphase chromosomes differ from one another in size and shape, and the absolute length of any one chromosome varies depending on the stage of mitosis in

which it was fixed. However, the relative position of the centromere is constant, which means that that the ratio of the lengths of the two arms is constant for each chromosome.

Centromere position and arm ratios can assist in identifying specific pairs of chromosomes, but inevitably several or many pairs of chromosomes appear identical by these criteria. The ability to identify specific chromosomes with certainty was revolutionized by discovery that certain dyes would produce reproducible patterns of bands when used to stain chromosomes.

Chromosome banding has since become a standard and indispensable tool for cytogenetic analysis, and several banding techniques have been developed:

- *Q banding*: chromosomes are stained with a fluorescent dye such as quinacrine
- *G banding*: produced by staining with Giemsa after digesting the chromosomes with trypsin
- *C banding:* chromosomes are treated with acid and base, then stained with Giesma stain

Each of these techniques produces a pattern of dark and light (or fluorescent versus non-fluorescent) bands along the length of the chromosomes. Importantly, each chromosome displays a unique banding pattern, analogous to a "bar code", which allows it to be reliably differentiated from other chromosomes of the same size and centromeric position.

5.3. Chromosome Abnormalities

Although chromosome abnormalities can be very complex there are two basic types: *numerical* and *structural*. Both types can occur simultaneously.

5.3.1. Numerical abnormalities

Numerical abnormalities involve the loss and/or gain of a whole chromosome or chromosomes and can include both autosomes and sex chromosomes. Generally chromosome loss has a greater effect on an individual than does chromosome gain although these can also have severe consequences.

Cells which have lost a chromosome are *monosomy* for that chromosome while those with an extra chromosome show *trisomy* for the chromosome involved. Nearly all autosomal monosomies die shortly after conception and only a few trisomy conditions survive to full term. The most common autosomal numerical abnormality is Down Syndrome or trisomy-21.

Trisomies for chromosomes 13 and 18 may also survive to birth but are more severely affected than individuals with Down Syndrome. Curiously, a condition called *triploidy* in which there is an extra copy of every chromosome (69 total), can occasionally survive to birth but usually die in the newborn period.

Another general rule is that loss or gain of an autosome has more severe consequences than loss or gain of a sex chromosome. The most common sex chromosome abnormality is monosomy of the X chromosome (45,X) or Turner Syndrome.

Another fairly common example is Klinefelter Syndrome (47,XXY). Although there is substantial variation within

each syndrome, affected individuals often lead fairly normal lives.

Occasionally an individual carries an extra chromosome which can't be identified by its banding pattern, these are called *marker* chromosomes. The introduction of FISH techniques has been a valuable tool in the identification of marker chromosomes.

5.3.2. Structural abnormalities

Structural abnormalities involve changes in the structure of one or more chromosomes. They can be incredibly complex but for the purposes of this discussion we will focus on the three of the more common types:

- *Deletions* involve loss of material from a single chromosome. The effects are typically severe since there is a loss of genetic material.
- *Inversions* occur when there are two breaks within a single chromosome and the broken segment flips 180° (inverts) and reattaches to form a chromosome that is structurally out-of-sequence. There is usually no risk for problems

to an individual if the inversion is of *familial* origin (has been inherited from a parent.) There is a slightly increased risk if it is a *de novo* (new) mutation due possibly to an interruption of a key gene sequence.

Although an inversion carrier may be completely normal, they are at a slightly increased risk for producing a chromosomally *unbalanced* embryo. This is because an inverted chromosome has difficulty pairing with it's normal homolog during meiosis, which can result in gametes containing unbalanced derivative chromosomes if an unequal cross-over event occurs.

- *Translocations* involve exchange of material between two or more chromosomes. If a translocation is *reciprocal* (balanced) the risk for problems to an individual is similar to that with inversions: usually none if *familial* and slightly increased if *de novo*.

Problems arise with translocations when gametes from a balanced parent are formed which do not

contain both translocation products. When such a gamete combines with a normal gamete from the other parent the result is an *unbalanced* embryo which is partially monosomic for one chromosome and partially trisomic for the other.

Numerical and structural abnormalities can be further divided into two main categories:

a. *constitutional*, those you are born with; and
b. *acquired*, those that arise as secondary changes to other diseases such as cancer.

Sometimes individuals are found who have both normal and abnormal cell lines. These people are called *mosaics* and in the vast majority of these cases the abnormal cell line has a numerical chromosome abnormality. Structural mosaics are extremely rare. The degree to which an individual is clinically affected usually depends on the percentage of abnormal cells.

These are just some of the more common abnormalities encountered by a Cytogenetic Laboratory. Because the number of abnormal possibilities is almost infinite, a

Cytogeneticist must be trained to detect and interpret virtually any chromosome abnormality that can occur.

5.4. Types of Chromatin

Chromatin is the name that describes nuclear material that contains the genetic code. In fact, the code is stored in individual units called "chromosomes". Two types of chromatin can be described as follows:

Heterochromatin

- This is the condensed form of chromatin organization.
- It is seen as dense patches of chromatin. Sometimes it lines the nuclear membrane, however, it is broken by clear areas at the pores so that transport is allowed. Sometimes, the heterochromatin forms a "cartwheel" pattern.
- Abundant heterochromatin is seen in resting, or reserve cells such as small lymphocytes (memory cells) waiting for exposure to a foreign antigen. Heterochromatin is considered transcriptionally inactive.

- Heterochromatin stains more strongly and is a more condensed chromatin.

Euchromatin

- Euchromatin is threadlike, delicate.
- It is most abundant in active, transcribing cells. Thus, the presence of euchromatin is significant because the regions of DNA to be transcribed or duplicated must uncoil before the genetic code can be read.
- Euchromatin stains weakly and is more open (less condensed).
- Euchromatin remains dispersed (uncondensed) during Interphase, when RNA transcription occurs.
- As the cell differentiates, the proportion of heterochromatin to euchromatin increases, reflecting increased specialization of the cell as it matures.

5.5. Codominant alleles

- Codominant alleles occur when rather than expressing an intermediate phenotype, the heterozygotes express both homozygous phenotypes.

- An example is in human ABO blood types, the heterozygote AB type manufactures antibodies to both A and B types. Blood Type A people manufacture only anti-B antibodies, while type B people make only anti-A antibodies.
- Codominant alleles are both expressed.
- Heterozygotes for codominant alleles fully express both alleles. Blood type AB individuals produce both A and B antigens. Since neither A nor B is dominant over the other and they are both dominant over O they are said to be codominant.

5.6. Incomplete dominance

- Incomplete dominance is a condition when neither allele is dominant over the other. The condition is recognized by the heterozygotes expressing an intermediate phenotype relative to the parental phenotypes.
- If a red flowered plant is crossed with a white flowered one, the progeny will all be pink. When pink is crossed with pink, the progeny are 1 red, 2 pink, and 1 white.

5.7. Multiple alleles

- Many genes have more than two alleles (even though any one diploid individual can only have at most two alleles for any gene), such as the ABO blood groups in humans, which are an example of multiple alleles.
- Multiple alleles result from different mutations of the same gene. Coat color in rabbits is determined by four alleles. Human ABO blood types are determined by alleles A, B, and O. A and B are codominants which are both dominant over O. The only possible genotype for a type O person is OO

5.8. Epistasis

- Epistasis is the term applied when one gene interferes with the expression of another (as in the baldness).
- It was reported that a different phenotypic ratio in sweet pea than could be explained by simple Mendelian inheritance. This ratio is 9:7 instead of the 9:3:3:1 one would expect of a dihybrid cross between heterozygotes. Of the two genes (C and P),

when either is homozygous recessive (cc or pp) that gene is epistatic to (or hides) the other. To get purple flowers one must have both C and P alleles present.

5.9. Environment and Gene Expression

- Phenotypes are always affected by their environment. In buttercup (*Ranunculus peltatus*), leaves below water-level are finely divided and those above water-level are broad, floating, photosynthetic leaf-like leaves.

- Expression of phenotype is a result of interaction between genes and environment. Siamese cats and Himalayan rabbits both animals have dark colored fur on their extremities. This is caused by an allele that controls pigment production being able only to function at the lower temperatures of those extremities. Environment determines the phenotypic pattern of expression.

5.10. Polygenic Inheritance

- Polygenic inheritance is a pattern responsible for many features that seem simple on the surface.
- Many traits such as height, shape, weight, color, and metabolic rate are governed by the cumulative effects of many genes.
- Polygenic traits are not expressed as absolute or discrete characters, as was the case with Mendel's pea plant traits. Instead, polygenic traits are recognizable by their expression as a gradation of small differences (a continuous variation). The results form a bell shaped curve, with a mean value and extremes in either direction.
- Height in humans is a polygenic trait, as is color in wheat kernels. Height in humans is not discontinuous. If you line up the entire class a continuum of variation is evident, with an average height and extremes in variation (very short (vertically challenged) and very tall [vertically enhanced]).
- Traits showing continuous variation are usually controlled by the additive effects of two or more

separate gene pairs. This is an example of polygenic inheritance. The inheritance of each gene follows Mendelian rules.

Usually polygenic traits are distinguished by

1. Traits are usually quantified by measurement rather than counting.
2. Two or more gene pairs contribute to the phenotype.
3. Phenotypic expression of polygenic traits varies over a wide range.

Human polygenic traits include

1. Height
2. Systemic Lupus Erythematus
3. Weight
4. Eye Color
5. Intelligence
6. Skin Color
7. Many forms of behavior

5.11. Pleiotropy

- Pleiotropy is the effect of a single gene on more than one characteristic. Examples :

1) The "frizzle-trait" in chickens. The primary result of this gene is the production of defective feathers. Secondary results are both good and bad; good include increased adaptation to warm temperatures, bad include increased metabolic rate, decreased egg-laying, changes in heart, kidney and spleen.

2) Cats that are white with blue eyes are often deaf, white cats with a blue and an yellow-orange eye are deaf on the side with the blue eye.

3) Sickle-cell anemia is a human disease originating in warm lowland tropical areas where malaria is common. Sickle-celled individuals suffer from a number of problems, all of which are pleiotropic effects of the sickle-cell allele.

5.12. Human Chromosome Abnormalities

- Chromosome abnormalities include inversion, insertion, duplication, and deletion. These are types of mutations.
- Since DNA is information, and information typically has a beginning point, an inversion would produce an inactive or altered protein. Likewise deletion or duplication will alter the gene product.
- A common abnormality is caused by nondisjunction, the failure of replicated chromosomes to segregate during Anaphase II. A gamete lacking a chromosome cannot produce a viable embryo. Occasionally a gamete with n+ 1 chromosome can produce a viable embryo.
- In humans, nondisjunction is most often associated with the 21st chromosome, producing a disease known as Down's syndrome (also referred to as trisomy 21).
- Sufferers of Down's syndrome suffer mild to severe mental retardation, short stocky body type, large tongue leading to speech difficulties, and (in those

who survive into middle-age), a propensity to develop Alzheimer's Disease.

- Ninety-five percent of Down's cases result from nondisjunction of chromosome 21. Occasional cases result from a translocation in the chromosomes of one parent.
- Remember that a translocation occurs when one chromosome (or a fragment) is transferred to a non-homologous chromosome. The incidence of Down's Syndrome increases with age of the mother, although 25% of the cases result from an extra chromosome from the father.

Sex-chromosome abnormalities may also be caused by nondisjunction of one or more sex chromosomes. Any combination (up to XXXXY) produces maleness. Males with more than one X are usually underdeveloped and sterile. XXX and XO women are known, although in most cases they are sterile.

- Chromosome deletions may also be associated with other syndromes such as Wilm's tumor.
- Prenatal detection of chromosomal abnormalities is accomplished chiefly by amniocentesis. A thin

needle is inserted into the amniotic fluid surrounding the fetus (a term applied to an unborn baby after the first trimester). Cells are withdrawn have been sloughed off by the fetus, yet they are still fetal cells and can be used to determine the state of the fetal chromosomes, such as Down's Syndrome and the sex of the baby after a karyotype has been made.

5.12.1. Human Allelic Disorders (Recessive)

Albinism, the lack of pigmentation in skin, hair, and eyes, is also a Mendelian human trait.

- Homozygous recessive (aa) individuals make no pigments, and so have face, hair, and eyes that are white to yellow.
- For heterozygous parents with normal pigmentation (Aa), two different types of gametes may be produced: A or a. From such a cross 1/4 of the children could be albinos. The brown pigment melanin cannot be made by albinos. Several mutations may cause albinism:

1) the lack of one or another enzyme along the melanin-producing pathway; or

2) the inability of the enzyme to enter the pigment cells and convert the amino acid tyrosine into melanin.

Phenylketonuria (PKU) is recessively inherited disorder whose sufferers lacks the ability to synthesize an enzyme to convert the amino acid phenylalanine into tyrosine Individuals homozygous recessive for this allele have a buildup of phenylalanine and abnormal breakdown products in the urine and blood.

- The breakdown products can be harmful to developing nervous systems and lead to mental retardation. 1 in 15,000 infants suffers from this problem. PKU homozygotes are now routinely tested for in most states. If you look closely at a product containing Nutra-sweet artificial sweetener, you will see a warning to PKU sufferers since phenylalanine is one of the amino acids in the sweetener. PKU sufferers are placed on a diet low in phenylalanine, enough for metabolic needs but not enough to cause the buildup of harmful intermediates.

Tay-Sachs Disease is an autosomal recessive resulting in degeneration of the nervous system. Symptoms manifest after birth. Children homozygous recessives for this allele rarely survive past five years of age.

- Sufferers lack the ability to make the enzyme N-acetyl-hexosaminidase, which breaks down the GM2 ganglioside lipid. This lipid accumulates in lysosomes in brain cells, eventually killing the brain cells.

Sickle-cell anemia is an autosomal recessive. Nine-percent of US blacks are heterozygous, while 0.2% are homozygous recessive.

- The recessive allele causes a single amino acid substitution in the beta chains of hemoglobin. When oxygen concentration is low, sickling of cells occurs.
- Heterozygotes make enough "good beta-chain hemoglobin" that they do not suffer as long as oxygen concentrations remain high, such as at sea-level.

5.12.2. Human Allelic Disorders (Dominant)

Autosomal dominants are rare, although they are (by definition) more commonly expressed.

Huntington's disease is an autosomal dominant resulting in progressive destruction of brain cells. If a parent has the disease, 50% of the children will have it (unless that parent was homozygous dominant, in which case all children would have the disease).

- The disease usually does not manifest until after age 30, although some instances of early onset phenomenon are reported among individuals in their twenties.

Polydactly is the presence of a sixth digit. In modern times the extra finger has been cut off at birth and individuals do not know they carry this trait..

- One of the wives of Henry VIII had an extra finger. In certain southern families the trait is also more common. The extra digit is rarely functional and definitely causes problems buying gloves, let alone fitting them on during a murder trial.

Muscular dystrophy is a term encompassing a variety of muscle wasting diseases.

- The most common type, Duchenne Muscular Dystrophy (DMD), affects cardiac and skeletal muscle, as well as some mental functions. DMD is an X-linked recessive occurring in 1 in 3500 newborns. Most sufferers die before their 20th birthday.

5.13. Cytogenetics

- Cytogenetics is the study of chromosomes and the related disease states caused by abnormal chromosome number and/or structure.
- Cytogenetics involves the study of human chromosomes in health and disease.
- Chromosome studies are an important laboratory diagnostic procedure in:
    - prenatal diagnosis,
    - certain patients with mental retardation and multiple birth defects,
    - patients with abnormal sexual development, and

- some cases of infertility or multiple miscarriages.
- the study and treatment of patients with malignancies and hematologic disorders.

- A variety of tissue types can be used to obtain chromosome preparations. Some examples include peripheral blood, bone marrow, amniotic fluid, and products of conception.
- Virtually all routine clinical Cytogenetic analyses are done on chromosome preparations that have been treated and stained to produce a banding pattern specific to each chromosome. This allows for the detection of subtle changes in chromosome structure.
- Although specific techniques differ according to the type of tissue used, the basic method for obtaining chromosome preparations is as follows:

    - Sample log-in and initial setup.
    - Tissue culture (feeding and maintaining cell cultures).
    - Addition of a mitotic inhibitor to arrest cells at metaphase. Harvest cells. This step is very

important in obtaining high quality preparations. It involves exposing the cells to a hypotonic solution followed by a series of fixative solutions. This causes the cells to expand so the chromosomes will spread out and can be individually examined.

- Stain chromosome preparations to detect possible numerical and structural changes.
- The most common staining treatment is called G-banding.
- A variety of other staining techniques are available to help identify specific abnormalities.
- Once stained metaphase chromosome preparations have been obtained they can be examined under the microscope.
- Typically 15-20 cells are scanned and counted with at least 5 cells being fully analyzed.
- During a full analysis each chromosome is critically compared band-for-band with it's homolog. It is necessary to examine this many cells in order to detect clinically significant mosaicism.

- Following microscopic analysis, either photographic or computerized digital images of the best quality metaphase cells are made.
- Each chromosome can then be arranged in pairs according to size and banding pattern into a karyotype. The karyotype allows the Cytogeneticist to even more closely examine each chromosome for structural changes. A written description of the karyotype which defines the chromosome analysis is then made.

Review Questions

1. What is Chromatin? Euchromatin? Heterochromatin?
2. What functionally can Eucharomatin do that Heterochromatin can not do?
3. Why are Telomeres absent from prokaryotic chromosomes?
4. What is the basic structure of a Telomere? Correlate this structure with the 2 functions of a Telomere.
5. Discuss the different types of chromosome
6. List the different types chromosome abnormalities

CHAPTER SIX

LINKAGE

Specific learning objectives

At the end of this chapter, students are expected to

- ⇒ describe methods used for analysis of linkage
- ⇒ explain the mechanism underlying linkage
- ⇒ describe role of Linkage in genetic make up
- ⇒ describe how linkage between genes or between genes and markers can be established in human populations
- ⇒ discuss risk assessment of X-linked recessive, autosomal recessive and autosomal dominant disorders using linked markers.
- ⇒ discuss the limitations of a marker analysis.
- ⇒ describe genetic recombination and discuss its effects on genetic analysis and testing.

6.0. Introduction

- Linkage occurs when genes are on the same chromosome. When genes occur on the same chromosome, they are inherited as a single unit. Genes inherited in this way are called Linked.
- Genes are located on specific regions of a certain chromosome, termed the gene locus (plural: loci). A gene therefore is a specific segment of the DNA molecule.
- Linkage groups are invariably the same number as the pairs of homologous chromosomes an organism possesses. Recombination occurs when crossing-over has broken linkage groups, as in the case of the genes for wing size and body color that Morgan studied.
- Chromosome mapping was originally based on the frequencies of recombination between alleles.

In dihybrid testcrosses for frizzle and white in chickens, Hutt (1931) obtained:

> frizzled is dominant over normal (if one combines slightly and extremely frizzled).

white is dominant over colored.

P_1: White, Normal Colored, Frizzle

F_1: White, Frizzle

Testcross: White, Frizzle (F_1) x Coloured,Normal

Counts in testcross 1 (Hutt 1931)			
	White	Coloured	Total
Frizzled	18	63	81
Normal	63	13	76
	81	76	157

Note the marginal counts are in the 1:1 ratio we expect, but there is deviation in the main table from 1:1:1:1. This deviation is due to *linkage* between the two genes. The *percent recombination* is 100*(18+13)/157 = 19.7%. Under independent assortment the percent recombination should be 50%.

After mating another set of chickens of exactly the same genotypes however, the following counts were made,

Counts in testcross 2 (Hutt 1933)			
	White	Coloured	
Frizzled	15	2	17
Normal	4	12	16
	19	14	33

In the first testcross, the Frizzled and Coloured phenotypes seemed to cosegregate, but the reverse is seen in the second cross. This is what is referred to as *repulsion* of the dominant traits (frizzled and white) in the first case, and *coupling* in the second. The percentage deviation from 1:1:1:1 seems to be about the same in each table, but in opposite directions. Actually, we always ignore the sign, and calculate the recombination in this table as 100*(4+2)/33=18.2%.

If one examines a large number of genes in such a fashion in any organism, sets of genes are always linked together, while assorting independently (recombination 50%) with respect to members of other *linkage groups*. It

was realised in the 1920s that each linkage group corresponds to a chromosome.

6.1. Mapping

If one can arrange testcrosses for triple (or higher order) heterozygotes and recessives (a *three-point cross*), the recombination can be calculated for the three pairs of genes. The data will look like this example:

Trait A is controlled by a gene with alleles **A** and **a**, **A** dominant to **a**

Trait B is controlled by a gene with alleles **B** and **b**, **B** dominant to **b**

Trait C is controlled by a gene with alleles **C** and **c**, **C** dominant to **c**

Testcross is AaBbCc x abc/abc

Data from three-point cross of corn (colourless, shrunken, waxy) due to Stadler.

	Progeny Phenotype	Count
1	A B C	17959
2	a b c	17699
3	A b c	509
4	a B C	524
5	A B c	4455
6	a b C	4654
7	A b C	20
8	a B c	12
	Total tested	45832

The table deviates drastically from the expected 1:1:1:1:1:1:1:1, so linkage is being observed.

ABC and *abc* are the two commonest phenotypes, and are "reciprocal classes", so the heterozygote parent's phase was **ABC**/**abc**, rather than **AbC**/**aBc** etc. Recombination events between **A** and **B** are calculated from the marginal **AB** table and so forth,

Marginal AB table created by collapsing across the two levels of C.			
	A	a	
B	22414	536	22950
b	529	22353	22882
	22943	22889	45832

c_{AB} = 100*(529+536)/45832 = 2.3% c_{AC} c_{AB} + c_{BC}

c_{BC} = 100*(4455+12+4654+20)/45832 = 19.9%

c_{AC} = 100*(509+4455+524+4654)/45832 = 22.1%

When similar experiments are carried out involving larger numbers of loci from the same linkage group, it becomes obvious that the set of pairwise recombination percentages suggest strongly that the genes are ordered in a linear fashion, with recombination acting as the distance between them.

The *linkage map* that one constructs using *recombination distance* turns out to correspond to the *physical map* of genes along the linear structure of the chromosome. Recombination is the "phenotypic" effect

of crossover or chiasma formation between homologous chromosomes, whereby they exchange segments of DNA.

Abbreviated linkage map of maize chromosome 9 (Brookhaven National Laboratory 1996).

Locus	**Coord**
csu95a	0.00
c1 colored aleurone1	27.90 * A
sh1 shrunken1	31.60 * B
bz1 bronze1	35.20
wx1 waxy1	55.30 * C
acp1 acid phosphatase1	64.30
sus1 sucrose synthase1	75.40
hsp18a 18 kda heat shock protein18a	78.00
csh2c(cdc2)	144.60

Positions on a linkage map are *loci*. Since "gene" can be taken to mean the different gene forms (alleles), or the factor controlling a phenotype, geneticists often refer to the latter as the locus *sh1*, rather than the gene *sh1*.

6.2. Double Crossovers

If all three markers are in the same linkage group, that is, on the same chromosome, then we can observe an ABC/abc undergoing two recombination events, one between A and B, and another between B and C, to give AbC/aBc. This is what is going on in cells 7 and 8 of the earlier example. If we had been looking only at dihybrid test cross data, then we would not be able to detect these *double recombinants*.

One notices that double recombinants are not very common, so the effect on the estimates of the percent recombination is not large. The corollary of this is that most chromosomes will experience only zero or one recombinants. The estimated double recombination rate does add to our estimate of the distance between the more distant loci (A and C in the example). Trow's formula states that: $c_{AC} = c_{AB}+c_{BC}-2c_{AB}c_{BC}$

6.3. Interference

The term *interference* refers to the fact that recombination seems to be suppressed close to a first

recombination event. The *coincidence coefficient* is the ratio of the observed number of double recombinants to the expected number.

For a given distance between two loci, one can estimate the number of double recombinants that one would expect. At a trivial level, imagine three loci, each 10% recombination distance apart. Then we would expect in 1% of cases that a double recombinant would occur (one in each interval). The rate of double recombinants is usually less than this expected value.

The expected frequency of double cross overs is thus is the product of the observed frquencies of the ingle crosses overs. Interferences is calculated as

$I = 1-c$

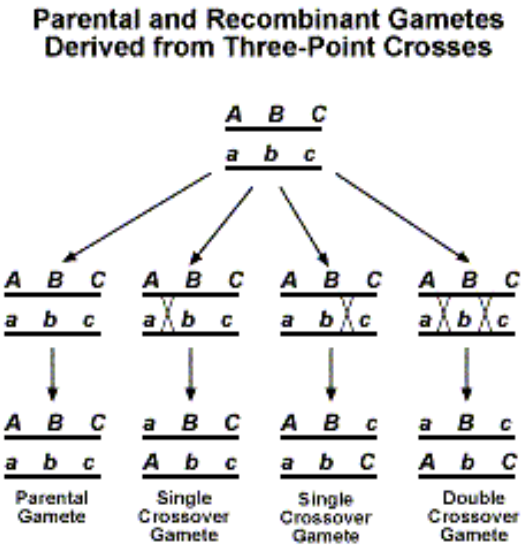
where, I = Index of interference

c = Coefficent of coincidence

$$c = \frac{\text{Observed frequency of Double Cross overs}}{\text{Expected frequency of Double Cross overs}}$$

6.4. Deriving Linkage Distance and Gene Order from Three-Point Crosses

By adding a third gene, we now have several different types of crossing over products that can be obtained. The following figure shows the different recombinant products that are possible.



- Now if we were to perform a testcross with F_1, we would expect a 1:1:1:1:1:1:1:1 ratio. As with the two-point analyzes described above, deviation from this expected ratio indicates that linkage is occurring.

- The best way to become familiar with the analysis of three-point test cross data is to go through an example. We will use the arbitrary example of genes *A*, *B*, and *C.*
- First make a cross between individuals that are *AABBCC* and *aabbcc*. Next the F_1 is testcrossed to an individual that is *aabbcc*.
- We will use the following data to determine the gene order and linkage distances. As with the two-point data, we will consider the F_1 gamete composition.

Genotype	**Observed**	**Type of Gamete**
ABC	390	Parental
abc	374	Parental
AbC	27	Single-crossover between genes *C* and *B*
aBc	30	Single-crossover between genes *C* and *B*
ABc	5	Double-crossover
abC	8	Double-crossover
Abc	81	Single-crossover between genes *A* and *C*
aBC	85	Single-crossover between genes *A* and *C*
Total	**1000**	

- The best way to solve these problems is to develop a systematic approach.
- First, determine which of the the genotypes are the parental gentoypes. The genotypes found most frequently are the parental genotypes. From the table it is clear that the *ABC* and *abc* genotypes were the parental genotypes.
- Next we need to determine the order of the genes. Once we have determined the parental genotypes, we use that information along with the information obtained from the double-crossover. The double-crossover gametes are always in the lowest frequency. From the table the *ABc* and *abC* genotypes are in the lowest frequency.
- The next important point is that a double-crossover event moves the middle allele from one sister chromatid to the other.
- This effectively places the non-parental allele of the middle gene onto a chromosome with the parental alleles of the two flanking genes.
- We can see from the table that the *C* gene must be in the middle because the recessive *c* allele is now on the same chromosome as the *A* and *B* alleles,

and the dominant *C* allele is on the same chromosome as the recessive *a* and *b* alleles.

- Now that we know the gene order is *ACB*, we can go about determining the linkage distances between *A* and *C*, and *C* and *B*.
- The linkage distance is calculated by dividing the total number of recombinant gametes into the total number of gametes. This is the same approach we used with the two-point analyses that we performed earlier. What is different is that we must now also consider the double-crossover events. For these calculations we include those double-crossovers in the calculations of both interval distances.
- So the distance between genes *A* and *C* is 17.9 cM [100*((81+85+5+8)/1000)], and the distance between *C* and *B* is 7.0 cM [100*((27+30+5+8)/1000)].
- Now let's try a problem from *Drosophila*, by applying the principles we used in the above example. The following table gives the results we will analyze.

Genotype	Observed	Type of Gamete
v cv$^+$ *ct*$^+$	580	Parental
v$^+$ *cv ct*	592	Parental
v cv ct$^+$	45	Single-crossover between genes *ct* and *cv*
v$^+$ *cv*$^+$ *ct*	40	Single-crossover between genes *ct* and *cv*
v cv ct	89	Single-crossover between genes *v* and *ct*
v$^+$ *cv*$^+$ *ct*$^+$	94	Single-crossover between genes *v* and *ct*
v cv$^+$ *ct*	3	Double-crossover
v$^+$ *cv ct*$^+$	5	Double-crossover
Total	**1448**	

Step 1: Determine the parental genotypes.

The most abundant genotypes are the partenal types. These genotypes are *v cv*$^+$ *ct*$^+$ and *v*$^+$ *cv ct*. What is different from our first three-point cross is that one parent did not contain all of the dominant alleles and the other all of the recessive alleles.

Step 2: Determine the gene order

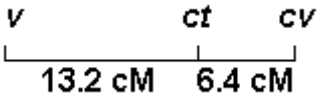
- To determine the gene order, we need the parental genotypes as well as the double crossover geneotypes.
- As we mentioned above, the least frequent genotypes are the double-crossover geneotypes. These geneotypes are *v cv*$^+$ *ct* and *v*$^+$ *cv ct*$^+$.
- From this information we can determine the order by asking the question: In the double-crossover genotypes, which parental allele is not associated with the two parental alleles it was associated with in the original parental cross.

- From the first double crossover, *v cv*$^+$ *ct*, the *ct* allele is associated with the *v* and *cv*$^+$ alleles, two alleles it was not associated with in the original cross. Therefore, *ct* is in the middle, and the gene order is *v ct cv.*

Step 3: Determing the linkage distances.

- *v* - *ct* distance caluculation. This distance is derived as follows: 100*((89+94+3+5)/1448) = 13.2 cM
- *ct* - *cv* distance calculation. This distance is derived as follows: 100*((45+40+3+5)/1448) = 6.4 cM

Step 4: Draw the map.



- Three-point crosses also allows one to measure **interference** (**I**) among crossover events within a given region of a chromosome. Specifically, the amount of double crossover gives an indication if interference occurs.
- The concept is that given specific recombination rates in two adjacent chromosomal intervals, the rate of double-crossovers in this region should be equal to the product of the single crossovers.
- In the *v ct cv* example described above, the recombination frequency was 0.132 between genes

v and *ct*, and the recombination frequency between *ct* and*cv* was 0.064.

- Therefore, we would expect 0.84% [100*(0.132 x 0.64)] double recombinants. With a sample size of 1448, this would amount to 12 double recombinants. We actually only detected 8.
- To measure interference, we first calculate the **coefficient of coincidence** (**c.o.c.**) which is the ratio of observed to expected double recombinants. Interference is then calculated as 1 - c.o.c. The formula is as follows:

$$I = 1 - c.o.c. = 1 - \left[\frac{\text{Observed \# of double recombinants}}{\text{Expected \# of double recombinants}}\right]$$

For the *v ct cv*data, the interference value is 33% [100*(8/12)].

- Most often, interference values fall between 0 and 1. Values less than one indicate that interference is occurring in this region of the chromosome.

Review Questions

1. How linkage occurs?

2. Why 2nd law of Mendel does not apply as a result of linkage?When analyzing a segregation ratio of phenotypes in one populaton, what result suggests that two genes are linked on the same chromosome?

3. Two genes can be coupling or repulsion phase on a parental chromosome. What is the difference between the two?

4. In Drosophila, b^+ is the allele for normal body color and at the same gene *b* is the allele for black body color. A second gene controls wing shape. The shape can be either normal (vg^+) or vestigial (*vg*). A cross is made between a homozygous wild type fly and fly with black body and vestigial wings. The offspring were then mated to black body, vesitigial winged flies. The following segregtion ratio was observed:

Phenotype	# Observed
Wild Type	405
Normal, vestigial	85
Black, normal	100
Black, vestigial	410

Are these two genes linked? How did you come to this conclusion? What calculation would you perform to confirm you conclusion?

5. What is the relationship between recombination frequency and genetic distance?

6. How do you recognize double cross progeny when analyzing the segregation data of three genes in a population?

7. How is linkage determined in humans? What information and assumptions are used in calculating linkage in humans?

CHAPTER SEVEN

PEDIGREE ANALYSIS

Specific learning objectives

At the end of this chapter, students are expected to

- ⇒ differentiate symbols used for human pedigree analysis
- ⇒ describe modes of inheritance
- ⇒ explain autosomal dominant and recessive
- ⇒ determine the type of Mendelian inheritance from a pedigree
- ⇒ describe features of patterns of inheritance seen in pedigrees
- ⇒ identify the recurrence risk for individuals in pedigrees.
- ⇒ describe the genetic basis of mitochondrial diseases and the expected inheritance patterns for mitochondrial traits;

⇒ describe the genetic mechanisms which result in uniparental disomy and the clinical consequences

7.1. Introduction

A pedigree is a diagram of family relationships that uses symbols to represent people and lines to represent genetic relationships. These diagrams make it easier to visualize relationships within families, particularly large extended families. Pedigrees are often used to determine the mode of inheritance (dominant, recessive, etc.) of genetic diseases.

If more than one individual in a family is afflicted with a disease, it is a clue that the disease may be inherited. A doctor needs to look at the family history to determine whether the disease is indeed inherited and, if it is, to establish the mode of inheritance. This information can then be used to predict recurrence risk in future generations.

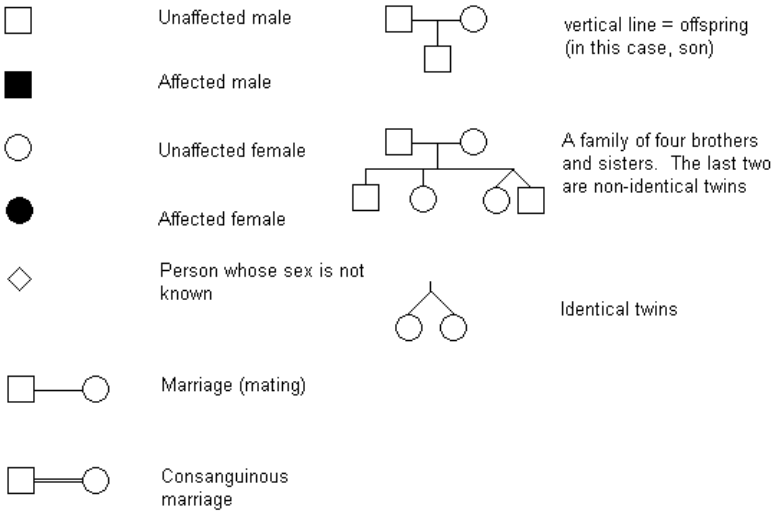
A basic method for determining the pattern of inheritance of any trait (which may be a physical attribute like eye color or a serious disease like Marfan

syndrome) is to look at its occurrence in several individuals within a family, spanning as many generations as possible. For a disease trait, a doctor has to examine existing family members to determine who is affected and who is not. The same information may be difficult to obtain about more distant relatives, and is often incomplete.

In a pedigree, squares represent males and circles represent females. Horizontal lines connecting a male and female represent mating. Vertical lines extending downward from a couple represent their children. Subsequent generations are therefore written underneath the parental generations and the oldest individuals are found at the top of the pedigree.

If the purpose of a pedigree is to analyze the pattern of inheritance of a particular trait, it is customary to shade in the symbol of all individuals that possess this trait.

Symbols Used to Draw Pedigree



- Generations are numbered from the top of the pedigree in uppercase. Roman numerals, I, II, III etc.
- Individuals in each generation are numbered from the left in Arab numerals as subscripts, III_1, III_2, III_3 etc.

7.2. Modes of inheritance

Most human genes are inherited in a Mendelian manner. It is usually unaware of the existence unless a variant form is present in the population which causes

an abnormal (or at least different) phenotype. One can follow the inheritance of the abnormal phenotype and deduce whether the variant allele is dominant or recessive.

Using genetic principles, the information presented in a pedigree can be analyzed to determine whether a given physical trait is inherited or not and what the pattern of inheritance is. In simple terms, traits can be either dominant or recessive.

A dominant trait is passed on to a son or daughter from only one parent. Characteristics of a dominant pedigree are:

1) Every affected individual has at least one affected parent;

2) Affected individuals who mate with unaffected individuals have a 50% chance of transmitting the trait to each child; and

3) Two affected individuals may have unaffected children.

Recessive traits are passed on to children from both parents, although the parents may seem perfectly "normal." Characteristics of recessive pedigrees are:

1) An individual who is affected may have parents who are not affected;

2) All the children of two affected individuals are affected; and

3) In pedigrees involving rare traits, the unaffected parents of an affected individual may be related to each other.

Penetrance and expressivity

Penetrance is the probability that a disease will appear in an individual when a disease-allele is present. For example, if all the individuals who have the disease-causing allele for a dominant disorder have the disease, the allele is said to have 100% penetrance. If only a quarter of individuals carrying the disease-causing allele show symptoms of the disease, the penetrance is 25%. Expressivity, on the other hand, refers to the range of symptoms that are possible for a given disease. For

example, an inherited disease like Marfan syndrome can have either severe or mild symptoms, making it difficult to diagnose.

Non-inheritedtraits

Not all diseases that occur in families are inherited. Other factors that can cause diseases to cluster within a family are viral infections or exposure to disease-causing agents (for example, asbestos). The first clue that a disease is not inherited is that it does not show a pattern of inheritance that is consistent with genetic principles (in other words, it does not look anything like a dominant or recessive pedigree).

7.3. Autosomal dominant

A dominant condition is transmitted in unbroken descent from each generation to the next. Most mating will be of the form M/m x m/m, i.e. heterozygote to homozygous recessive. Therefore, it is expected that every child of such a mating to have a 50% chance of receiving the mutant gene and thus of being affected. A typical pedigree might look like this (Figure 3.2):

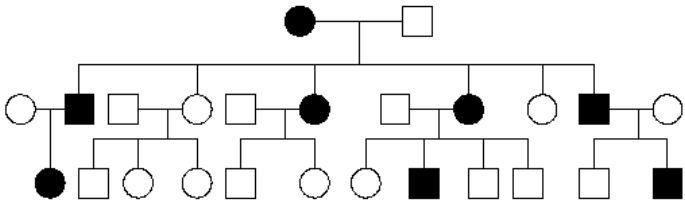


Fig.15. A typical pedigree

Examples of autosomal dominant conditions include:

- *Tuberous sclerosis*,

- *neurofibromatosis* and many other cancer causing mutations such as *retinoblastoma*

7.4. Autosomal recessive

A recessive trait will only manifest itself when homozygous. If it is a severe condition it will be unlikely that homozygotes will live to reproduce and thus most occurences of the condition will be in matings between two heterozygotes (or carriers).

An autosomal recessive condition may be transmitted through a long line of carriers before, by ill chance two carrier's mate. Then there will be a ¼ chance that any

child will be affected. The pedigree will therefore often only have one 'sibship' with affected members.

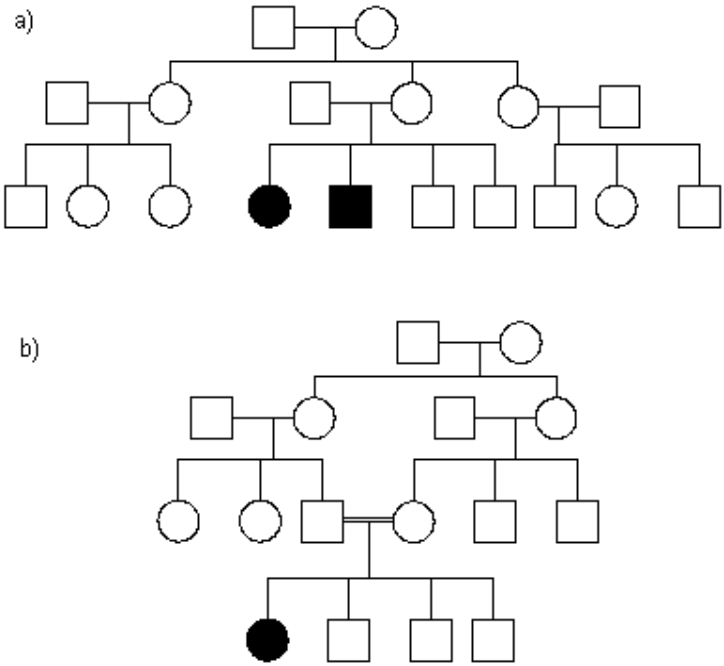


Fig.16. a) A 'typical' autosomal recessive pedigree, and b) an autosomal pedigree with inbreeding:

If the parents are related to each other, perhaps by being cousins, there is an increased risk that any gene present in a child may have two alleles identical by

descent. The degree of risk that both alleles of a pair in a person are descended from the same recent common ancestor is the degree of inbreeding of the person.

Let us examine b) in the figure above. Considering any child of a first cousin mating, one can trace through the pedigree the chance that the other allele is the same by common descent.

Let us consider any child of generation IV, any gene which came from the father, III_3 had a half chance of having come from grandmother II_2, a further half chance of being also present in her sister, grandmother II_4 a further half a chance of having been passed to mother III_4 and finally a half chance of being transmitted into the same child we started from. A total risk of ½ x ½ x ½ x ½ = 1/16

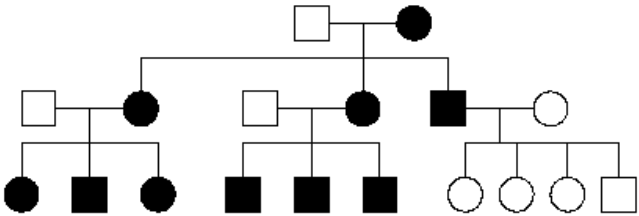


Fig.17. Maternal and paternal alleles and their breeding

This figure, which can be thought of as either

- the chance that both maternal and paternal alleles at one locus are identical by descent, or

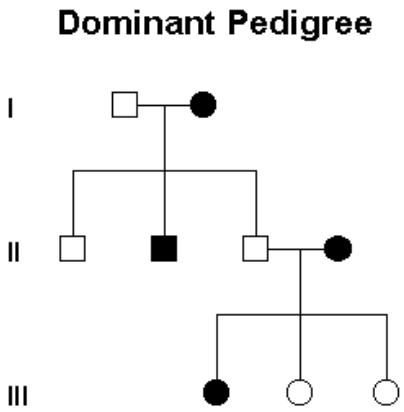
- the proportion of all the individual's genes that are homozygous because of identity by common descent, is known as the coefficient of inbreeding and is usually given the symbol F.

- Once phenotypic data is collected from several generations and the pedigree is drawn, careful analysis will allow you to determine whether the trait is dominant or recessive. Here are some rules to follow.

For those traits exhibiting dominant gene action:

- affected individuals have at least one affected parent
- the phenotype generally appears every generation
- two unaffected parents only have unaffected offspring

The following is the pedigree of a trait contolled by dominant gene action.

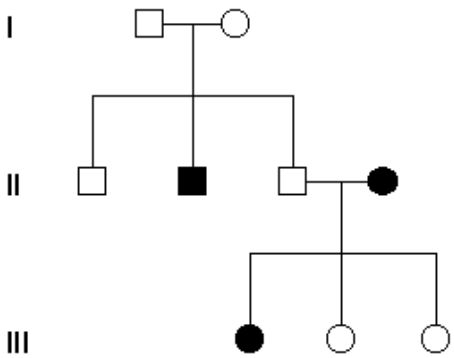


And for those traits exhibiting recessive gene action:

- unaffected parents can have affected offspring
- affected progeny are both male and female

The following is the pedigree of a trait contolled by recessive gene action.

Recessive Pedigree



7.5. Mitochondrial inheritance

Mitochondria are cellular organelles involved in energy production and conversion. They have a small amount of their own mitochondrial DNA (mtDNA). Though it is a relatively small portion of our total DNA, it is still subject to mutation and several diseases associated with mutations in mtDNA have been found. The inheritance patterns of mtDNA are unique. Mitochondrial DNA is inherited maternally.

Each person inherits the mtDNA of their mother, but none of their father's. This is because the relatively large ovum has many copies of mitochondrial DNA but the sperm has very few and these are lost during fertilization. Due to this unique feature of mitochondrial DNA inheritance, there are some constraints on the inheritance patterns of mitochondrial DNA disorders. These include:

- All children of affected males will not inherit the disease.
- All children of affected females will inherit it.

- An example of this type of disease is Leber's optic atrophy, a progressive loss of vision in the central visual field due to degeneration of the optic nerve.

- There are relatively few human genetic diseases caused by mitochondrial mutations but, because of their maternal transmission, they have a very distinctive pattern of inheritance.
- A mitochondrial inheritance pedigree is that all the children of an affected female but none of the children of an affected male will inherit the disease.

7.6. Uniparental disomy

Although it is not possible to make a viable human embryo with two complete haploid sets of chromosomes from the same sex parent it is sometimes possible that both copies of a single chromosome may be inherited from the same parent (along with no copies of the corresponding chromosome from the other parent.)

Rare cases of cystic fibrosis (a common autosomal recessive disease) have occurred in which one parent was a heterozygous carrier of the disease but the second parent had two wild type alleles. The child had received two copies of the mutant chromosome 7 from the carrier parent and no chromosome 7 from the unaffected parent.

Review Questions

1. What is pedigree analysis?

2. List the possible modes of inheritance?

3. What is autosomal recessive?

4. What is autosomal dominant?

5. Why mitochondrial inheritance is maternal?

CHAPTER EIGHT

NUCLEIC ACID STRUCTURE AND FUNCTION

Specific Learning objectives:

At the end of this chapter students are expected to:

- ⇒ Know the general structure of nucleic acids
- ⇒ Understand the phosphodiester bonds that join nucleosides together to form polynucleotides. Relate the direction of writing a DNA sequence to the polarity of the DNA chain.
- ⇒ Know how to apply nomenclature and shorthand conventions for DNA and RNA to draw polynucleotide structures.
- ⇒ Know the major structural features of the Watson-Crick DNA double helix.
- ⇒ Relate the specificity of pairing of adenine with thymine and cytosine with guanine to the duplex

(double-stranded) structure of DNA and to its replication.

⇒ Describe features of DNA and RNA

8.0. Introduction

Although genes are composed of DNA, DNA is for the most part an information storage molecule. That information is released or realized through the process of gene expression (namely transcription, RNA processing, and translation).

The process of converting the information contained in a DNA segment into proteins begins with the synthesis of mRNA molecules containing anywhere from several hundred to several thousand ribonucleotides, depending on the size of the protein to be made. Each of the 100,000 or so proteins in the human body is synthesized from a different mRNA that has been transcribed from a specific gene on DNA.

Genes are typically thought of as encoding RNAs that in turn produce proteins, but some RNAs are functional

themselves (e.g. rRNA, tRNA, snRNAs); thus some genes only encode RNAs, not proteins.

The transcribed strand of DNA is sometimes called the positive, plus, or sense strand. The template strand for the mRNA is sometimes called the negative, minus, or antisense strand.

Because genes are found in both orientations within a chromosome, one strand of the chromosome is not the coding strand for all genes; terms such as “transcribed strand” make sense only on a local basis, when considering the DNA region immediately encoding a particular gene.

8.1. Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is the material of which genes are made. This had not been widely accepted until 1953 when J.D. Watson and F.H, Crick proposed a structure for DNA which accounted for its ability to self--replicate and to direct the synthesis of proteins. All living cells (both prokaryotic and eukaryotic) contain double stranded DNA as their genetic material.

DNA is composed of a series of polymerized nucleotides, joined by phosphodiester bonds between the 5' and 3' carbons of deoxyribose units. DNA forms a double helix with these strands, running in opposite orientations with respect to the 3' and 5' hydrozxy groups.

The double helix structure is stabilized by base pairing between the nucleotides, with adenine and thymine forming two hydrogen bonds, and cytosine and guanine forming three.

Base + Sugar = nucleoside

Base + Sugar + Phosphate Group = nucleotide

Attached to each sugar residue is one of the four essentially planar nitrogenic organic bases: Adenine A, Cytosine C, Guanine G, Thymine T, The plane of each base is essentially perpendicular to the helix axis. Encoded in the order of the bases along a strand is the hereditary information.

The two strands coil about each other so that all the bases project inward towards the helix axis. The two

strands are held together by hydrogen bonds linking each base projecting from one backbone to its complementary base projecting from another backbone. The base A always binds to T and C always binds to G. This complementary pairing allows DNA to serve as a template for its own replication.

Fig.17. Comparison of Thymidine and Uracil

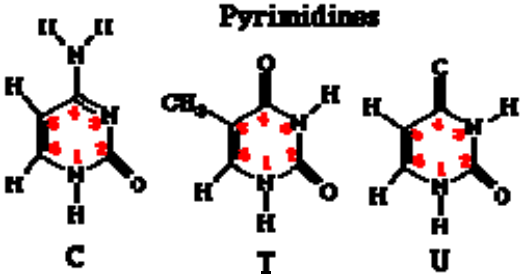


Fig.18. Comparison of Ribose and Deoxyribose sugars

Linking any two sugar residues is an -O--P--O-, a phosphate bridge between the 3' carbon atom of one of the sugars and the 5' carbon atom of the other sugar. Note that in solution DNA is negatively charged due to the presence of the phosphate group.

Because deoxyribose has an asymmetric structure, the ends of each strand of a DNA fragment are different. At one end the terminal carbon atom in the backbone is the 5' carbon atom of the terminal sugar (the carbon atom that lies outside the planar portion of the sugar); and at the other end it is the 3' carbon atom (one that lies within the planar portion of the sugar).

Double helical DNA in cells is an exceptionally long and stiff polymer. The winding and unwinding of the double helix for replication and transcription in the constrained intracellular space available to it, makes for the topological and energetic problems of DNA Supercoiling.

DNA in a circular form is often supercoiled. Negatively supercoiled DNA is in a more compact shape than

relaxed DNA and is partially unwound, facilitating interactions with enzymes such as polymerases.

Positive supercoiling results in the same space conservation as negative supercoiling, but makes DNA harder to work with. Negative supercoiling facilitates the separation of strands for replication, recombination, and transcription, and is therefore the preferred form for most natural DNA molecules.

Two enzymes work to maintain supercoiling in DNA:

1) Topoisomerases relax supercoiled DNA, and
2) DNA gyrase introduces supercoiling.

Topoisomerases are proteins which can catalyze the passage of one ("Type 1") or both ("Type II") DNA strands through a neighboring DNA segment, winding or unwinding DNA. Type II enzymes are unique in their ability to catenate and decatenate interlocked DNA circles.These enzymes are crucial for replication and segregation of chromosomes.

Topoisomerases work by cleaving one or both strands of DNA, passing a segment of DNA through the break,

and resealing the gap. The reaction to create supercoiled DNA requires an input in energy.

DNA gyrase uses the hydrolysis of ATP as a source of free energy for the insertion of negative supercoils in DNA. DNA is wrapped around the enzyme, and both strands are cleaved when ATP binds to the complex. As with the topoisomerase depicted above, the 5' ends remains bound to specific tyrosine residues of the enzyme, important so that any supercoils which are already present won't be lost.

This activity is an important process; several antibiotics exert their effects on this system, inhibiting prokaryotic enzymes more than eukaryotic ones. Novobiocin blocks ATP binding to DNA gyrase, while nalidixic acid and ciproflaxin interfere with the cleavage and joining of the strands.

8.2. Ribonucleic acid

RNA is similar to DNA but differs in several respects.
1. It is shorter
2. It is single stranded (with few exception: few virus)

3. It is nuclear and cytoplasmic
4. It has ribose
5. It has uracil rather than thymine. The other bases are the same.

There are three basic types of RNA:

1. *Messenger RNA (mRNA):* relatively long strands that encode the information from a single gene (DNA). It is the template for protein synthesis**.** This is the product of transcription. An mRNA is an RNA that is translated into protein. mRNAs are very short-lived compared to DNA.

► In prokaryotic cells a primary transcript is used directly as an mRNA (often times before it is even completely transcribed).

► In eukaryotic cells a primary transcript is **processed** before being exported from the nucleus as an mRNA:

- A **5'CAP** of 7-methyl guanosine is added.
- A **poly (A) tail** is added to the 3' end of the transcript.

- **Introns** (intervening sequences) must be cut from the transcript by a process known as **RNA splicing**.

In prokaryotes

- They are only around for a few minutes.
- Continuous synthesis of protein requires a continuous synthesis of mRNA. This helps the prokaryotic cell respond quickly to a fluctuating environment and fluctuating needs.
- The mRNA of prokaryotic cells is **polycistronic** (one transcript can code for several different proteins).

In eukaryotic cells

- The mRNA are stable for 4-24 hrs.
- The mRNA of eukaryotic cells is **monocistronic** (each transcript only encodes a single protein)

2. *ribosomal RNA (rRNA)*: Ribosomes are composed of rRNA and protein. The rRNA forms base pairs with the nucleotides of mRNA during translation (protein synthesis).

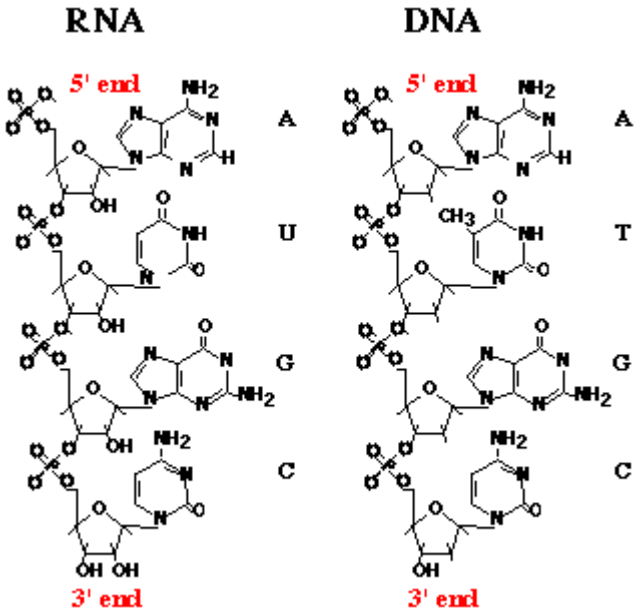
3. *transfer RNA (tRNA):* short (90 nucleotides) RNA molecules responsible for translating nucleic acid language to protein language. In other words the "adapter" molecule that converts nucleic acid sequence to protein sequence.

8.3. Chemical differences between DNA & RNA

- ► Both RNA and DNA are composed of repeated units. The repeating units of RNA are ribonucleotide monophosphates and of DNA are 2'-deoxyribonucleotide monophosphates.
- ► Both RNA and DNA form long, unbranched polynucleotide chains in which different purine or pyrimidine bases are joined by N-glycosidic bonds to a repeating sugar-phosphate backbone.
- ► The chains have a polarity. The sequence of a nucleic acid is customarily read from 5' to 3'. For example the sequence of the RNA molecule is AUGC and of the DNA molecule is ATGC
- ► The base sequence carries the information, i.e. the sequence ATGC has different information that AGCT even though the same bases are involved.

Consequences of RNA/DNA chemistry

- The DNA backbone is more stable, especially to alkaline conditions. The 2' OH on the RNA forms 2'3'phosphodiester intermediates under basic conditions which breaks down to a mix of 2' and 3' nucleoside monophosphates. Therefore, the RNA polynucleotide is unstable.
- The 2' deoxyribose allows the sugar to assume a lower energy conformation in the backbone. This helps to increase the stability of DNA polynucleotides.
- Cytidine deamination to Uridine can be detected in DNA but not RNA because deamination of Cytidine in DNA leads to Uridine not Thymidine. Uridine bases in DNA are removed by a specific set of DNA repair enzymes and replaced with cytidine bases.



- The role of DNA is long-term information storage. Thus DNA can be looked upon as a chemical information storage medium. All such media have certain common properties:
- The molecule must be able to carry information:
- The molecule must be able to hold information, without this property it is useless.
- The molecule must be readable:
- The information in the medium must be able to be used for some purpose. It is no use putting

information into a storage medium if the information cannot be retrieved.

- ► The molecule must be stable and secure:
- ► The information storage medium must be passed from generation to generation. Thus the molecule must be able to remain essentially unchanged for many generations.

- ► The role of RNA is three-fold: as a structural molecule, as an information transfer molecule, as an information decoding molecule
- ► RNA molecules read and interpret the information in DNA. RNA molecules are key players in the reactions that turn information into useful work.

8.4. DNA Replication

Before a cell can divide, it must duplicate its entire DNA. In eukaryotes, this occurs during S phase of the cell cycle.

The Steps:

- A portion of the double helix is unwound by a helicase.
- A molecule of a DNA polymerase binds to one strand of the DNA and begins moving along it in the 3' to 5' direction, using it as a template for assembling a leading strand of nucleotides and reforming a double helix. In eukaryotes, this molecule is called DNA polymerase delta (δ).

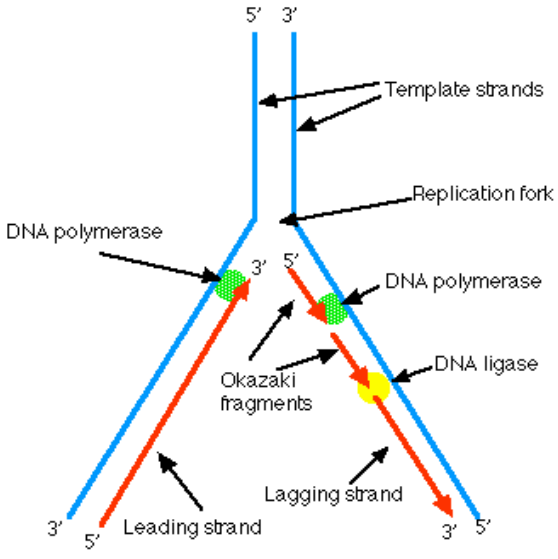


Fig.19. DNA Replication

- Because DNA synthesis can only occur 5' to 3', a molecule of a second type of DNA polymerase (epsilon, ε, in eukaryote) binds to the other template strand as the double helix opens. This molecule must synthesize discontinuous segments of polynucleotides (called Okazaki fragments). Another enzyme, DNA ligase I then stitches these together into the lagging strand.

When the replication process is complete, two DNA molecules — identical to each other and identical to the original — have been produced. Each strand of the original molecule has

- remained intact as it served as the template for the synthesis of
- a complementary strand.

This mode of replication is described as semi-conservative: one-half of each new molecule of DNA is old; one-half new. Watson and Crick had suggested that this was the way the DNA would turn out to be replicated.

8.4.1. Replication in Prokaryotes

The single molecule of DNA that is the E. coli genome contains 4.7 x 10^6 nucleotide pairs. DNA replication begins at a single, fixed location in this molecule, the replication origin, proceeds at about 1000 nucleotides per second, and thus is done in no more than 40 minutes. And thanks to the precision of the process (which includes a "proof-reading" function), the job is done with only about one incorrect nucleotide for every 10^9 nucleotides inserted. In other words, more often than not, the E. coli genome (4.7 x 10^6) is copied without error!

Replication strategies of bacteriophage

Studies of DNA replication in bacteriophage have been very valuable because of the insights that have been obtained into replication strategies, mechanisms and enzymology. However, the faithful and accurate replication of a genome is easily accomplished only if the genome is circular and is made of double-stranded DNA. If the genome is linear or if it is single-stranded or

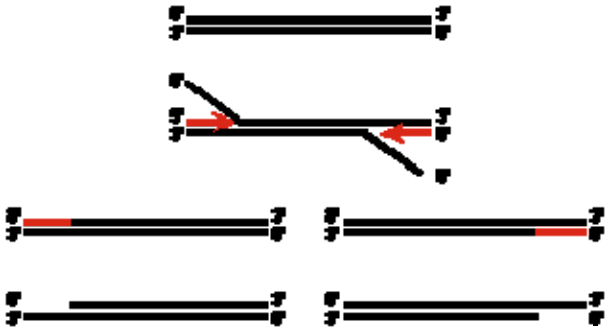
if it is made of RNA then special strategies are called for.

LINEAR GENOMES

If the genome is linear then it will progressively shorten with each round of replication unless the organism adopts a strategy for dealing with this problem.

The fact that DNA polymerase requires an RNA primer coupled with the fact that DNA polymerases are capable of synthesizing DNA only in the 5' -> 3' direction poses a critical problem for the replication of linear DNA molecules. Simply put - it is impossible to synthesize two exact copies of a parental molecule under these enzymological constraints. The problem boils down to one of **how do you fill in the ends?**

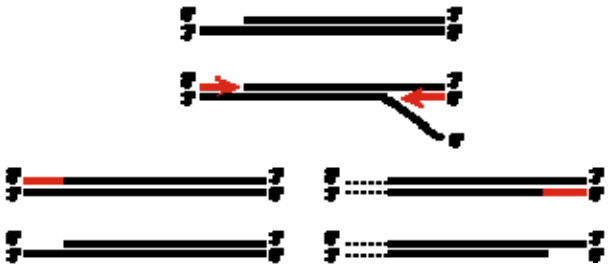
Consider the following:



Assume that replication of the linear molecule is initiated by the synthesis of RNA primers (red arrows) at each end. In the above example, these serve as primers for leading strand synthesis which copies each of the two parental strands. The end results are two daughter molecules each with an RNA-DNA hybrid polynucleotide chain (line 3).

However, when these RNA primers are removed, we are left with two molecules with single-stranded ends. These ends cannot be repaired or copied by DNA polymerase because of their polarity. Remember that no known DNA polymerase works in a 3' -> 5' direction.

If we now try and follow another round of replication using one of the original two daughters as the new parent, we see the full scope of the problem with replicating linear genomes:



After another round of replication, we recover one molecule that is identical to the parent but the other is not - it is shorter and has lost some genetic material.

So the problem with linear genomes is that they will progressively shorten with each round of replication unless some strategy is adopted to prevent this from occuring.

Two different strategies are described below for overcoming this problem: bacteriophage lambda

circularises its chromosome; bacteriophage T7 forms concatemers.

OTHER GENOMES

If the genome is either a ssDNA or an RNA genome then the organism must use special enymes or strategies or some combination of the two in order to replicate. An example is discussed below in the replication of bacteriophage M13 which has ssDNA genome.

Rolling Circle Replication

Replication via theta forms is not the only method by which circular molecules can replicate their genetic information. Another method is **rolling circle replication**, though it generates linear copies of a genome rather than circular copies.

Consider a circular molecule of double-stranded DNA with a nick in one of the two phosphodiester backbones. As long as there is a free 3' OH end, this can serve as a template for DNA polymerase. When the 3' OH end is extended, the 5' end can be displaced in a manner

analogous to the **strand displacement** reaction. Synthesis on this strand is also analogous to **leading strand** synthesis. The displaced strand can, in turn, serve as an template for replication as long as a suitable primer is available. Synthesis on this strand is analogous to **lagging strand** synthesis.

If synthesis continues in this manner, the consequence of this mechanism of replication can be the production of **concatemer** copies of the circular molecule. As a result, multiple copies of a genome are produced.

A rolling circle mode of replication is seen both during replication of **bacteriophage lambda** where rapid production of many copies of the genome is desired, and in the replication of **bacteriophage M13** where only a single copy is produced each time.

Replication of Bacteriophage Lambda

Bacteriophage lambda contains a linear dsDNA genome. However, the ends of the genomic DNA are single-stranded and are **cohesive**, i.e. they are

complementary to one another. The two cohesive ends - known as ***cos*** sites - are 12 nt in length.

After adsorption of the phage to the bacterial cell surface and injection into the cell, the chromosome circularizes by means of these complementary cohesive ends. This helps to protect it from degradation by bacterial exonucleases. Circularization is also an essential step if bacteriophage lambda chooses a lysogenic mode of growth.

Bacteriophage lambda replicates in two stages.

Early replication

Bacteriophage lambda initially replicates by means of **theta** form intermediates. The origin of replication (***ori***) is located within the **O** gene, whose product is required for replication. The gene **P** product is also required for replication.

The gene **O** product has a function analogous to that of **DnaA**. It binds to repeated sequences at the origin and initiates melting of the two strands nearby. The gene **P** product has a function analogous to that of **DnaC**. It

helps **DnaB** to bind to the "melted" DNA. Thereafter, the other components of a bacterial replisome can bind and replication ensues.

This mode of replication continues for 5 - 15 minutes after replication.

Late replication

After 15 minutes, bacteriophage lambda switches to replication by a rolling circle mechanism. It is not known what causes the switch from one mechanism to the other.

As concatemers are synthesized, they must be processed into linear molecules. This occurs by the action of **Terminase** which consists of two protein subunits coded by the lambda ***A*** and ***Nu1*** genes. Gene ***A*** codes for a 74 kDa protein; ***Nu1*** codes for a 21 kDa protein. **Terminase** recognizes the ***cos*** sites (in its double-stranded form) and cleaves them to generate new cohesive ends.

In vivo, processing of the concatemers also requires some of the other capsid proteins and there are length constraints on the amount of DNA that can be packaged. After the first ***cos*** site has been recognized, the second one must be located within 75% to 105% of the unit length of the phage chromosome.

The ability of the capsid to measure the amount of DNA that is packaged as well as recognizing specific sites is an important factor in the use of bacteriophage lambda as a cloning vector. Bacteriophage lambda derived cloning vectors can only be used to clone DNA fragments that are less than 15 kb in size (the actual size depends on the specific vector).

Replication of Bacteriophage M13

Bacteriophage M13 (and other filamentous phage like it) has a circular ssDNA molecule in the capsid. When the phage attaches to an *E. coli* cell, this molecule is injected into the cell where most of it is coated with **single-strand binding protein** (**SSB**). Since bacteriophageM13 does not code for its own DNA polymerase, it must use the host cell machinery in order

to replicate. It is, therefore, constrained by the requirements of the host cell replication machinery. Although its ssDNA genome is a perfect template for DNA synthesis, it is not such a suitable template either for RNA synthesis by either RNA polymerase or by primase.

However, although most of the genome is single-stranded, one part of it forms a double-stranded hairpin. This region somehow can serve as a promoter for the host cell RNA polymerase, which transcribes a short RNA primer. Transcription also disrupts the hairpin. DNA PolIII can then take over and synthesizes a dsDNA molecule.

This dsDNA molecule is known as **RFI** - **replicative form I**.

Further replication of **RFI** does not proceed by means of theta intermediates but by a type of rolling circle replication. The **gp2 endonuclease**, which is encoded by the phage gene 2, nicks the **RFI** DNA at a specific site (+ strand origin). **Rolling circle replication** now occurs with displacement of a single strand.

Concatemers are not formed; rather the gp2 endonuclease cleaves a second time after one complete copy has been synthesized. Thus the products of this one round of replication are a ssDNA circular molecule (the displaced strand - ligated into a circle) and a dsDNA **RFI** molecule. The circular ssDNA molecule can now be duplicated by repeating this entire sequence.

In order to synthesize the ssDNA strands that are to be packaged into the capsid, the displaced ssDNA molecules must be coated with a single strand binding protein - **gp5** - which is coded by the phage gene 5. These molecules are then packaged into new phage capsids.

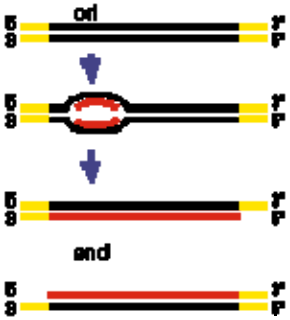
The fact that the life-cycle of filamentous phage such as M13 includes both a ssDNA phase and a dsDNA phase has been very useful for molecular biologists. Cloning vectors based on M13 have been developed which allow one to clone small DNA fragments and propagate them as phage particles. The dsDNA form permits routine cloning operations. The ssDNA form is ideally suited for the Sanger sequencing protocol and for many protocols for site-directed mutagenesis.

Replication of Bacteriophage T7

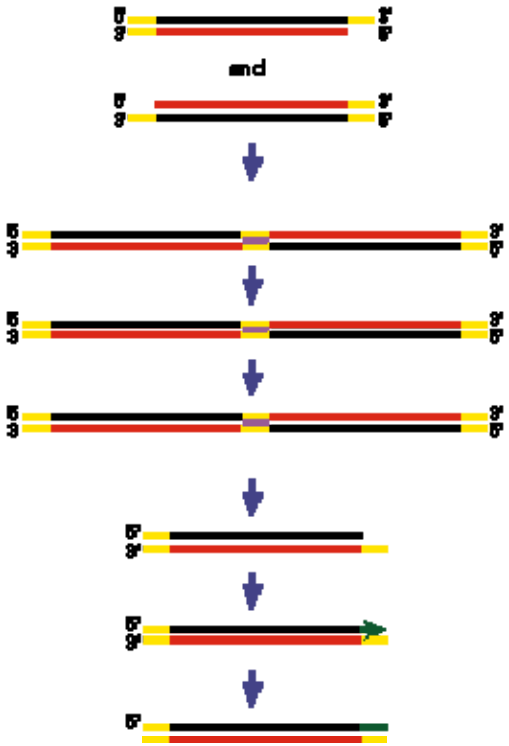
Bacteriophage T7 has a linear dsDNA genome, 39,937 bp in size. Replication initiates at a site located approx. 5900 bp from the left end of the phage and proceeds bidirectionally.

Although this specific case is different from the one drawn above, the problem is exactly the same. You should draw a diagram of the replication of T7 using pencil and paper. You will see that a bidirectional model of replication will not result in two complete daughter molecules in this case either.

The solution to the problem of replicating T7 lies in the left and right ends of the genome. The first 160 bp at the left end are identical with the final 160 bp at the right end. It is this **terminal redundancy** that is the key to its replication. The following cartoon shows the products of one round of replication. Although that the extent of the single-stranded region is identical to that of the terminal repeat in this picture, this does not need to be the case.



The ssDNA at the right end (3' end after synthesis) of one T7 chromosome is able to anneal with ssDNA at the left end (also a 3' end after synthesis) of another. The remaining gaps can then be filled by **DNA polymerase** and ligated by **DNA ligase**. The resulting dimeric molecule can be cleaved in two again - but this time generating two 5' overhangs on each daughter, which are now able to act as templates for normal 5' -> 3' synthesis:



T7 encodes its own **DNA ligase** (gene 1.3), **SSB** (gene 2.5) and **DNA polymerase** (gene 5).

8.4.2. Replication in Eukaryotes

► Eukaryotic DNA replication is clearly a much more complex process than bacterial DNA replication. We

discussed four aspects of eukaryotic DNA replication.

- ► In eukaryotes, the process of DNA replication is the same as that of the bacterial/prokaryotic DNA replication with some minor modifications. In eukaryotes, the DNA molecules are larger than in prokaryotes and are not circular; there are also usually multiple sites for the initiation of replication.
- ► Thus, each eukaryotic chromosome is composed of many replicating units or repliconsstretches of DNA with a single origin of replication. In comparison, the E. coli chromosome forms only a single replication fork. In eukaryotes, these replicating forks, which are numerous all along the DNA, form "bubbles" in the DNA during replication.
- ► The replication fork forms at a specific point called autonomously replicating sequences (ARS). The ARS contains a somewhat degenerate 11-bp sequences called the origin replication element (ORE). The ORE is located adjacent to an 80-bp AT rich sequence that is easy to unwind

- ► The average human chromosome contains 150 x 10^6 nucleotide pairs which are copied at about 50

base pairs per second. The process would take a month (rather than the hour it actually does) but for the fact that there are many places on the eukaryotic chromosome where replication can begin.

► Replication begins at some replication origins earlier in S phase than at others, but the process is completed for all by the end of S phase.

8.5. Control of Replication

When a cell in G_2 of the cell cycle is fused with a cell in S phase, the DNA of the G_2 nucleus does not begin replicating again even though replication is proceeding normally in the S-phase nucleus. Not until mitosis is completed, can freshly-synthesized DNA be replicated again.

Two control mechanisms have been identified — one positive and one negative. This redundancy probably reflects the crucial importance of precise replication to the integrity of the genome.

In order to be replicated, each origin of replication must be bound by:

- An Origin Recognition Complex of proteins (ORC). These remain on the DNA throughout the process.
- Accessory proteins called licensing factors. These accumulate in the nucleus during G_1 of the cell cycle. They include:
    - CDC-6 and CDT-1, which bind to the ORC and are essential for coating the DNA with
    - MCM proteins. Only DNA coated with MCM proteins (there are 6 of them) can be replicated.

- Once replication begins in S phase,
    - CDC-6 and CDT-1 leave the ORCs (the latter by ubiquination and destruction in proteasomes).
    - The MCM proteins leave in front of the advancing replication fork.

G_2 nuclei also contain at least one protein — called geminin — that prevents assembly of MCM proteins on freshly-synthesized DNA (probably by sequestering Cdt1).

As the cell completes mitosis, geminin is degraded so the DNA of the two daughter cells will be able to respond to licensing factors and be able to replicate their DNA at the next S phase.

Some cells deliberately cut the cell cycle short allowing repeated S phases without completing mitosis and/or cytokinesis. This is called endoreplication. How these cells regulate the factors that normally prevent DNA replication if mitosis has not occurred is still being studied.

8.6. DNA Ligation

Ligase requires precisely positioned 5'PO4 and 3'OH groups to catalyse phosphodiester bond formation.

The best and probably the most common physiological substrate for the reaction is a double helix with a single break in one of the two phosphodiester backbones in which free 5'-PO4 and 3'-OH groups are held in close proximity by stacked bases which remain hydrogen bonded to the intact DNA strand.

Free DNA ends can be ligated only when base stacking or both stacking and hydrogen bonding interactions can create transient pseudo-continuous DNA double helices with structures similar to that described above.

The stability, and therefore probability of finding such structures depends on the strength and number of interactions between the free ends. Complementary single stranded ends (cohesive ends) with either 5' or 3' overhangs such as those formed by the action of some restriction enzymes or the lambda terminase protein are better substrates than blunt ended molecules.

The probability of finding ligatable complexes between the free ends increases with increasing concentration of substrate when the ends in question are on different molecules. For unimolecular circularization reactions, the end-joining probabilities are determined by the length of the intervening DNA and are indepedent of concentration.

This cyclization probability can be thought of as an effective concentration of one end in the vicinity in the other. The effective concentration is equivalent to the

bulk concentration of ends when dimerization and cyclization are equally probable.

In Summary,

- DNA polymerases catalyze the synthesis of DNA.
- The reaction involves a nucleophilic attack by the 3'-hydroxyl group on the innermost phosphorous atom of the nucleotide triphosphate. Pyrophosphate is the leaving group.
- The synthesis reaction occurs in the 5'->3' direction - new bases are added at the 3' end of the growing chain.
- DNA polymerase I has three activities:
    - 5'->3' DNA synthesis
    - 3'->5' exonuclease - used as an error corrector to check the last base of the chain to ensure accuracy.
    - 5'->3' nuclease - used to remove bases (especially the RNA primer) ahead of synthesis occurring in the same direction.

- DNA synthesis occurs at replication forks. Synthesis begins at an origin of replication and proceeds in a bidirectional manner.
- DNA polymerase III is the primary enzyme for DNA replication in *E. coli*.
- Leading the synthesis is the helicase enzyme, which unwinds the DNA strands. This introduces positive supercoils into the DNA which must be relieved by DNA gyrase. The single stranded DNA is protected by binding to a single stranded binding protein.
- A primase synthesizes a short strand of RNA (about 5 nucleotides), because DNA polymerase requires a primer annealed to the template strand.
- The polymerase proceeds down the helix, directly synthesizing one strand in the 5'->3' direction - the leading strand. The other strand loops around and through the polymerase, and is synthesized in short, Okazaki fragments in the 5'->3' direction - the lagging strand.

- DNA polymerase I removes the RNA primers from the Okazaki fragments, replacing them with DNA.
- DNA ligase seals the breaks that are left after DNA polymerase I finishes.

Review Questions

1. What is a DNA nucleotide? What are its 3 components? What is the “backbone” of a DNA (or RNA) molecule?

2. 2.How do DNA strands join together to form a “double helix”? Which part of the DNA double helix is covalently bonded, and which is hydrogen-bonded? What is the consequence for DNA structure and function?

3. What are the 4 bases that make up the 4 nucleotides found in DNA? What is the base-pairing rule?

4. Who originally worked out the structure of the DNA molecule? Who provided the x-ray diffraction data, and who actually built the models?

5. What is semi-conservative DNA replication? How does it work to ensure that the new generation receives DNA molecules that are identical to the original ‘parents’? What is the main enzyme used to replicate (or synthesize) DNA?

6. Describe the 3 main mechanisms of DNA 'proofreading' and repair. How are they different, and how are they similar? What would be the effect of having no repair mechanisms? Relate this to a) the existence of genetic disorders; b) evolution.

7. What are the substrates and reaction mechanism for synthesis of 3'-5' phosphodiester bonds by DNA polymerase I (Pol I)?

8. What are the functions of the template and primer?

9. Outline the stps for bacteriophage replication

10. List the different methods used for bacteriophage replication

CHAPTER NINE

DNA DAMAGE, REPAIR AND MUTUATION

Specific Learning Objectives

At the end of this chapter students are expected

- ⇨ list causes of DNA damage
- ⇨ describe Mechanisms used to repair damage DNA
- ⇨ explain mutation

9.0. Introduction

DNA in the living cell is subject to many chemical alterations (a fact often forgotten in the excitement of being able to do DNA sequencing on dried and/or frozen specimens. If the genetic information encoded in the DNA is to remain uncorrupted, any chemical changes must be corrected. A failure to repair DNA produces a mutation.

The recent publication of the human genome has already revealed 130 genes whose products participate in DNA repair. More will probably be identified soon.

9.1. Agents that Damage DNA

- Certain wavelengths of radiation
    - ionizing radiation such as gamma rays and x-rays
    - Ultraviolet rays, especially the UV-C rays (~260 nm) that are absorbed strongly by DNA but also the longer-wavelength UV-B that penetrates the ozone shield .
- Highly-reactive oxygen radicals produced during normal cellular respiration as well as by other biochemical pathways.
- Chemicals in the environment
    - many hydrocarbons, including some found in cigarette smoke
    - some plant and microbial products, e.g. the aflatoxins produced in moldy peanuts
- Chemicals used in chemotherapy, especially chemotherapy of cancers

9.2. Types of DNA Damage

1. All four of the bases in DNA (A, T, C, G) can be covalently modified at various positions.
    - One of the most frequent is the loss of an amino group ("deamination") — resulting, for example, in a C being converted to a U.
2. Mismatches of the normal bases because of a failure of proofreading during DNA replication.
    - Common example: incorporation of the pyrimidine U (normally found only in RNA) instead of T.
3. Breaks in the backbone.
    - Can be limited to one of the two strands (a single-stranded break, SSB) or
    - on both strands (a double-stranded break (DSB).
    - Ionizing radiation is a frequent cause, but some chemicals produce breaks as well.
4. Crosslinks Covalent linkages can be formed between bases
    - on the same DNA strand ("intrastrand") or
    - on the opposite strand ("interstrand").

Several chemotherapeutic drugs used against cancers crosslink DNA

9.3. Repairing Damaged Bases

Damaged or inappropriate bases can be repaired by several mechanisms:

- Direct chemical reversal of the damage
- Excision Repair, in which the damaged base or bases are removed and then replaced with the correct ones in a localized burst of DNA synthesis. There are three modes of excision repair, each of which employs specialized sets of enzymes.
    1. Base Excision Repair (BER)
    2. Nucleotide Excision Repair (NER)
    3. Mismatch Repair (MMR)

9.3.1. Direct Reversal of Base Damage

Perhaps the most frequent cause of point mutations in humans is the spontaneous addition of a methyl group (CH_3-) (an example of alkylation) to Cs followed by

deamination to a T. Fortunately, most of these changes are repaired by enzymes, called glycosylases, that remove the mismatched T restoring the correct C. This is done without the need to break the DNA backbone (in contrast to the mechanisms of excision repair described below).

Some of the drugs used in cancer chemotherapy ("chemo") also damage DNA by alkylation. Some of the methyl groups can be removed by a protein encoded by our *MGMT* gene. However, the protein can only do it once, so the removal of each methyl group requires another molecule of protein.

This illustrates a problem with direct reversal mechanisms of DNA repair: they are quite wasteful. Each of the myriad types of chemical alterations to bases requires its own mechanism to correct. What the cell needs are more general mechanisms capable of correcting all sorts of chemical damage with a limited toolbox. This requirement is met by the mechanisms of excision repair.

9.3.2. Base Excision Repair (BER)

The steps and some key players:

1. removal of the damaged base (estimated to occur some 20,000 times a day in each cell in our body!) by a DNA glycosylase. We have at least 8 genes encoding different DNA glycosylases each enzyme responsible for identifying and removing a specific kind of base damage.
2. removal of its deoxyribose phosphate in the backbone, producing a gap. We have two genes encoding enzymes with this function.
3. replacement with the correct nucleotide. This relies on DNA polymerase beta, one of at least 11 DNA polymerases encoded by our genes.
4. ligation of the break in the strand. Two enzymes are known that can do this; both require ATP to provide the needed energy.

9.3.3. Nucleotide Excision Repair (NER)

NER differs from BER in several ways.

- It uses different enzymes.

- Even though there may be only a single "bad" base to correct, its nucleotide is removed along with many other adjacent nucleotides; that is, NER removes a large "patch" around the damage.

The steps and some key players:

1. The damage is recognized by one or more protein factors that assemble at the location.
2. The DNA is unwound producing a "bubble". The enzyme system that does this is Transcription Factor IIH, TFIIH, (which also functions in normal transcription).
3. Cuts are made on both the 3' side and the 5' side of the damaged area so the tract containing the damage can be removed.
4. A fresh burst of DNA synthesis — using the intact (opposite) strand as a template — fills in the correct nucleotides. The DNA polymerases responsible are designated polymerase delta and epsilon.
5. A DNA ligase covalent binds the fresh piece into the backbone.

Xeroderma Pigmentosum (XP): It is a rare inherited disease of humans which, among other things, predisposes the patient to

- pigmented lesions on areas of the skin exposed to the sun and
- an elevated incidence of skin cancer.

It turns out that XP can be caused by mutations in any one of several genes — all of which have roles to play in NER. Some of them:

- *XPA*, which encodes a protein that binds the damaged site and helps assemble the other proteins needed for NER.
- *XPB* and *XPD*, which are part of TFIIH. Some mutations in XPB and XPD also produce signs of premature aging. [Link]
- *XPF*, which cuts the backbone on the 5' side of the damage
- *XPG*, which cuts the backbone on the 3' side.

9.3.4. Mismatch Repair (MMR)

Mismatch repair deals with correcting mismatches of the normal bases; that is, failures to maintain normal Watson-Crick base pairing (A•T, C•G)

It can enlist the aid of enzymes involved in both base-excision repair (BER) and nucleotide-excision repair (NER) as well as using enzymes specialized for this function.

- Recognition of a mismatch requires several different proteins including one encoded by *MSH2*.
- Cutting the mismatch out also requires several proteins, including one encoded by *MLH1*.

Mutations in either of these genes predispose the person to an inherited form of colon cancer. So these genes qualify as tumor suppressor genes.

Synthesis of the repair patch is done by the same enzymes used in NER: DNA polymerase delta and epsilon.

Cells also use the MMR system to enhance the fidelity of recombination; i.e., assure that only homologous regions of two DNA molecules pair up to crossover and recombine segments (e.g., in meiosis).

9.4. Repairing Strand Breaks

Ionizing radiation and certain chemicals can produce both single-strand breaks (SSBs) and double-strand breaks (DSBs) in the DNA backbone.

Single-Strand Breaks (SSBs): Breaks in a single strand of the DNA molecule are repaired using the same enzyme systems that are used in Base-Excision Repair (BER).

Double-Strand Breaks (DSBs): There are two mechanisms by which the cell attempts to repair a complete break in a DNA molecule:

- Direct joining of the broken ends. This requires proteins that recognize and bind to the exposed ends and bring them together for ligating. They would prefer to see some complementary nucleotides but can proceed without them so this

type of joining is also called Nonhomologous End-Joining (NHEJ).

- Errors in direct joining may be a cause of the various translocations that are associated with cancers.
- Examples:
    - Burkitt's lymphoma
    - the Philadelphia chromosome in chronic myelogenous leukemia (CML)
    - B-cell leukemia

9.5. Mutations

In the living cell, DNA undergoes frequent chemical change, especially when it is being replicated (in S phase of the eukaryotic cell cycle). Most of these changes are quickly repaired. Those that are not result in a mutation. Thus, mutation is a failure of DNA repair.

Fig. 20. Hemoglobin sequence

9.5.1. Single-base substitutions

A single base, say an A, becomes replaced by another. Single base substitutions are also called point mutations. (If one purine [A or G] or pyrimidine [C or T] is replaced by the other, the substitution is called a transition. If a purine is replaced by a pyrimidine or vice-versa, the substitution is called a transversion.)

9.5.2. Missense mutations

With a missense mutation, the new nucleotide alters the codon so as to produce an altered amino acid in the protein product.

EXAMPLE: sickle-cell disease The replacement of A by T at the 17th nucleotide of the gene for the beta chain of hemoglobin changes the codon GAG (for glutamic acid) to GTG (which encodes valine). Thus the 6th amino acid in the chain becomes valine instead of glutamic acid. Another example: **Patient A** with **cystic fibrosis**.

9.5.3. Nonsense mutations

With a nonsense mutation, the new nucleotide changes a codon that specified an amino acid to one of the STOP codons (TAA, TAG, or TGA). Therefore, translation of the messenger RNA transcribed from this mutant gene will stop prematurely. The earlier in the gene that this occurs, the more truncated the protein product and the more likely that it will be unable to function.EXAMPLE: **Patient B.**

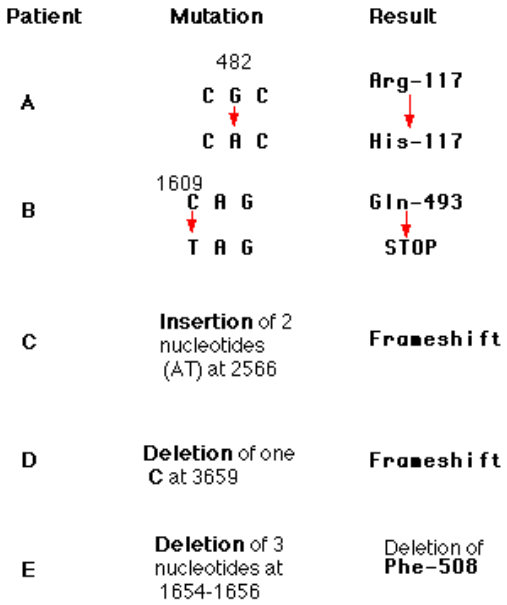


Fig.21 Effects of mutation

9.5.4. Silent mutations

Most amino acids are encoded by several different codons. For example, if the third base in the **TCT** codon for **serine** is changed to any one of the other three bases, serine will still be encoded. Such mutations are said to be silent because they cause no change in their product and cannot be detected without sequencing the gene (or its mRNA).

9.5.5. Splice-site mutations

The removal of intron sequences, as pre-mRNA is being processed to form mRNA, must be done with great precision. Nucleotide signals at the splice sites guide the enzymatic machinery. If a mutation alters one of these signals, then the intron is not removed and remains as part of the final RNA molecule. The translation of its sequence alters the sequence of the protein product.

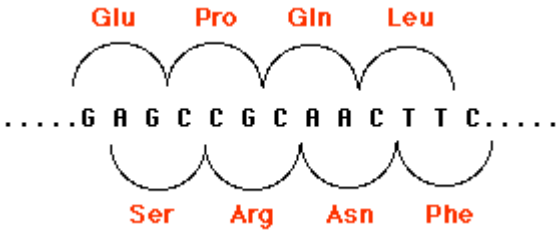


Fig.22 Frame shift

9.6. Insertions and Deletions (Indels)

Extra base pairs may be added (insertions) or removed (deletions) from the DNA of a gene. The number can range from one to thousands. Collectively, these mutations are called indels.

Indels involving one or two base pairs (or multiples thereof) can have devastating consequences to the gene because translation of the gene is "frameshifted". This figure shows how by shifting the reading frame one nucleotide to the right, the same sequence of nucleotides encodes a different sequence of amino acids. The mRNA is translated in new groups of three nucleotides and the protein specified by these new codons will be worthless. Scroll up to see two other examples (Patients C and D).

Frameshifts often create new stop codons and thus generate nonsense mutations. Perhaps that is just as well as the protein would probably be too garbled anyway to be useful to the cell.

Indels of three nucleotides or multiples of three may be less serious because they preserve the reading frame (see Patient E above).

However, a number of inherited human disorders are caused by the insertion of many copies of the same triplet of nucleotides. Huntington's disease and the fragile X syndrome are examples of such trinucleotide repeat diseases.

Fragile X Syndrome:Several disorders in humans are caused by the inheritance of genes that have undergone insertions of a stretch of identical codons repeated over and over. A locus on the human X chromosome contains such a stretch of nucleotides in which the triplet CGG is repeated (CGGCGGCGGCGG, etc.). The number of CGGs may be as few as 5 or as many as 50 without causing a harmful phenotype (these repeated nucleotides are in a noncoding region of the gene).

Even 100 repeats usually cause no harm. However, these longer repeats have a tendency to grow longer still from one generation to the next (to as many as 4000 repeats).

This causes a constriction in the X chromosome, which makes it quite fragile. Males who inherit such a chromosome (only from their mothers, of course) show a number of harmful phenotypic effects including mental retardation. Females who inherit a fragile X (also from their mothers; males with the syndrome seldom become fathers) are only mildly affected.

9. 7. Duplications

Duplications are a doubling of a section of the genome. During meiosis, crossing over between sister chromatids that are out of alignment can produce one chromatid with an duplicated gene and the other (not shown) having two genes with deletions. In the case shown here, unequal crossing over created a second copy of a gene needed for the synthesis of the steroid hormone aldosterone.

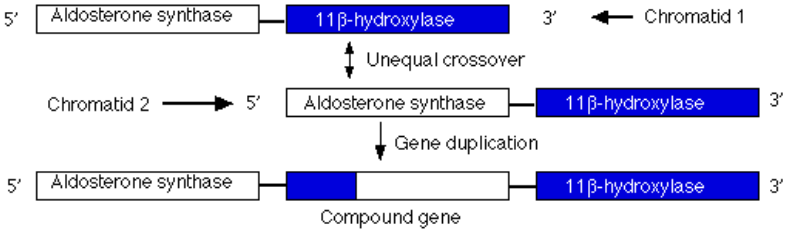


Fig. 23 Genome Duplication

However, this new gene carries inappropriate promoters at its 5' end (acquired from the 11-beta hydroxylase gene) that cause it to be expressed more strongly than the normal gene. The mutant gene is dominant: all members of one family (through four generations) who inherited at least one chromosome carrying this duplication suffered from high blood pressure and were prone to early death from stroke.

Gene duplication has occurred repeatedly during the evolution of eukaryotes. Genome analysis reveals many genes with similar sequences in a single organism. Presumably these paralogous genes have arisen by repeated duplication of an ancestral gene.

Such gene duplication can be beneficial.

- Over time, one of the duplicates can acquire a new function. This can provide the basis for adaptive evolution.
- But even while two paralogous genes are still similar in sequence and function, their existence provides redundancy ("belt and suspenders"). This may be a major reason why knocking out genes in yeast, "knockout mice", etc. so often has such a mild effect on the phenotype. The function of the knocked out gene can be taken over by a paralog.
- After gene duplication, random loss — or inactivation — of one of these genes at a later time in
    - one group of descendants
    - different from the loss in another group

  could provide a barrier (a "post-zygotic isolating mechanism") to the two groups interbreeding. Such a barrier could cause speciation: the evolution of two different species from a single ancestral species.

9.8. Translocations

Translocations are the transfer of a piece of one chromosome to a nonhomologous chromosome. Translocations are often reciprocal; that is, the two nonhomologues swap segments.

Translocations can alter the phenotype is several ways:

- the break may occur within a gene destroying its function
- translocated genes may come under the influence of different promoters and enhancers so that their expression is altered. The translocations in Burkitt's lymphoma are an example.
- the breakpoint may occur within a gene creating a hybrid gene. This may be transcribed and translated into a protein with an N-terminal of one normal cell protein coupled to the C-terminal of another. The Philadelphia chromosome found so often in the leukemic cells of patients with chronic myelogenous leukemia (CML) is the

result of a translocation which produces a compound gene (bcr-abl).

9.9. Frequency of Mutations

Mutations are rare events. This is surprising. Humans inherit 3 x 10^9 base pairs of DNA from each parent. Just considering single-base substitutions, this means that each cell has 6 billion (6 x 10^9) different base pairs that can be the target of a substitution.

Single-base substitutions are most apt to occur when DNA is being copied; for eukaryotes that means during S phase of the cell cycle.

No process is 100% accurate. Even the most highly skilled typist will introduce errors when copying a manuscript. So it is with DNA replication. Like a conscientious typist, the cell does proofread the accuracy of its copy. But, even so, errors slip through.

It has been estimated that in humans and other mammals, uncorrected errors (= mutations) occur at the rate of about 1 in every 50 million (5 x 10^7) nucleotides

added to the chain. But with 6 x 10^9 base pairs in a human cell, that mean that each new cell contains some 120 new mutations.

How can we measure the frequency at which phenotype-altering mutations occur? In humans, it is not easy.

- First we must be sure that the mutation is newly-arisen. (Some populations have high frequencies of a particular mutation, not because the gene is especially susceptible, but because it has been passed down through the generations from a early "founder".
- Recessive mutations (most of them are) will not be seen except on the rare occasions that both parents contribute a mutation at the same locus to their child.
- This leaves us with estimating mutation frequencies for genes that are inherited as
    - autosomal dominants
    - X-linked recessives; that is, recessives on the X chromosome which will be

expressed in males because they inherit only one X chromosome.

Some Examples (expressed as the frequency of mutations occurring at that locus in the gametes)

- Autosomal dominants
    - Retinoblastoma in the RB gene [Link]: about 8 per million (8×10^{-6})
    - Osteogenesis imperfecta in one or the other of the two genes that encode Type I collagen: about 1 per 100,000 (10^{-5})
    - Inherited tendency to polyps (and later cancer) in the colon. in a tumor suppressor gene (APC)~10^{-5}
- X-linked recessives
    - Hemophilia A ~3×10^{-5} (the Factor VIII gene)
    - Duchenne Muscular Dystrophy (DMD) >8×10^{-5} (the dystrophin gene) Why should the mutation frequency in the dystrophin gene be so much larger than most of the others? It's probably a matter of size. The dystrophin gene stretches over 2.3×10^{6}

base pairs of DNA. This is almost 0.1% of the entire human genome! Such a huge gene offers many possibilities for damage.

9.10. Measuring Mutation Rate

The frequency with which a given mutation is seen in a population (e.g., the mutation that causes cystic fibrosis) provides only a rough approximation of mutation rate — the rate at which fresh mutations occur — because of historical factors at work such as

- natural selection (positive or negative)
- drift
- founder effect

In addition, most methods for counting mutations require that the mutation have a visible effect on the phenotype. Thus

- mutations in noncoding DNA
- mutations that produce

    - synonymous codons (encode the same amino acid)
    - or, sometimes, new codons that encode a chemically-similar amino acid
- mutations which disrupt a gene whose functions are redundant; that is, can be compensated for by other genes will not be seen.

Review Questions

1. What are the causes of DNA damage

2. How damaged DNA is repaired

3. Which types of DNA damges are reversible

CHAPTER TEN
GENE TRANSFER IN BACTERIA

Specific LEARNING OBJECTIVES

At the end of this chapter student are expected

- ⇒ to describe the different methods of gene transfer
- ⇒ to explain the role of gene transfer for bacterial existence

10.0. Introdction

Gene transfer describes the introduction of genetic information into a cell from another cell. This process occurs naturally in both bacteria and eukaryotes, and may be termed horizontal genetic transmission to distinguish it from the trans formation of genetic information from parent to offspring, which is vertical genetic transmission.

Bacteria reproduce by the process of binary fission. In this process, the chromosome in the mother cell is replicated and a copy is allocated to each of the daughter cells. As a result, the two daughter cells are genetically identical. If the daughter cells are always identical to the mother, how are different strains of the same bacterial species created? The answer lies in certain events that change the bacterial chromosome and then these changes are passed on to future generations by binary fission.In this chapter, you will explore some of the events that result in heritable changes in the genome: genetic transfer and recombination, plasmids and transposons.

10.1. Conjugation

Bacterial coajugation involves the transfer of genetic information form one cell to another while the cells are in physical contact.

The ability to transfer DNA by conjugation is conferred by a conjugative plasmid, which is a self-transmissible

element which encodes all the functions required to transfer a copy of itself to another cell by conjugation.

The fertility (f) factor or transfer factor, is an extrachromosomal molecule that encodes the information necessary for conjugation.

Conjugation involves two cell types:

A. Donors, which posses the F-factor and reterrred to as F^+, and
B. Recipients, which lacks the F-factor and are referred to as F^-.

The R-factor contains the genes for the specialized pilus, called sex pilus, used in conjugation for other surface structures involved interactions with F^- cells.

- The f- factor is self transmissible once it is passed to an f=cell, the recipient cell becomes F^+ and is able to pass the fertility factor to another F^- cell. This is the means by which bacteria acquire multiple resistance to antibacterial agents.

- Bacteria with F-factor in plasmid form are called F^+.

Higher frequency recombinant (Hfr) cells

- When plasmid containing the F-factor is integrated into bacterial chromosome, the cells are referred to as Hfr cells. This because they facilitate high frequency of recombination between chromosomal marker of donor and recipient origin.

Hfr conjugation

Donors: Hfr bacteria perform as donors during conjugation. One strand of the chromosome copy is transferred to the recipient F^- cells, while the other strand remains in the Hfr cell. The donor remains unchanged genetically.

Recipients: F^- cells receive chromosomal fragments, the size which depends on the time conjugation allowed to persist. The limiting factor for gene transfer is the stability of the

bond between the sex pilus and the pilus receptor.

The recipients' cell of an Hfr conjugation usually remains F^-.

Transfer

Once contact has been established, the DNA of conjugative plasmid is mobilized (prepared for transfer). Only one strand of DNA is transferred. In the recipnent cell, the single- strand red DNA is used as a template to generate double stranded molecule.

In the donor cell, the remaining single strand also used as a template to replace the transferred strand. Conjugation is thus semi-conservative process.

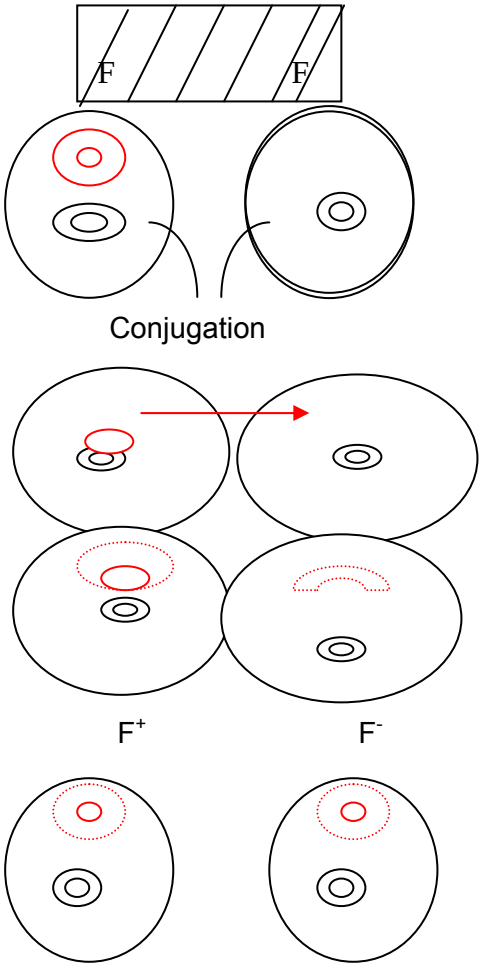


Fig.24. Gene transfer during conjugation between $F^+$and F^-. A conjugation bridge is formed between the donor and recipient cells. A single DNA strand is transferred to the recipient complementary DNA synthesis occurs on the strand remaining in the donor. Once the F-factor has been transferred, the cells separate.

One strand of F factor DNA moves in to recipient cell.

Complementary DNA synthesis of both strands of F-factor

Synthesis of complementary strands completed cells separated as F^+

F′ Factors

- F factor plasmid with some chromosomal material, which occur as a result of imprecise exicision, is called an F^1 factor.
- An F′ factor can conjugate. The recipient becomes F′, and information passed across is part chromosome and part plasmid.

10.2. Transformation

Transformation involves the uptake of naked DNA from the surrounding medium by a recipient cell and the recombination of genetic elements which change the genotype of the recipient cell.

Griffith's Experiment

- The transformation process was first demonstrated in 1928 by Frederick Griffith.
- Griffith experimented on *Streptococcus pneumoniae*, a bacteria that causes pneumonia in mammals.

- When he examined colonies of the bacteria on petri plates, he could tell that there were two different strains.
  - The colonies of one strain appeared smooth.
    - Later analysis revealed that this strain has a polysaccharide capsule and is virulent, that it, it causes pneumonia.
  - The colonies of the other strain appeared rough.
    - This strain has no capsules and is avirulent.
  - When Griffith injected living encapsulated cells into a mouse, the mouse died of pneumonia and the colonies of encapsulated cells were isolated from the blood of the mouse.
  - When living nonencapsulated cells were injected into a mouse, the mouse remained healthy and the colonies of nonencapsulated cells were isolated from the blood of the mouse.

- Griffith then heat killed the encapsulated cells and injected them into a mouse.
  - The mouse remained healthy and no colonies were isolated.
  - The encapsulated cells lost the ability to cause the disease.
- However, a combination of heat-killed encapsulated cells and living nonencapsulated cells did cause pneumonia and colonies of living encapsulated cells were isolated from the mouse.
  - How can a combination of these two strains cause pneumonia when either strand alone does not cause the disease?
  - If you guessed the process of transformation you are right!
  - The living nonencapsulated cells came into contact with DNA fragments of the dead capsulated cells.

- The genes that code for thr capsule entered some of the living cells and a crossing over event occurred.
- The recombinant cell now has the ability to form a capsule and cause pneumonia.
- All of the recombinant's offspring have the same ability.
- That is why the mouse developed pneumonia and died.

Regulation of transformation it depends on two variables:

- The competence of the recipient bacterium, &
- The qualities of the transforming DNA.

A. Competence is the ability of bacteria to take up DNA.

- Transformable bacteria become competent only under certain growth condition.

B. Qualities of transforming DNA

a. Homogeny

- In Gram- negative organisms, the specificity of the competence proteins is such that only

homologous or very similar DNA will be taken up by a competent bacteria.

- In gram-positive bacteria, the uptake is less restrictive, but if the DNA is not homologous it will not integrate fast enough and will e digested by end nucleases.

b. Double-strandenness

- Is required because one stranded is degraded as the other strand is brought in. degradation of one strand may provide the energy that is necessary for the entery of the surviving strand.

c.High mol.wt.($>10^7$)

- Increases the chance of integration, which may take place even if portions of the DNA are attacked by endonucleases before integration is completed.

Process of transfer

1. Revessible association of DNA to the cell wall is mediated by an ionic interaction between DNA and the cell wall of competent organism.

- This type of association occurs in bacteria all the time, but if the cell is not competent the association is tenuous and the DNA is released and adsorbed else where
- Artificial competence can be induced by treating bacteria with calcium chloride. Calcium chloride alters cell membrane permeability, enabling the uptake of DNA by cells that are normally incapable of DNA adsorption.
- Artificial competence allows transformation to be used as the basis for most recombinant DNA techniques.

2. Reversible association of the DNA and the inner cell membrane is established following transport of the DNA through the cell wall.
    1. Resistance to entracellular DNA are occurs as a consequence of conformational changes that take place after the DNA binds irreversibly to the cell membrane.
    2. Entry of DNA into cytoplasm
        - DNA enters the cytoplasm as a single strand

3. Integration in chromosomal DNA requires homology regions and involves displacement of one chromosomal srtand, recombination of the invading strand, elimination of the remaining chromosomal segment, and duplication of the invading strand.

Applications to bacterial gene mapping

- Transformation is a good chromosome mapping tool because transformed cells acquire different segments of DNA.
- By determining how frequently two given characterstics are simultaeoulsy acquired (the closer the genes, the most likely that both will be include in the same DNA piece), an idea about the location of corresponding genes in the chromosome is generated.

10.3. Transduction

- In transduction, DNA is transferred from one cell to another by means of bacterial viruses, also known as bacteriophages.

- Bacteriophage can interact with bacteria intwo ways:

1. Virulent (lytic) infection eventually destroys the host bacterium.
2. Template (lysogenic) infection is characterized by the integration of viral DNA into bacterial chromosome.

- The bacteria acquires a new set of genes: those of integrated phages (prophages)
- Transduction can occur in two ways:

1. Generalized transduction

- In generalized transduction, chromosomal or plasmid DNA accidentally become packaged into phage heads instead of the phage genome.
- Generalized transduction can be used for mapping the bacterial chromosome, following the same principles involved in mapping by transformation.

Properties of generalized transducing partice

a/ They carry all host DNA or plasmid DNA, but no phage DNA

b/ They can't replicate where these viruses infect another host cell, they inject purely

chromosomal DNA from their former hosts. They are no functional viruses, just vessel carry in a piece of bacterial DNA.

c/ The generalized trnasducing phage can carry any part of the host chromosome

2. Specialized transduction

- It takes place when a prophage contained in lysogenized bacterium replicates.
- Just as F^1 plasmid are generated, a specialized transducing virus is generated when the cutting enzyme make a mistake.

Properties of specialized transducing particie

1. They can't replicate
2. They carry hybrid DNA 9i.e. part phage DNA and part bacterial chromosome DNA (i.e part phage DNA and part bacterial chromosome DNA)

- Unlike generalized transduction, specialized transduction is not a good mapping tool.

10.4.Transposons

- Transposons (Transposable Genetic Elements) are pieces of DNA that can move from one location on the chromosome another, from plasmid to chromosome or vice versa or from one plasmid to another.
- The simplest transposon is an insertion sequence.
    - An insertion sequence contains only one gene that codes frotransposase, the enzyme that catalyzes transposition.
    - The transposase gene is flanked by two DNA sequences called inverted repeats because that two regions are upside-down and backward to each other.
- Transposase binds to these regions and cuts DNA to remove the gene.
- Yhe transposon can enter a number of locations.
    - When it invades a gene it usually inactivates the gene by interrupting the coding sequence and the protein that the gene codes for.

    - Luckil, transposition occurs rarely and is comparable to spontaneous mutation rates in bacteria.
- Complex transposons consist of one or more genes between two insertion sequences.
- The gene, coding for antibiotic resistance, for example, is carried along with the transposon as it inserts elsewhere.
- It could insert in a plasmid and be passed on to other bacteria by conjugation.

10.5. Recombination

- Genetic recombination refers to the exchange between two DNA molecules.
    - It results in new combinations of genes on the chromosome.
- You are probably most familiar with the recombination event known as crossing over.
    - In crossing over, two homologous chromosomes (chromosomes that contain the same sequence of genes but can have different alleles) break at corresponding points, switch fragments and rejoin.

  - The result is two recombinant chromosomes.
- In bacteria, crossing over involves a chromosome segment entering the cell and aligning with its homologous segment on the bacterial chromosome.
- The two break at corresponding point, switch fragments and rejoin.
- The result, as before, is two recombinant chromosomes and the bacteria can be called a recombinant cell.
- The recombinant pieces left outside the chromosome will eventually be degraded or lost in cell division.

10.6. Plasmids

- Plasmids are genetic elements that can also provides a mechanism for genetic change.
- Plasmids, as we discussed previously, are small, circular pieces of DNA that exist and replicate separately from the bacterial chromosome.
- We have already seen the importance of the F plasmid for conjugation, but other plasmids of equal importance can also be found in bacteria.
- One such plasmid is the R plasmid.

- Resistance or R plasmids carry genes that confer resistance to certain antibiotics. A R plasmid usually has two types of genes:
    1. R-determinant: resistance genes that code for enzymes that inactivate certain drugs
    2. RTF (Resistance Transfer Factor): genes for plasmid replication and conjugation.
- Without resistance genes for a particular antibiotic, a bacterium is sensitive to that antibiotic and probably destroyed by it.
- But the presence of resistance genes, on the other hand, allows for their transcription and translation into enzymes that make the drug inactive.
- Resistance is a serious problem. The widespread use of antibiotics in medicine and agriculture has lead to an increasingnumber of resistant strain pathogens.
- These bacteria survive in the presence of the antibiotic and pass the resistance genes on to future generations.
- R plasmids can also be transferred by conjugation from one bacterial cell to another, further increasing numbers in the resistant population.

Reviwe Questions

1. Compare and contrast the following terms

    - Conjugation
    - Transformation
    - Transduction
    - Transposition
    - Recombination

CHAPTER ELEVEN

TRANSCRIPTION AND TRANSLATION

Specific learning objectives

At the end of this chapter students are expected to

- ⇨ List the steps of transcription
- ⇨ Describe the central dogma
- ⇨ Enumerates properties of genetic code and how is it translated to protein
- ⇨ explain the major features of ribosome structure and function
- ⇨ describe the differences of prokaryotes and eukaryotes in terms of translation and transcription

11.0. Introduction

The majority of genes are expressed as the proteins they encode. The process occurs in two steps:

- **Transcription = DNA → RNA**
- **Translation = RNA → protein**

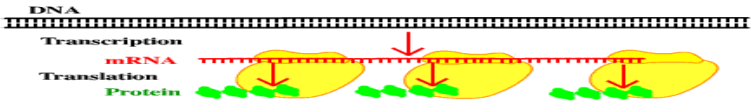


Fig. 25 Transcription and translation

Gene Transcription: DNA → RNA
DNA serves as the template for the synthesis of RNA much as it does for its own replication.

The Steps

- Some 50 different protein **transcription factors** bind to **promoter** sites, usually on the 5′ side of the gene to be transcribed.

- An enzyme, an **RNA polymerase**, binds to the complex of transcription factors.
- Working together, they open the DNA double helix.
- The RNA polymerase proceeds down one strand moving in the 3′ → 5′ direction.
- In eukaryotes, this requires — at least for protein-encoding genes — that the nucleosomes in front of the advancing RNA polymerase (RNAP II) be removed. A complex of proteins is responsible for this. The same complex replaces the nucleosomes after the DNA has been transcribed and RNAP II has moved on.
- As the RNA polymerase travels along the DNA strand, it assembles **ribo**nucleotides (supplied as triphosphates, e.g., ATP) into a strand of RNA.
- Each ribonucleotide is inserted into the growing RNA strand following the rules of base pairing. Thus for each C encountered on the DNA strand, a G is inserted in the RNA; for each G, a C; and for each T, an A. However, each A on the DNA guides the insertion of the pyrimidine uracil (**U**,

from uridine triphosphate, UTP). There is no T in RNA.

- Synthesis of the RNA proceeds in the 5′ → 3′ direction.
- As each nucleoside triphosphate is brought in to add to the 3′ end of the growing strand, the two terminal phosphates are removed.
- When transcription is complete, the transcript is released from the polymerase and, shortly thereafter, the polymerase is released from the DNA.

Note that at any place in a DNA molecule, either strand may be serving as the template; that is, some genes "run" one way, some the other (and in a few remarkable cases, the same segment of double helix contains genetic information on both strands!). In all cases, however, RNA polymerase proceeds along a strand in its 3′ → 5′ direction.

11.1 Transcription

Transcription is the synthesis of RNA from a DNA Template. A single gene (DNA) is transcribed using only

one of the two DNA strands, the coding strand. Its complement, the silent strand, is not used.

The two DNA strands of the gene move apart to provide access by RNA polymerase. This enzyme attaches to the initiation site at the 3' end of the coding strand of the gene (DNA). The enzyme moves along the coding strand, inserting the appropriate RNA nucleotides in place as dictated by the nucleotide sequence of the gene.

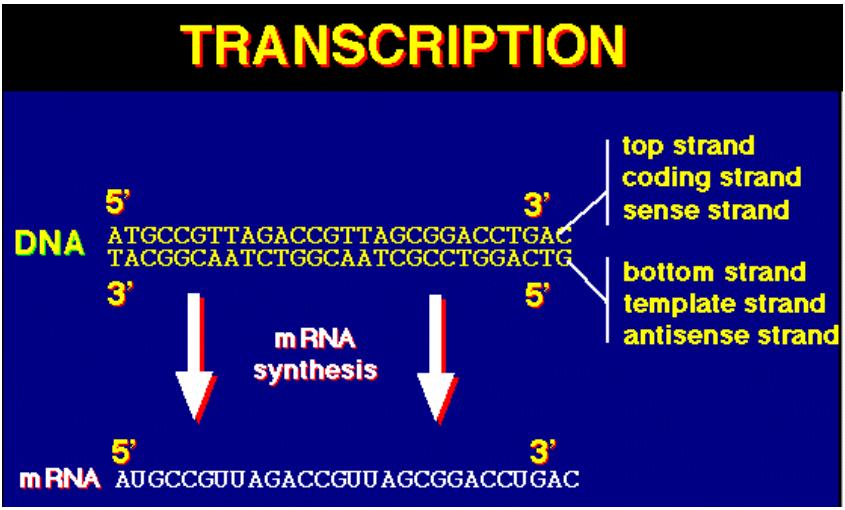


Fig.26 Transcription

Transcription consists of 3 steps:

1. **Initiation**: occurs at an initiation site at the 3' end of the gene. The initiation site is part of a larger promoter. Each gene has its own promoter. The promoter consists of a TATA box of about 100 nucleotides, mostly T and A, and an initiation sequence. The TATA box is upstream of the initiation sequence. Proteinaceous transcription factors attach to the promoter and help the polymerase find and attach to the initiation site. RNA polymerase attaches to the initiation sequence.

2. **Elongation**: RNA polymerase moves along, unwinding one turn of the double helix at a time thus exposing about 10 bases. New RNA nucleotides are added to the 3' end of the growing mRNA molecule at a rate of about 60 sec^{-1}. The double helix reforms behind the enzyme. Many RNA polymerase molecules can transcribe simultaneously (remember, the gene is hundreds of thousands of nucleotides long). Only one DNA strand, the coding strand, is transcribed. The other strand is not used (silent). But, which strand is coding and which silent varies

from gene to gene. The reading direction is 3' to 5' along the coding strand.

3. **Termination:** When the RNA polymerase reaches the 5' end of the coding strand of the gene it encounters a termination site, usually AATAAA. Here the enzyme falls off the coding strand and releases the pre mRNA strand (which must now be modified).

This has produced a strand of pre mRNA which contains many areas of nonsense known as introns interspersed between useful areas known as exons. This pre RNA must be modified to remove the introns. These nonsense areas are faithful transcriptions of similar nonsense areas of the gene (DNA). Most genes have introns, but why is not known.

11.2. Translation

Translation is the process of converting the mRNA codon sequences into an amino acid sequence.Genes determine phenotype through the synthesis of proteins.The next step is translation in which the

nucleotide sequence of the mRNA strand is translated into an amino acid sequence. This is accomplished by tRNA and ribosomes. The amino acid sequence is encoded in the nucleotide sequence. This code is a (nearly) universal one that is now known in its entirety.

The initiator codon (AUG) codes for the amino acid N-formylmethionine (f-Met). No transcription occurs without the AUG codon. f-Met is always the first amino acid in a polypeptide chain, although frequently it is removed after translation. The intitator tRNA/mRNA/small ribosomal unit is called the initiation complex. The larger subunit attaches to the initiation complex. After the initiation phase the message gets longer during the elongation phase.

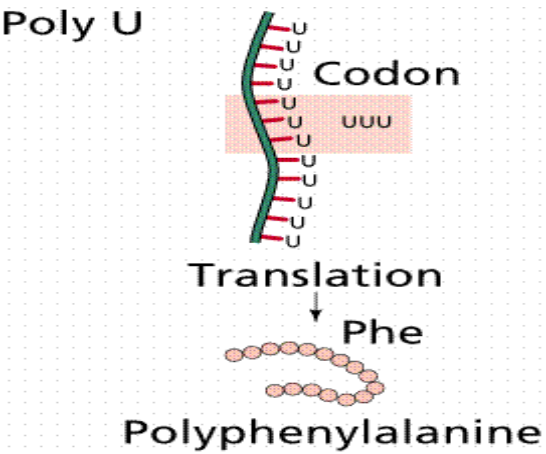


Fig.27. Steps in breaking the genetic code: the deciphering of a poly-U mRNA.

11.3. Triplet Code

What is the code by which the nucleotide sequences encode protein sequences? How can 4 nucleotides be used to specify 20 amino acids?

1. Could we let each nucleotide equal one amino acid? Well, there are 4 nucleotides and 20 amino acids so that probably wouldn't work. If, for example, we let C = proline, this would allow us to encode only $4^1 = 4$ amino acids which is 16 short of the number needed.

2. Would double nucleotides work? That would be 4^2 = 16, which is closer but still not sufficient. eg let CC = proline.

3. If we try 3 letter nucleotide words we have plenty of words to specify all the amino acids and lots left over for redundancy and even punctuation. 4^3 = 64. e.g. let CCC = proline. It turns out that the nucleotide code words are indeed composed of three nucleotides and this is known as the triplet code.

The genetic code consists of 61 amino-acid coding codons and three termination codons, which stop the process of translation. The genetic code is thus redundant (degenerate in the sense of having multiple states amounting to the same thing), with, for example, glycine coded for by GGU, GGC, GGA, and GGG codons. If a codon is mutated, say from GGU to CGU, is the same amino acid specified?

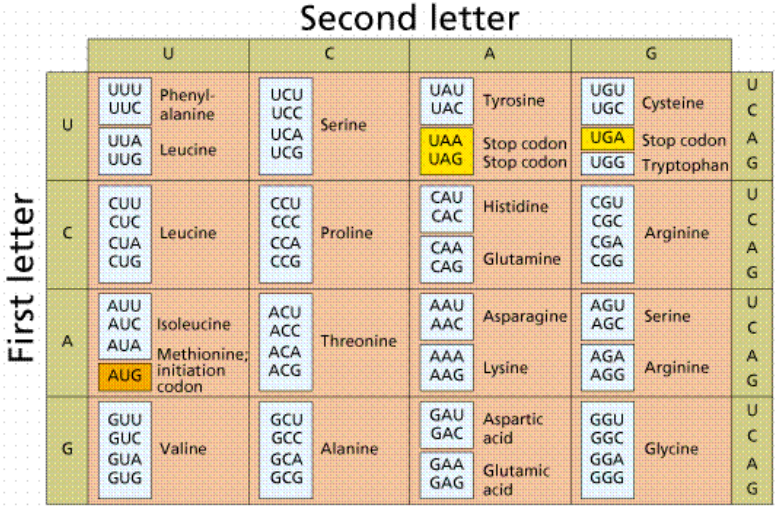


Fig.28. The genetic code.

- Each three letter sequence of mRNA running from 5' to 3' is known as a codon and almost all of them specify an amino acid (a few are punctuation). Note that a codon is a feature of the mRNA, not the DNA.
- Codon dictionaries are available; in fact there is one in your text and another in your lab manual. They are easy to use but you must remember they are for codons, i.e. mRNA and you can't look up DNA triplets in them. It would be an easy

matter to make a dictionary for DNA but it would be different.

Use a codon dictionary to translate the codon CCG to its amino acid (pro). Now do UUG (ile).

What would be the amino acid sequence specified by the gene (DNA) sequence

3'-TAGCATGAT-5'?

First transcribe the gene to mRNA and get **5'-AUCGUACUA-3'**

There are three codons here and they translate to **ile-val-leu**

- ► mRNA strands usually begin with an AUG sequence which means start and end with UAA, UAG, or UGA (stop).
- ► AUG always means methionine. But all nucleotide sequences begin with AUG so it also means START. As a consequence, all polypeptides begin with methionine, at least

initially. The initial methionine is trimmed off in most polypeptides later.

- ► Note that the reading direction is 5' to 3'. Note also the importance of reading frame. It is essential that the ribosome begin reading at exactly the right position in the nucleotide sequence in order to create the desired protein.

11.4. Transfer RNA

Ttansfer RNA (tRNA) molecules are small, about 90 nucleotides in length with most of the bases paired internally with other bases in the same molecule. This internal base pairing holds the molecule in a cloverleaf shape.

Three of the base pairs are exposed however and are not involved in hydrogen bonding with any bases in the tRNA molecule. These three can pair (hydrogen bond) with a codon of the mRNA molecule and are known as an anticodon. The anticodon sequence is 3' to 5'. Note that the anticodon sequence is the same as the DNA sequence (except with U instead of T). (You can't use a

codon dictionary to translate (directly) anticodons to amino acids.)

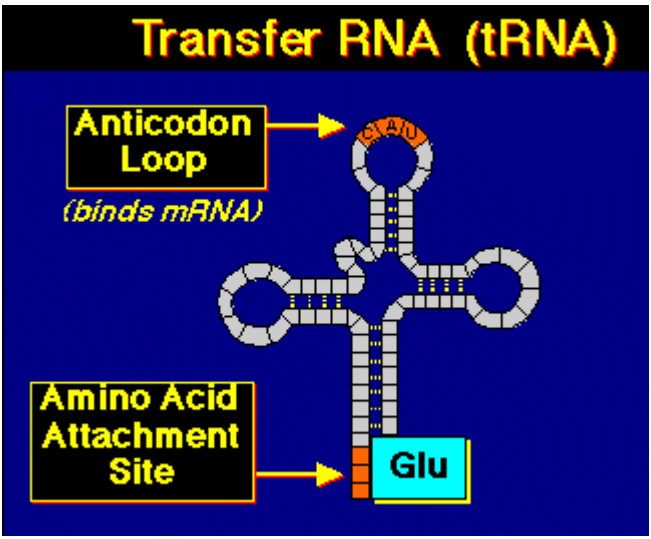


Fig. 29 transfer RNA

One end (3') of the tRNA is attached to a specific amino acid. The same amino acid is always associated with any given anticodon.

tRNA molecules are linked to their appropriate amino acid by the mediation of a specific aminoacyl-tRNA

synthetase enzyme which recodnizes the tRNA molecule and puts the correct amino acid at the 3' terminal. In all tRNA molecules the 3' tetrminus sequence is 3'-ACC-5'. There is a different aminoacyl synthetase for each tRNA/amino acid combination. The enzyme also activates the tRNA with an ATP molecule. (The synthetase actually recognizes a part of the tRNA molecule other than the anticodon.)

The mRNA molecule moves to the cytoplasm through the nuclear pores. In the cytoplasm there are tRNA molecules, amino acids, aminoacyl synthetase molecules, and the large and small ribosome subunits.

The ribosome subunits (40S and 60S) are separate until translation begins. They are composed of rRNA and protein. It is the responsibility of the ribosome to coordinate the matching of the correct tRNA anticodons with the mRNA codons. Each ribosome has a binding site for mRNA and 2 binding sites for tRNA. The P site holds the correct tRNA molecule and the A site holds the next tRNA molecule.

11.5. Function of Ribosomes

- ► The ribosome serves as the site of protein synthesis. mRNAs, tRNAs, and amino acids are brought together.
- ► On the ribosome, the mRNA fits between the two subunits (the interactions are stabilized by interchain hydrogen bonding). The tRNAs occupy a site on the large ribosomal subunit.
- ► The ribosome attaches to the mRNA at or near the 5'end.

    - In prokaryotes there is a ribosome binding site near the 5'end of the mRNA.
    - In eukaryotes, the ribosome first attaches at the 5'CAP (7-methyl guanosine).

The ribosome then moves along the mRNA in the 5' to 3' direction, one codon at a time.

11.6. The Central Dogma

The DNA is divided into segments, called genes, each of which is hundreds of thousands of base pairs long.

Each gene is the recipe for one polypeptide and specifies the sequence of amino acids in the polypeptide. One gene, one polypeptide. The nucleotide sequence is a code for an amino acid sequence. Since the DNA controls the synthesis of proteins, hence enzymes, it controls cell chemistry (including the synthesis of all other molecules such as carbohydrates, nucleotides, DNA, RNA, and lipids) and hence determines what a cell can and cannot do. For example it determines if the iris cells can produce brown pigment (melanin) and if pancreas cells can produce insulin.

The **central dogma** of molecular biology is

genotype (DNA) ------> transcription -----> translation----> protein ----->phenotype

- DNA codes for the production of DNA (replication) and of RNA (transcription).
- RNA codes for the production of protein (translation).
- Genetic information is stored in a linear message on nucleic acids. We use a shorthand notation to write a DNA sequence:

- 5'-AGTCAATGCAAGTTCCATGCAT....

- A gene determine the sequence of amino acids in proteins. We use a shorthand notation to write a protein sequence:

NH2-Met-Gln-Cys-Lys-Phe-Met-His.... (or a one letter code: M Q C K F M H)

Information flow (with the exception of reverse transcription) is from DNA to RNA via the process of transcription, and thence to protein via translation.

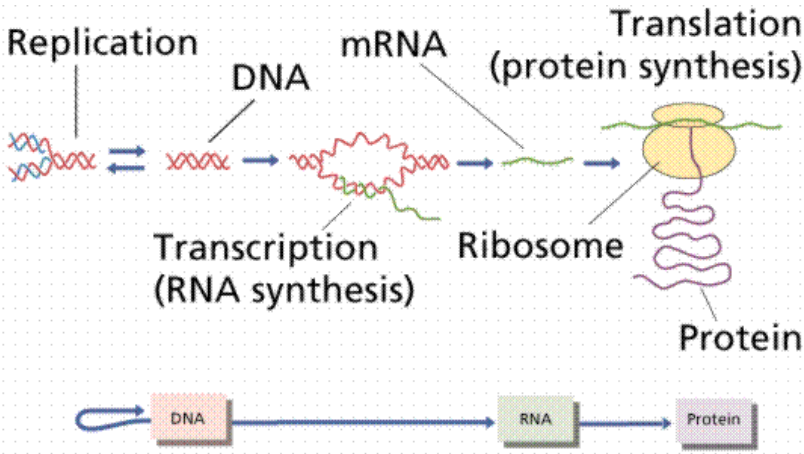


Fig. 30.The central dogma.

11.7. Protein Synthesis

- ► RNA Links the Information in DNA to the Sequence of Amino Acids in Protein
- ► Ribonucleic acid (RNA) was discovered after DNA. DNA, with exceptions in chloroplasts and mitochondria, is restricted to the nucleus (in eukaryotes, the nucleoid region in prokaryotes). RNA occurs in the nucleus as well as in the cytoplasm (also remembers that it occurs as part of the ribosomes that line the rough endoplasmic reticulum).
- ► Transcription is the making of an RNA molecule from a DNA template. Translation is the construction of an amino acid sequence (polypeptide) from an RNA molecule. Although originally called dogma, this idea has been tested repeatedly with almost no exceptions to the rule being found (save retroviruses).
- ► Messenger RNA (mRNA) is the blueprint for construction of a protein. Ribosomal RNA (rRNA) is the construction site where the protein is made. Transfer RNA (tRNA) is the truck

delivering the proper amino acid to the site at the right time.

- ► RNA has ribose sugar instead of deoxyribose sugar. The base uracil (U) replaces thymine (T) in RNA. Most RNA is single stranded, although tRNA will form a "cloverleaf" structure due to complementary base pairing.
- ► New tRNAs bring their amino acids to the open binding site on the ribosome/mRNA complex, forming a peptide bond between the amino acids. The complex then shifts along the mRNA to the next triplet, opening the A site. The new tRNA enters at the A site. When the codon in the A site is a termination codon, a releasing factor binds to the site, stopping translation and releasing the ribosomal complex and mRNA.
- ► Often many ribosomes will read the same message, a structure known as a polysome forms. In this way a cell may rapidly make many proteins.

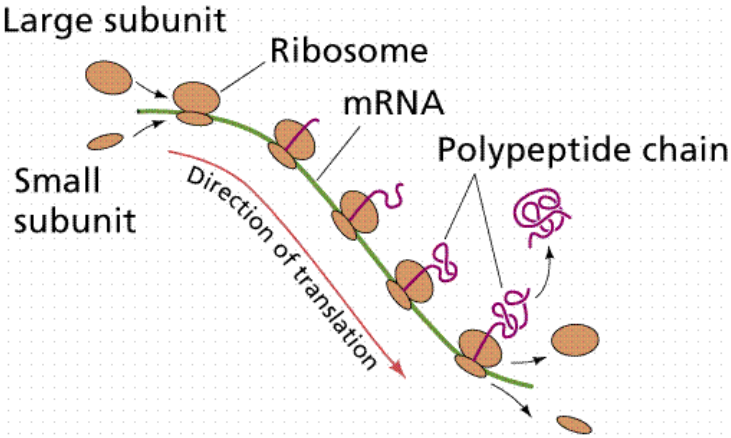


Fig.31 A polysome

Review Questions

1. Dedescribe the steps of transcription and translation.

2. What is the “one-gene-one polypeptide” hypothesis? Relate Garrod’s idea of an “inborn error of metabolism” to a defective enzyme in a metabolic pathway. (no testing on details on Beadle-Tatum experiments but you should read this section to understand the concept)

3. Name the 3 major processes in “central dogma” of molecular biology. What enzymes or “adapters” are required for each process? Where does each take place in a cell? Relate this to “flow of information” in gene expression. What extra step is needed for retroviruses?

CHAPTER TWELVE
CONTROL OF GENE EXPRESSION

Specific Learning objectives

At the end this chapter, student will be able to describe*:*

- ⇒ Control of gene expression in Prokaryotes
- ⇒ The *lac* Operon of *E. coli*
- ⇒ The *trp* Operon of *E. coli*
- ⇒ Control of Gene Expression in Eukaryotes
- ⇒ Structural Motifs in Eukaryotic Transcription Factors

12.0. Introduction

Gene expression is expensive, inappropriate gene expression can be harmful to cells/organisms, the proper expression of the phenotype of an organism is dependent upon expression and lack of expression of genes at appropriate times and in appropriate cells/places.

The controls that act on gene expression (i.e., the ability of a gene to produce a biologically active protein) are much more complex in eukaryotes than in prokaryotes. A major difference is the presence in eukaryotes of a nuclear membrane, which prevents the simultaneous transcription and translation that occurs in prokaryotes. Whereas, in prokaryotes, control of transcriptional initiation is the major point of regulation, in eukaryotes the regulation of gene expression is controlled nearly equivalently from many different points.

Control of gene expression basically occurs at two levels, prior to transcription and post-transcriptionally. There are four primary levels of control of gene activity:

1. **Transcriptional control** in nucleus determines which structural genes are transcribed and rate of transcription; includes organization of chromatin and transcription factors initiating transcription.

   Transcription is controlled by DNA-binding proteins called transcription factors; operons have not been found in eukaryotic cells. Group of transcription

factors binds to a promoter adjacent to a gene; then the complex attracts and binds RNA polymerase.

Transcription factors are always present in cell and most likely they have to be activated in some way (e.g., regulatory pathways involving kinases or phosphatases) before they bind to DNA.

2. **Posttranscriptional control** occurs in nucleus after DNA is transcribed and preliminary mRNA forms.This may involve differential processing of preliminary mRNA before it leaves the nucleus. Speed with which mature mRNA leaves nucleus affects ultimate amount of gene product.

   Posttranscriptional control involves differential processing of preliminary mRNA before it leaves the nucleus and regulation of transport of mature mRNA. Differential excision of introns and splicing of mRNA can vary type of mRNA that leaves nucleus.

   Evidence of different patterns of mRNA splicing is found in cells that produce neurotransmitters, muscle regulatory proteins, antibodies. Speed of transport of

mRNA from nucleus into cytoplasm affects amount of gene product realized. There is difference in length of time it takes various mRNA molecules to pass through nuclear pores.

3. **Translational control** occurs in cytoplasm after mRNA leaves nucleus but before protein product. Life expectancy of mRNA molecules can vary, as well as their ability to bind ribosomes. The longer an active mRNA molecule remains in the cytoplasm, the more product is produced.

   Mature mRNA has non-coding segments at 3' cap and 5' poly-A tail ends; differences in these segments influence how long the mRNA avoids being degraded.

   Prolactin promotes milk production by affecting the length of time mRNA persists and is translated.

   Estrogen interferes with action of ribonuclease; prolongs vitellin production in amphibian cells.

4. **Post-translational control** takes place in the cytoplasm after protein synthesis. Polypeptide

products may undergo additional changes before they are biologically functional. A functional enzyme is subject to feedback control; binding of an end product can change the shape of an enzyme so it no longer carries out its reaction.

Some proteins are not active after translation; polypeptide product has to undergo additional changes before it is biologically functional.

12.1. Gene Control in Prokaryotes

In bacteria, genes are clustered into operons: gene clusters that encode the proteins necessary to perform coordinated function, such as biosynthesis of a given amino acid. RNA that is transcribed from prokaryotic operons is polycistronic a term implying that multiple proteins are encoded in a single transcript.

In bacteria, control of the rate of transcriptional initiation is the predominant site for control of gene expression. As with the majority of prokaryotic genes, initiation is controlled by two DNA sequence elements that are approximately 35 bases and 10 bases, respectively,

upstream of the site of transcriptional initiation and as such are identified as the -35 and -10 positions. These 2 sequence elements are termed promoter sequences, because they *promote* recognition of transcriptional start sites by RNA polymerase.
The consensus sequence are for the -35 position is TTGACA, and - 10 position, TATAAT.

The -10 position is also known as the Pribnow-box. These promoter sequences are recognized and contacted by RNA polymerase.

The activity of RNA polymerase at a given promoter is in turn regulated by interaction with accessory proteins, which affect its ability to recognize start sites. These regulatory proteins can act both positively (activators) and negatively (repressors).

The accessibility of promoter regions of prokaryotic DNA is in many cases regulated by the interaction of proteins with sequences termed operators. The operator region is adjacent to the promoter elements in most operons and in most cases the sequences of the operator bind a

repressor protein. However, there are several operons in *E. coli* that contain overlapping sequence elements, one that binds a repressor and one that binds an activator.

Two major modes of transcriptional regulation function in bacteria (*E. coli*) to control the expression of operons. Both mechanisms involve repressor proteins. One mode of regulation is exerted upon operons that produce gene products necessary for the utilization of energy; these are:

- catabolite-regulated operons, and
- The other mode regulates operons that produce gene products necessary for the synthesis of small biomolecules such as amino acids.

Expression from the latter class of operons is attenuated by sequences within the transcribed RNA.

A classic example of a catabolite-regulated operon is the lac operon, responsible for obtaining energy from □-galactosides such as lactose. A classic example of an attenuated operon is the trp operon, responsible for the biosynthesis of tryptophan.

12.2. The *lac* Operon

Several gene codes for an enzyme in same metabolic pathway and are located in sequence on chromosome; expression of structural genes controlled by same regulatory genes. Operon is structural and regulatory genes that function as a single unit; it includes the following:

1) A **regulator gene** is located outside the operon; codes for a repressor protein molecule.
2) A **promoter** is a sequence of DNA where RNA polymerase attaches when a gene is transcribed.
3) An **operator** is a short sequence of DNA where repressor binds, preventing RNA polymerase from attaching to the promoter.
4) **Structural genes** code for enzymes of a metabolic pathway; are transcribed as a unit.

Lactose, milk sugar, is split by the enzyme □-galactosidase. This enzyme is inducible, since it occurs in large quantities only when lactose, the substrate on which it operates, is present. Conversely, the enzymes

for the amino acid tryptophan are produced continuously in growing cells unless tryptophan is present. If tryptophan is present the production of tryptophan-synthesizing enzymes is repressed.

If *E. coli* is denied glucose and given lactose instead, it makes three enzymes to metabolize lactose. These three enzymes are encoded by three genes. One gene codes for b-galactosidase that breaks lactose to glucose and galactose. A second gene codes for a permease that facilitates entry of lactose into the cell. A third gene codes for enzyme transacetylase, which is an accessory in lactose metabolism. The three genes are adjacent on chromosome and under control of one promoter and operator.

The regulator gene codes for a *lac* operon repressor protein that binds to the operator and prevents transcription of the three genes. When *E. coli* is switched to medium containing an allolactose, lactose binds to the repressor, the repressor undergoes a change in shape that prevents it from binding to the operator. Because the repressor is unable to bind to the operator, the promoter is able to bind to RNA

polymerase, which carries out transcription and produces the three enzymes. An inducer is any substance, lactose in the case of the *lac* operon, that can bind to a particular repressor protein, preventing the repressor from binding to a particular operator, consequently permitting RNA polymerase to bind to the promoter, causing transcription of structural genes.

The *lac* operon (see diagram below) consists of one regulatory gene (the *i* gene) and three structural genes (*z*, *y*, and *a*). The *i* gene codes for the repressor of the *lac* operon. The *z* gene codes for □-galactosidase (□-gal), which is primarily responsible for the hydrolysis of the disaccharide, lactose into its monomeric units, galactose and glucose. The *y* gene codes for permease, which increases permeability of the cell to □-galactosides. The *a* gene encodes a transacetylase.

During normal growth on a glucose-based medium, the *lac* repressor is bound to the operator region of the *lac* operon, preventing transcription. However, in the presence of an inducer of the *lac* operon, the repressor protein binds the inducer and is rendered incapable of interacting with the operator region of the operon. RNA

polymerase is thus able to bind at the promoter region, and transcription of the operon ensues.

The *lac* operon is repressed, even in the presence of lactose, if glucose is also present. This repression is maintained until the glucose supply is exhausted. The repression of the *lac* operon under these conditions is termed catabolite repression and is a result of the low levels of cAMP that result from an adequate glucose supply. The repression of the *lac* operon is relieved in the presence of glucose if excess cAMP is added. As the level of glucose in the medium falls, the level of cAMP increases. Simultaneously there is an increase in inducer binding to the *lac* repressor. The net result is an increase in transcription from the operon.

The ability of cAMP to activate expression from the *lac* operon results from an interaction of cAMP with a protein termed CRP (for cAMP receptor protein). The protein is also called CAP (for catabolite activator protein). The cAMP-CRP complex binds to a region of the *lac* operon just upstream of the region bound by RNA polymerase and that somewhat overlaps that of the repressor binding site of the operator region. The

binding of the cAMP-CRP complex to the *lac* operon stimulates RNA polymerase activity 20-to-50-fold.

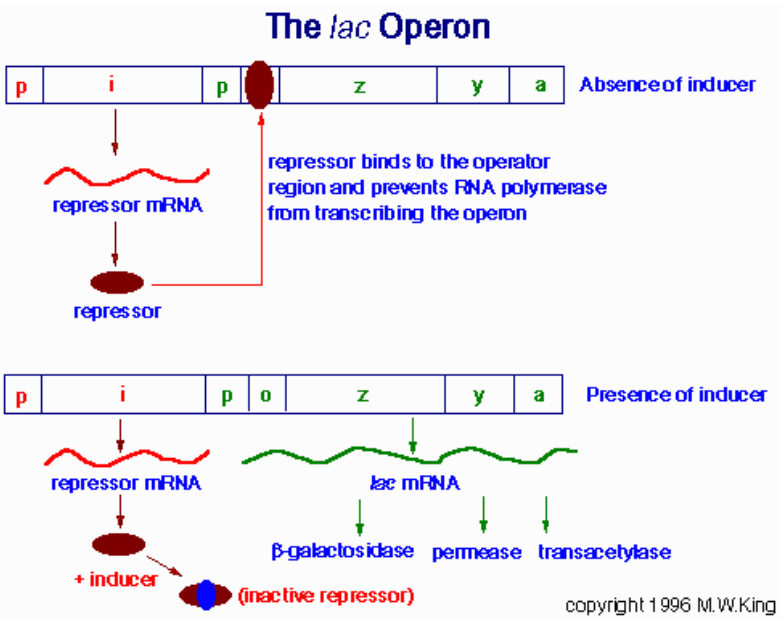


Fig. 32.Regulation of the *lac* operon in *E. coli*. The repressor of the operon is synthesized from the *i* gene. The repressor protein binds to the operator region of the operon and prevents RNA polymerase from transcribing the operon. In the presence of an inducer (such as the natural inducer, allolactose) the repressor is inactivated by interaction with

the inducer. This allows RNA polymerase access to the operon and transcription proceeds. The resultant mRNA encodes the □-galactosidase, permease and transacetylase activities necessary for utilization of □-galactosides (such as lactose) as an energy source. The *lac* operon is additionally regulated through binding of the cAMP-receptor protein, CRP (also termed the catabolite activator protein, CAP) to sequences near the promoter domain of the operon. The result is a 50 fold enhancement of polymerase activity.

Bacteria do not require same enzymes all the time; they produce just enzymes needed at the moment. In 1961, French microbiologist Francis Jacob and Jacques Monod proposed operon model to explain regulation of gene expression in prokaryotes; they received a Nobel prize for this.

Further Control of the *lac* Operon

When glucose is absent, cyclic AMP (cAMP) accumulates. Cytosol contains catabolite activator

protein (CAP). When cAMP binds to CAP, the complex attaches to the lac promoter. Only then does RNA polymerase bind to the promoter.

When glucose is present, there is little cAMP in the cell. CAP is inactive and the lactose operon does not function maximally. CAP affects other operons when glucose is absent. This encourages metabolism of lactose and provides backup system for when glucose is absent. Negative Versus Positive Control Active repressors shut down activity of an operon; they are negative control. CAP is example of positive control; when molecule is active, it promotes activity of operon. Use of both positive and negative controls allows cells to fine-tune its control of metabolism.

12.3. The *trp* Operon

The operon model of prokaryotic gene regulation was proposed by Fancois Jacob and Jacques Monod. Groups of genes coding for related proteins are arranged in units known as operons. An operon consists of an operator, promoter, regulator, and structural

genes. The regulator gene codes for a repressor protein that binds to the operator, obstructing the promoter (thus, transcription) of the structural genes. The regulator does not have to be adjacent to other genes in the operon. If the repressor protein is removed, transcription may occur.

Operons are either inducible or repressible according to the control mechanism. Seventy-five different operons controlling 250 structural genes have been identified for *E. coli*. Both repression and induction are examples of negative control since the repressor proteins turn off transcription.

Jacob and Monod found some operons in *E. Coli* exist in the on rather than the off condition. This prokaryotic cell produces five enzymes to synthesize the amino acid tryptophan. If tryptophan is already present in medium, these enzymes are not needed. In the *trp* operon, the regulator codes for a repressor that usually is unable to attach to the operator; the repressor has a binding site for tryptophan (if tryptophan is present, it binds to the repressor). This changes the shape of the repressor that now binds to the operator. The entire unit is called a

repressible operon; tryptophan is the corepressor. Repressible operons are involved in anabolic pathways that synthesize substances needed by cells.

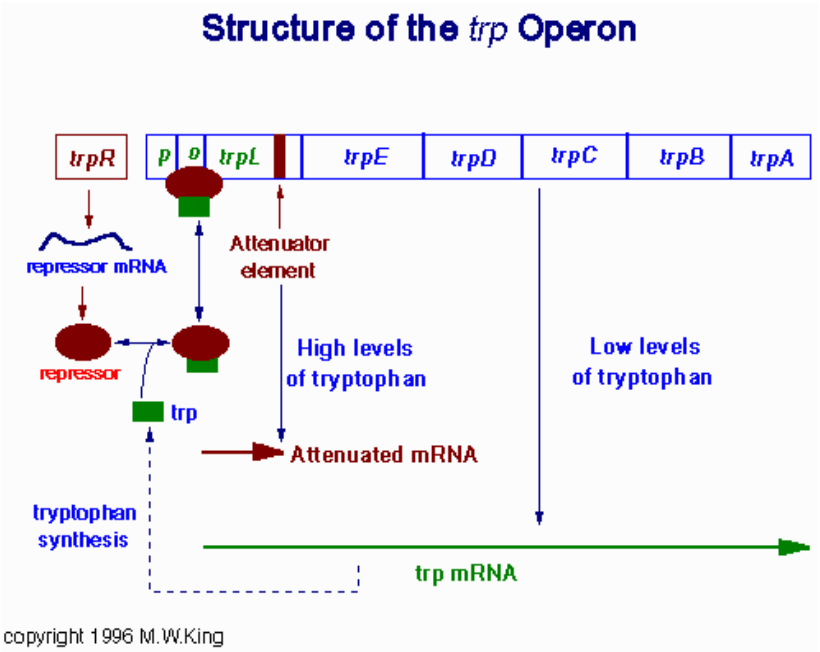
The *trp* operon (see Fig 10.2. below) encodes the genes for the synthesis of tryptophan. This cluster of genes, like the *lac* operon, is regulated by a repressor that binds to the operator sequences. The activity of the *trp* repressor for binding the operator region is enhanced when it binds tryptophan; in this capacity, tryptophan is known as a corepressor. Since the activity of the *trp* repressor is enhanced in the presence of tryptophan, the rate of expression of the *trp* operon is graded in response to the level of tryptophan in the cell.

Expression of the *trp* operon is also regulated by attenuation. The attenuator region, which is composed of sequences found within the transcribed RNA, is involved in controlling transcription from the operon after RNA polymerase has initiated synthesis. The attenuators of sequences of the RNA are found near the 5' end of the RNA termed the leader region of the RNA. The leader sequences are located prior to the start of the coding region for the first gene of the operon (the

trpE gene). The attenuator region contains codons for a small leader polypeptide, that contains tandem tryptophan codons. This region of the RNA is also capable of forming several different stable stem-loop structures.

Depending on the level of tryptophan in the cell---and hence the level of charged trp-tRNAs---the position of ribosomes on the leader polypeptide and the rate at which they are translating allows different stem-loops to form. If tryptophan is abundant, the ribosome prevents stem-loop 1-2 from forming and thereby favors stem-loop 3-4. The latter is found near a region rich in uracil and acts as the transcriptional terminator loop as described in the RNA synthesis page. Consequently, RNA polymerase is dislodged from the template.

The operons coding for genes necessary for the synthesis of a number of other amino acids are also regulated by this attenuation mechanism. It should be clear, however, that this type of transcriptional regulation is not feasible for eukaryotic cells.



12.4. Gene Control in Eukaryotes

The primary RNA transcripts are co-linear with their DNA templates, but in eukaryotes, the primary transcripts can be subsequently modified to produce a complete mRNA that is not entirely co-linear with its DNA template.

A quick look at the generation of functional mRNA in a typical eukaryote:

1. The first step is the recruitment of RNA polymerase to a specific chromosomal location and with a specific directionality. RNA polymerases by themselves are relatively non-specific enzymes that in purified form can initiate and elongate essentially randomly on a DNA template, but in vivo they are very specific enzymes, directed to specific sequences (promoters) by accessory general transcription factors and promoter-specific DNA-binding proteins. The transcription step is most frequently the rate-limiting step in the gene expression pathway.
2. RNA polymerase elongates through the entire coding region and beyond, producing an RNA that is elongated in the 5' to 3' orientation. The elements that recruit polymerase are different in prokaryotes and eukaryotes.
3. In eukaryotes transcription of protein-coding genes usually begins about 25 nucleotides downstream of a TATA box, which is the only strongly conserved element in pol II-dependent promoters.
4. The transcription start site is not the same as the translational start site.

5. The promoter is not transcribed; it is not part of the transcript.
6. Upon transcription termination (actually during the process of transcription), the primary transcript becomes rapidly modified, by addition of a 7-methyl guanosine “cap” at the 5’ end of the RNA, the addition of a polyA+ tail at its 3’ end, and by the removal of introns in a process called splicing.

7. The landmarks contained within a typical primary transcript are:
    - The capped 5’ end
    - 5’ untranslated region (UTR)
    - AUG initiation codon (usually the first AUG in the RNA)
    - First exon
    - Intron bordered by a splice donor and splice acceptor and containing a splice branchpoint
    - Additional exons and introns
    - Final exon, teminated by a STOP codon (UAG, UGA, or UAA)
    - 3' UTR
    - polyA+ tail

8. An open reading frame (ORF) is the series of codons in the final mRNA that will result in the translation of a protein, from the initiator AUG to the STOP codon. The ORF therefore does not constitute either the entire mRNA or the entire gene; 5' and 3' untranslated sequences can have important roles, and the promoter (or combined regulatory regions) should be considered part of the gene.
9. Proteins are translated from the mRNA template from the 5' to 3' orientation of the RNA, with the 5' end of the mRNA encoding the N-terminus of the protein, through to the 3' end of the mRNA encoding the C-terminus of the protein, with codon-anti-codon base pairing between the mRNA and charged tRNAs being responsible for insertion of the correct amino acids.

In eukaryotic cells, the ability to express biologically active proteins comes under regulation at several points:

1. **Chromatin Structure:** The physical structure of the DNA, as it exists compacted into chromatin, can

affect the ability of transcriptional regulatory proteins (termed **transcription factors**) and RNA polymerases to find access to specific genes and to activate transcription from them. The presence of the histones and CpG methylation most affect accessibility of the chromatin to RNA polymerases and transcription factors.

2. Transcriptional Initiation: This is the most important mode for control of eukaryotic gene expression (see below for more details). Specific factors that exert control include the strength of promoter elements within the DNA sequences of a given gene, the presence or absence of enhancer sequences (which enhance the activity of RNA polymerase at a given promoter by binding specific transcription factors), and the interaction between multiple activator proteins and inhibitor proteins.

3. **Transcript Processing and Modification:** Eukaryotic mRNAs must be capped and polyadenylated, and the introns must be accurately removed (see RNA Synthesis Page). Several genes have been identified that undergo tissue-specific

patterns of alternative splicing, which generate biologically different proteins from the same gene.

4. **RNA Transport:** A fully processed mRNA must leave the nucleus in order to be translated into protein.

5. **Transcript Stability:** Unlike prokaryotic mRNAs, whose half-lives are all in the range of 1--5 minutes, eukaryotic mRNAs can vary greatly in their stability. Certain unstable transcripts have sequences (predominately, but not exclusively, in the 3'-non-translated regions) that are signals for rapid degradation.

6. **Translational Initiation:** Since many mRNAs have multiple methionine codons, the ability of ribosomes to recognize and initiate synthesis from the correct AUG codon can affect the expression of a gene product. Several examples have emerged demonstrating that some eukaryotic proteins initiate at non-AUG codons. This phenomenon has been known to occur in *E. coli* for quite some time, but

only recently has it been observed in eukaryotic mRNAs.

7. **Post-Translational Modification:** Common modifications include glycosylation, acetylation, fatty acylation, disulfide bond formations, etc.

8. **Protein Transport:** In order for proteins to be biologically active following translation and processing, they must be transported to their site of action.

9. **Control of Protein Stability:** Many proteins are rapidly degraded, whereas others are highly stable. Specific amino acid sequences in some proteins have been shown to bring about rapid degradation.

12.5. Control of Eukaryotic Transcription Initiation

Initiation of transcription is the most important step in gene expression. Without the initiation of transcription, and the subsequent transcription of the gene into mRNA by RNA polymerase, the phenotype controlled by the

gene will not be seen. Therefore in depth studies have revealed much about what is needed for transcription to begin.

The control of transcription is an integrated mechanism involving cis-acting sequences and trans-acting factors. Cis-acting sequence usually lies 5' of the transcriptional start site. These sequences are the substrate for trans-acting factors. These factors bind to the cis-acting sequences and prepare the DNA in their vicinity for transcription. Because the trans-acting factors are proteins, they must also be encoded by genes. And these genes may also be controlled the interaction of cis-acting sequences and trans-acting factors. This interplay between genes and their cis-acting sequences and trans-acting factors is a cascade of genetic events.

Transcription of the different classes of RNAs in eukaryotes is carried out by three different polymerases (see RNA Synthesis Page). RNA pol I synthesizes the rRNAs, except for the 5*S* species. RNA pol II synthesizes the mRNAs and some small nuclear RNAs (snRNAs) involved in RNA splicing. RNA pol III synthesizes the 5*S* rRNA and the tRNAs. The vast

majority of eukaryotic RNAs are subjected to post-transcriptional processing.

The most complex controls observed in eukaryotic genes are those that regulate the expression of RNA pol II-transcribed genes, the mRNA genes. Almost all eukaryotic mRNA genes contain a basic structure consisting of coding exons and non-coding introns and basal promoters of two types and any number of different transcriptional regulatory domains (see diagrams below). The basal promoter elements are termed CCAAT-boxes (pronounced cat) and TATA-boxes because of their sequence motifs. The TATA-box resides 20 to 30 bases upstream of the transcriptional start site and is similar in sequence to the prokaryotic Pribnow-box (consensus $TATA^{T}/_{A}A^{T}/_{A}$, where $^{T}/_{A}$ indicates that either base may be found at that position).

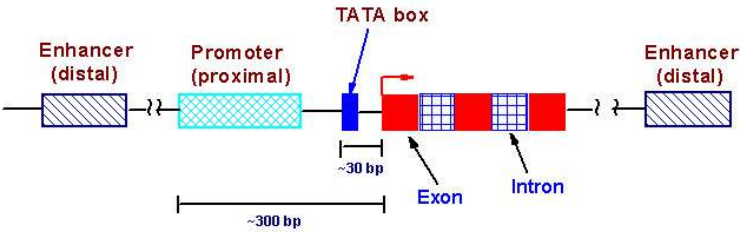


Fig. 33 Typical structure of a eukaryotic mRNA gene

Numerous proteins identified as TFIIA, B, C, etc. (for transcription factors regulating RNA pol II), have been observed to interact with the TATA-box. The CCAAT-box (consensus $GG^{T}/_{C}CAATCT$) resides 50 to 130 bases upstream of the transcriptional start site. The protein identified as C/EBP (for CCAAT-box/Enhancer Binding Protein) binds to the CCAAT-box element.

The are many other regulatory sequences in mRNA genes, as well, that bind various transcription factors (see diagram below). Theses regulatory sequences are predominantly located upstream (5') of the transcription initiation site, although some elements occur downstream (3') or even within the genes themselves. The number and type of regulatory elements to be found varies with each mRNA gene. Different combinations of

transcription factors also can exert differential regulatory effects upon transcriptional initiation. The various cell types each express characteristic combinations of transcription factors; this is the major mechanism for cell-type specificity in the regulation of mRNA gene expression.

Negative control [repression, repressor protein]

The binding of a specific protein (*repressor protein*) to DNA at a point that interferes with the action of RNA polymerase on a specific gene is a form of *negative control* of protein synthesis.

This interference with RNA polymerase activity is termed *repression* (of the action of RNA polymerase) and the consequent to this lack of gene expression a gene is described as *repressed*. Action results in lack of activity**.** This form of *control of gene expression* is called *negative control* because the controlling action results in an absence of activity.

Positive control [activation, activator protein]

In contrast to negative control, very often a specific gene requires the binding of a specific protein (an *activating protein*) in order to acheive RNA polymerase binding and gene expression.

This type of *control of gene expression* is termed *activation* since in its absence the gene is not *active* (i.e., is not expressed). Action results in activity.This type of control is also termed *positive control* in the sense that the action of the *activating protein* results in a positive action: gene expression.

12.6. Transcription and Processing of mRNA

The process of transcription in eukaryotes is similar to that in prokaryotes, although there are some differences. Eukaryote genes are not grouped in operons as are prokaryote genes. Each eukaryote gene is transcribed separately, with separate transcriptional controls on each gene. Whereas prokaryotes have one type of RNA polymerase for all types of RNA,

eukaryotes have a separate RNA polymerase for each type of RNA. One enzyme for mRNA-coding genes such as structural proteins. One enzyme for large rRNAs. A third enzyme for smaller rRNAs and tRNAs.

Prokaryote translation begins even before transcription has finished, while eukaryotes have the two processes separated in time and location (remember the nuclear envelope). After eukaryotes transcribe an RNA, the RNA transcript is extensively modified before export to the cytoplasm. A cap of 7-methylguanine (a series of an unusual base) is added to the 5' end of the mRNA; this cap is essential for binding the mRNA to the ribosome. A string of adenines (as many as 200 nucleotides known as poly-A) is added to the 3' end of the mRNA after transcription. The function of a poly-A tail is not known, but it can be used to capture mRNAs for study. Introns are cut out of the message and the exons are spliced together before the mRNA leaves the nucleus. There are several examples of identical messages being processed by different methods, often turning introns into exons and vice-versa. Protein molecules are attached to mRNAs that are exported, forming ribonucleoprotein particles (mRNPs) which may help in

transport through the nuclear pores and also in attaching to ribosomes.

Some features/similarities that are important for practical experimental reasons:

Promoters are required both in bacteria and eukaryotes, although bacterial and eukaryotic promoters are not interchangeable the polymerases and co-factors have evolved to recognize different elements. To express in bacteria you generally need a bacterial promoter; eukaryotic promoters don't work efficiently in *E. coli*.

The genetic code is generally the same in bacteria and eukaryotes (usually taken for granted). One exception…mitochondria have their own genetic code, and mitochondrially-translated genes won't generate the same protein in bacteria! Furthermore, because the usage of specific codon frequency differs between organisms, translation of some eukaryotic ORFs might not occur efficiently in bacteria due to codon frequency problems.

Another thing we need to remember is that translation is not the end of the line in producing a functional protein;

post-translational modifications can be important for protein function (and are frequently of great regulatory interest), but those modifications might not take place in bacteria.

Usual features of prokaryotic genes

- polycistronic mRNAs (single RNA with multiple ORFs)
- operons (chromosomal localization of genes into functional groups)
- no splicing

Usual features of eukaryotic genes

- Monocistronic (one mRNA encodes one gene)
- ribosome scanning (though there are internal ribosomal entry sites)
- often spliced (though some eukaryotes have few spliced genes and perhaps all eukaryotes have at least a few unspliced genes)

Review Questions

1. Describe the roles of cis-acting sequences and trans-acting factors in the control of eukaryotic gene expression.

2. What transcription factors are required for the successful transcription of eukaryotic DNA by RNA polymerase II?

3. Describe the relationship between the promoter, CCAAT box, GC box, enhancers and silencers.

4. What specific role might methylation play in the control of eukaryotic gene expression?

5. How was it determined that trans-acting factors have two functional domains?

6. How does *E. coli* know how to turn on when glucose is available?

7. What is an operon? What is the main purpose of gene expression regulation via operons? Why is this particularly appropriate for prokaryotic organisms? Why is it NOT appropriate for regulation of genes

involved in developmental processes, eg embryogenesis in higher organisms?

8. What are structural genes? What are regulator genes? Why are the products of regulator genes in operons usually negatively-acting? What does "negatively-acting" mean?

9. What genetic mutations are the most convincing that the products of regulator genes are negatively-acting?

10. What is the Repressor? What is the Operator? What is an Attenuator?

11. What is "constitutive expression"?

12. What is an Inducer? What is a Co-Repressor? What is Allosterism? How is allosterism important in operon theory?

13. Which of these types of processes are characterized by Inducers of the operon? Co-Repressors of the operon?

14. In what ways are the two classes of operons similar to each other? in what ways are they different?

CHAPTER THIRTEEN

RECOMBINANT DNA TECHNOLOGY

Specific Learning Objectives

At the end of this chapter students are expected to

- ⇒ list enzymes used in molecular Biology
- ⇒ differentiate role of each enzymes Understand how to identify the gene responsible for a phenotype or disease.
- ⇒ describe how to study the functions of a known gene.
- ⇒ list the steps involved in a cloning experiment.
- ⇒ explain the importance of genetics for understanding and treating human diseases.
- ⇒ convey how genetics is currently used to investigate fundamental biological processes in the common model genetic organisms,

13.0. Introduction

While the period from 1900 to the Second World War has been called the "golden age of genetics".

Recombinant DNA technology allows us to manipulate the very DNA of living organisms and to make conscious changes in that DNA. Prokaryote genetic systems are much easier to study and better understood than are eukaryote systems.

Recombinant DNA technology is also known as Genetic engineering. Genetic engineering is an umbrella term which can cover a wide range of ways of changing the genetic material - the DNA code - in a living organism. This code contains all the information, stored in a long chain chemical molecule, which determines the nature of the organism - whether it is an amoeba, a pine tree, a robin, an octopus, a cow or a human being - and which characterizes the particular individual.

There are various means of manipulating DNA and there are various means of transferring DNA to a recipient cell (e.g., transformation,). Additionally, there

are various things that one can do with the DNA that has been transferred to a recipient cell. Note that the transferred DNA may be from the same species or from a different species than the recipient. Such successfully transferred DNA is said to be cloned.

13.1. Uses of Genetic Engineering

But why should we do this manipulation, be it within or across species? The purposes of doing genetic engineering are many and various. A range of them are listed below. These include :

- to repair a genetic "defect" (as with the current early trials of gene therapy in humans),
- to enhance an effect already natural to that organism (e.g. to increase its growth rate),
- to increase resistance to disease or external damage (e.g. crops - blight, cold or drought),
- to enable it to do something it would not normally do :
- e.g. getting a micro-organism to produce human insulin for diabetics, or a sheep to produce a

human blood-clotting protein in her milk, in both cases a transgenic method,

- e.g. getting a tomato to ripen without going squashy - this can be done simply by taking one of its own genes, turning its "pattern" upside down and putting it back again!

13.2. Basic Tools of Genetic Engineering

The basic tools of genetic engineering are

1. Cloning vectors:- Which can be used to deliver the DNA sequences in to receptive bacteria and amplify the desired sequence;
2. Restriction enzymes:
   Restriction enzymes recognize a specific palindromic sequence of make a staggered cut, which generates sticky ends, or blund cut, which generate blunt ends.
   Which are used to cleave DNA reproducibly at defined sequences and
3. DNA ligases
   Ligation of the vector with the DNA fragments generates a molecule capable of replicating the inserted sequence called recombinant DNA

The total No of recombinant vectors obtained when cloning chromosomal DNA is known as genomic library between there should be at least one representative of each gene in the library .

Genetic engineering represents a number of methods employed to

(i) manipulate DNA outside of cells

(ii) place manipulated DNA back into cells

(iii) manipulate that DNA following its incorporation back into cells

13.3 Enzymes in molecular Biology

Many enzymes have been isolated from prokaryotes which can be used to modify DNA and/or RNA in vitro reactions.

These include:-Nucleases, RNases, DNA Polymerase, RNA polymerases, Kinase and Ligase.

I. Nulease

There are two types of nucleases.

A. Exonucleases -

- Cleaves nucleotide one at a time from the end of a nucleic acid

B. Endonucleases

- Cleaved bonds within a contiguous molecule of nucleic acid.

1. Restriction enzymes

- Endonucleases:- cleave dsDNA,
  - do not cleave ssDNA or RNA
- Each recognizes a specific nucleotide sequence, often palindromic.

Palindromic is a state where both strands have the same nucleotide sequence but in anti-parallel directions

1. 6- cutters - recognize 6 nucleotide sequences
   - less frequent sites within DNA, in 4100bp
   e.g. Bamtel GI GATCC
   EcoRI GIAA TTC
2. 4- cutters - recognize 4 nucleotide sequences
   - More frequent, in 256bp
   e.g- Sau3A GATC,
   HaeIII GGCC

II. RNases

- Endonucleases cleave at specific ribonucleotides

A. RNAse A cleaves ssRNA at pyrimidine nucleotides (C,U)

5^1 A-G-Cl-G-A-Ul-GC-ACl-A^3 1 + RNase A

A-G-C G-A-U G-C A-C A

B. RNase T1 cleaves ssRNA at G nucleotides

5 A-Gl-C-G-A-U-Gl-A-C-A^3 lRNAse Tl

AG C-G AUG A-C-A

III. DNA polymerases

-They synthesize DNA polymerizes requirement for all DNA polymerases

-Template strand (DNA or RNA)

- Primer with a free 3^1-OH group

2.1. DNA polymerase (Holoenzyme)

- The common source is E.coli
- They have the following activities: a. $5^1 \rightarrow 3^1$ polymerase

b. $5^1 \rightarrow 3^1$ exanuclease

c. $3^1 \rightarrow 5^1$ exanuclease

2.2. **Klenow**

- Is on E.coli DNA pol I with $5^1 \rightarrow 3^1$ exonuclease activity removed
- Its activity include: $5^1 \rightarrow 3^1$ polymerase

a. $3^1 \rightarrow 5^1$ exonuclease

- They are used: for

a. Filling in &/or labeling recessed 3^1 ends created by restriction enzyme digestion
b. Random primer labeling
c. Second strand cDNA synthesis is cDNA cloning
d. DNA sequencing using the dideoxy system

1.3 T_4 DNA polymerase

- It is obtained from T_4 plasmid infected E.coli

a. It has similar activities as to that of klenow except that $3^1 \rightarrow 5^1$= exonuclease activity is 200 α
b. $5^1 \rightarrow 3^1$ polymerase

- It is used for end labeling of recessed or $\rightarrow$

1.4 T 7 DNA polymerase

- It is obtained from T 7 –infected E.coli
- It is highly possessive $5^1 \rightarrow 3^1$ polymerase activity (much better than klenow)

- It has no $3^1 \rightarrow 5^1$ exonuclease activity
- Its activity is $5^1 \rightarrow 3^1$ polymerase
- It is used for DNA sequencing using the dideoxy system

1.5 Tag DNA polymerase

- It is obtained from a thermophilic bacterium found in hot springs called thermos aquatic
- It is heat stable polymerase that will synthesis DNA at elevated temperatures.
- It is used for:

a. DNA synthesis in PCR reactions
b. Dideoxy DNA sequencing through regions of high 2^0 structures (i.e, regions destabilized by high temperature. Cyclic DNA sequencing of tow abundance DNA

1.6 Reverse transcriptase

- It is obtained from either

a. Arian myeloblastosiv virus, or
b. Moloney murine levkemia virus

- it is an RNA dependent DNA polymerase
- its activity include:

a. $5^1 \rightarrow 3^1$ DNA polymerase
b. exoribonvclease 9specifically degrades RNA in a DNA: RNA hybrid.

- It is used to transcribe first cDNA in cDNA cloning.

IV.RNA polymerases

- They are isolated from various bacteriophages including

a. Phage SP6
b. Phage T7
c. Phage T3

- Its activity is to synthesize ssRNA from a DNA template
- Its uses include to synthesize sequence specific RNA probes which is used for:

a. Labeling 5^1 ends of synthetic oligonuceoties
b. Phosphorylating synthetic linkers and other synthetic DNA fragments lacking a 3^1- prior to ligation

V. T_4 polynucleotide kinase

- It is obtained from T_4 plasmid infected E.coli
- It activities include:

b. Transfer of γ- phosphate of AJP to a 5^1-OH terminals of DNA or RNA

c. Exchange the γ- phosphate of $xP^{3R}ATP$ with 5^1= terminal phosphate of DNA in the presence of excess ADP

VI .T_4 DNA ligase

- It is produced by T_4 plasmid infected
- Its activity is catalyzes formation of phosphodiester bond between adjacent 3^1- OH and 5^1-p terminal in DNA
- Its use include

a. to join together DNA molecules with complementary or blunt ends in DNA cloning

b. to region nicked DNA

VII. Alkaline phosphatase

- It is produced from either Bacteria (BAP) or calf intestine (CIAP)

- Its activity is removing the 5^1- phosphate from DNA or RNA
- Its uses include
    1. Remove 5^1 –phosphate from DNA prior to labeling
    2. Removing 5^1- phosphate from vector DNA to prevent self-ligation during cloning

Polymerases are used in molecular genetics in many ways:

1. Nick translation
2. In vitro transcription
3. cDNA synthesis
4. End labeling
5. Random hexamer-primed synthesis
6. Polymerase chain reaction

Each of the above can be used to label oligo and polynucleotides. In addition, the nucleic acids can be chemically derivatized to provide tagged molecules.

13.4. DNA manipulation

13.4.1. DNA manipulation outside of cells

The key to manipulating DNA outside of cells is the existence of enzymes known as restriction endonucleases. Restriction endonucleases cut DNA only at specific nucleotide sequences and thus are tools by which DNA may be cut at specific locations. Thus, a specific gene may be cut out of an organism's genome. Further techniques allow one to specifically change the nucleotide sequence of the isolated gene.

1. The *restriction* part of the name derives from the actual use of these enzymes by the bacteria that make them: restricting the replication of bacteriophages (by chewing up the bacteriophage DNA)
2. The *nuclease* part of the name means these enzymes cut DNA
3. The *endo* part of the name means that they cut DNA in the middle of double helix strands (rather

than chewing DNA up from the ends, i.e., as do *exo*nucleases)

13.4.2. DNA transfer to recipient cell (vector)

- To transfer manipulated DNA back into a cell, one typically first inserts the DNA into a vector
- A vector may be a plasmid (transformation) or a bacteriophage chromosome (transduction) or both
- The vector or plasmid are opened up (cut) using specific restriction endonucleases
- The isolated gene is then inserted into this opening
- An additional enzyme, DNA ligase, then covalently attaches the gene into the vector, thus making gene and vector into one double helix
- The vector may then be transduced or transformed into a recipient cell
- Within that cell the vector is allowed to replicate
- Often these vectors also contain antibiotic-resistance genes which, in the presence of the appropriate antibiotic, allow only those cells that have successfully received the vector to replicate

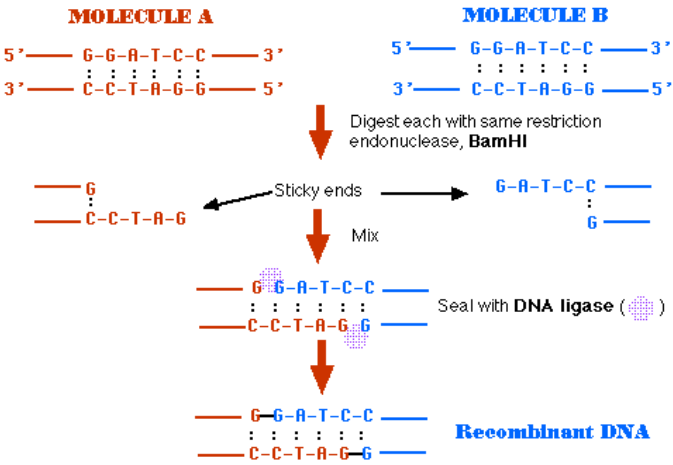
13.4.3. DNA manipulation within the recipient cell

- Once the DNA is in a recipient cell, things can be done with it
- One thing that can be done is to allow the introduced gene to express (e.g., produce a new protein), thus changing the phenotype of the recipient cell
- A second thing that can be done is the gene product (a protein) can be overly expressed so that the resulting relatively high concentration of protein can be purified and either used for a specific purpose or employed for the characterization of the protein (which often is far easier given a relative abundance of protein)
- A third thing that can be done is the inserted gene may be sequenced using DNA sequencing techniques; sequencing permits further characterization as well as further manipulation of the gene
- The inserted gene may serve as a source of DNA for further cloning of the gene (e.g., to place in vectors having different properties, so that relatively large

concentrations of the gene sequence may be manipulated outside of the cell, etc.)

13.5. Making a Recombinant DNA: An Overview

Recombinant DNA is DNA that has been created artificially. DNA from two or more sources is incorporated into a single recombinant molecule.



- Treat DNA from both sources with the same restriction endonuclease (BamHI in this case).

- BamHI cuts the same site on both molecules

5' GGATCC 3' 3' CCTAGG 5'

- The ends of the cut have an overhanging piece of single-stranded DNA.
- These are called "sticky ends" because they are able to base pair with any DNA molecule containing the complementary sticky end.
- In this case, both DNA preparations have complementary sticky ends and thus can pair with each other when mixed.
- DNA ligase covalently links the two into a molecule of recombinant DNA.

To be useful, the recombinant molecule must be replicated many times to provide material for analysis, sequencing, etc. Producing many identical copies of the same Cloning

13.6. Cloning

Cloning is a group of replicas of all or part of a macromolecule (such as DNA or an Antibody).In gene

(DNA) cloning a particular gene is copied (cloned). Cloning often gets referred to in the same breath as genetic engineering, but it is not really the same. In genetic engineering, one or two genes are typically changed from amongst perhaps 100,000. Cloning essentially copies the entire genetic complement of a nucleus or a cell, depending on which method is used.

Cloning in vivo can be done in

- unicellular prokaryotes like E. coli
- unicellular eukaryotes like yeast and
- in mammalian cells grown in tissue culture.

Cloning can be done in vitro,

- by polymerase chain reaction (PCR).

In every case, the recombinant DNA must be taken up by the cell in a form in which it can be replicated and expressed. This is achieved by incorporating the DNA in a vector.

A number of viruses (both bacterial and of mammalian cells) can serve as vectors. But here let us examine an

example of cloning using E. coli as the host and a plasmid as the vector.

13.6.1. Transforming E. coli

Treatment of E. coli with the mixture of religated molecules will produce some colonies that are able to grow in the presence of both ampicillin and kanamycin.

- A suspension of E. coli is treated with the mixture of religated DNA molecules.
- The suspension is spread on the surface of agar containing both ampicillin and kanamycin.
- The next day, a few cells — resistant to both antibiotics — will have grown into visible colonies containing billions of transformed cells.
- Each colony represents a clone of transformed cells.

However, E. coli can be simultaneously transformed by more than one plasmid, so we must demonstrate that the transformed cells have acquired the recombinant plasmid.

Electrophoresis of the DNA from doubly-resistant colonies (clones) tells the story.

- Plasmid DNA from cells that acquired their resistance from a recombinant plasmid only show only the 3755-bp and 1875-bp bands (Clone 1, lane 3).
- Clone 2 (Lane 4) was simultaneous transformed by religated pAMP and pKAN. (We cannot tell if it took up the recombinant molecule as well.)
- Clone 3 (Lane 5) was transformed by the recombinant molecule as well as by an intact pKAN.

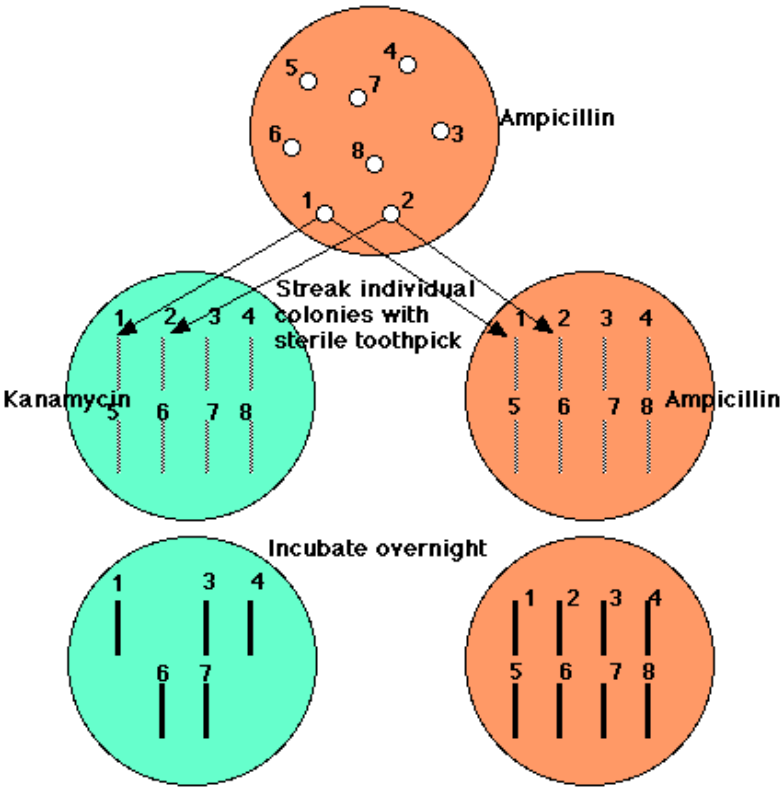


Fig.34 Transforming *E.coli*

13.6.2. Cloning other Genes

The recombinant vector described above could itself be a useful tool for cloning other genes. Let us assume that within its kanamycin resistance gene (kan^r) there is a

single occurrence of the sequence: 5' GAATTC 3' 3' CTTAAG 5'

This is cut by the restriction enzyme EcoRI, producing sticky ends. If we treat any other sample of DNA, e.g., from human cells, with EcoRI, fragments with the same sticky ends will be formed. Mixed with EcoRI-treated plasmid and DNA ligase, a small number of the human molecules will become incorporated into the plasmid which can then be used to transform E. coli.

But how to detect those clones of E. coli that have been transformed by a plasmid carrying a piece of human DNA?

The key is that the EcoRI site is within the kan^r gene, so when a piece of human DNA is inserted there, the gene's function is destroyed.

All E. coli cells transformed by the vector, whether it carries human DNA or not, can grow in the presence of ampicillin. But E. coli cells transformed by a plasmid carrying human DNA will be unable to grow in the presence of kanamycin.

So,

- Spread a suspension of treated E. coli on agar containing ampicillin only
- grow overnight
- with a sterile toothpick transfer a small amount of each colony to an identified spot on agar containing kanamycin
- (do the same with another ampicillin plate)
- Incubate overnight

All those clones that continue to grow on ampicillin but fail to grow on kanamycin (here, clones 2, 5, and 8) have been transformed with a piece of human DNA.

13.6. 3. Cloning Vectors

Cloning vector **-** a DNA molecule that carries foreign DNA into a host cell, replicates inside a bacterial (or yeast) cell and produces many copies of itself and the foreign DNA

The molecular analysis of DNA has been made possible by the cloning of DNA. The two molecules that are

required for cloning are the DNA to be cloned and a cloning vector.

Vectors are specialized plasmids, phages or hybrids which have been developed to make the construction of recombinant libraries and the identification of individual recombinant clones simpler.

Three features of all cloning vectors

1. sequences that permit the propagation of itself in bacteria (or in yeast for YACs)
2. a cloning site to insert foreign DNA; the most versatile vectors contain a site that can be cut by many restriction enzymes
3. a method of selecting for bacteria (or yeast for YACs) containing a vector with foreign DNA; uually accomplished by selectable markers for drug resistance

13.6.3.1. Types of Cloning Vectors

DNA libraries are stratified by type of genome as well as type of vector. Classification by type of vector is equivalent to classification by size. There are libraries

which contain the entire chromosomes (YAC libraries) as well as the libraries containing pieces of 30kb long. Cloning vectors:

- must be relatively small molecules for convenience of manipulation.
- must be capable of prolific replication in a living cell, thereby enabling the amplification of the inserted donor fragment.
- must be convenient restriction sites that can be used for insertion of the DNA to be cloned. Generally, we would like to see a unique restriction site because then the insert can be specifically targeted to one site in the vector.
- It's also desirable that there be a mechanism for easy identification, recovery and sequencing of recombinant molecule. There are numerous cloning vectors in current use and the choice between them usually depends on the size of DNA fragment to be cloned.

All vectors have the following properties:

- Replicate autonomovsly in E.coli

- Easily separated and purified from bacterial chromosomal DNA
- Contains non-essential regions of DNA in which foreign DNA can be inserted.

The following is a list of vectors which are being used for library creation.

1. **Plasmid** - an autonomous, an extra chromosomal circular DNA molecule that autonomously replicates inside the bacterial cell;

- cloning limit: 100 to 10,000 base pairs or 0.1-10 kb.

Plasmids are characterized by ability to replicate prolifically. They produce double--stranded fragments. The typical size of insert is about 3,500 bps long. In theory plasmids can hold up to 20kb; however, they easily lose inserts of such size. It often carries antibiotic resistance genes. A useful plasmid vector must:

- be small
- has an efficient origin of replication - relaxed= 200-1000 copies

- stringent = 1-10 copies
- carry one or more selectable markers to allow identification of transform ants.
- Unique restriction enzyme sites (RE) to facilitate ease of cloning.
- Plasmid vectors include; P^{uc}, P^{BR322} and PGEM

Features

- Relatively small, 4.36kb. This size provides easy purification from E.coli cells
- Relaxed origin of replication gives high copy number of plasmids within each bacterium (200-1000 plasmid copies)
- Tow antibiotic resistance genes, (AMP) and Tetracycline (Tet)
- Unique pst1 site within the AMpiclin gene and unique BamHI side within the tetracycline gene.

  This unique restriction target sites are useful in cloning

2. Phage - derivatives of bacteriophage lambda; linear DNA molecules, whose region can be replaced with foreign DNA without disrupting its life cycle;

- cloning limit: 8-20 kb

2.1.λ-phage

λ-phages can accept inserts of about 10-15 kb long and is characterized by good ``packaging'' in the sense that it's very unlikely to lose its insert.

2.2. Bacteriophage M13

M13 is extensively used by LLNL in the context of shot-gun sequencing. The size of inserts is about 1,500 bps, i.e. the fragments grown in M13 are ready to be sequenced. M13 contains single--stranded inserts, i.e. there is not going to be a denaturing step in preparation of M13 insert for sequencing.

The disadvantage of this particular vector is a large cloning bias. M13 is prone to loosing (refusing to amplify with) certain types of sequences. I.e. if only M13 were to be used, certain sequences in genome would never be discovered.

3. **Cosmids** - an extra chromosomal circular DNA molecule that combines features of plasmids and phage;

   - cloning limit - 35-50 kb.

They are plasmids which contain lambda packaging signals (cos sites). Clones can be size selected (39-52 kb) by packaging *in vitr*o. Colonies carrying transduced DNA are selected by their antibiotic resistance, and maintained through plasmid replication origins.

The small size of plasmid genes required for selection and maintenance (3kb) compared to lambda genes required for propagation as a phage (>20kb) makes more room for cloned DNA.

The idea behind its creation was to combine ``good'' properties of plasmids and --phages. In particular, cosmids

- It rarely lose inserts,
- It replicate intensely and,
- it can hold inserts of about 45kb in size.

4. Bacterial Artificial Chromosomes (BAC)

- Bacterial Artificial Chromosomes is based on bacterial mini-F plasmids

- Its Cloning limit: 75-300 kb.

BACs are based on F DNA and can carry very large DNA fragments (> 200kb) very stably at low copy number.

BACs are amplified in bacteria as the name suggests.; however, the average size is about 100kb.

BACs inserts can be manipulated with standard plasmid technology and they form fewer chimeras than YACs.

In principle, BAC is used like a plasmid. We construct BACs that carry DNA from humans or mice or wherever, and we insert the BAC into a host bacterium. As with the plasmid, when we grow that bacterium, we replicate the BAC as well. Huge pieces of DNA can be easily replicated using BACs - usually on the order of 100-400 kilobases (kb). Using BACs, scientists have cloned (replicated) major chunks of human DNA.

5. Yeast Artificial Chromosomes (YAC) - an artificial chromosome that contains telomeres, origin of replication, a yeast centromere, and a selectable marker for identification in yeast cells;

- YACs can hold huge inserts, up to and beyond 1,000kb, i.e. the entire chromosomes are being replicated in this vector. However, the problem with YAC libraries is that they contain up to if *chimeras* and it's extremely laborious to separate chimeras from ``real'' recombinants.
- A chimera is a recombinant molecule in which non-contiguous donor fragments are being joined together. They are linear plasmids that replicate in yeast.

They differ mainly in the amount of DNA that can be inserted. Each has some of the following features:

- sequences for maintenance and propagation of the vector and insert
- (required)
- polylinker sites for rapid cloning
- selection system for identifying those with inserts

- expression cassettes (high constitutive or inducible promoter systems)
- epitope tag/cassettes for detection and/or rapid purification of the
- expressed proteins

13.6.4. General Steps of Cloning with Any Vector

1. Prepare the vector and DNA to be cloned by digestion with restriction enzymes to generate complementary ends
2. Ligate the foreign DNA into the vector with the enzyme DNA ligase
3. Introduce the DNA into bacterial cells (or yeast cells for YACs) by transformation select cells containing foreign DNA by screening for selectable markers (usually drug resistance)

13.7. Cloning DNA

We want to prepare *Subcloned DNA templates*, i.e. grow multiple copies of a given piece of DNA for further subcloning and/or sequencing. The following properties

are desirable in a subcloning procedure, in the order of importance. Note that the second property is relevant only when the next step is sequencing rather that further subcloning.

1. To be able to represent *all* regions of genome (i.e. with minimal or no cloning bias)
2. Produce sequence ready DNA templates (i.e. we don't want to have difficulties with preparing the template for sequencing)

As a result we want to end up with a Genomic Library: a collection of DNA clones that covers the entire genome. We should also be able to order clones along the genome, i.e. to determine the relative positions of the clones (Physical mapping).

Suppose we are presented with a given genome and we are after its base composition. Its DNA is also referred to as a foreign or *donor* DNA.

- The idea is that we are going to create *recombinant* DNA by cutting the donor DNA, inserting a given fragment into a small replicating

molecule (*vector*) which, under certain conditions, will then *amplify* the fragment along with itself, resulting in a molecular *clone* of the inserted DNA molecule.

- The vector molecules with their inserts are called *recombinant DNA* because they represent combinations of DNA from the donor genome with vector DNA from a completely different source (generally a bacterial plasmid or a virus).
- The recombinant DNA structure is then used to transform bacterial cells and it's common for single recombinant vector molecules to find their way into individual bacterial cells. Bacterial cells are then plated and allowed to grow into colonies.
- An individual transformed cell with a single recombinant vector will divide into a colony with millions of cells all carrying the same recombinant vector. Therefore an individual colony represents a very large population of identical DNA inserts and this population is called a *DNA clone*.

Procedures:

1. Isolating DNA

The first step is to isolate donor and vector DNA. The procedure used for obtaining vector DNA depends on the nature of the vector.

Bacterial plasmids are commonly used vectors and these must be purified away from the bacterial genomic DNA. One of the possible protocols is based on the observation that that at a specific alkaline pH, bacterial genomic DNA denatures but plasmids do not. Subsequent neutralization precipitates the genomic DNA but the plasmids stay in the solution.

2. Cutting DNA

The discovery and characterization of *restriction enzymes* made the recombinant DNA technology possible. Restriction enzymes are produced by bacteria as defense mechanism against phages. In other words they represent bacteria immune system.

The enzymes inactivate the phage by cutting up its DNA at the *restriction sites*. Restriction sites are specific target sequences which are palindromic and this is one of many features that makes them suitable for DNA manipulation.

Any DNA molecule will contain the restriction enzyme target just by chance and therefore may be cut into defined fragments of size suitable for cloning.

Different methods are used at different stages of subcloning at different laboratories.

3. Joining DNA

Donor DNA and vector DNA are digested with the same restriction enzyme and mixed in a test tube in order to allow the ends to join to each other and form recombinant DNA. At this stage the sugar-phosphate backbones are still not complete at two positions at each junction.

However, the fragments can be linked permanently by the addition of the enzyme DNA ligase, which creates phosphodiester bonds at the joined ends to make a

continuous DNA molecule. One of the problems of free availability of sticky ends in solution is that the cut ends of a molecule can rejoin rather than form recombinant DNA.

In order to combat the problem, the enzyme terminal transferase is added. It catalyzes the addition of nucleotide ``tails'' to the 3' ends of DNA chain. Thus, ddA (dideoxiadenine) molecules are added to, say vector DNA fragments and dT molecules are added to donor DNA fragments, only chimeras can form. Any single stranded gaps created by restriction cleavage are filled by DNA polymerase 1 and the joins subsequently sealed by DNA ligase.

4. Amplifying Recombinant DNA

Recombinant plasmid DNA is introduced into host cells by transformation. In the host cell, the vector will replicate in the normal way, but now the donor DNA is automatically replicated along with the vector. Each recombinant plasmid that enters a cell will form multiple copies of itself in that cell. Subsequently, many cycles of cell-division will occur and the recombinant vector will

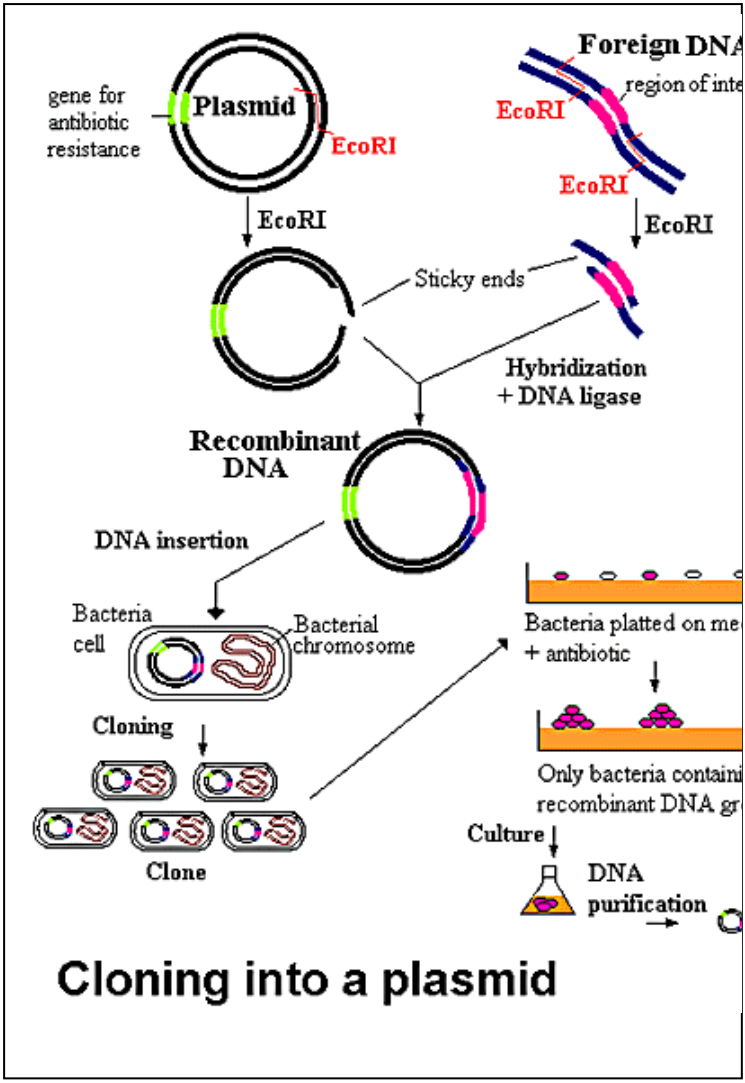
undergo more rounds of replication. The resulting colony of bacteria will contain billions of copies of the single donor DNA insert. This set of amplified copies of a single donor DNA is the DNA clone.

13.8. Cloning into a Plasmid

By fragmenting DNA of any origin (human, animal, or plant) and inserting it in the DNA of rapidly reproducing foreign cells, billions of copies of a single gene or DNA segment can be produced in a very short time. DNA to be cloned is inserted into a plasmid (a small, self-replicating circular molecule of DNA) that is separate from chromosomal DNA. When the recombinant plasmid is introduced into bacteria, the newly inserted segment will be replicated along with the rest of the plasmid.

Many diseases are caused by gene alterations. Our understanding of genetic diseases was greatly increased by information gained from DNA cloning. In DNA cloning, a DNA fragment that contains a gene of interest is inserted into a cloning vector or plasmid.

The plasmid carrying genes for antibiotic resistance, and a DNA strand, which contains the gene of interest, are both cut with the same restriction endonuclease


Foreign DNA
region of inte
gene for antibiotic resistance
Plasmid
EcoRI
EcoRI
EcoRI
EcoRI
EcoRI
Sticky ends
Hybridization + DNA ligase
Recombinant DNA
DNA insertion
Bacteria cell
Bacterial chromosome
Cloning
Clone
Bacteria platted on me
+ antibiotic
Only bacteria containi
recombinant DNA gr
Culture
DNA purification
Cloning into a plasmid

The plasmid is opened up and the gene is freed from its parent DNA strand. They have complementary "sticky ends." The opened plasmid and the freed gene are mixed with DNA ligase, which reforms the two pieces as recombinant DNA.

Plasmids + copies of the DNA fragment produce quantities of recombinant DNA.

This recombinant DNA stew is allowed to transform a bacterial culture, which is then exposed to antibiotics. All the cells except those which have been encoded by the plasmid DNA recombinant are killed, leaving a cell culture containing the desired recombinant DNA.

DNA cloning allows a copy of any specific part of a DNA (or RNA) sequence to be selected among many others and produced in an unlimited amount. This technique is the first stage of most of the genetic engineering experiments: production of DNA libraries, PCR, DNA sequencing, et al.

13.9. Expression and Engineering of Macromolecules

A detailed understanding of the function of proteins and nucleic acids requires experiments performed in vitro on purified components. The tools of molecular biology provide not only access to large quantities of homogeneous macromolecules to study, but the ability to modify those molecules and identify variants with interesting new properties.

I. Expression

- **DNA** Synthesis

  Automated DNA synthesis machines allow one to obtain single stranded molecules of any specific or randomized sequence up to 100 nucleotides long. Cycle times are a few minutes per base.

- **RNA** *In vitro* transcription

The favored route to obtaining specific RNA molecules is transcription of synthetic or natural DNA molecules in vitro by T7 RNA polymerase. This single subunit enzyme is simpler, more specific and more active than E. coli enzyme. Repeated rounds of RNA synthesis are obtained from either duplex DNA with desired sequences downstream from T7 promoter or, minimally, a single stranded DNA template with a DS promoter.

- **Proteins**, **Peptide synthesis**

Peptides are not as easy to synthesize as DNA, not all sequences soluble. Cycle times are on the order of an hour per amino acid. It is generally not possible to synthesize whole proteins. Most peptides are used to model protein fragments or to generate antibodies.

Expression in Ecoli

Up to 50% of total soluble protein can often be obtained. Possible problems include solubility (formation of inclusion bodies), lethality

(expression should be tightly regulated to avoid adverse affects during growth and maintenance of the strain), sensitivity of foreign proteins to E. coli proteases. Often complicated DNA constructions are required to align foreign gene with proper transcription and translation signals, can be avoided by making protein fusion to well expressed E. coli protein.

E. coli promoters (lac,tac, *lambda* PL and PR) are regulated by repression.

- lac operator is repressed up to 10,000 fold (inducible by IPTG)
- *lambda* promoters are induced by inactivation of temperature sensitive repressor.
- T7 Promoters {Studier et al Meth Enz 185 60 (1990)} Regulated by availability of T7 RNA polymerase. T7 transcription out competes host polymerase for nucleotides, making expression selective for gene of interest.

II. Manipulation

Directed manipulation of DNA sequences (via synthetic DNA) are the preferred route to modifying protein and RNA sequences as well. Since production of RNA is primarily in vitro cloning of the variant DNA sequences is usually not required.

III. Selections

The ability to generate large populations of molecules of different sequence coupled with a method for linking their replication to their ability to function creates new opportunities to identify functional molecules that may never have arisen in nature. Comparing these novel macromolecules to each other and to their natural counterparts may enable us to overcome the limitations imposed by the patchy, historical sampling of possible sequences which occurred during evolution and achieve new insights on the rules underlying the function of proteins and nucleic acids.

13.10. Creating mutations

Genetics relies upon mutations: you either have to isolate mutant forms of the gene in vivo that cause some interesting phenotype, or one can create mutations in any cloned gene in vitro. The ability to create any mutation in any cloned DNA at will (and test its activity in vivo) revolutionized genetic analysis. Mutations can be created in cloned DNA by:

A. Random mutagenesis

1. Chemical mutagensis (e.g. hydroxylamine)
2. Passing them through mutator (DNA-repair-deficient) strains of bacteria
3. Transposon insertions

B. Site-directed mutagenesis

1. Deletion analysis
2. Oligo-directed mutagenesis
3. PCR

Site-directed mutations can be point mutations, insertions, or deletions

These mutagenesis techniques are independent of the organism; any cloned DNA can be mutagenized regardless of its origin.

C.Primer directed Mutagenesis

Any gene can be modified using oligonucleotide primers, so long as it has been cloned in a single strand producing vector and its sequence is known. Lots of tricks have been developed to overcome the inherent inefficiency of having both mutant and parental alleles in the heteroduplex DNA. Most inactivate the parental strand either chemically or biologically to ensure survival of the modifications.

1. Obtain gene of interest on single stranded circular DNA.
2. Anneal mismatched primer (mutation can be substitution , insertion or deletion.) with 6 to 10 complimentary bases on either side of the mutation.
3. Extend the primer with DNA polymerase.
4. Ligate the product strand to make a covalently closed circle.

5. After transformation of E. coli, strands segregate to give parental and mutant genes.

D. Cassette mutagenesis

A double strand synthetic DNA cassette with ends complimentary to existing restriction sites is required.

E. Combinatorial mutagenesis

With either technology, mixtures of DNA bases can replace pure solutions at selected positions in the sequence to be synthesized. The result is a population of individual molecules, each with a different sequence. Upon transformation of E. coli individual DNA molecules give rise to pure clones. In this way many different variants can be obtained in a single experiment. If the randomization was such that small number of variants are expected

Libraries: contain a collection of all or part of the genome within an appropriate vector

Genomic library: all parts of the genome are represented equally (in theory)

cDNA library: each gene's representation within the library is proportional to its expression level. Only transcribed regions are present in cDNA libraries

Libraries can be screened for specific DNA sequences in several ways:

1. screen for specific sequences by hybridization with DNA or RNA probes
2. screen / select for biological activity in vivo
3. screen for biochemical activity (e.g. antibody reactivity) in vitro

Some recombinant DNA products being used in human therapy:

Using procedures like this, many human genes have been cloned in E. coli or in yeast. This has made it possible — for the first time — to produce unlimited amounts of human proteins in vitro. Cultured cells (E. coli, yeast, mammalian cells) transformed with the human gene are being used to manufacture:

- insulin for diabetics
- factor VIII for males suffering from hemophilia A

- factor IX for hemophilia B
- human growth hormone (GH)
- erythropoietin (EPO) for treating anemia
- three types of interferons
- several interleukins
- granulocyte-macrophage colony-stimulating factor (GM-CSF) for stimulating the bone marrow after a bone marrow transplant
- tissue plasminogen activator (TPA) for dissolving blood clots
- adenosine deaminase (ADA) for treating some forms of severe combined immunodeficiency (SCID)
- angiostatin and endostatin for trials as anti-cancer drugs
- parathyroid hormone
- leptin
- hepatitis B surface antigen (HBsAg) to vaccinate against the hepatitis B virus

Review Questions

1. What are the four key discoveries that led to the Recombinant DNA Technology revolution?

2. Cloning Experiments are the source of the word "recombinant" in Recombinant DNA Technology. Explain why.

3. Cloning experiments are often used to isolate genes.

4. Why is this difficult to do with mammalian genomes?

5. What is a "complementation assay" is the cloning, for example, of a bacterial gene?

6. What is Reverse Genetics? What are the steps in a Reverse Genetics approach to cloning a gene?

7. Design a strategy to clone the E. coli origin of DNA replication.

8. Once a gene is isolated via cloning, how would you proceed in analysis of the cloned DNA?

9. What is a cDNA?

10. What is cDNA cloning?

11. Why are four sequencing reactions performed in Sanger sequencing?

12. What is a Sequencing Gel? How does it differ from a standard R.fragment agarose gel?

13. What types of mutants would you isolate? Why?

14. What three major classes of E. coli dna mutants have been isolated? Which step in DNA replication would you expect the gene products of each of these three gene classes to be involved?

15. How have the E. coli dna mutants been used to purify DNA replication proteins?

16. What features of a Type II R.enzyme make possible cloning experiments?

17. What is a Restriction Enzyme?

18. What features of a Type II Restriction Enzyme are important for recombinant DNA work?

19. What is a "sticky end", and why is it said that some R.Enzymes generate "sticky ends"?

20. Why is this enzyme called a "6 cutter"?

21. Why is this 6-bp sequence called a "nucleotide palindrome"?

22. What is the sticky end generated?

23. In the following DNA sequence, draw both strands of the product molecules following NcoI digestion:

CHAPTER FOURTEEN
DNA SEQUENCING

Specific learning Objectives

At the end of this chapter students are expected to:

- ⇒ List importance of sequencing
- ⇒ Describe the Methods used for sequencing

14.0. Introduction

DNA sequencing, first devised in 1975, has become a powerful technique in molecular biology, allowing analysis of genes at the nucleotide level. DNA sequencing is the determination of the precise sequence of nucleotides in a sample of DNA. For this reason, this tool has been applied to many areas of research including:

- For example, the polymerase chain reaction (PCR), a method which rapidly produces numerous copies of a desired piece of DNA,

requires first knowing the flanking sequences of this piece.

- Another important use of DNA sequencing is identifying restriction sites in plasmids. Knowing these restriction sites is useful in cloning a foreign gene into the plasmid.
- Before the advent of DNA sequencing, molecular biologists had to sequence proteins directly; now amino acid sequences can be determined more easily by sequencing a piece of cDNA and finding an open reading frame.
- In eukaryotic gene expression, sequencing has allowed researchers to identify conserved sequence motifs and determine their importance in the promoter region.
- Furthermore, a molecular biologist can utilize sequencing to identify the site of a point mutation. These are only a few examples illustrating the way in which DNA sequencing has revolutionized molecular biology.

DNA sequencing can be used to determine the nucleotide sequence of any region of a DNA strand. The sequencing techniques make use of DNA's capability to

replicate itself. Segments of DNA of the region of interest are allowed to replicate *in vitro* (in glass as opposed to *in vivo*, in life).

Replication will occur in vitro if the following are present:

a. DNA segment

b. deoxyribonucleotides (dATP, dGTP, dCTP, dTTP)

c. DNA polymerase

d. Primer

Remember that replication takes place from 3' to 5' along the original strand and this means new nucleotides are added to the 3" end of the growing strand.

Remember that this involves the –OH group on the #3 carbon of the deoxyribose molecule. There must be a free –OH on the # 3 carbon.

Normally, DNA polymerase would move along the segment following the rules for base pairing and would add the appropriate deoxyribonucleotides at each base

and at the 3" end of the growing strand. We provide polymerase with a few dideoxyribonucleotides (ddATP, ddTTP, ddGTP, ddCTP).

These nucleotides have no –OH on their #2 or #3 carbons. Should one of these be used by polymerase, it would result in the termination of the growing strand because another nucleotide cannot be attached to it.

For example: The following crick strand would be made from the watson templete by adding, one at a time, the appropriate nucleotides tot he growing chain.

5’ TACCTGACGTA 3’ crick (growing strand made from watson templete)

3’ ATGGACTGCAT 5’ watson

Bit if we provided some ddATP, in addition to the usual dATP and polymerase happened to use one of them the first time it needed an ATP we would get the following

5’ TA-stop

3’ ATGGACTGCAT 5’ watson

Replication could not proceed past the second nucleotide because of the ddATP at that position. ddATP has no –OH on the #3 carbon so the CTP which should come next cannot be attached. The chain stops at two nucleotides. This would not happen every time, only when polymerase happened to pick up a ddATP instead of a dATP. Remember, both are present in the system.

Another time, polymerase might put a dATP in place at the second position and then continue in normal fashion to the 7th position and then pick up a ddATP. We would get a longer, but still not complete, chain as follows

5’ TACCTG-stop

3’ ATGGACTGCAT 5’ watson

We would get a mixture of replicated chains of various lengths. Each T in the original strand (crick) would provide a possibility of termination. The mixture would include some complete watson strands in which no ddATP was used, and some shorter strands, of various lengths, in which ddATP was used at the #2 and #7

positions. (5’ TA-5’ , 5’ TACCTG- , and 5’ TACCTGACGTA)

We could do the same thing using ddGTP, ddCTP, and ddTTP. In each case we would get mixtures of chains of various lengths.

Our task now is to separate each of our four mixtures into its component chains. For this we will use a technique known as gel electrophoresis.

DNA molecules are negatively charged and will migrate toward a positive electrical pole. DNA strands can migrate though the gel used in electrophoresis but they are impeded by it. The larger (longer) the chain, the slower is its progress through the gel.

We place a little of our mixture in a well at one end of a strip of gel and then attach electrodes to opposite ends of the gel strip with the negative pole being located at the well end (origin) of the strip.

The small fragments (like 5’ TA) move the fastest through the gel on their way to the positive pole. The largest chains (such as 5’ TACCTGACGTA) are the

slowest and do not get far from the origin. This technique separates the mixture into its components with the smallest fragments being closest to the positive electrode and the largest being closest to the origin (negative electrode). Intermediate size chains are in between and are positioned according to their size.

We do this with each of our 4 mixtures (one for each nucleotide), putting each in a different lane of the gel. The gel would thus have four lanes. Interpreting the resulting gels is simply a matter of reading from the positive end of the gel (5') to the negative (3').

14.1. Sanger Method for DNA Sequencing

Dideoxynucleotide sequencing represents only one method of sequencing DNA. It is commonly called Sanger sequencing since Sanger devised the method dideoxy method.

DNA is synthesized from four deoxynucleotide triphosphates. The top formula shows one of them: deoxythymidine triphosphate (dTTP). Each new

nucleotide is added to the 3′ -OH group of the last nucleotide added.

The dideoxy method gets its name from the critical role played by synthetic nucleotides that lack the -OH at the 3′ carbon atom (red arrow). A dideoxynucleotide (dideoxythymidine triphosphate — ddTTP — is the one shown here) can be added to the growing DNA strand but when it is, chain elongation stops because there is no 3′ -OH for the next nucleotide to be attached to. For this reason, the dideoxy method is also called the chain termination method.

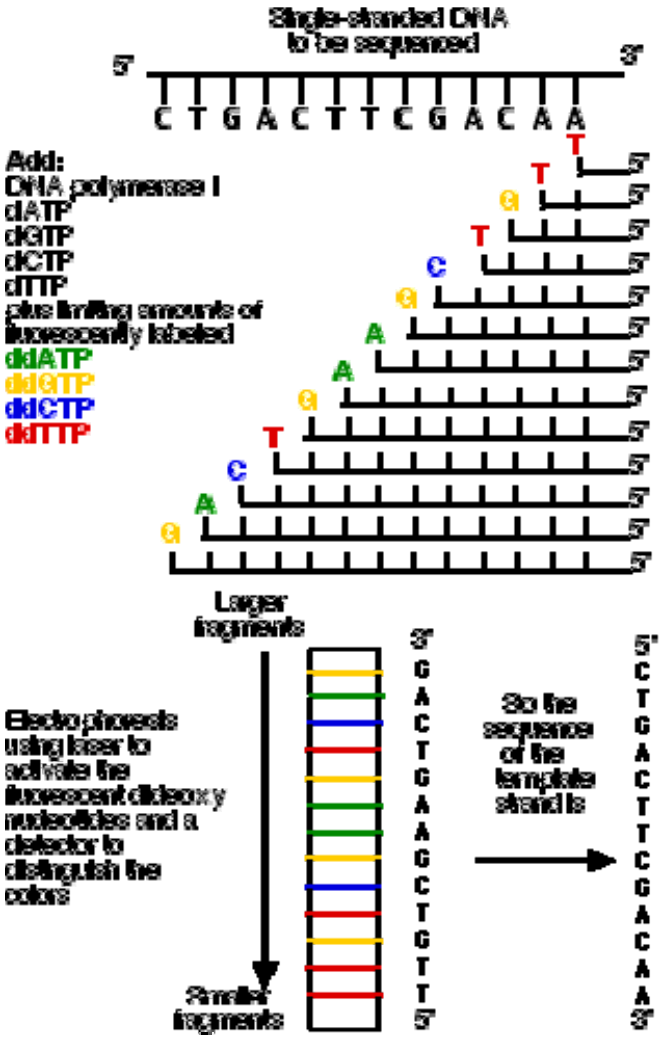


Fig.35. Dideoxy method of sequencing

14.1.1. Procedure

The DNA to be sequenced is prepared as a single strand. This template DNA is supplied with

- a mixture of all four **normal** (deoxy) nucleotides in ample quantities
    - dATP
    - dGTP
    - dCTP
    - dTTP
- a mixture of all four **dideoxy**nucleotides, each present in limiting quantities and each labeled with a "tag" that fluoresces a different color:
    - **ddATP**
    - **ddGTP**
    - **ddCTP**
    - **ddTTP**
- DNA polymerase I

Because all four normal nucleotides are present, chain elongation proceeds normally until, by chance, DNA polymerase inserts a dideoxy nucleotide (shown as colored letters) instead of the normal deoxynucleotide

(shown as vertical lines). If the ratio of normal nucleotide to the dideoxy versions is high enough, some DNA strands will succeed in adding several hundred nucleotides before insertion of the dideoxy version halts the process.

At the end of the incubation period, the fragments are separated by length from longest to shortest. The resolution is so good that a difference of one nucleotide is enough to separate that strand from the next shorter and next longer strand. Each of the four dideoxynucleotides fluoresces a different color when illuminated by a laser beam and an automatic scanner provides a printout of the sequence.

In order to perform the sequencing, one must first convert double stranded DNA into single stranded DNA. This can be done by denaturing the double stranded DNA with NaOH.

A Sanger reaction consists of the following:

- a strand to be sequenced (one of the single strands which was denatured using NaOH),

- DNA primers (short pieces of DNA that are both complementary to the strand which is to be sequenced and radioactively labelled at the 5' end), a mixture of a particular ddNTP (such as ddATP) with its normal dNTP (dATP in this case), and
- the other three dNTPs (dCTP, dGTP, and dTTP). The concentration of ddATP should be 1% of the concentration of dATP. The logic behind this ratio is that after DNA polymerase is added, the polymerization will take place and will terminate whenever a ddATP is incorporated into the growing strand. If the ddATP is only 1% of the total concentration of dATP, a whole series of labeled strands will result (Figure 1). Note that the lengths of these strands are dependent on the location of the base relative to the 5' end.

This reaction is performed four times using a different ddNTP for each reaction. When these reactions are completed, a polyacrylamide gel electrophoresis (PAGE) is performed. One reaction is loaded into one lane for a total of four lanes (Figure 2). The gel is transferred to a nitrocellulose filter and autoradiography is performed so

that only the bands with the radioactive label on the 5' end will appear. In PAGE, the shortest fragments will migrate the farthest. Therefore, the bottom-most band indicates that its particular dideoxynucleotide was added first to the labeled primer.

In Figure 2, for example, the band that migrated the farthest was in the ddATP reaction mixture. Therefore, ddATP must have been added first to the primer, and its complementary base, thymine, must have been the base present on the 3' end of the sequenced strand. One can continue reading in this fashion. Note in Figure 2 that if one reads the bases from the bottom up, one is reading the 5' to 3' sequence of the strand complementary to the sequenced strand. The sequenced strand can be read 5' to 3' by reading top to bottom the bases complementary to the those on the gel.

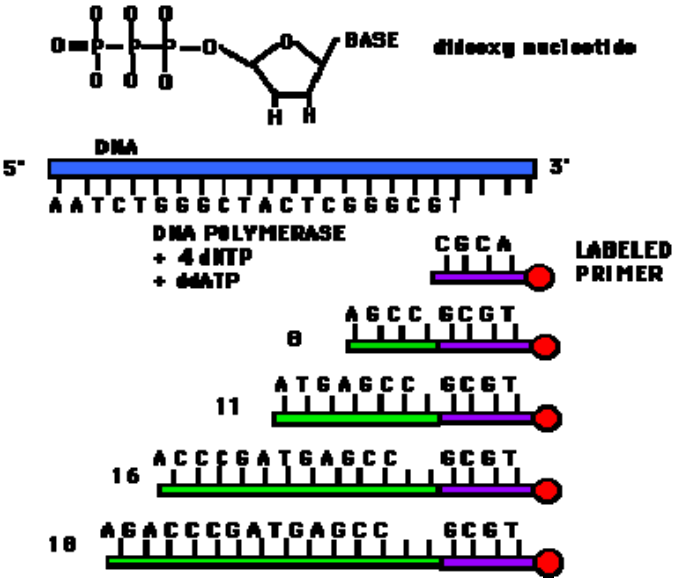


Fig. 36 The structure of a dideoxynucleotide (notice the H atom attached to the 3' carbon). Also depicted in this figure are the ingredients for a Sanger reaction. Notice the different lengths of labeled strands produced in this reaction.

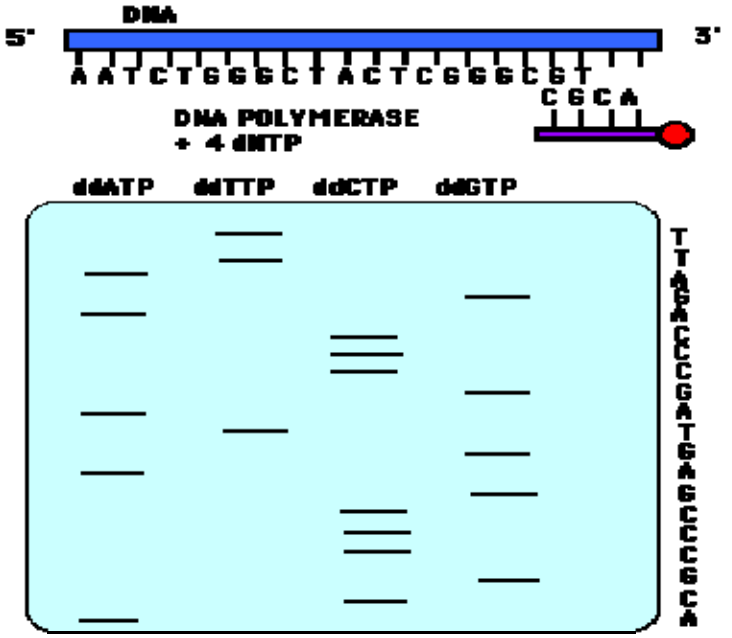


Fig.37. A representation of an acrylamide sequencing gel.

Notice that the sequence of the strand of DNA complementary to the sequenced strand is 5' to 3' ACGCCCGAGTAGCCCAGATT while the sequence of the sequenced strand, 5' to 3', is AATCTGGGCTACTCGGGCGT.

DNA sequencing reactions are just like the PCR reactions for replicating DNA. The reaction mix includes the template DNA, free nucleotides, an enzyme (usually a variant of Taq polymerase) and a 'primer' - a small piece of single-stranded DNA about 20-30 nt long that can hybridize to one strand of the template DNA.

The reaction is initiated by heating until the two strands of DNA separate, then the primer sticks to its intended location and DNA polymerase starts elongating the primer. If allowed to go to completion, a new strand of DNA would be the result. If we start with a billion identical pieces of template DNA, we'll get a billion new copies of one of its strands.

Now the key to this is that most of the nucleotides are regular ones, and just a fraction of them are dideoxy nucleotides....

14.1.3. Putting all four deoxynucleotides into the picture:

Well, OK, it's not so easy reading just C's, as you perhaps saw in the last figure. The spacing between the

bands isn't all that easy to figure out. Imagine, though, that we ran the reaction with *all four* of the dideoxy nucleotides (A, G, C and T) present, and with *different* fluorescent colors on each. NOW look at the gel we'd get (at left). The sequence of the DNA is rather obvious if you know the color codes ... just read the colors from bottom to top: TGCGTCCA-(etc).

14.2. An Automated sequencing

- **Principles of automated fluorescent DNA sequencing**
- Automated fluorescent sequencing utilizes a variation of the Sanger chain-termination protocol developed over 20 years ago. In this method, the DNA to be sequenced acts as a template molecule to which a short, complementary oligonucleotide (primer) will anneal to begin enzymatic extension and amplification of a specific region of double-stranded DNA.
- The newly created fragments will be complementary to the template DNA. This process takes place during the cycle sequencing reaction, a process that each sample we receive must undergo in order to

become amplified and fluorescently labeled for detection on our sequencers.

- In this cycle sequencing reaction, template and primer are combined together with a reaction mixture composed of dNTPs, fluorescently labeled ddNTPs, Amplitaq FS polymerase enzyme and buffer.
- The cycle sequencing reaction is composed of three steps - denaturation, annealing and extension - and takes place in a thermal cycler, an instrument that allows for controlled heating and cooling of our reactions.
- These steps are repeated for 35 cycles to ensure sufficient amplification of the labeled DNA, and takes about 3 1/2 hours to complete.
- During the denaturation step, which occurs at 96ºC, the double-stranded template DNA is first separated into single-stranded molecules. At the annealing stage, the temperature is lowered to 50ºC so that the small primer molecules can find their complementary regions on the now single-stranded template DNA and hybridize, or anneal, correctly.
- The temperature is then raised to 60ºC (extension step) to allow the Taq polymerase enzyme to begin

incorporation of nucleotides into growing chains of newly created fragments that are complementary to the single-stranded template DNA. These extension products begin at the end of the primer and extend in the 3' direction.

- Chain termination occurs during this extension step. Our reaction mixture contains a mixture of deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs), at concentrations that create a statistical probability that a dideoxynucleotide will be incorporated instead of a deoxynucleotide at each nucleotide position in the newly generated fragments. When a dNTP is incorporated, the new fragment will continue to grow.
- A ddNTP contains a hydrogen atom instead of a hydroxyl group at its 3' end and cannot participate in further extension. Therefore, when a ddNTP is incorporated, further chain elongation is blocked and this results in a population of truncated products of varying lengths. When separated by electrophoresis, a "ladder" of these truncated products will form, each differing in size by one nucleotide, with the smallest terminated fragments running fastest on the gel.

- There are four different ddNTPs that correspond to each of the four DNA nucleotides, and each ddNTP has a different color. Thus, each truncated fragment will contain a fluorescently labeled ddNTP at its 3' end, and the sequencing ladder will be composed of colored bands.
- The sequence can then be determined by correlating the color of a band on the gel with its specific ddNTP, and the order in which they ran on the gel.
- So, the first and smallest band visualized will correspond to the first labeled nucleotide incorporated immediately adjacent to the primer.
- The second band will be the fragments that consist of 1 unlabeled dNTP and 1 labeled ddNTP that terminated those particular growing chains.
- The third band will be made up of 2 unlabeled dNTPS followed by 1 labeled ddNTP, and so on up the gel. Once the sample has been amplified and labeled, it must be electrophoresed for separation of the labeled fragments and their visualization. As mentioned before, the ddNTPs are fluorescently labeled. The attached dyes are energy transfer dyes

and consist of a fluorescein energy donor dye linked to an energy acceptor dichlororhodamine dye.

- ► This energy transfer system is much more sensitive than a single dye system and allows us to use less DNA for detection and, in addition, allows us to now sequence very large DNA molecules, such as BACs, PACs and even some bacterial genomic DNA, that previously less sensitive methods were unable to manage.
- ► These dye-labeled fragments are loaded onto the sequencers and during electrophoresis, migrate either through the polyacrylamide gel or liquid polymer and are separated based on their size.
- ► Towards the end of the gel or the capillary, they pass through a region that contains a read window, behind which a laser beam passes back and forth behind the migrating samples.
- ► This laser excites the fluorescent dyes attached to the fragments and they then emit light at a wavelength specific for each dye.
- ► This emitted light is separated according to wavelength by a spectrograph onto a cooled, charge-coupled device, or CCD camera, so that all

four fluorescent emissions can be detected by one laser pass.

The data collection software collects these light intensities from the CCD camera at particular wavelength bands, or virtual filters, and stores them onto the sequencer's computer as digital signals for processing. The analysis software then interprets the fluorescent intensity at each data point and assigns its base call interpretations.

14.3. Shotgun Sequencing

Shotgun sequencing is a method for determining the sequence of a very large piece of DNA. The basic DNA sequencing reaction can only get the sequence of a few hundred nucleotides.

For larger ones (like BAC DNA), we usually fragment the DNA and insert the resultant pieces into a convenient vector (a plasmid, usually) to replicate them. After we sequence the fragments, we try to deduce from them the sequence of the original BAC DNA.

To sequence a BAC, we take millions of copies of it and chop them all up *randomly*. We then insert those into plasmids and for each one we get, we grow lots of it in bacteria and sequence the insert. If we do this to enough fragments, eventually we'll be able to reconstruct the sequence of the original BAC based on the overlapping fragments we've sequenced!

Review Questions

Understand the essential features of a Sanger DiDeoxy Sequencing experiment.

Understand the essential features of a Maxam-Gilbert Chemical Sequencing experiment.

What is a DiDeoxynucleotide? Be able to draw each of the 4 dideoxynucleotides.

Why are diddeoxynucleotides called "chain terminators"?

Why are four sequencing reactions performed in Sanger sequencing?

What is a Sequencing Gel? How does it differ from a standard R.fragment agarose gel?

What are the two most common assay methods used to detect DNA fragments in a sequencing gel?

What is a "nested set" of DNA fragments? Why is this concept important is understanding DNA sequencing methodologies?

Understand how to determine the sequence of a DNA fragment by "reading" the sequencing gel.

How does a Maxam-Gilbert Chemical sequencing experiment differ from a Sanger DiDeoxy sequencing experiment? How is it similar?

CHAPTER FIFTEEN

MOLECULAR TECHNIQUES

Specific learning objectives

At the end of this chapter students are expected to

- ⇒ describe molecular techniques including Northern blotting, souther blotting,
- ⇒ to differentiate the different types of blottings
- ⇒ Describe the essential features of a Polymerase Chain Reaction
- ⇒ List Applications of PCR
- ⇒ to describe uses of RFLP
- ⇒ to describe the importance of DNA typing
- ⇒ Eexplain the principle of RFLp

15.1. Electrophoresis

Electrophoresis is the migration of charged molecules in solution in response to an electric field. Their rate of migration depends on the strength of the field; on the

net charge, size and shape of the molecules and also on the ionic strength, viscosity and temperature of the medium in which the molecules are moving.

As an analytical tool, electrophoresis is simple, rapid and highly sensitive. It is used analytically to study the properties of a single charged species, and as a separation technique.

15.1.1. Support Matrices

Generally the sample is run in a support matrix such as paper, cellulose acetate, starch gel, agarose or polyacrylamide gel. The matrix inhibits convective mixing caused by heating and provides a record of the electrophoretic run: at the end of the run, the matrix can be stained and used for scanning, autoradiography or storage.

In addition, the most commonly used support matrices - agarose and polyacrylamide - provide a means of separating molecules by size, in that they are porous gels. A porous gel may act as a sieve by retarding, or in some cases completely obstructing, the movement of

large macromolecules while allowing smaller molecules to migrate freely. Because dilute agarose gels are generally more rigid and easy to handle than polyacrylamide of the same concentration, agarose is used to separate larger macromolecules such as nucleic acids, large proteins and protein complexes.

Polyacrylamide, which is easy to handle and to make at higher concentrations, is used to separate most proteins and small oligonucleotides that require a small gel pore size for retardation.

15.1.2. Separation of Proteins and Nucleic Acids

Proteins are amphoteric compounds; their nett charge therefore is determined by the pH of the medium in which they are suspended. In a solution with a pH above its isoelectric point, a protein has a nett negative charge and migrates towards the anode in an electrical field. Below its isoelectric point, the protein is positively charged and migrates towards the cathode. The net charge carried by a protein is in addition independent of its size - ie: the charge carried per unit mass (or length, given proteins and nucleic acids are linear

macromolecules) of molecule differs from protein to protein.

At a given pH therefore, and under non-denaturing conditions, the electrophoretic separation of proteins is determined by both size and charge of the molecules.

Nucleic acids however, remain negative at any pH used for electrophoresis and in addition carry a fixed negative charge per unit length of molecule, provided by the PO4 group of each nucleotide of the the nucleic acid. Electrophoretic separation of nucleic acids therefore is strictly according to size.

15.1.3. Separation of Proteins under Denaturing conditions

Sodium dodecyl sulphate (SDS) is an anionic detergent which denatures proteins by "wrapping around" the polypeptide backbone - and SDS binds to proteins fairly specifically in a mass ratio of 1.4:1. In so doing, SDS confers a negative charge to the polypeptide in proportion to its length - ie: the denatured polypeptides become "rods" of negative charge cloud with equal

charge or charge densities per unit length. It is usually necessary to reduce disulphide bridges in proteins before they adopt the random-coil configuration necessary for separation by size: this is done with 2-mercaptoethanol or dithiothreitol. In denaturing SDS-PAGE separations therefore, migration is determined not by intrinsic electrical charge of the polypeptide, but by molecular weight.

15.1.4. Determination of Molecular Weight

This is done by SDS-PAGE of proteins - or PAGE or agarose gel electrophoresis of nucleic acids - of known molecular weight along with the protein or nucleic acid to be characterised. A linear relationship exists between the logarithm of the molecular weight of an SDS-denatured polypeptide, or native nucleic acid, and its *Rf*. The Rf is calculated as the ratio of the distance migrated by the molecule to that migrated by a marker dye-front. A simple way of determining relative molecular weight by electrophoresis *(M*r) is to plot a standard curve of distance migrated vs. log10MW for known samples, and read off the log*Mr* of the sample after measuring distance migrated on the same gel.

15.1.5. Continuous and Discontinuous Buffer Systems

There are two types of buffer systems in electrophoresis, continuous and discontinuous. A continuous system has only a single separating gel and uses the same buffer in the tanks and the gel. In a discontinuous system, a non-restrictive large pore gel, called a stacking gel, is layered on top of a separating gel called a resolving gel. Each gel is made with a different buffer, and the tank buffers are different from the gel buffers. The resolution obtained in a discontinuous system is much greater than that obtained with a continuous system (read about this in any textbook).

15.1.6. Assembling gel apparatus

Assemble two glass plates (one notched) with two side spacers, clamps, grease, etc. as shown by demonstrators or instructions. Stand assembly upright using clamps as supports, on glass plate. Pour some **pre-heated 1% agarose** onto glass plate, place

assembly in pool of agarose: this seals the bottom of the assembly.

15.2. Complementarity and Hybridization

Molecular techniques use one of several forms of complementarity to identify the macromolecules of interest among a large number of other molecules. *Complementarity* is the sequence-specific or shape-specific molecular recognition that occurs when two molecules bind together. For example: the two strands of a DNA double-helix bind because they have complimentary sequences.

Complementarity between a probe molecule and a target molecule can result in the formation of a probe-target complex. This complex can then be located if the probe molecules are tagged with radioactivity or an enzyme. The location of this complex can then be used to get information about the target molecule.

In solution, hybrid molecular complexes (usually called hybrids) of the following types can exist (other combinations are possible):

1) DNA-DNA. A single-stranded DNA (ssDNA) probe molecule can form a double-stranded, base-paired hybrid with a ssDNA target if the probe sequence is the reverse complement of the target sequence.

2) DNA-RNA. A single-stranded DNA (ssDNA) probe molecule can form a double-stranded, base-paired hybrid with an RNA (RNA is usually a single-strand) target if the probe sequence is the reverse complement of the target sequence.

3) Protein-Protein. An antibody probe molecule (antibodies are proteins) can form a complex with a target protein molecule if the antibody's antigen-binding site can bind to an epitope (small antigenic region) on the target protein. In this case, the hybrid is called an 'antigen-antibody complex' or 'complex' for short.

There are two important features of hybridization:

1) Hybridization reactions are specific - the probes will only bind to targets with complimentary sequence

(or, in the case of antibodies, sites with the correct 3-d shape).

2) Hybridization reactions will occur in the presence of large quantities of molecules similar but not identical to the target. That is, a probe can find one molecule of target in a mixture of zillions of related but non-complementary molecules.

These properties allow you to use hybridization to perform a molecular search for one DNA molecule, or one RNA molecule, or one protein molecule in a complex mixture containing many similar molecules.

These techniques are necessary because a cell contains tens of thousands of genes, thousands of different mRNA species, and thousands of different proteins. When the cell is broken open to extract DNA, RNA, or protein, the result is a complex mixture of the entire cell's DNA, RNA, or protein. It is impossible to study a specific gene, RNA, or protein in such a mixture with techniques that cannot discriminate on the basis of sequence or shape. Hybridization techniques allow you

to pick out the molecule of interest from the complex mixture of cellular components and study it on its own.

15.3. Blots

Blots are membranes such as nitrocellulose or coated nylon to which nucleic acids have been permanently bound. Blot hybridizations with specific nucleic acid probes provide critical information regarding gene expression and genome structure. The most common blot applications used in modern laboratories are Northern blots, Southern blots and dot/slot blots. Regardless of the type of blot, the principles of probe synthesis, hybridization, washing and detection are the same.

Blots are named for the target molecule.

Southern Blot: DNA cut with restriction enzymes - probed with radioactive DNA.

Northern Blot: RNA - probed with radioactive DNA or RNA.

Western Blot: Protein - probed with radioactive or enzymatically-tagged antibodies.

The formation of hybrids in solution is of little experimental value - if you mix a solution of DNA with a solution of radioactive probe, you end up with just a radioactive solution. You cannot tell the hybrids from the non-hybridized molecules. For this reason, you must first physically separate the mixture of molecules to be probed on the basis of some convenient parameter.

These molecules must then be immobilized on a solid support, so that they will remain in position during probing and washing. The probe is then added, the non-specifically bound probe is removed, and the probe is detected. The place where the probe is detected corresponds to the location of the immobilized target molecule.

In the case of Southern, Northern, and Western blots, the initial separation of molecules is done on the basis of molecular weight.

In general, the process has the following steps, detailed below:

- Gel electrophoresis
- Transfer to Solid Support
- Blocking
- Preparing the Probe
- Hybridization
- Washing
- Detection of Probe-Target Hybrids

15.3.1. Gel Electrophoresis

This is a technique that separates molecules on the basis of their size. First, a slab of gel material is cast. Gels are usually cast from agarose or poly-acrylamide. These gels are solid and consist of a matrix of long thin molecules forming sub-microscopic pores. The size of the pores can be controlled by varying the chemical composition of the gel. The gel is cast soaked with buffer.

The gel is then set up for electrophoresis in a tank holding buffer and having electrodes to apply an electric field.

The pH and other buffer conditions are arranged so that the molecules being separated carry a net (-) charge so that they will me moved by the electric field from left to right. As they move through the gel, the larger molecules will be held up as they try to pass through the pores of the gel, while the smaller molecules will be impeded less and move faster. This results in a separation by size, with the larger molecules nearer the well and the smaller molecules farther away.

Note that this separates on the basis of size, not necessarily molecular weight. For example, two 1000 nucleotide RNA molecules, one of which is fully extended as a long chain (**A**); the other of which can base-pair with itself to form a hairpin structure (**B**):

As they migrate through the gel, both molecules behave as though they were solid spheres whose diameter is the same as the length of the rod-like molecule. Both have the same molecular weight, but because **B** has

secondary (2') structure that makes it smaller than **A**, **B** will migrate faster than **A** in a gel. To prevent differences in shape (2' structure) from confusing measurements of molecular weight, the molecules to be separated must be in a long extend rod conformation - no 2' structure. In order to remove any such secondary or tertiary structure, different techniques are employed for preparing DNA, RNA and protein samples for electrophoresis.

Preparing DNA for Southern Blots

DNA is first cut with restriction enzymes and the resulting double-stranded DNA fragments have an extended rod conformation without pre-treatment.

Preparing RNA for Northern Blots

Although RNA is single-stranded, RNA molecules often have small regions that can form base-paired secondary structures. To prevent this, the RNA is pre-treated with formaldehyde.

Preparing Proteins for Western Blots

Proteins have extensive 2' and 3' structures and are not always negatively charged. Proteins are treated with the detergent SDS (sodium dodecyl sulfate) which removes 2' and 3' structure and coats the protein with negative charges.

If these conditions are satisfied, the molecules will be separated by molecular weight, with the high molecular weight molecules near the wells and the low molecular weight molecules far from the wells. The distance migrated is roughly proportional to the log of the inverse of the molecular weight (the log of 1/MW).

Gels are normally depicted as running vertically, with the wells at the top and the direction of migration downwards. This leaves the large molecules at the top and the smaller molecules at the bottom. Molecular weights are measured with different units for DNA, RNA, and protein:

- DNA: Molecular weight is measured in base-pairs, or bp, and commonly in kilobase-pairs (1000bp), or kbp.

- RNA: Molecular weight is measured in nucleotides, or nt, and commonly in kilonucleotides (1000nt), or knt. [Sometimes, bases, or b and kb are used.]
- Protein: Molecular weight is measured in Daltons (grams per mole), or Da, and commonly in kiloDaltons (1000Da), or kDa.

On most gels, one well is loaded with a mixture of DNA, RNA, or protein molecules of known molecular weight. These 'molecular weight standards' are used to calibrate the gel run and the molecular weight of any sample molecule can be determined by interpolating between the standards.

Different stains and staining procedures are used for different classes of macromolecules:

Staining DNA

DNA is stained with ethidium bromide (EtBr), which binds to nucleic aids. The DNA-EtBr complex fluoresces under UV light.

Staining RNA

RNA is stained with ethidium bromide (EtBr), which binds to nucleic aids. The RNA-EtBr complex fluoresces under UV light.

Staining Protein

Protein is stained with Coomassie Blue (CB). The protein-CB complex is deep blue and can be seen with visible light.

15.3.2. Transfer to Solid Support

After the DNA, RNA, or protein has been separated by molecular weight, it must be transferred to a solid support before hybridization. Hybridization does not work well in a gel. This transfer process is called blotting and is why these hybridization techniques are called blots.

Usually, the solid support is a sheet of nitrocellulose paper sometimes called a filter because the sheets of nitrocellulose were originally used as filter paper, although other materials are sometimes used. DNA,

RNA, and protein stick well to nitrocellulose in a sequence-independent manner.

The DNA, RNA, or protein can be transferred to nitrocellulose in one of two ways:

1) Electrophoresis, which takes advantage of the molecules' negative charge:
2) Capillary blotting, where the molecules are transferred in a flow of buffer from wet filter paper to dry filter paper:

In a Southern Blot, the DNA molecules in the gel are double-stranded, so they must be made single stranded in order for the probe to hybridize to them. To do this, the DNA is transferred using a strongly alkaline buffer, which causes the DNA strands to separate - this process is called denaturation - and bind to the filter as single-stranded molecules.

RNA and protein are run in the gels in a state that allows the probe to bind without this pre-treatment.

15.3.3. Blocking

At this point, the surface of the filter has the separated molecules on it, as well as many spaces between the lanes, etc., where no molecules have yet bound. If one added the probe directly to the filter now, the probe would stick to these blank parts of the filter, like the molecules transferred from the gel did. This would result in a filter completely covered with probe which would make it impossible to locate the probe-target hybrids.

For this reason, the filters are soaked in a blocking solution which contains a high concentration of DNA, RNA, or protein. This coats the filter and prevents the probe from sticking to the filter itself. During hybridization, we want the probe to bind only to the target molecule.

15.3.4. Preparing the Probe

15.3.4.1. Radioactive DNA probes for Southerns and Northerns

The objective is to create a radioactive copy of a double-stranded DNA fragment. The process usually begins

with a restriction fragment of a plasmid containing the gene of interest. The plasmid is digested with particular restriction enzymes and the digest is run on an agarose gel. Since a plasmid is usually less than 20 kbp long, this results in 2 to 10 DNA fragments of different lengths. If the restriction map of the plasmid is known, the desired band can be identified on the gel. The band is then cut out of the gel and the DNA is extracted from it. Because the bands are well separated by the gel, the isolated DNA is a pure population of identical double-stranded DNA fragments.

The DNA restriction fragment (template) is then labeled by Random Hexamer Labeling.:

1) The template DNA is denatured - the strands are separated - by boiling.

2) A mixture of DNA hexamers (6 nucleotides of ssDNA) containing all possible sequences is added to the denatured template and allowed to base-pair. They pair at many sites along each strand of DNA.

3) DNA polymerase is added along with dATP, dGTP, dTTP, and radioactive dCTP. Usually, the phosphate bonded to the sugar (the a-phosphate, the one that is incorporated into the DNA strand) is synthesized from phosphorus-32 (32P), which is radioactive.

4) The mixture is boiled to separate the strands and is ready for hybridization.

This produces a radioactive single-stranded DNA copy of both strands of the template for use as a probe.

15.3.4.2. Radioactive Antibodies for Westerns

Antibodies are raised by injecting a purified protein into an animal, usually a rabbit or a mouse. This produces an immune response to that protein. Antibodies isolated from the serum (blood) of that rabbit will bind to the protein used for immunization. These antibodies are protein molecules and are not themselves radioactive.

They are labeled by chemically modifying the side chains of tyrosines in the antibody with iodine-125 (125I), which is radioactive. A set of enzymes catalyzes the following reaction:

antibody-tyrosine + 125I- + H2O2 ---------> H2O + 125iodo-tyrosine-antibody

15.3.4.3. Enzyme-conjugated Antibodies for Westerns

Antibodies against a particular protein are raised as above and labeled by chemically cross-linking the antibody molecules to molecules of an enzyme. The resulting antibody-enzyme conjugate is still able to bind to the target protein.

15.3.5. Hybridization

In all three blots, the labeled probe is added to the blocked filter in buffer and incubated for several hours to allow the probe molecules to find their targets.

15.3.6. Washing

After hybrids have formed between the probe and target, it is necessary to remove any probe that is on the filter that is not stuck to the target molecules. Because the nitrocellulose is absorbent, some of the probe soaks into the filter and must be removed. If it is not removed,

the whole filter will be radioactive and the specific hybrids will be undetectable.

To do this, the filter is rinsed repeatedly in several changes of buffer to wash off any un-hybridized probe.

In Southerns and Northerns, hybrids can form between molecules with similar but not necessarily identical sequences (For example, the same gene from two different species.). This property can be used to study genes from different organisms or genes that are mutated. The washing conditions can be varied so that hybrids with differing mismatch frequencies are maintained. This is called 'controlling the 'stringency' - the higher the wash temperature, the more stringent the wash, the fewer mismatches per hybrid are allowed.

15.3.7. Detecting the Probe-Target Hybrids

At this point, you have a sheet of nitrocellulose with spots of probe bound wherever the probe molecules could form hybrids with their targets. The filter now looks like a blank sheet of paper - you must now detect where the probe has bound.

15.3.7.1. Autoradiography

If the probe is radioactive, the radioactive particles that it emits can expose X-ray film. If you press the filter up against X-ray film and leave it in the dark for a few minutes to a few weeks, the film will be exposed wherever the probe bound to the filter. After development, there will be dark spots on the film wherever the probe bound.

15.3.7.2. Enzymatic Development

If an antibody-enzyme conjugate was used as a probe, this can be detected by soaking the filter in a solution of a substrate for the enzyme. Usually, the substrate produces an insoluble colored product (a chromogenic substrate) when acted upon by the enzyme. This produces a deposit of colored product wherever the probe bound.

15.4. The Polymerase Chain Reaction

Introduction

Polymerase chain reaction is a technique, which is used to amplify the number of copies of a specific region of DNA, in order to produce enough DNA to be adequately tested. This technique can be used to identify with a very high-probability, disease-causing viruses and/or bacteria, a deceased person, or a criminal suspect. In order to use PCR, one must already know the exact sequences which flank (lie on either side of) both ends of a given region of interest in DNA (may be a gene or any sequence).

One of the most important and profound contributions to molecular biology is the advent of the polymerase chain reaction (PCR). PCR, one of the most significant advances in DNA and RNA-based technologies, is a powerful tool enabling us to detect a single genome of an infectious agent in any body fluid with improved accuracy and sensitivity. Many infectious agents which

are missed by routine cultures, serological as-says, DNA probes, and Southern blot hybridizations can be detected by PCR. Therefore, PCR-based tests are best suited for the clinical and epidemiological investigation of pathogenic bacteria and viruses.

The introduction of PCR in the late 1980's dominated the market because it was superior to all previously used culture techniques and the more recently developed DNA probes and kits. PCR based tests are several orders of magnitude more sensitive than those based on direct hybridization with the DNA probe. PCR does not depend on the ability of an organism to grow in culture. Furthermore, PCR is fast, sensitive and capable of copying a single DNA sequence of a viable or non-viable cell over a billion times within 3-5 hours.

The sensitivity of the PCR test is also based on the fact that PCR methodology requires only 1-5 cells for detection, whereas a positive culture requires an inoculum equivalent to about 1000 to 5000 cells, making PCR the most sensitive detection method available.

How PCR Works

PCR is an in vitro method for amplifying a selected nucleic acid sequence. To target the amplification to a specific DNA segment, two primers bearing the complementary sequences that are unique to the target gene are used. These two primers hybridize to opposite strands of the target DNA, thus enabling DNA polymerase to extend the sequence between them. Each cycle produces a complementary DNA strand to the target gene. Consequently, the product of each cycle is doubled, generating an exponential increase in the overall number of copies synthesized.

In addition to PCR rapidly becoming a major tool in the diagnostic repertoire for infectious disease, it promises to play a role in the diagnosis and monitoring of cancer, in clinical genetics, and in forensics. In amplification methods, a target nucleic acid (DNA or RNA) isolated from tissue or fluids is amplified enzymatically. The amplified product, the amplicon, is detected either by hybridization with a homologous probe or by direct visualization after enzyme immunoassay. Amplification methods are analogous to culture of bacteria, in which a

few organisms replicate on a plate until visible colonies are formed. The tar-get sequence may, with suitable choice of technique, be either DNA or RNA, and as potentially as few as a single target molecule can be amplified and detected. Amplification technique has been used extensively in viral diagnosis for a wide range of pathogens. But nucleic acid amplification has particular value in retroviral studies, since latent, unculturable viruses can be detected easily.

These methods are also useful in epidemiologic analysis, since suitable primers may be used to discriminate between strains of the same organism. The speed, sensitivity, and specificity of amplification techniques allow rapid, direct diagnosis of diseases which formerly could be diagnosed only slowly, indirectly (e.g., by serology), at great expense, or not at all.

The Significance of Amplification Techniques

Nucleic acid amplification and serological techniques for the diagnosis of infectious disease have complementary characteristics. Because amplification methods directly detect minute quantities of pathogen genetic material,

they can pro-vide acute phase diagnosis with high sensitivity without the need to await antibody formation. Amplification methods will detect a pathogen only if nucleic acid from that organism is actually present, so confusion with infections in the distant past is unlikely.

Serology may also be used to determine whether a patient has been exposed to a pathogen, regardless of whether the infecting organism is actually present, whereas amplification methods require the presence of the organism. An antibody response also pro-vides information on the pathogenicity or invasive-ness of an organism such as Legionella sp. which may be normally present in the environment and thus contain clinical specimens.

Amplification is unlikely to discriminate between colonization and infection; therefore, immunoserology should be used in conjunction with PCR technology. Amplification techniques provide adjunct methods for special cases rather than replacement of existing technology. For pathogens that fail to grow in vitro, such as M. leprae or T. gondii, for pathogens that grow slowly, such as M. tuberculosis, and for extremely

hazardous pathogens such as HIV or Francisella tularensis, amplification methods may provide significant savings in time and effort or re-duce hazard to personnel.

PCR is used widely in:

- Molecular cloning
- Pathogen detection
- Genetic engineering
- Mutagenesis
- Genetics, producing molecular markers
- Forensics

15.4.1. Principle of the PCR

The purpose of a PCR (Polymerase Chain Reaction) is to make a huge number of copies of a gene. This is necessary to have enough starting template for sequencing. There are three major steps in a PCR, which are repeated for 30 or 40 cycles.

A PCR consists of multiple cycles of annealing of the oligonucleotides, synthesis of a DNA chain, denaturation of the synthesized DNA. This is done on an automated

cycler, which can heat and cool the tubes with the reaction mixture in a very short time.

1. Denaturation at 94°C :

During the denaturation, the double strand melts open to single stranded DNA, all enzymatic reactions stop.

60ºC is hot enough to denature an ordinary DNA polymerase; however the DNA polymerase used in PCR comes from *Thermus aquaticus* (*Tac* for short) which lives in hot springs and consequently is a DNA polymerase that is adapted to high temperatures

2. Annealing at 54°C :

The primers are jiggling around, caused by the Brownian motion. Ionic bonds are constantly formed and broken between the single stranded primer and the single stranded template. The more stable bonds last a little bit longer (primers that fit exactly) and on that little piece of double stranded DNA (template and primer), the polymerase can attach and starts copying the template.

Once there are a few bases built in, the ionic bond is so strong between the template and the primer, that it does not break anymore.

3. Extension at 72°C:

This is the ideal working temperature for the polymerase. The primers, where there are a few bases built in, already have a stronger ionic attraction to the template than the forces breaking these attractions.

Primers that are on positions with no exact match get loose again (because of the higher temperature) and don't give an extension of the fragment.

The bases (complementary to the template) are coupled to the primer on the 3' side (the polymerase adds dNTP's from 5' to 3', reading the template from 3' to 5' side, bases are added complementary to the template).

Using automated equipment, each cycle of replication can be completed in less than 5 minutes. After 30 cycles, what began as a single molecule of DNA has been amplified into more than a billion copies (2^{30} = 1.02 x 10^9).

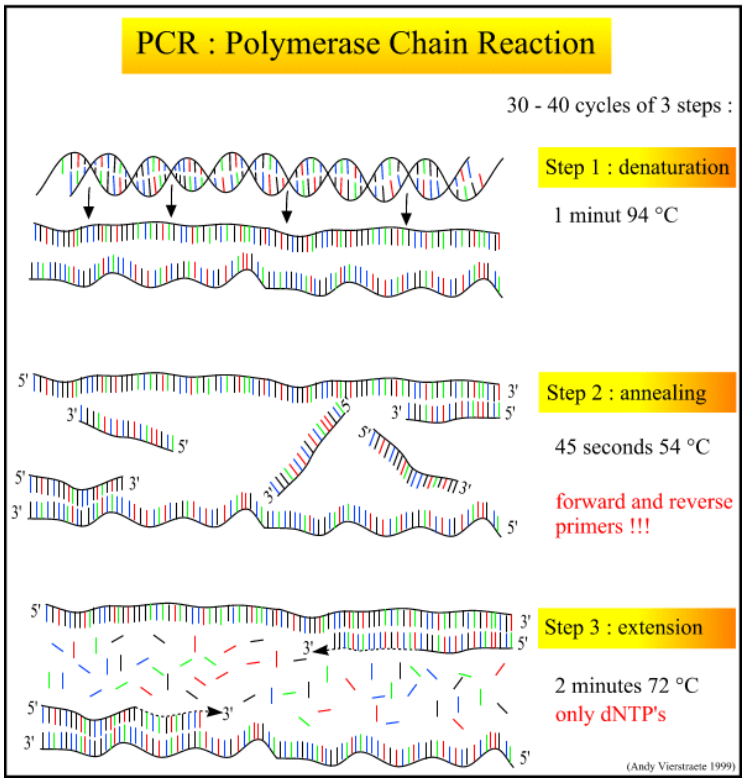


Fig.38. The steps in PCR

Because both strands are copied during PCR, there is an **exponential** increase of the number of copies of the gene. Suppose there is only one copy of the wanted gene before the cycling starts, after one cycle, there will

be 2 copies, after two cycles, there will be 4 copies, three cycles will result in 8 copies and so on.

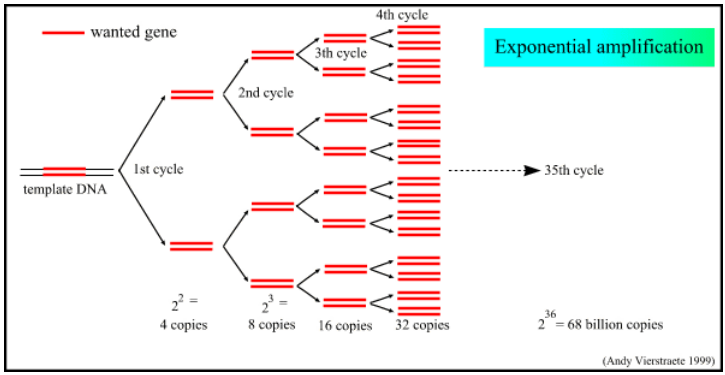


Fig. 39 The exponential amplification of the gene in PCR.

15.4.2. The procedure

- In order to perform PCR, one must know at least a portion of the sequence of the DNA molecule that one wish to replicate.
- then synthesize primers: short oligonucleotides that are precisely complementary to the sequence at the 3' end of each strand of the DNA you wish to amplify.
- The DNA sample is heated to separate its strands and mixed with the primers.

- If the primers find their complementary sequences in the DNA, they bind to them.
- Synthesis begins (as always 5' -> 3') using the original strand as the template.
- The reaction mixture must contain
    - all four deoxynucleotide triphosphates (dATP, dCTP, dGTP, dTTP)
    - a DNA polymerase. It helps to use a DNA polymerase that is not denatured by the high temperature needed to separate the DNA strands.
- Polymerization continues until each newly-synthesized strand has proceeded far enough to contain the site recognized by the other primer.
- Now one has two DNA molecules identical to the original molecule.
- take these two molecules, heat them to separate their strands, and repeat the process.
- Each cycle doubles the number of DNA molecules.

Using automated equipment, each cycle of replication can be completed in less than 5 minutes. After 30 cycles, what began as a single molecule of DNA has

been amplified into more than a billion copies (2^{30} = 1.02 x 10^9). With PCR, it is routinely possible to amplify enough DNA from a single hair follicle for DNA typing. Some workers have successfully amplified DNA from a single sperm cell.

The PCR technique has even made it possible to analyze DNA from microscope slides of tissue preserved years before. However, the great sensitivity of PCR makes contamination by extraneous DNA a constant problem.

Before the PCR product is used in further applications, it has to be checked if:

a. **There is a product formed.**
   Though biochemistry is an exact science, not every PCR is successful. There is for example a possibility that the quality of the DNA is poor, that one of the primers doesn't fit, or that there is too much starting template
b. **The product is of the right size**
   It is possible that there is a product, for example a band of 500 bases, but the expected gene should be

1800 bases long. In that case, one of the primers probably fits on a part of the gene closer to the other primer. It is also possible that both primers fit on a totally different gene.

c. **Only one band is formed**. As in the description above, it is possible that the primers fit on the desired locations, and also on other locations. In that case, you can have different bands in one lane on a gel.

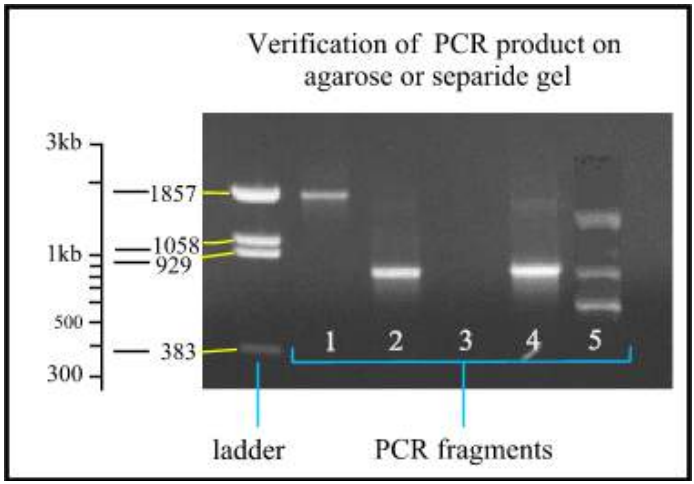


Fig.40. Verification of the PCR product on gel.

The ladder is a mixture of fragments with known size to compare with the PCR fragments. Notice that the distance between the different fragments of the ladder is logarithmic. Lane 1: PCR fragment is approximately 1850 bases long. Lane 2 and 4: the fragments are approximately 800 bases long. Lane 3: no product is formed, so the PCR failed. Lane 5: multiple bands are formed because one of the primers fits on different places.

Variations - starting or ending with SS nucleic acids

- rt-PCR starts with RNA and reverse transcriptase
- PCR Sequencing - single strands generated by an excess of one primer.

Fidelity

- Primer annealing (hybridization) specificity depends on:
    - primer length
    - match to template sequences (degenerate primers)
    - GC content of the primer

    - heterogeneity of the DNA sample
- Polymerase reaction
    - Intrinsic enzymatic fidelity, proofreading ability.
    - Solution conditions, dNTP pools

- The availability of thermostable polymerases and the design of thermal cycler machines made PCR widely accessible.
- Synthesis times can be manipulated to guarantee just the synthesis of shorter DNAs.
- To amplify RNA rather than DNA sequences, a cDNA copy of the RNA can be made by reverse transcription prior to PCR. The combination is called RT-PCR.
- A diverse variety of PCR's have been devised.
- PCR has been made quantitative by the use of fluorescence that can be read at each cycle. There are several strategies: Taqman, Molecular Beacons, dual oligo binding. PCR is extremely sensitive. Contamination is a constant worry.

Real-Time PCR: The TaqMan Method

The advent of Polymerase Chain Reaction (PCR) by Kary B. Mullis in the mid-1980s revolutionized molecular biology as we know it. PCR is a fairly standard procedure now, and its use is extremely wide-ranging. At its most basic application, PCR can amplify a small amount of template DNA (or RNA) into large quantities in a few hours. This is performed by mixing the DNA with primers on either side of the DNA (forward and reverse), *Taq* polymerase (of the species *Thermus aquaticus*, a thermophile whose polymerase is able to withstand extremely high temperatures), free nucleotides (dNTPs for DNA, NTPs for RNA), and buffer.

The temperature is then alternated between hot and cold to denature and reanneal the DNA, with the polymerase adding new complementary strands each time. In addition to the basic use of PCR, specially designed primers can be made to ligate two different pieces of DNA together or add a restriction site, in addition to many other creative uses. Clearly, PCR is a procedure that is an integral addition to the molecular

biologist's toolbox, and the method has been continually improved upon over the years. (Purves, et al. 2001)

Fairly recently, a new method of PCR quantification has been invented. This is called "real-time PCR" because it allows the scientist to actually view the increase in the amount of DNA as it is amplified. Several different types of real-time PCR are being marketed to the scientific community at this time, each with their advantages. This web site will explore one of these types, TaqMan® real-time PCR, as well as give an overview of the other two types of real-time PCR, molecular beacon and SYBR® Green.

How TaqMan® works:

TaqMan® utilizes a system that is fairly easy to grasp conceptually. First, we must take a look at the TaqMan® probe.

The probe consists of two types of fluorophores, which are the fluorescent parts of reporter proteins (Green Fluorescent Protein (GFP) has an often-used fluorophore). While the probe is attached or unattached

to the template DNA and before the polymerase acts, the quencher (Q) fluorophore (usually a long-wavelength colored dye, such as red) reduces the fluorescence from the reporter (R) fluorophore (usually a short-wavelength colored dye, such as green). It does this by the use of Fluorescence (or Förster) Resonance Energy Transfer (FRET), which is the inhibition of one dye caused by another without emission of a proton. The reporter dye is found on the 5' end of the probe and the quencher at the 3' end.

Once the TaqMan® probe has bound to its specific piece of the template DNA after denaturation (high temperature) and the reaction cools, the primers anneal to the DNA. *Taq* polymerase then adds nucleotides and removes the Taqman® probe from the template DNA. This separates the quencher from the reporter, and allows the reporter to give off its emit its energy. This is then quantified using a computer. The more times the denaturing and annealing takes place, the more opportunities there are for the Taqman® probe to bind and, in turn, the more emitted light is detected.

Quantification:

The specifics in quantification of the light emitted during real-time PCR are fairly involved and complex.

Other Real-Time PCR Methods:

There are two other types of real-time PCR methods, the molecular beacon method and the SYBR® Green method. The molecular beacon method utilizes a reporter probe that is wrapped around into a hairpin. It also has a quencher dye that must be in close contact to the reporter to work. An important difference of the molecular beacon method in comparison to the TaqMan® method is that the probe remains intact throughout the PCR product, and is rebound to the target at every cycle. The SYBR® Green probe was the first to be used in real-time PCR. It binds to double-stranded DNA and emits light when excited. Unfortunately, it binds to any double-stranded DNA which could result in inaccurate data, especially compared with the specificity found in the other two methods.

15.5. Restriction Fragment Length Polymorphisms

Restriction fragment length polymorphism (RFLP), is a method by which mutations in or different alleles of genes are recognized using restriction enzymes and electrophoresis.

Different alleles have different patterns of restriction sites and therefore produce a different pattern of bands on a gel.

RFLP, its name is a bit imprecise, but here's what it means. Basically, a RFLP is a band in a gel or in a southern blot produced from a gel.

RFLPs have provided valuable information in many areas of biology, including:

- Screening human DNA for the presence of potentially deleterious genes
- Providing evidence to establish the innocence of or a probability of the guilt of, a crime suspect by DNA "fingerprinting".

A RFLP is something that one makes from the genome, not something that exists on its own. Therefore, some RFLPs are produced from:

- DNA sequences in genes (both introns and exons),
- controlling regions like promoters, and
- the bulk of DNA, which seems to have no function at all.

In fact, most RFLPs used in criminal work have no function at all, but, like other RFLPs, they can be used by an investigator to identifiy individual DNA, to map genes or to follow their passage from one generation to the next.

Polymorphisms in the lengths of particular restriction fragments can be used as molecular markers on the genetic chromosome. Polymorphisms are inherited differences found among the individuals in a population.

Restriction enzymes cut DNA at precise points producing:

- a collection of DNA fragments of precisely defined length.
- These can be separated by electrophoresis, with the smaller fragments migrating farther than the larger fragments.
- One or more of the fragments can be visualized with a "probe" — a molecule of single-stranded DNA that is
    - complementary to a run of nucleotides in one or more of the restriction fragments and is
    - radioactive (or fluorescent).

If probes encounter a complementary sequence of nucleotides in a test sample of DNA, they bind to it by Watson-Crick base pairing and thus identify it.

Fig. 41. Hemoglobin sequence

Case 1: Screening for the sickle-cell gene

Sickle cell anemia is a genetic disease in which both genes in the patient encode the amino acid valine (Val) in the sixth position of the beta chain ($beta^S$) of the hemoglobin molecule. "Normal" beta chains ($beta^A$) have glutamic acid at this position. The only difference between the two genes is the substitution of a T for an A in the middle position of codon 6. This

- converts a GAG codon (for Glu) to a GTG codon for Val and
- abolishes a sequence (CTGAGG, which spans codons 5, 6, and 7) recognized and cut by one of the restriction enzymes.

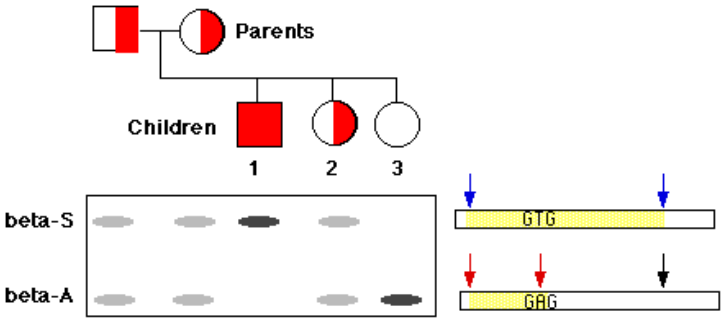


Fig. 42.The pedigree of a family whose only son has sickle-cell disease

When the normal gene ($beta^A$) is digested with the enzyme and the fragments separated by electrophoresis, the probe bind to a short fragment (between the GAG and left arrows).However, the enzyme cannot cut the sickle-cell gene at this site, so the probe attaches to a much larger fragment (between the above two arrows).

The figure shows the pedigree of a family whose only son has sickle-cell disease. Both his father and mother were heterozygous (semifilled box and circle respectively) as they had to be to produce an afflicted child (solid box). The electrophoresis patterns for each member of the family are placed directly beneath them.

Note that the two homozygous children (1 and 3) have only a single band, but these are more intense because there is twice as much DNA in them.

In this example, a change of a single nucleotide produced the RFLP. This is a very common cause of RFLPs and now such polymorphisms are often referred to as single nucleotide polymorphisms.

How can these tools be used?

By testing the DNA of prospective parents, their genotype can be determined and their odds of producing an afflicted child can be determined. In the case of sickle-cell disease, if both parents are heterozygous for the genes, there is a 1 in 4 chance that they will produce a child with the disease. Amniocentesis and chorionic villus sampling make it possible to apply the same techniques to the DNA of a fetus early in pregnancy. The parents can learn whether the unborn child will be free of the disease or not. They may choose to have an abortion rather than bring an afflicted child into the world.

15.5.1. Screening for a RFLP "marker"

If a particular RFLP is usually associated with a particular genetic disease, then the presence or absence of that RFLP can be used to counsel people about their risk of developing or transmitting the disease.

The assumption is that the gene they are really interested in is located so close to the RFLP that the presence of the RFLP can serve as a surrogate for the disease gene itself. But people wanting to be tested cannot simply walk in off the street. Because of crossing over, a particular RFLP might be associated with the mutant gene in some people, with its healthy allele in others. Thus it is essential to examine not only the patient but as many members of the patient's family as possible.

The most useful probes for such analysis bind to a unique sequence of DNA; that is, a sequence occurring at only one place in the genome. Often this DNA is of unknown, if any, function. This can actually be helpful as this DNA has been free to mutate without harm to the

owner. The probe will hybridize (bind to) different lengths of digested DNA in different people depending on where the enzyme cutting sites are that each person has inherited. Thus a large variety of alleles (polymorphisms) may be present in the population. Some people will be homozygous and reveal a single band; others (e.g., all the family members shown below) will be heterozygous with each allele producing its band.

The pedigree shows the inheritance of a RFLP marker through three generations in a single family. A total of 8 alleles (numbered to the left of the blots) are present in the family. The RFLPs of each member of the family are placed directly below his (squares) or her (circles) symbol and RFLP numbers.

If, for example, everyone who inherited RFLP 2 also has a certain inherited disorder, and no one lacking RFLP 2 has the disorder, we deduce that the gene for the disease is closely linked to this RFLP. If the parents decide to have another child, prenatal testing could reveal whether that child was apt to come down with the disease.

But note that crossing over during gamete formation could have moved the RFLP to the healthy allele.

So, the greater the distance between the RFLP and the gene locus, the lower the probability of an accurate diagnosis.

15.6. DNA FINGERPRINTING

- It seems certain that if one could read the entire sequence of DNA in each human, one would never find two that were identical unless the samples were from identical siblings; i.e., derived from a single zygote. So each person's DNA is as unique as a fingerprint.
- Like the fingerprints that came into use by detectives and police labs during the 1930s, each person has a unique DNA fingerprint. Unlike a conventional fingerprint that occurs only on the fingertips and can be altered by surgery.
- DNA fingerprint is the same for every cell, tissue, and organ of a person. It cannot be altered by any known treatment. Consequently, DNA fingerprinting is rapidly becoming the primary method for

identifying and distinguishing among individual human beings.

- An additional application of DNA fingerprint technology is the diagnosis of inherited disorders in adults, children, and unborn babies. The technology is so powerful that, for example, even the blood-stained clothing of Abraham Lincoln could be analyzed for evidence of a genetic disorder called Marfan's Syndrome.

Making DNA Fingerprints

DNA fingerprinting is a laboratory procedure that requires six steps:

1: Isolation of DNA. DNA must be recovered from the cells or tissues of the body. Only a small amount of tissue - like blood, hair, or skin - is needed. For example, the amount of DNA found at the root of one hair is usually sufficient.

2: Cutting, sizing, and sorting. Special enzymes called restriction enzymes are used to cut the DNA at specific places. For example, an enzyme called EcoR1, found in bacteria, will cut DNA only when the sequence GAATTC occurs. The DNA pieces are

sorted according to size by a sieving technique called electrophoresis. The DNA pieces are passed through a gel made from seaweed agarose (a jelly-like product made from seaweed). This technique is the biotechnology equivalent of screening sand through progressively finer mesh screens to determine particle sizes.

3: Transfer of DNA to nylon. The distribution of DNA pieces is transferred to a nylon sheet by placing the sheet on the gel and soaking them overnight.

4-5: Probing. Adding radioactive or colored probes to the nylon sheet produces a pattern called the DNA fingerprint. Each probe typically sticks in only one or two specific places on the nylon sheet.

6: DNA fingerprint. The final DNA fingerprint is built by using several probes (5-10 or more) simultaneously. It resembles the bar codes used by grocery store scanners.

Uses of DNA Fingerprints

DNA fingerprints are useful in several applications of human health care research, as well as in the justice system.

Diagnosis of Inherited Disorders

DNA fingerprinting is used to diagnose inherited disorders in both prenatal and newborn babies in hospitals around the world. These disorders may include cystic fibrosis, hemophilia, Huntington's disease, familial Alzheimer's, sickle cell anemia, thalassemia, and many others.

Early detection of such disorders enables the medical staff to prepare themselves and the parents for proper treatment of the child. In some programs, genetic counselors use DNA fingerprint information to help prospective parents understand the risk of having an affected child. In other programs, prospective parents use DNA fingerprint information in their decisions concerning affected pregnancies.

Developing Cures for Inherited Disorders

Research programs to locate inherited disorders on the chromosomes depend on the information contained in DNA fingerprints. By studying the DNA fingerprints of relatives who have a history of some particular disorder, or by comparing large groups of people with and without the disorder, it is possible to identify DNA patterns

associated with the disease in question. This work is a necessary first step in designing an eventual genetic cure for these disorders.

Biological Evidence

FBI and police labs around the U.S. have begun to use DNA fingerprints to link suspects to biological evidence - blood or semen stains, hair, or items of clothing - found at the scene of a crime. Since 1987, hundreds of cases have been decided with the assistance of DNA fingerprint evidence.

Another important use of DNA fingerprints in the court system is to establish paternity in custody and child support litigation. In these applications, DNA fingerprints bring an unprecedented, nearly perfect accuracy to the determination.

Personal Identification

Because every organ or tissue of an individual contains the same DNA fingerprint, the U.S. armed services have just begun a program to collect DNA fingerprints from all personnel for use later, in case they are needed to identify casualties or persons missing in action. The DNA method will be far superior to the dogtags, dental records, and blood typing strategies currently in use.

Review Questions

1. What are the possible method for detecting probes
2. PCR is a combination of what two types of procedures?
3. What is the role of DNA denaturation in PCR? of renaturation? of polymerization?
4. How are these procedures combined into a single cycle procedure?
5. Why is use of heat-resistant DNA polymerases such as Taq polymerase useful in PCR?
6. What is Taq polymerase?
7. Why does the amount of DNA increase exponentially as a function of number of PCR cycles?
8. PCR can be used to prepare the nested set of fragments for a Sanger dideoxy sequencing experiment. In this PCR application, only one Primer is used and ddNTPs are included as well as the dNTPs.

9. Describe more completely how you would do this experiment.In this experiment, the amount of DNA increases arithmetically rather than exponentially. Why is this the case?

10. What are the differences between northern blotting and southern blotting?

11. What are the bases for using blotting techniques?

12. What are the uses of blotting techniques?

GLOSSARY OF TERMS COMMONLY USED IN MOLECULAR BIOLOGY

AGAROSE GEL ELECTROPHORESIS - A method for separating nucleic acids (DNA or RNA) within a gel made of agarose in a suitable buffer under the influence of an electrical field. Suitable for separation of large fragments of nucleic acid, separation is based primarily upon the size of the nucleic acid.

ALLELE - One of several alternate forms of a gene occupying a given locus on a chromosome or plasmid.

AMINO ACIDS - The 20 basic building blocks of proteins, consisting of the basic formula NH2-CHR-COOH, where "R" is the side chain which defines the amino acid:

AMINO TERMINUS - Refers to the NH2 end of a *peptide* chain (by custom drawn at the left of a protein sequence)

AMPLIFICATION - Refers to the production of additional copies of a chromosomal sequence, found either as intrachromosomal or extrachromosomal DNA. Also refers to the in vitro process in the *polymerase chain reaction*.

AMPLIMER - Region of DNA sequence which is amplified during a *PCR* reaction and which is defined by a pair of *PCR primers* (these *primer* pairs are sometimes called amplimers).

NCHOR SEQUENCE - A hydrophobic amino acid sequence which fixes a segment of a newly synthesized, *translocating* protein within the lipid bilayer membrane of the endoplasmic reticulum.

ANNEAL - See HYBRIDIZATION.

ANTISENSE STRAND (OR PRIMER) - Refers to the RNA or DNA strand of a *duplex* molecule which is *complementary* to that encoding a *polypeptide*. More specifically, the DNA strand which serves as *template* for the synthesis of RNA and which is complementary to it. "Antisense *oligonucleotides*" *hybridize* to *mRNA*, and are used to prime *cDNA* synthesis.

ASSEMBLED EPITOPE - See CONFORMATIONAL EPITOPE.

AUTORADIOGRAPHY - A process to detect radioactively labeled molecules (which usually have been separated in an *SDS-PAGE* or *agarose gel*) based on their ability to create an image on photographic or X-ray film.

AVIDIN - A glycoprotein which binds to *biotin* with very high affinity (Kd = 10-15).

BACK MUTATION - Reverse the effect of a point or frame-shift mutation that had altered a gene; thus it restores the wild-type phenotype (see REVERTANT).

BACTERIOPHAGE - A virus that infects bacteria; often simply called a phage. The phages which are most often used in molecular biology are the *E. coli* viruses lambda, M13 and T7.

BASE - The *purine* or *pyrimidine* component of a *nucleotide*; often used to refer to a *nucleotide* residue within a nucleic acid chain.

BASE PAIR - One pair of *complementary nucleotides* within a *duplex* strand of a nucleic acid. Under Watson-Crick rules, these pairs consist of one *pyrimidine* and one *purine*: i.e., C-G, A-T (DNA) or A-U (RNA). However, "noncanonical" base pairs (e.g., G-U) are common in RNA *secondary structure*.

BIOTIN - A coenzyme which is essential for carboxylation reactions (see AVIDIN).

BLUNT END - A terminus of a *duplex* DNA molecule which ends precisely at a base pair, with no *overhang* (unpaired nucleotide) in either strand. Some but not all *restriction endonucleases* leave blunt ends after cleaving DNA. Blunt-ended DNA can be *ligated* nonspecifically to other blunt-ended DNA molecules (compare with STICKY END). 5'-->3' NNNCCC GGGNNN *Sma*I cut, no overhang NNNGGG CCCNNN 3'<--5'

bp - "base pair"

BOX - Refers to a short nucleic acid *consensus sequence* or *motif* that is universal within kingdoms of organisms. Examples of DNA boxes are the Pribow box (TATAAT) for *RNA polymerase*, the Hogness box (TATA) that has a similar function in eukaryotic organisms, and the homeo box. RNA boxes have also been described, such as Pilipenko's Box-A motif that may be involved in ribosome binding in some viral RNAs.

C TERMINUS - See CARBOXYL TERMINUS.

CARBOXYL TERMINUS - Refers to the COOH end of a *peptide* chain (by custom drawn at the right of a protein sequence)

cDNA - Complementary DNA. A DNA molecule which was originally copied from an RNA molecule by *reverse transcription*. The term "cDNA" is commonly used to describe double-stranded DNA which originated from a single-stranded RNA molecule, even though only one strand of the DNA is truly complementary to the RNA.

cDNA LIBRARY - A collection of *cDNA* fragments, each of which has been cloned into a separate *vector* molecule.

CAP - A 7-methyl guanosine residue linked 5' to 5' through a triphosphate bridge to the 5' end of eukaryotic *mRNAs*; facilitates initiation of *translation*.

CHAIN TERMINATOR - See DIDEOXYNUCLEOTIDE.

CHAPERONE PROTEINS - A series of proteins present in the endoplasmic reticulum which guide the proper folding of secreted proteins through a complex series of binding and release reactions.

CHROMOSOME WALKING - The sequential isolation of *clones* carrying overlapping sequences of DNA which span large regions of a chromosome. Overlapping regions of clones can be identified by *hybridization*.

CLONE - Describes a large number of cells, viruses, or molecules which are identical and which are derived from a single ancestral cell, virus or molecule. The term can be used to refer to the process of isolating single cells or viruses and letting them proliferate (as in a *hybridoma* clone, which is a "biological clone"), or the process of isolating and replicating a piece of DNA by recombinant DNA techniques ("molecular clone"). The use

of the word as a verb is acceptable for the former meaning, but not necessarily the latter meaning.

CIS - As used in molecular biology, an interaction between two sites which are located within the same molecule. However, a *cis*-acting protein can either be one which acts only on the molecule of DNA from which it was expressed, *or* a protein which acts on itself (e.g., self-proteolysis).

CISTRON - A nucleic acid segment corresponding to a polypeptide chain, including the relevant *translational start (initiation)* and *stop (termination) codons*.

CODON - A nucleotide triplet (sequence of three nucleotides) which specifies a specific *amino acid*, or a *translational* start or stop.

CODON BIAS - The tendency for an organism or virus to use certain codons more than others to encode a particular amino acid. An important

determinant of codon bias is the guanosine-cytosine (GC) content of the genome. An organism that has a relatively low G+C content of 30% will be less likely to have a G or C at the third position of a codon (wobble position) than a A or T to specify an amino acid that can be represented by more than one codon.

COMPETENT - Bacterial cells which are capable of accepting foreign extra-chromosomal DNA. There are a variety of processes by which cells may be made competent.

COMPLEMENTARY - See BASE PAIR.

CONFORMATIONAL EPITOPE - An epitope which is dependent upon folding of a protein; amino acid residues present in the antibody binding site are often located at sites in the primary sequence of the protein which are at some distance from each other. The vast majority of B-cell (antibody binding) epitopes are conformational.

CONSENSUS SEQUENCE - A linear series of nucleotides, commonly with gaps and some degeneracy, that define common features of homologous sequences or recognition sites for proteins that act on or bind to nucleic acids.

CONSERVATIVE SUBSTITUTION - A nucleotide mutation which alters the amino acid sequence of the protein, but which causes the substitution of one amino acid with another which has a side chain with similar charge/polarity characteristics (see AMINO ACID). The size of the side chain may also be an important consideration. Conservative mutations are generally considered unlikely to profoundly alter the structure or function of a protein, but there are many exceptions (see NONCONSERVATIVE SUBSTITUTION).

CONSERVED - Similar in structure or function.

CONTIG - A series of two or more individual DNA sequence determinations that overlap. In a sequencing project the contigs get larger and larger until the gaps between the contigs are filled in.

COSMID - A genetically-engineered *plasmid* containing bacteriophage lambda packaging signals and potentially very large pieces of inserted foreign DNA (up to 50 kb) which can be replicated in bacterial cells. Cosmid cloning allows for isolation of DNA fragments which are larger than those which can be cloned in conventional plasmids.

DATABASE SEARCH - Once an open *reading frame* or a partial amino acid sequence has been determined, the investigator compares the sequence with others in the databases using a computer and a search algorithm. This is usually done in a protein database such as PIR or Swiss-Prot. Nucleic acid sequences are in GenBank and EMBL databases. The

search algorithms most commonly used are BLAST and FASTA.

DEGENERACY - Refers to the fact that multiple different *codons* in *mRNA* can specify the same *amino acid* in an encoded protein.

DENATURATION - With respect to nucleic acids, refers to the conversion from double-stranded to the single-stranded state, often achieved by heating or alkaline conditions. This is also called "melting" DNA. With respect to proteins, refers to the disruption of *tertiary* and *secondary* structure, often achieved by heat, detergents, chaotropes, and sulfhydryl-reducing agents.

DENATURING GEL - An *agarose* or *acrylamide gel* run under conditions which destroy *secondary* or *tertiary* protein or RNA structure. For protein, this usually means the inclusion of 2-ME (which reduces disulfide bonds between cysteine residues) and SDS and/or urea in an acrylamide gel. For RNA, this usually

means the inclusion of formaldehyde or glyoxal to destroy higher ordered RNA structures. In DNA sequencing gels, urea is included to denature dsDNA to ssDNA strands. In denaturing gels, macromolecules tend to be separated on the basis of size and (to some extent) charge, while shape and oligomerization of molecules are not important. Contrast with NATIVE GEL.

DEOXYRIBONUCLEASE (DNase) - An enzyme which specifically catalyzes the hydrolysis of DNA.

DEOXYRIBONUCLEOTIDE - *nucleotides* which are the building blocks of DNA and which lack the 2' hydroxyl moiety present in the ribonucleotides of RNA.

DIDEOXYRIBONUCLEOTIDE - A *nucleotide* which lacks both 3' and 2' hydroxyl groups. Such dideoxynucleotides can be added to a growing nucleic acid chain, but do not then present a 3' -OH group which can support further propagation of the nucleic acid chain.

Thus such compounds are also called "chain terminators", and are useful in DNA and RNA sequencing reactions (see DEOXYRIBO-NUCLEOTIDE).

DIDEOXY SEQUENCING - Enzymatic determination of DNA or RNA sequence by the method of Sanger and colleagues, based on the incorporation of chain terminating *dideoxynucleotides* in a growing nucleic acid strand copied by *DNA polymerase* or *reverse transcriptase* from a DNA or RNA *template*. Separate reactions include dideoxynucleotides containing A,C, G, or T bases. The reaction products represent a collection of new. labeled DNA strands of varying lengths, all terminating with a dideoxynucleotide at the 3' end (at the site of a complementary base in the template nucleic acid), and are separated in a polyacrylamide/urea gel to generate a sequence "ladder". This method is more commonly used than "Maxam-Gilbert" (chemical) sequencing.

DIRECT REPEATS - Identical or related sequences present in two or more copies in the same orientation in the same molecule of DNA; they are not necessarily adjacent.

DNA LIGASE - An enzyme (usually from the T4 bacteriophage) which catalyzes formation of a phosphodiester bond between two adjacent bases from double-stranded DNA fragments. RNA ligases also exist, but are rarely used in molecular biology.

DNA POLYMERASE - A polymerase which synthesizes DNA (see POLYMERASE).

DNase - see DEOXYRIBONUCLEASE.

DOT BLOT - DNA or RNA is simply spotted onto nitrocellulose or nylon membranes, denatured and hybridized with a *probe*. Unlike *Southern* or *northern blots*, there is no separation of the target DNA or RNA by electrophoresis (size), and thus potentially much less specificity.

DOWNSTREAM - Identifies sequences proceeding farther in the direction of *expression*; for example, the coding region is downstream from the *initiation codon*, toward the 3' end of an mRNA molecule. Sometimes used to refer to a position within a protein sequence, in which case downstream is toward the *carboxyl* end which is synthesized after the *amino* end during translation.

ds - "double-stranded"

DUPLEX - A nucleic acid molecule in which two strands are *base paired* with each other.

ELECTROPORATION - A method for introducing foreign nucleic acid into bacterial or eukaryotic cells that uses a brief, high voltage DC charge which renders the cells permeable to the nucleic acid. Also useful for introducing synthetic peptides into eucaryotic cells.

END LABELING - The technique of adding a radioactively labeled group to one end (5' or 3' end) of a DNA strand.

ENDONUCLEASE - Cleaves bonds within a nucleic acid chain; they may b especific for RNA or for single-stranded or double-stranded DNA. A restriction enzyme is a type of endonuclease.

ENHANCER - A eukaryotic *transcriptional* control element which is a DNA sequence which acts at some distance to enhance the activity of a specific *promoter* sequence. Unlike promoter sequences, the position and orientation of the enhancer sequence is generally not important to its activity.

ETHIDIUM BROMIDE - Intercalates within the structure of nucleic acids in such a way that they fluoresce under UV light. Ethidium bromide staining is commonly used to visualize RNA or DNA in agarose gels placed on UV light boxes. Proper precautions are required, because the ethidium bromide is highly

mutagenic and the UV light damaging to the eyes. Ethidium bromide is also included in cesium chloride gradients during ultracentrifugation, to separate *supercoiled* circular DNA from linear and *relaxed* circular DNA.

EVOLUTIONARY CLOCK - Defined by the rate at which mutations accumulate within a given gene.

EXON The portion of a gene that is actually translated into protein (see INTRON, SPLICING).

EXONUCLEASE - An enzyme which hydroylzes DNA beginning at one end of a strand, releasing nucleotides one at a time (thus, there are 3' or 5' exonucleases)

EXPRESSION - Usually used to refer to the entire process of producing a protein from a gene, which includes *transcription*, *translation, post-translational modification* and possibly transport reactions.

EXPRESSION VECTOR - A plasmid or phage designed for production of a polypeptide from inserted foreign DNA under specific controls. Often an i*nducer* is used. The vector always provides a promoter and often the *transcriptional start site*, *ribosomal binding sequence*, and *initiation codon*. In some cases the product is a *fusion protein*.

FOOTPRINTING - A technique for identifying the site on a DNA (or RNA) molecule which is bound by some protein by virtue of the protection afforded *phosphodiester bonds* in this region against attack by *nuclease* or nucleolytic compounds.

FRAMESHIFT MUTATION - A mutation (deletion or insertion, never a simple substitution) of one or more *nucleotides* but never a multiple of 3 nucleotides, which shortens or lengthens a trinucleotide sequence representing a *codon*; the result is a shift from one *reading frame* to another reading frame. The amino acid sequence of the protein downstream of the

mutation is completely altered, and may even be much shorter or longer due to a change in the location of the first *termination (stop) codon*: Asn Tyr Thr Asn Leu Gly His Wild-type polypeptide AAU UAC ACA AAU UUA GGG CAU mRNA Asn Thr Gln Ile STOP Mutant polypeptide | Deletion of A from mRNA creates frame-shift mutant

FUSION PROTEIN - A product of recombinant DNA in which the foreign gene product is juxtaposed ("fused") to either the *carboxyl-terminal* or *amino-terminal* portion of a polypeptide encoded by the vector itself. Use of fusion proteins often facilitates expression of otherwise lethal products and the purification of recombinant proteins.

GEL SHIFT - A method by which the interaction of a nucleic acid (DNA or RNA) with a protein is detected. The mobility of the nucleic acid is monitored in an agarose gel in the presence and absence of the protein: if the protein binds to the nucleic acid, the complex

migrates more slowly in the gel (hence "gel shift"). A "supershift" allows determination of the specific protein, by virtue of a second shift in mobility that accompanies binding of a specific antibody to the nucleic acid-protein complex.

GENE - Generally speaking, the *genomic* nucleotide sequence that codes for a particular polypeptide chain, including relevant *transcriptional* control sequences and *introns* (if a eukaryote). However, the term is often loosely used to refer to only the relevant coding sequence.

GENE CONVERSION - The alteration of all or part of a gene by a homologous donor DNA that is itself not altered in the process.

GENOME - The complete set of genetic information defining a particular animal, plant, organism or virus.

GENOMIC LIBRARY - A DNA library which contains DNA fragments hopefully representing each region of the genome of an organism, virus, etc, cloned into individual vector molecules for subsequent selection and amplification. The DNA fragments are usually very small in size compared with the genome. Such libraries are *cDNA libraries* when prepared from RNA viruses.

GENOTYPE - The genetic constitution of an organism; determined by its nucleic acid sequence. As applied to viruses, the term implies a group of evolutionarily related viruses possessing a defined degree of nucleotide sequence relatedness.

GLYCOPROTEIN - A *glycosylated* protein.

GLYCOSYLATION - The covalent addition of sugar moities to N or O atoms present in the side chains of certain amino acids of certain proteins, generally occuring within the Golgi apparatus during secretion of a protein.

HAIRPIN - A helical (duplex) region formed by base pairing between adjacent (inverted) complementary sequences within a single strand of RNA or DNA.

HETERODUPLEX DNA - Generated by base pairing between complementary single strands derived from different parental *duplex* molecules; heteroduplex DNA molecules occur during genetic *recombination* in vivo and during *hydridization* of different but related DNA strands in vitro. Since the sequences of the two strands in a heteroduplex differ, the molecule is not perfectly base-paired; the *melting* temperature of a heteroduplex DNA is dependent upon the number of mismatched base pairs..

HOMOLOGOUS RECOMBINATION - The exchange of sequence between two related but different DNA (or RNA) molecules, with the result that a new "chimeric" molecule is created. Several mechanisms may result in

recombination, but an essential requirement is the existence of a region of *homology* in the recombination partners. In DNA recombination, breakage of single strands of DNA in the two recombination partners is followed by joining of strands present in opposing molecules, and may involve specific enzymes. Recombination of RNA molecules may occur by other mechanisms.

HOMOLOGY - Indicates similarity between two different *nucleotide* or *amino acid* sequences, often with potential evolutionary significance. It is probably better to use more quantitative and descriptive terms such as nucleotide "identity" or, in the case of proteins, amino acid "identity" or "relatedness" (the latter refers to the presence of amino acids residues with similar polarity/charge characteristics at the same position within a protein).

HYBRIDIZATION - The process of *base pairing* leading to formation of *duplex* RNA or DNA or RNA-DNA molecules.

HYBRIDOMA - A *clone* of plasmacytoma cells which secrete a monoclonal antibody; usually produced by fusion of peripheral or splenic plasma cells taken from an immunized mouse with an immortalized murine plasmacytoma cell line (fusion partner), followed by *cloning* and *selection* of appropriate antibody-producing cells.

IMMUNOBLOT - See WESTERN BLOT.

IMMUNOPRECIPITATION - A process whereby a particular protein of interest is isolated by the addition of a specific antibody, followed by centrifugation to pellet the resulting immune complexes. Often, staphylococcal proteins A or G, bound to sepharose or some other type of macroscopic particle, is added to the reaction mix to increase the size and ease collection of the complexes. Usually, the

precipitated protein is subsequently examined by *SDS-PAGE*.

INDUCER - A small molecule, such as IPTG, that triggers gene transcription by binding to a regulator protein, such as LacZ.

INITIATION CODON - The *codon* at which translation of a polypeptide chain is initiated. This is usually the first AUG triplet in the *mRNA* molecule from the 5' end, where the ribosome binds to the cap and begins to scan in a 3' direction. However, the surrounding sequence context is important and may lead to the first AUG being bypassed by the scanning ribosome in favor of an alternative, downstream AUG. Also called a "start codon". Occasionally other codons may serve as initiation codons, e.g. UUG.

INSERT - Foreign DNA placed within a vector molecule.

INSERTION SEQUENCE - A small bacterial transposon carrying only the genetic functions involved in

transposition. There are usually *inverted repeats* at the ends of the insertion sequence.

INTRON - Intervening sequences in eukaryotic genes which do not encode protein but which are *transcribed* into RNA. Removed from *pre-mRNA* during nuclear *splicing* reactions.

INVERTED REPEATS - Two copies of the same or related sequence of DNA repeated in opposite orientation on the same molecule (contrast with DIRECT REPEATS). Adjacent inverted repeats constitute a *palindrome*.

IN VITRO TRANSLATION - See RETICULOCYTE LYSATE.

kb - 'kilobase"

KILOBASE - Unit of 1000 nucleotide bases, either RNA or DNA.

KINASE - See PHOSPHORYLATION.

KLENOW FRAGMENT - The large fragment of E. coli DNA polymerase I which lacks 5' -> 3' exonuclease activity. Very useful for sequencing reactions, which proceed in a 5' -> 3' fashion (addition of nucleotides to templated free 3' ends of primers).

KNOCK-OUT - The excision or inactivation of a gene within an intact organism or even animal (e.g., "knock-out mice"), usually carried out by a method involving *homologous recombination*.

LIBRARY - A set of cloned fragments together representing with some degree of redundancy the entire genetic complement of an organism (see cDNA LIBRARY, GENOMIC LIBRARY).

LIGASE - See DNA LIGASE.

LIGATION - See DNA LIGASE.

LINEAR EPITOPE - An epitope formed by a series of amino acids which are adjacent to each other within the primary structure of the protein. Such epitopes can be successfully modelled by synthetic peptides, but comprise only a small proportion of all epitopes. The minimal epitope size is about 5 amino acid residues. Also called a sequential epitope.

LINKAGE - The tendency of genes to be inherited together as a result of their relatively close proximity on the same chromosome, or location on the same plasmid.

LINKER - A short *oligodeoxyribonucleotide*, usually representing a specific *restriction endonuclease recognition sequence*, which may be *ligated* onto the termini of a DNA molecule to facilitate cloning. Following the ligation reaction, the product is digested with

the endonuclease, generating a DNA fragment with the desired *sticky* or *blunt ends*.

MELTING - The dissociation of a duplex nucleic acid molecule into single strands, usually by increasing temperature. See DENATURATION.

MISSENSE MUTATION - A nucleotide mutation which results in a change in the amino acid sequence of the encoded protein (contrast with SILENT MUTATION).

MONOCLONAL ANTIBODY - An antibody with very specific and often unique binding specificity which is secreted by a biologically cloned line of plasmacytoma cells in the absence of other related antibodies with different binding specificities. Differs from *polyclonal antibodies*, which are mixed populations of antibody molecules such as may be present in a serum specimen, within which many

different individual antibodies have different binding specificities.

MOTIF - A recurring pattern of short sequence of DNA, RNA, or protein, that usually serves as a recognition site or active site. The same motif can be found in a variety of types of organisms.

mRNA - A cytoplasmic RNA which serves directly as the source of code for protein synthesis. See TRANSLATION.

MULTICISTRONIC MESSAGE - An *mRNA transcript* with more than one *cistron* and thus encoding more than one *polypeptide*. These generally do not occur in eukaryotic organisms, due to differences in the mechanism of translation initiation.

MULTICOPY PLASMIDS - Present in bacteria at amounts greater than one per chromosome. Vectors for cloning DNA are usually multicopy; there are sometimes advantages in using a single copy plasmid.

MULTIPLE CLONING SITE - An artificially constructed region within a *vector* molecule which contains a number of closely spaced *recognition sequences* for *restriction endonucleases*. This serves as a convenient site into which foreign DNA may be inserted.

N TERMINUS - See AMINO TERMINUS.

NATIVE GEL - An electrophoresis gel run under conditions which do not denature proteins (i.e., in the absence of SDS, urea, 2-mercaptoethanol, etc.).

NESTED PCR - A very sensitive method for amplfication of DNA, which takes part of the product of a single *PCR* reaction (after 30-35 cycles), and subjects it to a new round of PCR using a

different set of *PCR primers* which are nested within the region flanked by the original primer pair (see **POLYMERASE CHAIN REACTION**).

NICK - In *duplex* DNA, this refers to the absence of a *phosphodiester bond* between two adjacent *nucleotides* on one strand.

NICK TRANSLATION - A method for introducing labeled *nucleotides* into a double-stranded DNA molecule which involves making small *nicks* in one strand with DNase, and then repairing with DNA polymerase I.

NONCONSERVATIVE SUBSTITUTION - A mutation which results in the substitution of one *amino acid* within a polypeptide chain with an amino acid belonging to a different polarity/charge group (see AMINO ACIDS, CONSERVATIVE MUTATION)

NONSENCE CODON - See STOP CODON.

NONSENSE MUTATION - A change in the sequence of a nucleic acid that causes a *nonsense* (*stop* or *termination*) *codon* to replace a codon representing an amino acid.

NONTRANSLATED RNA (NTR) - The segments located at the 5' and 3' ends of a mRNA molecule which do not encode any part of the polyprotein; may contain important translational control elements.

NORTHERN BLOT - RNA molecules are separated by electrophoresis (usually in an agarose gel) on the basis of size, then transferred to a solid-phase support (nitrocellulose paper or suitable other membrane) and detected by *hybridization* with a labeled probe (see **SOUTHERN BLOT, WESTERN BLOT).**

NUCLEOSIDE - The composite sugar and *purine or pyrimidine base* which are present in *nucleotides* which are the basic building blocks of DNA and RNA. Compare with NUCLEOTIDE: Nucleoside = Base + Sugar

NUCLEOTIDE - The composite phosphate, sugar, and *purine* or *pyrimidine base* which are the basic building blocks of the nucleic acids DNA and RNA. The five nucleotides are adenylic acid, guanylic acid (contain *purine bases*), and cytidylic acid, thymidylic acid, and uridylic acid (contain *pyrimidine bases*). Nucleotide = Base + Sugar + Phosphate (1, 2, or 3)

OLIGODEOXYRIBONUCLEOTIDE - A short, single-stranded DNA molecule, generally l5-50 *nucleotides* in length, which may be used as a *primer* or a *hybridization probe*. Oligodeoxyribonucleotides are synthesized chemically under automated conditions.

OLIGONUCLEOTIDE - See OLIGODEOXYRIBONUCLEOTIDE.

ONCOGENE - One of a number of genes believed to be associated with the malignant transformation of cells; originally identified in certain oncogenic retroviruses (*v-onc*) but also present in cells (*c-onc*). See **PROTO-ONCOGENE**.

OPEN READING FRAME - A region within a reading frame of an mRNA molecule that potentially encodes a *polypeptide*; and which does not contain a *translational stop codon* (see READING FRAME).

OPERATOR - The site on DNA at which a *repressor* protein binds to prevent *transcription* from initiating at the adjacent *promoter*.

OPERON - A complete unit of bacterial gene expression and regulation, including the structural gene or genes, regulator gene(s), and control elements in DNA recognized by regulator gene products(s).

ORIGIN - A site within a DNA sequence of a chromosome, plasmid, or non-integrated virus at which replication of the DNA is initiated.

OVERHANG - A terminus of a *duplex* DNA molecule which has one or more unpaired nucleotides in one of the two strands (hence either a 3' or

5' overhang). Cleavage of DNA with many restriction endonucleases leaves such overhangs (see STICKY END).

PACKAGE - In recombinant DNA procedures, refers to the step of incorporation of *cosmid* or other lambda *vector* DNA with an *insert* into a *phage* head for transduction of DNA into host.

PALINDROMIC SEQUENCE - A nucleotide sequence which is the same when read in either direction, usually consisting of adjacent inverted repeats. *Restriction endonuclease recognition sites* are palindromes: 5'-->3' GAATTC *Eco*RI recognition site CTTAAG 3'<--5'

PCR - See POLYMERASE CHAIN REACTION

PEPTIDE - A chain formed by two or more *amino acids* linked through *peptide bonds*: dipeptide = two *amino acids*, oligopeptide = small number of *amino acids*, etc.

PEPTIDE - A molecule formed by peptide bonds covalently linking two or more *amino acids*. Short peptides (generally less than 60 amino acid residues, and usually only half that length) can be chemically synthesized by one of several different methods; larger peptides (more correctly, *polypeptides*) are usually *expressed* from recombinant DNA.

PEPTIDE BOND - A covalent bond between two *amino acids*, in which the carboxyl group of one amino acid (X1--COOH) and the amino group of an adjacent amino acid (NH2--X2) react to form X1-CO-NH-X2 plus H2O.

PHAGE - See BACTERIOPHAGE.

PHENOTYPE - The appearance of other characteristics of an organism resulting from the interaction of its genetic constitution with the environment.

PHOSPHATASE, ALKALINE - An enzyme which catalyzes the hydrolysis of phosphomonoesters of the 5' nucleotides. Used to dephosphorylate (remove phosphate groups from) the 5' ends of DNA or RNA molecules, to facilitate 5' end-labeling with 32P added back by T4 polynucleotide kinase; or to dephosphorylate the 5' ends of DNA molecules to prevent unwanted *ligation* reactions during cloning.

PHOSPHODIESTER BOND - The covalent bond between the 3' hydroxyl in the sugar ring of one nucleotide and the 5' phosphate group of the sugar ring of the adjacent nucleotide residue within a nucleic acid:

5'-Ribose- 3' - O - P(O)2 - O - 5' -Ribose - 3' - etc.

PHOSPHORYLATION - The addition of a phosphate monoester to a macromolecule, catalyzed by a specific kinase enzyme. With respect to proteins, certain *amino acid* side chains

(serine, threonine, tyrosine) are subject to phosphorylation catalyzed by protein kinases; altering the phosphorylation status of a protein may have dramatic effects on its biologic properties, and is a common cellular control mechanism. With respect to DNA, 5' ends must be phosphorylated for *ligation*.

PLASMID - An extrachromosomal,usually circular, double-stranded DNA which is capable of *replication* within a cell, and which usually contains and expresses genes encoding resistance to antibiotics. By strict definition, a plasmid is not essential to the life of the cell.

POINT MUTATION - A single nucleotide substitution within a gene; there may be several point mutations within a single gene. Point mutations do not lead to a shift in reading frames, thus at most cause only a single amino acid substitution (see FRAMESHIFT MUTATION).

POLY-A TRACK - A lengthy adenylic acid polymer (RNA) which is covalently linked to the 3' end of newly synthesized *mRNA* molecules in the nucleus. Function not POLYMERASE CHAIN REACTION (PCR) - A DNA amplification reaction involving multiple (30 or more) cycles of *primer annealing, extension*, and denaturation, usually using a heat-stable DNA *polymerase* such as *Taq polymerase*. Paired primers are used, which are complementary to opposing strands of the DNA and which flank the area to be amplified. Under optimal conditions, single DNA sequence can be amplified a million-fold.

POLYMORPHISM - Variation within a DNA or RNA sequence.

POLYNUCLEOTIDE KINASE - Enzyme which catalyzes the transfer of the terminal phosphate of ATP to 5' hydroxyl termini of polynucleotides, either DNAor RNA. Usually derived from T4 bacteriophage.

POLYPEPTIDE - See PEPTIDE.

pre-mRNA - An RNA molecule which is *transcribed* from chromosomal DNA in the nucleus of eukaryotic cells, and subsequently processed through *splicing* reactions to generate the mRNA which directs protein synthesis in the cytoplasm.

PRIMARY STRUCTURE -Refers to the sequence of *amino acid residues* or *nucleotides* within protein or nucleic acid molecules, respectively (also see SECONDARY and TERTIARY STRUCTURE).

PRIMER - An *oligonucleotide* which is *complementary* to a specific region within a DNA or RNA molecule, and which is used to prime (initiate) synthesis of a new strand of complementary DNA at that specific site, in a reaction or series of reactions catalyzed by a *DNA polymerase*. The newly synthesized DNA strand will contain the primer at its 5' end. Typically, primers are chemically

synthesized oligonucleotides 15-50 nucleotides in length, selected on the basis of a known sequence. However, "random primers" (shorter oligonucleotides, about 6 nucleotides in length, and comprising all possible sequences) may be used to prime DNA synthesis from DNA or RNA of unknown sequence.completely known, but probably serves to enhance stability of the RNA. Is frequently used to select mRNA for cloning purposes by *annealing* to a column containing a matrix bound to poly-uridylic acid.

POLYACRYLAMIDE GEL (PAGE) - Used to separate proteins and smaller DNA fragments and *oligonucleotides* by electrophoresis. When run under conditions which denature proteins (i.e., in the presence of 2-mercaptoethanol, SDS, and possibly urea), molecules are separated primarily on the basis of size.

POLYCLONAL ANTIBODY - See MONOCLONAL ANTIBODY.

POLYMERASE - An enzyme which catalyzes the addition of a *nucleotide* to a nucleic acid molecule. There are a wide variety of RNA and DNA polymerases which have a wide range of specific activities and which operate optimally under different conditions. In general, all polymerases require *templates* upon which to build a new strand of DNA or RNA; however, DNA polymerases also require a *primer* to initiate the new strand, while RNA polymerases start synthesis at a specific *promoter* sequence.

POST-TRANSLATIONAL MODIFICATION - Modifications made to a polypeptide molecule after its initial synthesis, this includes proteolytic cleavages, *phosphorylation, glycosylation*, carboxylation, addition of fatty acid moieties, etc.

PRIMER EXTENSION - A reaction in which DNA is *reverse transcribed* from an RNA *template* to which a specific oligonucleotide *primer* has been *annealed*. The new cDNA product is an

extension of the primer, which is synthesized at the 3' end of the primer in a direction extending toward the 5' end of the RNA. This reaction is useful for exploring the extreme 5' end of RNA molecules.

PROBE - Usually refers to a DNA or RNA molecule which has been labeled with 32P or with *biotin*, to facilitate its detection after it has specifically *hybridized* with a target DNA or RNA sequence. However, the term may also refer to antibody probes used in *western* blots.

PROCESSING - With respect to proteins, generally used to refer to proteolytic *post-translational modifications* of a polypeptide. In the case of RNA, processing may involve the addition of a 5' *cap* and 3' *poly-A* tracks as well as *splicing* reactions in the nucleus.

PROCESSIVITY - The extent to which an RNA or DNA *polymerase* adhers to a *template* before dissociating, determines the average length

(in kilobases) of the newly synthesized nucleic acid strands. Also applies to the action of *exonucleases* in digesting from the ends to the middle of a nucleic acid.

PROMOTER - A specific sequence within a double-stranded DNA molecule that is recognized by an RNA *polymerase*, which binds to it and uses it to begin transcribing the DNA *template* into a new RNA. The location and orientation of the *promoter* within a DNA molecule determines the start site of the new RNA. Other proteins (e.g. transcriptional activators such as *sigma factor*) are usually required for an RNA polymerase to recognize a promoter (see TRANSCRIPTION).

PROTO-ONCOGENE - A cellular *oncogene*-like sequence which is thought to play a role in controlling normal cellular growth and differentiation.

PSEUDOGENE - Inactive but stable components of the genome which derived by duplication and

mutation of an ancestral, active gene. Pseudogenes can serve as the donor sequence in *gene conversion* events.

PSEUDOREVERTANT - A mutant virus or organism which has recovered a wildtype phenotype due to a second-site mutation (potentially located in a different region of the genome, or involving a different polypeptide) which has eliminated the effect of the initial mutation.

PSEUDOKNOT - A feature of RNA *tertiary structure*; best visualized as two overlapping *stem-loops* in which the loop of the first stem-loop participates as half of the stem in the second stem-loop.

PURINE BASES - Adenine (A) or Guanine (G) (see NUCLEOTIDE).

PULSED-FIELD GEL ELECTROPHORESIS (PFGE) - Separation of large (>50 kb) pieces of DNA, including complete chromosomes and

genomes, by rapidly alternating the direction of electrophoretic migration in agarose gels.

PYRIMIDINE BASES - Cytosine (C), Thymine (T) or Uracil (U) (see NUCLEOTIDE).

READING FRAME - Refers to a polypeptide sequence potentially encoded by a single-stranded *mRNA*. Because *codons* are nucleotide triplets, each mRNA has 3 reading frames (each nucleotide can participate in 3 codons, at the 1st, 2nd, and 3rd base position). Duplex DNA strands have 6 reading frames, 3 in each strand (see OPEN READING FRAME):

AlaSerProLeuVal . . 1st reading frame
ProAlaProTERTrp . . 2nd reading frame: TER = Stop
GlnProProSerGly . . 3rd reading frame

GCCAGCCCCCUAGTGGG... Nucleotide sequence of mRNA

RECOGNITION SEQUENCE - A specific *palindromic sequence* within a double-stranded DNA molecule which is recognized by a *restriction endonuclease*, and at which the restriction endonuclease specifically cleaves the DNA molecule.

RECOMBINATION - See HOMOLOGOUS RECOMBINATION.

RECOMBINATION-REPAIR - A mode of filling a gap in one strand of duplex DNA by retrieving a homologous single strand from another duplex. Usually the underlying mechanism behind *homologous recombination* and *gene conversion*.

RELAXED DNA - See SUPERCOIL.

REPLICATION - The copying of a nucleic acid molecule into a new nucleic acid molecule of similar type (i.e., DNA --> DNA, or RNA --> RNA).

REPORTER GENE - The use of a functional enzyme, such as beta-galactosidase, luciferase, or

chloramphenicol acetyltransderase, downstream of a gene, promoter, or translational control element of interest, to more easily identify successful introduction of the gene into a host and to measure transcription and/or translation.

REPRESSION - Inhibition of transcription (or translation) by the binding of a repressor protein to a specific site on DNA (or *mRNA*).

RESIDUE - As applied to proteins, what remains of an *amino acid* after its incorporation into a peptide chain, with subsequent loss of a water molecule (see PEPTIDE BOND).

RESTRICTION ENDONUCLEASE - A bacterial enzyme which recognizes a specific *palindromic sequence (recognition sequence)* within a double-stranded DNA molecule and then catalyzes the cleavage of both strands at that site. Also called a restriction enzyme. Restriction endonucleases may generate

either *blunt* or *sticky ends* at the site of cleavage.

RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP) - Variations in the lengths of fragments of DNA generated by digestion of different DNAs with a specific *restriction endonuclease*, reflecting genetic variation (*polymorphism*) in the DNAs.

RESTRICTION FRAGMENTS - DNA fragments generated by digestion of a DNA preparation with one or more *restriction endonucleases*; usually separated by *agarose gel electrophoresis* and visualized by *ethidium bromide* staining under UV light (or alternatively subjected to *Southern blot analysis*).

RESTRICTION MAP - A linear array of sites on a particular DNA which are cleaved by various selected *restriction endonucleases*.

RESTRICTION SITE - See RECOGNITION SEQUENCE.

RETICULOCYTE LYSATE - A lysate of rabbit reticulocytes, which has been extensively digested with micrococcal nuclease to destroy the reticulocyte *mRNA*s. With the addition of an exogenous, usually synthetic, mRNA, *amino acids* and a source of energy (ATP), the translational machinery of the reticulocyte (*ribosomes*, eukaryotic translation factors, etc.) will permit in vitro *translation* of the added mRNA with production of a new *polypeptide*. This is only one of several available in vitro translation systems.

REVERSE TRANSCRIPTASE - A *DNA polymerase* which copies an RNA molecule into single-stranded cDNA; usually purified from retroviruses.

REVERSE TRANSCRIPTION - Copying of an RNA molecule into a DNA molecule.

REVERTANT - See BACK MUTATION.

RIBONUCLEASE (RNase) - An enzyme which catalyzes the hydrolysis of RNA. There are many different RNases, some of the more important include:

RNase A Cleaves ssRNA 3' of pyrimidines
RNase T1 Cleaves ssRNA at guanasine nucleotides
RNase V1 Cleaves dsRNA (helical regions)
RNase H Degrades the RNA part of RNA:DNA hybrids.

RIBOSOMAL BINDING SEQUENCE (Shine-Dalgarno sequence) - In prokaryotic organisms, part or all of the polypurine sequence AGGAGG located on *mRNA* just upstream of an AUG *initiation codon*; it is complementary to the sequence at the 3' end of 16S rRNA; and involved in binding of the ribosome to *mRNA*. The *internal ribosomal entry site* found in some viruses may be an analogous eukaryotic genetic element.

RIBOSOME - A complex ribonucleoprotein particle (eukaryotic ribosomes contain 4 RNAs and at least 82 proteins) which is the "machine" which translates *mRNA* into protein molecules. In eukaryotic cells, ribosomes are often in close proximity to the endoplasmic reticulum.

RIBOZYME - A catalytically active RNA. A good example is the hepatitis delta virus RNA which is capable of self-cleavage and self-ligation in the absence of protein enzymes.

RNA POLYMERASE - A polymerase which synthesizes RNA (see POLYMERASE).

RNA SPLICING - A complex and incompletly understood series of reactions occuring in the nucleus of eukaryotic cells in which *pre-mRNA transcribed* from chromosomal DNA is processed such that noncoding regions of the pre-mRNA (*introns*) are excised, and coding regions (*exons*) are covalently linked to produce an *mRNA* molecule ready for

transport to the cytoplasm. Because of splicing, eukaryotic DNA representing a gene encoding any given protein is usually much larger than the mRNA from which the protein is actually *translated*.

RNase - see RIBONUCLEASE

rRNA - Ribosomal RNA (four sizes in humans: 5S, 5.8S, 18S, and 28S); RNA component of the *ribosome*, which may play catalytic roles in *translation*.

RT/PCR REACTION - A series of reactions which result in RNA being copied into DNA and then amplified. A single *primer* is used to make single-stranded *cDNA* copies from an RNA *template* under direction of *reverse transcriptase*. A second primer *complementary* to this "first strand" cDNA is added to the reaction mix along with *Taq polymerase*, resulting in synthesis of double-stranded DNA. The reaction mix is then cycled (denaturation, *annealing* of primers,

extension) to amplify the DNA by conventional PCR.

RUNOFF TRANSCRIPT - RNA which has been synthesized from plasmid DNA (usually by a bacteriophage *RNA polymerase* such as *T7* or *SP6*) and which terminates at a specific 3' site because of prior cleavage of the plasmid DNA with a *restriction endonuclease*.

S1 NUCLEASE - An enzyme which digests single-stranded DNA or RNA

SDS-PAGE - Denaturing protein gel electrophoresis (see POLYACRYLAMIDE GEL ELECTROPHORESIS).

SECONDARY STRUCTURE - (also see PRIMARY and TERTIARY STRUCTURE) Local structure within a protein which is conferred by the nature of the side chains of adjacent *amino acids* (e.g., alpha helix, beta sheet, random coil); local structure within an RNA molecule which is conferred by *base pairing* of

nucleotides which are relatively closely positioned within the sequence (e.g., hairpins, stem-loop structures).

SELECTION - The use of particular conditions, such as the presence of ampicillin, to allow survival only of cells with a particular *phenotype*, such as production of beta-lactamase.

SEQUENCE POLYMORPHISM - See POLYMORPHISM.

SEQUENTIAL EPITOPE - See LINEAR EPITOPE.

SHOTGUN CLONING or SEQUENCING - Cloning of an entire genome or large piece of DNA in the form of randomly generated small fragments. The individual sequences obtained from the clones will be used to construct *contigs*.

SHUTTLE VECTOR - A small *plasmid* capable of *transfection* into both prokaryotic and eukaryotic cells.

SIDE CHAIN - See AMINO ACID.

SIGMA FACTOR - Certain small ancillary proteins in bacteria that increase the binding affinity of RNA polymerase to a promoter. Different sigma factors recognize different promoter sequences.

SIGNAL PEPTIDASE - An enzyme present within the lumen of the endoplasmic reticulum which proteolytically cleaves a secreted protein at the site of a *signal sequence.*

SIGNAL SEQUENCE - A hydrophobic amino acid sequence which directs a growing peptide chain to be secreted into the endoplasmic reticulum.

SILENT MUTATION - A nucleotide substitution (never a single deletion or insertion) which does not alter the amino acid sequence of an encoded protein due to the *degeneracy* of the genetic code. Such mutations usually involve the third base (*wobble* position) of *codons*.

SITE-DIRECTED MUTAGENESIS - The introduction of a mutation, usually a *point mutation* or an insertion, into a particular location in a cloned DNA fragment. This mutated fragment may be used to "*knock out*" a gene in the organism of interest by *homologous recombination*.

SITE-SPECIFIC RECOMBINATION - Occurs between two specific but not necessarily homologous sequences. Usually catalyzed by enzymes not involved in general or *homologous recombination*.

SOUTHERN BLOT - DNA is separated by electrophoresis (usually in *agarose gels*), then transferred to nitrocellulose paper or other suitable solid-phase matrix (e.g., nylon membrane), and denatured into single strands so that it can be *hybridized* with a specific *probe*. The Southern blot was developed by E.M. Southern, a molecular biologist in Edinburgh. *Northern* and *western* blots were given contrasting names to reflect

the different target substances (RNA and proteins, respectively) that are subjected in these procedures to electrophoresis, blotting and subsequent detection with specific probes.

SOUTHWESTERN BLOT - The binding of protein to a nucleic acid on a matrix similar to what is done for western, northern, and southern blots. This technique is used to identify DNA binding proteins and the recognition sites for these proteins.

SP6 RNA POLYMERASE - A bacteriophage *RNA polymerase* which is commonly used to transcribe *plasmid* DNA into RNA. The plasmid must contain an SP6 *promoter* upstream of the relevant sequence.

SPLICING - see RNA SPLICING.

ss - Single stranded.

START CODON - See INITIATION CODON.

STEM-LOOP - A feature of RNA *secondary structure*, in which two complementary, inverted sequences which are separated by a short-intervening sequence within a single strand of RNA base pair to form a '"stem" with a "loop" at one end. Similar to a *hairpin*, but these usually have very small loops and longer stems.

STICKY END - The terminus of a DNA molecule which has either a 3' or 5' overhang, and which typically results from a cut by a *restriction endonuclease*. Such termini are capable of specific ligation reactions with other termini which have complementary overhangs. A sticky end can be "blunt ended" either by the removal of an overhang, or a "filling in" reaction which adds additional nucleotides complementary to the overhang (see BLUNT END).

5'-->3'
NNNG AATTCNNN *Eco*RI cut, 5' overhang
NNNCTTAA GNNN

3'<--5'

5'-->3'
XXXAGCGC TNNN *Hae*II cut, 3' overhang
XXXT CGCGANNN
3'<--5'

STOP CODON - A *codon* (UAA, UAG, UGA) which terminates *translation*.

STREPTAVIDIN - A bacterial analog of egg white *avidin*.

STRINGENCY - The conditions employed for *hybridization* which determine the specificity of the *annealing* reaction between two single-stranded nucleic acid molecules. Increasingly stringent conditions may be reached by raising temperature or lowering ionic strength, resulting in greater specificity (but lower sensitivity) of the hybridization reaction.

SUPERCOIL - Double-stranded circular DNA which is twisted about itself. Commonly observed with *plasmids* and circular viral DNA genomes

(such as that of hepatitis B virus). A nick in one strand of the plasmid may remove the twist, resulting in a *relaxed*, circular DNA molecule. A complete break in the DNA puts the plasmid in a linear form. Supercoils, relaxed circular DNA, and linear DNA all have different migration properties in agarose gels, even though they contain the same number of base pairs.

T7 RNA POLYMERASE - A bacteriophage *RNA polymerase* which is commonly used to transcribe plasmid DNA into RNA. The plasmid must contain a T7 *promoter* upstream of the relevant sequence.

***Taq* POLYMERASE** - A *DNA polymerase* which is very stable at high temperatures, isolated from the thermophilic bacterium *Thermus aquaticus*. Very useful in *PCR* reactions which must cycle repetitively through high temperatures during the denaturation step.

TEMPLATE - A nucleic acid strand, upon which a *primer* has *annealed* and a nascent RNA stand is being extended.

TERMINATION CODON - See STOP CODON.

TERMINATOR - A sequence *downstream* from the 3' end of an *open reading frame* that serves to halt *transcription* by the RNA polymerase. In bacteria these are commonly sequences that are *palindromic* and thus capable of forming *hairpins*. Sometimes termination requires the action of a protein, such as Rho factor in *E. coli.*

TERTIARY STRUCTURE - (also see PRIMARY and SECONDARY STRUCTURE) Refers to higher ordered structures conferred on proteins or nucleic acids by interactions between *amino acid residues* or *nucleotides* which are not closely positioned within the sequence (primary structure) of the molecule.

Tm - The midpoint of the temperature range over which DNA is melted or denatured by heat; the temperature at which a *duplex* nucleic acid molecule is 50% *melted* into single strands, it is dependent upon the number and proportion of G-C *base pairs* as well as the ionic conditions. Often referred to as a measure of the thermal stability of a nucleic acid *probe*:target sequence hybrid.

TRANS - As used in molecular biology, an interaction that involves two sites which are located on separate molecules.

TRANSCRIPT - A newly made RNA molecule which has been copied from DNA.

TRANSCRIPTION - The copying of a DNA template into a single-stranded RNA molecule. The processes whereby the transcriptional activity of eukaryotic genes are regulated are complex, involve a variety of accessory transcriptional factors which interact with *promoters* and *polymerases*, and constitute

one of the most important areas of biological research today.

TRANSCRIPTION/TRANSLATION REACTION - An in vitro series of reactions, involving the synthesis (*transcription*) of an mRNA from a *plasmid* (usually with *T7* or *SP6 RNA polymerase*), followed by use of the mRNA to program *translation* in a cell-free system such as a rabbit *reticulocyte lysate*. The *polypeptide* product of translation in usually labelled with [35S]-methionine, and examined in an *SDS-PAGE* gel with or without prior *immunoprecipitation*. This series of reactions permits the synthesis of a polypeptide from DNA in vitro.

TRANSCRIPTIONAL START SITE - The nucleotide of a gene or cistron at which *transcription* (RNA synthesis) starts; the most common triplet at which transcription begins in *E. coli* is CAT. *Primer extension* identifies the transcriptional start site.

TRANSFECTION - The process of introducing foreign DNA (or RNA) into a host organism, usually a eukaryotic cell.

TRANSFORMATION - Multiple meanings. With respect to cloning of DNA, refers to the transformation of bacteria (usually to specific antibiotic resistance) due to the uptake of foreign DNA. With respect to eukaryotic cells, usually means conversion to less-restrained or unrestrained growth.

TRANSGENE - A foreign gene which has been introduced into the germ line of an animal species.

TRANSGENIC - An animal (usually a mouse) or plant into which a foreign gene has been introduced in the germ line. An example: transgenic mice expressing the human receptor for poliovirus are susceptible to human polioviruses.

TRANSITION - A nucleotide substitution in which one pyrimidine is replaced by the other pyrimidine, or one purine replaced by the other purine (e.g., A is changed to G, or C is changed to T) (contrast with TRANSVERSION) .

TRANSLATION - The process whereby *mRNA* directs the synthesis of a protein molecule; carried out by the *ribosome* in association with a host of translation initiation, elongation and termination factors. Eukaryotic genes may be regulated at the level of translation, as well as the level of *transcription*.

TRANSLOCATION - The process by which a newly synthesized protein is directed toward a specific cellular compartment (i.e, the nucleus, the endoplasmic reticulum).

TRANSPOSON - A transposable genetic element; certain sequence elements which are capable of moving from one site to another in a DNA molecule without any requirement for

sequence relatedness at the donor and acceptor sites. Many transposons carry antibiotic resistance determinants and have *insertion sequences* at both ends, and thus have two sets of *inverted repeats*.

TRANSPOSITION - The movement of DNA from one location to another location on the same molecule, or a different molecule within a cell.

TRANSVERSION - A nucleotide substitution in which a purine replaces a pyrimidine, or vice versa (e.g., A is changed to T, or T is changed to G) (see TRANSITION)

TRIPLET - A three-nucleotide sequence; a *codon*.

tRNA - Small, tightly folded RNA molecules which act to bring specific amino acids into *translationally* active *ribosomes* in a fashion which is dependent upon the *mRNA* sequence. One end of the tRNA molecule recognizes the nucleotide triplet which is the *codon* of the

mRNA, while the other end (when activated) is covalently linked to the relevant *amino acid*.

UNTRANSLATED RNA - See NONTRANSLATED RNA.

UPSTREAM - Identifies sequences located in a direction opposite to that of expression; for example, the bacterial *promoter* is upstream of the *initiation codon*. In an mRNA molecule, upstream means toward the 5' end of the molecule. Occasionally used to refer to a region of a polypeptide chain which is located toward the amino terminus of the molecule.

VECTOR - A *plasmid, cosmid*, bacteriophage, or virus which carried foreign nucleic acid into a host organism.

WESTERN BLOT - Proteins are separated by *SDS-PAGE*, then electrophoretically transferred to a solid-phase matrix such as nitrocellulose,

then probed with a labelled antibody (or a series of antibodies)

WILDTYPE - The native or predominant genetic constitution before mutations, usually referring to the genetic consitution normally existing in nature.

WOBBLE POSITION - The third base position within a *codon*, which can often (but not always) be altered to another nucleotide without changing the encoded amino acid (see DEGENERACY).