Engineering Applications in Genomics

Aniruddha Datta

Department of Electrical & Computer Engineering
Texas A&M University
College Station, TX 77843-3128

http://gsp.tamu.edu
Outline of Course

• Introduction
• Review of Organic Chemistry
• Energy Considerations in Biochemical Reactions
• Proteins
• DNA
• Transcription and Translation
• Chromosomes and Gene Regulation
• Genetic Variation
• DNA Technology
- Cell Division
- Cell Cycle Control, Cell Death and Cancer
- Expression Microarrays
- Classification
- Clustering
- Genetic Regulatory Networks
- Intervention
- External Intervention Based on Optimal Control Theory
Introduction
Cell

- A cell is the basic unit of life.
- All living creatures are made of cells.
- Typically 5-20 μm in diameter
Components of a typical plant cell

- Mitochondrion
- Nucleus (contains DNA)
- Chloroplast
- Golgi apparatus
- Endoplasmic Reticulum
- Plasma membrane
- Peroxisome
Organelle functions

Site for Energy Generation: sites where photosynthesis takes place

Synthesized components undergo appropriate modifications

Site for synthesis of many cellular components.

Organelles used to compartmentalize cellular reactions that release hydrogen peroxide.

Cytosol: cell stripped of organelles.
Two Broad Categories of Cells

- *Procaryotes* (or Procaryotic Cells): do not contain a nucleus and other organelles, e.g. bacteria.
- *Eucaryotes* (or Eucaryotic cells): contain a nucleus, e.g. human cells, yeast, etc.
Cell composition

• Cytosol: aqueous (watery) gel.

• Plasma membrane and other membranes: made of lipids.

• Lipids have a hydrophilic (water loving) part and a hydrophobic (water hating) part.

• Lipid Bilayer: basis of all cell membranes.
Lipid Bilayer and Vesicular Transport
Cytoskeleton and its Functions

- **Cytoskeleton**: provides structure to the interior of the cell.
- The cytoskeleton is made up of three types of proteins:
  (i) *actin filaments*: generate contractile forces, e.g. in muscle cells
  (ii) *microtubules*: provide (a) tracks along which different cell components can be moved and (b) distribute the DNA properly between the daughter cells at cell division.
  (iii) *intermediate filaments*: provide mechanical strength to the cell and its neighbors
Features Common to all Cells

- Cells vary enormously in size and function.
- But same underlying chemistry.
- The **genetic information in the DNA** is coded using the same blocks in all organisms.
- This information is interpreted using essentially the same machinery to produce proteins.
- Proteins made up of the same 20 blocks.
- This greatly simplifies the study of many aspects of molecular biology: not organism specific.
Some Model Organisms

- Model Organisms are organisms used for molecular biology studies.
- Model organisms must be simple and capable of quickly replicating themselves.
- For procaryotes, it is *E. Coli* bacteria: contain a few genes and reproduce every 20 minutes.
- For flowering plants, it is the plant *Arabidopsis*: produces thousands of offsprings in 8 to 10 weeks.
- For insects, it is the *Drosophila* or fruit fly.
- For mammals, it is rats and mice.
- For unicellular eucaryotes, it is the yeast.
Review of Organic Chemistry
• *Elements*: cannot be broken down or converted into other substances by chemical means.

• *Atom*: smallest particle of an element that still retains its distinctive chemical properties.

• *Molecules*: formed by two or more atoms of the same element or of different elements combining together.

• Subatomic particles: *protons* (mass = 1, charge = +1); *neutrons* (mass = 1, charge = 0); *electrons* (mass=0, charge = -1).

• Protons and neutrons reside in the nucleus

• Electrons revolve around the nucleus in certain orbits.
• **Atomic number**: number of protons in the nucleus of an atom.

• Examples: Hydrogen (H), atomic number = 1; Oxygen (O), atomic number = 8; Carbon (C), atomic number = 6; Nitrogen (N), atomic number = 7; Phosphorous (P), atomic number = 15; Sodium (Na), atomic number = 11; and Calcium (Ca), atomic number = 20.

• **Atomic weight**: total number of protons and neutrons in the nucleus of an atom.

• Example: hydrogen atomic weight = 1 (1 proton, 1 electron, no neutron).

• Atomic weight is also the mass of an atom relative to that of hydrogen atom.

• Atom is electrically neutral: number of protons = number of electrons.
• Number of neutrons in an atom of a particular element can vary.
• Example: carbon atoms usually have 6 protons and 6 neutrons (atomic weight = 12); carbon atoms possessing 6 protons and 8 neutrons (atomic weight = 14) exist.
• Isotopes: atoms of same element with varying numbers of neutrons.
• Usually one isotope of an element is the most stable one and the other isotopes radioactively decay towards it with time.
• The protons and neutrons are held together tightly in the nucleus.
• The electrons revolve around the nucleus in certain discrete orbits called shells.
Schematic diagram of an Atom

• Helium (He, atomic number = 2), Neon (Ne, atomic number = 10), Argon (Ar, atomic number = 18) are all inert gases.
Sodium

Electrovalent Bonds

Chlorine

Sodium ion, Na⁺

Chloride ion, Cl⁻
Electrovalent/Ionic Bonds

- extremely strong bonds held together in place by the forces of electrostatic attraction.

Sodium Chloride

NaCl

**Sodium ion, Na^+**

**Chloride ion, Cl^-**

Electrostatic attraction
Covalent Bonds

- Atoms can also achieve completely filled outermost shells by sharing pairs of electrons.
- Resulting Bonds are called **Covalent Bonds**.
- Examples:
  - (1) Two Hydrogen atoms (atomic no. = 1) can share a pair of electrons to form a hydrogen molecule \( \text{H}_2 \).
  - (2) Carbon (atomic no. = 6) has the electronic distribution 2+4. It can share its four outermost electrons with 4 hydrogen atoms forming methane, \( \text{CH}_4 \).
  - (3) Oxygen (atomic no. = 8) has the electronic distribution 2+6. It can share two of its outermost electrons with two hydrogen atoms forming water (\( \text{H}_2\text{O} \)).
• **Valence**: the number of electrons that an atom of an element must donate/accept/share to form a complete outermost electronic shell.

• Hydrogen has a valence of 1, carbon has a valence of 4, oxygen has a valence of 2, sodium and chlorine each has a valence of 1.

• Each carbon atom can share up to four pairs of electrons with other atoms.
Single and Double Bonds

- A **single bond** is a covalent bond where only one pair of electrons is shared e.g. ethane \((C_2H_6)\).

```
H   H
\(\text{H-C-C-H}\)
 \(\text{H-H}\) (Ethane)
```

- A **double bond** is formed when two pairs of electrons are shared between the participating atoms e.g. ethene \((C_2H_4)\).

```
H   H
\(\text{H-C=C-H}\)
 \(\text{H-H}\) (Ethene)
```

- **Stearic hindrance** due to double bonds.
Polar and Non-Polar molecules

- In a single covalent bond between different elements, the two atoms usually attract the shared electrons to different degrees.
- E.g., oxygen and nitrogen atoms attract electrons quite strongly while hydrogen attracts electrons quite weakly.
- In a hydrogen-oxygen covalent bond, the hydrogen atom acquires a positive charge while the oxygen atom acquires a negative charge.
- Overall molecule has an uneven charge distribution: polar molecules.
• For e.g., Covalent bond between oxygen and hydrogen O-H; bond between nitrogen and hydrogen.
• Bond between carbon and hydrogen C-H is non-polar.
• Polar covalent bonds are extremely important in biology (electrical forces).
• *Hydrogen bonding* between water molecules.
• Polar molecules readily dissolve in water (hydrophilic).
• Non-polar molecules do not dissolve in water (hydophobic).
• **Amphipathic** molecules have both hydrophobic and hydrophilic parts.
• Amphipathic molecules: building blocks for all cell membranes.
Some Chemical Bonds and Groups Commonly Encountered in Biological Molecules

- Compounds containing only Carbon and Hydrogen are called *hydrocarbons*.

- Examples: Methane and Ethane.

![Methane](image1.png) ![Methyl Group](image2.png) ![Ethane](image3.png) ![Ethyl Group](image4.png)
C-O compounds

• *Alcohols* are characterized by the presence of the hydroxyl (OH) group and have the general formula R-OH where R could be any organic group.

• R = CH$_3$, gives *Methyl Alcohol* (*Methanol*)

• R=C$_2$H$_5$ gives *Ethyl Alcohol* (*Ethanol*)
Aldehydes

- *Aldehydes* are characterized by the general formula

\[ R\overset{\text{C}}{\sim}H \overset{\text{O}}{\sim} \]

- When \( R = H \), we get *Formaldehyde* and when \( R = CH_3 \), we get *Acetaldehyde*. 
Ketones

- *Ketones* are characterized by the general formula $R_1 - C = C - R_2$.

*Carbonyl group* is present in both *aldehydes* and *ketones*. 
Carboxylic Acids

- *Carboxylic Acids* are characterized by the general formula

$$\text{R} - \text{C} - \text{OH}$$

*Carboxyl group*

- When $R = H$, we get *Formic Acid* and when $R = CH_3$, we get *Acetic Acid*. 
**Acid**

Substances that release hydrogen ions into solution are called *acids*.

Ex.

\[
\text{HCl} \rightarrow \text{H}^+ + \text{Cl}^-
\]

hydrochloric acid
Carboxylic acids

- Carboxylic acids are called acids because they \textit{partially} dissociate in solution to yield hydrogen ions:

\[
\begin{array}{c}
\text{R-C=O} \quad \text{H}^+ + \text{R-C=O} \\
\text{O} \quad \text{O}
\end{array}
\]

Because of this partial dissociation, organic acids are called \textit{weak acids} as opposed to say hydrochloric acid, which is called a \textit{strong acid}. 
**pH value**

- Hydrogen ion concentration of a solution is usually measured according to its pH value

\[
pH = -\log_{10}[H^+]
\]

\([H^+]\) = hydrogen ion concentration in moles per liter

Water pH=7

for *Acids* pH < 7 higher concentration of hydrogen ions than pure water

for *Bases* pH > 7 lower concentration of hydrogen ions than pure water
**Bases**

- *Bases* are defined as substances that reduce the number of hydrogen ions in aqueous solution.
- For instance, sodium hydroxide (NaOH) dissociates in solution to yield sodium ions and hydroxyl ions:
  \[
  \text{NaOH} \rightarrow \text{Na}^+ + \text{OH}^-
  \]
  hydroxyl ions then react with some of the hydrogen ions in the water, thereby reducing the number of hydrogen ions.

Similarly Ammonia (NH$_3$) is a base
\[
\text{NH}_3 + \text{H}^+ \rightarrow \text{NH}_4^+
\]

Ammonium ion
Acid-Base Reaction

• Important fact from *inorganic* chemistry is that an acid reacts with a base to produce a salt and water.
  
  \[
  \text{Ex. } \quad \text{NaOH} + \text{HCl} \rightarrow \text{NaCl} + \text{H}_2\text{O}
  \]

• In *organic* chemistry, when a carboxylic acid reacts with an alcohol, the reaction produces water and a compound called an *ester*.

```
\begin{align*}
\text{carboxylic acid} & \quad + \quad \text{Alcohol} & \quad \rightarrow \quad \text{ester} \\
R_1\text{C}=\text{O} & \quad + \quad \text{HO-R}_2 & \quad \rightarrow \quad R_1\text{C}-\text{O}-\text{R}_2 + \text{H}_2\text{O}
\end{align*}
```
C-N compounds

- **Amines**: organic compounds characterized by the presence of the amino \((NH_2)\) group

\[
\begin{array}{c}
\text{R} \quad \text{N} \quad \text{H} \\
\text{H}
\end{array}
\]

*weak base* \(R - NH_2 + H^+ \rightleftharpoons R - NH_3^+\)
Amine + Carboxylic acid

- An amine can react with a carboxylic acid to produce an *amide* and water.

\[
\begin{align*}
\text{R}_1 & \quad \text{C} & \quad \text{OH} \quad + \quad \text{H}_2\text{N} & \quad \text{R}_2 \\
\text{carboxylic acid} & \quad \text{amine} & \quad \text{amine} & \quad \text{amide}
\end{align*}
\]
Phosphates

- Phosphates are characterized by the presence of the inorganic phosphate ion, which is derived from phosphoric acid ($H_3PO_4$) by the loss of two hydrogen ions.

$$H_3PO_4 \xrightarrow{\text{loss of 2H}^+} \text{Phosphate ion}$$
Phosphate ion reaction examples

- The inorganic phosphate ion can react with an alcohol to produce a phosphate ester.

Phosphate ester bonds as shown above are important in the formation of nucleic acids such as DNA and RNA.
Phosphate ion reaction examples contd.

The inorganic phosphate ion can react with the carboxyl group of an organic acid to produce a *carboxylic-phosphoric acid anhydride*.

![Chemical structure of carboxylic-phosphoric acid anhydride]

The compound formed above is called an *anhydride*, since it results from the removal of a water molecule.
Phosphate ion reaction examples contd.

- Two inorganic phosphate ions can react together to produce a *phosphoanhydride* bond:

![Phosphoanhydride bond diagram](image)

Phosphoanhydride bonds are high energy bonds and the reversibility of the above reaction makes such bonds suitable for energy storage and transfer.
Building Blocks for Common Organic Molecules

Cells contain four major families of small organic molecules, or modular units, that are combined together to form the large macromolecules:

1. *Sugars*, which are the building blocks for more complex sugars and carbohydrates;
2. *Fatty acids*, which are the building blocks for fats, lipids and all cell membranes;
3. *Amino acids*, which are the building blocks for proteins; and
4. *Nucleotides*, which are the building blocks for nucleic acids such as DNA and RNA.
Sugars

• The simplest sugars (monosaccharides) are compounds with the general formula \((\text{CH}_2\text{O})_n\) where \(n\) is usually 3, 4, 5, 6 or 7.

• Sugars are also called *carbohydrates* since their general formula suggests that they are somehow built up from carbon and water.

• Monosaccharides usually occur as aldehydes or ketones.
5-carbon sugars: *pentoses*

- **Ribose (aldehyde):**
  - H\(\rightarrow\) C\(\equiv\) O
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H

- **Ribulose (ketone):**
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H\(\rightarrow\) C\(\rightarrow\) OH
  - H

6-carbon sugars: *hexoses*

- **Glucose (aldehyde)**
  
  - H\(\text{-}\)C\(\equiv\)O
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)

- **Fructose (ketone)**
  
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)C\(\text{-}\)OH
  - H\(\text{-}\)
• The 5-carbon sugar *ribose* and its derivative *deoxyribose* are important constituents of nucleic acids such as DNA and RNA

• 6-carbon sugar *glucose* serves as an important source of energy.
Ring formation in aqueous solution

In aqueous solution, the aldehyde or ketone group of a sugar molecule tends to react with a hydroxyl group of the same molecule, thereby closing the molecule into a ring.

Since the environment inside a cell is aqueous, it is this ring structure that we will be encountering over and over again in this course.
Deoxyribose

Replacing hydroxyl group at carbon number 2 in ribose with hydrogen produces *deoxyribose*, which is the sugar present in DNA.
Isomers

- In inorganic chemistry, the chemical formula of a compound usually determines a unique structural formula and associated properties.
- In organic chemistry, on the other hand, compounds having the same chemical formula may have totally different structural formulae and properties.
- Compounds which have the same chemical formula but different structural formulae are called isomers of each other.
- Organic chemistry is filled with instances of isomers whose different properties make them especially well suited for particular biological functions.
Isomer examples

• If we interchange the hydrogen and the hydroxyl group at carbon No.4 in glucose we obtain a different sugar called galactose.

• If we interchange the hydrogen and the hydroxyl groups at carbon No.2 in glucose, we obtain the sugar mannose.
Disaccharides

- The monosaccharide building blocks can be combined together to yield more complex sugars.
- Disaccharides are made up of two monosaccharides that are linked together in a *condensation* reaction.
Since there are many hydroxyl groups on each monosaccharide, two monosaccharides can link together in many different ways.

Due to the availability of additional hydroxyl groups on the disaccharide, more monosaccharides can get linked to it, producing chains and branches of various lengths.

Short chains are called **oligosaccharides**, while long chains are called **polysaccharides**.

An example of a polysaccharide is **glycogen**, which is made up entirely of glucose units linked together. Glycogen serves as an energy storage in animals.
• Condensation reactions are not unique to sugars.
• In fact, they are used to form proteins from amino acids and nucleic acids from nucleotides.
• The reverse reaction of condensation is called *hydrolysis* (splitting with water) and plays an important role in our digestion of carbohydrates and other foods.

Some specific examples of disaccharides resulting from linking together two monosaccharide units:

- glucose + glucose = maltose;
- glucose + galactose = lactose (the sugar found in milk);
- glucose + fructose = sucrose.
Fatty Acids

• A fatty acid molecule has two distinct regions:
  (i) a long hydrocarbon chain, which is hydrophobic and
  (ii) a carboxyl group, which behaves as a carboxylic acid, is ionized in solution and is extremely hydrophilic.

  Ex.

  • Palmitic acid (with 16 carbon atoms) is a *saturated* fatty acid.
  • Saturated fatty acids are characterized by the absence of double bonds between the carbon atoms.
Fatty Acid Ex. Contd.

- Oleic acid has 18 carbon atoms and is *unsaturated* since it has a double bond between a pair of carbon atoms.

\[
\text{Oleic acid (C}_{18}\text{)}
\]

- The **double bond creates stearic hindrance**, which is an important factor in determining the final shape of the molecule.
Triglycerides

- Fatty acids serve as concentrated food reserves in cells, as they can be broken down to produce about six times as much usable energy, weight for weight, as glucose.
- They are usually stored in cells as triacylglycerol molecules (triglycerides), which consist of three fatty acids joined to a glycerol backbone.

Structure of triglycerides

![Diagram of Glycerol and Triacylglycerol molecules]
Phospholipids

- Phospholipids are the major constituents of cell membranes.
- In phospholipids, two of the hydroxyl groups in glycerol are linked to fatty acids, while the third hydroxyl group is linked to phosphoric acid.

A typical phospholipid
Formation of a surface film and a micelle

- Since a fatty acid possesses a hydrophilic part and a hydrophobic part, it can form a surface film or small micelles (a globular aggregation of molecules).
Formation of Lipid Bilayer

- Because of similar reasons, phospholipids and glycolipids (lipids made up of two fatty acids for the hydrophobic part and one or more sugar residues for the hydrophilic part) form *self-sealing lipid bilayers* that are the basis for all cellular membranes.
Amino Acids

- Amino acids are the subunits of proteins — they possess both a carboxylic acid group and an amino group, both linked to a single carbon atom called the \textit{alpha-carbon}.
- Amino acids differ from each other because of the different side chains that are also attached to the \textit{alpha-carbon}.
- A typical amino acid has the following structural formula:

\[
\begin{align*}
\text{Amino group} & \quad \text{H}_2\text{N} \quad \underset{\alpha\text{-carbon}}{\text{C}} \quad \text{COOH} \quad (\text{carboxyl group}) \\
& \quad \text{R} \quad (\text{side chain group})
\end{align*}
\]

\(R\) is one of 20 different side chains.
Classification of 20 different side chains

1. **Basic side chains**: These side chains usually contain a nitrogen atom/ amino group which takes up a H+ ion in solution. Examples of amino acids with basic side chains are lysine, arginine and histidine.

2. **Acidic side chains**: These amino acids contain a carboxyl group in the side chain. Examples of acidic side chains are R=CH₂COOH corresponding to aspartic acid, and R=C₂H₄COOH corresponding to glutamic acid.

3. **Uncharged polar side chains**: These amino acids contain a hydrophilic side chain containing an amino group or a hydroxyl group. Examples of amino acids with uncharged polar side chains are asparagine, glutamine, serine, threonine and tyrosine.

4. **Non-polar side chains**: These side chains contain hydrocarbons and are usually hydrophobic or non-polar. Examples of amino acids with nonpolar side chains are alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan, glycine and cysteine.
A representative amino acid

Disulphide bonds (S-S bonds) can form between two cysteine side chains and can play an important role in determining the 3-dimensional shape of the protein containing these cysteines.
Amino acids can be linked to each other via peptide bonds which are essentially amide linkages.

The carboxyl group of one amino acid reacts with the amino group of the next amino acid and the two get linked together in a condensation reaction which, as usual, is accompanied by the removal of a water molecule.

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{H} & \quad \text{C} \quad \text{C} \quad \text{O} \\
\text{N} & \quad \text{C} \quad \text{C} \quad \text{OH}
\end{align*}
\]

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{H} & \quad \text{C} \quad \text{C} \quad \text{O} \\
\text{N} & \quad \text{C} \quad \text{C} \quad \text{OH}
\end{align*}
\]

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{H} & \quad \text{C} \quad \text{C} \quad \text{O} \\
\text{N} & \quad \text{C} \quad \text{C} \quad \text{OH}
\end{align*}
\]

\[
\begin{align*}
\text{H} & \quad \text{H} \\
\text{H} & \quad \text{C} \quad \text{C} \quad \text{O} \\
\text{N} & \quad \text{C} \quad \text{C} \quad \text{OH}
\end{align*}
\]
Amino acid characteristics

• Well-defined directionality: an amino group on one end and a carboxyl group on the other.

• Possible to attach additional amino acids to the existing chain via further condensation reactions.

• The final chain that results will also have an amino group at one end (called the *amino terminus* or *N-terminus*) and a carboxyl group at the other end (called the *carboxyl terminus* or *C-terminus*).
• The four atoms in the box form a rigid planar unit so that there is no rotation about the C-N bond.
• However, the two bonds linking the α-carbons to the rest of the polypeptide chain allow rotation so that long chains of amino acids are very flexible.
• Since polypeptide chains can be of any length and at each location we can have 1 out of 20 possible amino acids, there is an enormous variety of proteins that can be synthesized using peptide bonds between amino acids.
Nucleotides

- Nucleotides are the subunits of nucleic acids such as DNA and RNA.
- A nucleotide is a molecule made up of a nitrogen-containing compound linked to a 5-carbon sugar that carries one or more phosphate groups.
- The nitrogen containing ring compounds (bases) are of two types: *pyrimidines* and *purines*.
Nucleotide *base* categories

six-member ring,

Pyrimidine

nine-member ring.

Purine
Pyrimidines

Cytosine (C)  Uracil (U)  Thymine (T)
Purines

![Chemical structures of Adenine (A) and Guanine (G)]
Nucleotides Sugars

Ribose

Deoxyribose
Formation of the nucleotide **cytidine monophosphate**.
Nucleic Acids

- Nucleic acids are formed by stringing nucleotide units together.
- If the sugar is deoxyribose and the bases are A, G, C, T, we have *deoxyribonucleic acid* or DNA.
- If the sugar is ribose and the bases are A, G, C, U, then we have *ribonucleic acid* or RNA.

The formation of nucleic acids from nucleotides will be treated in more detail in a later part of the course.
Nucleotide functions

- In addition to serving as the building blocks for DNA and RNA, nucleotides have many other functions:
  (i) they can serve as store houses of chemical energy;
  (ii) they combine with other groups to form activated carrier molecules or coenzymes; and
  (iii) they are used as specific signaling molecules in the cell, e.g. cyclic AMP (cAMP).

Details will be discussed later in the course.
Recap

(i) Macromolecules (polysaccharides, proteins and nucleic acids) found in cells contain a specific sequence of units (sugars, amino acids and nucleotides); these macromolecules are formed by condensation reactions where a new unit is added to a growing chain by the expulsion of a water molecule.

(ii) Noncovalent bonds such as hydrogen bonds, hydrophobic forces, etc. make macromolecules assume specific shapes.

(iii) These specific shapes as well as the noncovalent bonds allow macromolecules to seek out their appropriate partners and undergo the required reactions with them.
Energy Considerations in Biochemical Reactions
We examine the factors that determine whether a particular biochemical reaction is feasible or not.

If it is feasible, what additional factors are needed to get the reaction started?
• The Universe is always tending to greater disorder.

• A unique characteristic of living things: to create and maintain order.

• Living things obey the basic laws of thermodynamics.
  ➢ 1st Law: Energy can neither be created nor destroyed. It can only be converted from one form to another.
  ➢ 2nd Law: The entropy of the universe can only increase. The entropy is a measure of disorder.

• When a cell creates and maintains order, the disorder in its environment increases.

• For the cell and its environment taken together, the disorder does increase.
• Living cells perform a never-ending stream of reactions.
• Each cell is like a tiny chemical factory.
• Thousands of reactions are performed per second.
• Many cellular reactions do not normally occur at biological temperatures.
• Such reactions are made possible by using catalysts called enzymes and coupling reactions together.
Types of Chemical Reactions in Cells

• Two main types of chemical reactions in cells:
  ➢ *Catabolic reactions*: foodstuffs are broken down into smaller molecules, generating useful energy.
  ➢ *Anabolic reactions*: the energy harnessed by catabolism is used for synthesis of other molecules.

• Catabolic and anabolic reactions together constitute the *metabolism* of the cell.
Concept of Free Energy

- **Free energy**: the amount of energy that is available to do useful work.
- A reaction is energetically favorable if it proceeds with a decrease in free energy.
- Cells generate biological order by coupling the biological order generating reactions (energetically unfavorable) to other heat generating reactions (energetically favorable).
- Releasing heat to the environment ensures that
  - the entropy of the cell and its surroundings increases.
  - biological order is created within the cell itself.
Some Common Biochemical Reactions in Cells

• *Photosynthesis* : Process by which plants trap energy directly from sunlight to produce sugars.

• *Cellular Respiration* : Oxidation of carbohydrates and sugars to produce useful energy.

• *Oxidation and Reduction* :
  - *Oxidation* reactions involve the addition of oxygen or the removal of hydrogen.
  - *Reduction* reactions involve the addition of hydrogen or the removal of oxygen.
Photosynthesis

- **Photosynthesis** occurs in the chloroplasts of plant cells.
- Overall reaction for photosynthesis:

  \[ nCO_2 + nH_2O + \text{sunlight} \rightarrow (\text{sugars}) \ (CH_2O)n + nO_2 + \text{heat energy} \]

- Photosynthesis: A two stage process.
  - Stage 1: Energy from sunlight is captured and transiently stored as chemical bond energy in specialized small molecules.
  - Stage 2: Energy carriers from the first stage are used to drive a process in which sugars are manufactured from carbon dioxide and water.
- The sugars produced are used:
  - as a source of chemical bond energy
  - as source of materials to make the many other organic molecules.
Cellular Respiration

• Cells obtain energy by the oxidation of organic molecules.
• For carbohydrates, the general reaction:
  \[(CH_2O) \text{ (sugars)} + O_2 \rightarrow CO_2 + H_2O + \text{energy.}\]
• Referred to as cellular respiration because oxygen is used up and carbon dioxide & water vapour are released.
• Also, a great deal of energy is released.
• The cell would be unable to store this much energy if this process occurred in a single step.
• Therefore, the reaction is carried out in several incremental steps.
• Each step results in the release of a small amount of energy.
• The cell can store this energy in specialized energy carriers.
• About 50% of the bond energy is stored in a retrievable form by the cell: highly efficient process.
Oxidation and Reduction

- Consider the reaction:
  \[(CH_2O) \text{ (sugar)} + O_2 \rightarrow CO_2 + H_2O\]
- From preceding definitions: the sugar molecule is oxidized while the oxygen molecule is reduced.
- Alternative definitions:
  - oxidation involves loss of electrons
  - reduction involves gain of electrons
- In a chemical reaction, the total number of electrons does not change.
- Whenever one compound gets oxidized, another compound gets reduced, i.e., oxidation and reduction occur simultaneously.
- Such reactions are called **Redox** reactions.
• The terms oxidation and reduction apply even when there is only a partial shift of electrons between atoms linked by a covalent bond.
• Such a shift occurs when a carbon atom is covalently bonded to an atom with a strong affinity for electrons. (Example: oxygen)
• The carbon atom gives up more than its equal share of electrons and forms a polar covalent bond.
• It is said to have been oxidized.
• The step by step oxidation of methane:

\[
\text{methane} \rightarrow \text{methanol} \rightarrow \text{formaldehyde} \rightarrow \text{formic acid} \rightarrow \text{carbon dioxide}
\]
Role of Enzymes

- Chemical reactions proceed only in a direction that leads to a loss of free energy. (from 2\textsuperscript{nd} law of thermodynamics)
- Such reactions are said to be \textit{energetically favorable}.
- Even such reactions do not occur spontaneously.
- The molecules require \textit{activation energy}, i.e., a kick over an energy barrier to get the reaction going.
- One way to try to overcome the energy barrier is to supply heat energy.
- Living cells have to operate within a narrow temperature range; such a mechanism may not work.
Role of Enzymes

• The necessary boost over the energy barrier is provided by a specialized class of proteins called *enzymes*.
• Enzymes are very effective catalysts, often speeding up reactions by a factor of as much as $10^{14}$.
• Each enzyme binds tightly to one or two molecules called *substrates*.
• This greatly reduces the activation energy of a particular chemical reaction.
• Once the substrates have reacted, the enzyme dissociates from the products and is free to bind additional substrate molecules.
Activation energy requirement in the absence (red) and presence (blue) of enzymes
Role of Enzymes (contd.)

- Enzymes are highly selective and each enzyme only catalyzes a particular reaction.
- By using different enzymes, a cell can regulate a number of reactions in a somewhat decoupled way.
- Enzymes find their substrates through rapid diffusion.
- Random collisions between the molecules in a cell result in almost immediate dissociation.
- When a substrate encounters the appropriate enzyme, immediate association takes place.
- Several weak bonds are formed linking the two of them together and facilitating the necessary reaction.
Feasibility of Chemical Reactions

• Chemical reactions are energetically favorable if they are accompanied by a decrease in the free energy, i.e., $\Delta G < 0$.
• Many reactions inside cells, e.g. the synthesis of complex carbohydrates from sugars are *energetically unfavorable*.
• Such reactions will occur only if they are coupled to an energetically favorable reaction such that the combined free energy change is negative.

• **Adenosine tri phosphate** (ATP) is a molecule that participates in many coupled reactions.
• ATP is capable of storing a large amount of energy in its phosphoanhydride bonds.
Adenosine Triphosphate (ATP)

- ATP is obtained by attaching two inorganic phosphate ions in cascade to the Adenosine Monophosphate (AMP) molecule.
- One or both of the phosphoanhydridre bonds can be readily hydrolyzed to yield Adenosine Diphosphate (ADP) or Adenosine Monophosphate (AMP) and both the reactions are energetically favourable.
An example of a coupled reaction

• For the following condensation reaction:
  \[
  \text{glucose} + \text{fructose} \rightarrow \text{sucrose}
  \]
  \[
  \Delta G^0 = 5.5 \text{ kcal/mole.}
  \]

• Since this number is positive, this reaction will not occur by itself.

• The hydrolysis of ATP is an energetically favourable reaction, having \( \Delta G^0 = -7.3 \text{kcal/mole:} \)
  \[
  \text{ATP} \rightarrow \text{ADP} + \text{P}_i.
  \]
A coupled reaction (contd.)

- The two reactions above can be coupled together as follows:
  - The ATP is readily hydrolyzed to ADP and the high energy phosphate ion is incorporated into location 1 on the glucose molecule to yield glucose-1-phosphate.
    \[
    \text{glucose} + \text{ATP} \rightarrow \text{glucose-1-P} + \text{ADP}
    \]
  - The glucose-1 phosphate then reacts with fructose to produce sucrose and the inorganic phosphate ion
    \[
    \text{glucose-1-P} + \text{fructose} \rightarrow \text{sucrose} + \text{P}_i
    \]
A coupled reaction (contd.)

- The free energy change for sequential reactions is additive.
- The standard free energy change for the overall reaction is \( \Delta G^0 = -1.8 \text{kcal/mole} \)
- The net result is that sucrose is made in a reaction driven by the hydrolysis of ATP.
Standard Free Energy Change

• The free energy change for a reaction depends on the concentration of the reactants.

• Consider the simple reaction below:

\[ A \rightarrow B \]

• The free energy change for this reaction is given by:

\[ \Delta G = \Delta G^0 + 0.616 \ln \frac{[B]}{[A]} \]

➢ \( \Delta G^0 \) = the standard free energy change
➢ \([B]\) = concentration of \( B \) in moles/liter
➢ \([A]\) = concentration of \( A \) in moles/liter

• If \([B] = [A]\) then

\[ \Delta G = \Delta G^0 = \text{standard free energy change for the reaction.} \]
Equilibrium Constant of a Reaction

- At equilibrium, the concentration of the reactants and products does not change; hence $\Delta G = 0$.
- If $[A_e]$ and $[B_e]$ denote the equilibrium concentrations of $[A]$ and $[B]$ respectively, then:

$$\Delta G^0 = -0.616 \ln \frac{[B_e]}{[A_e]}.$$

- The quantity $K := [B_e] / [A_e]$ is called the equilibrium constant of the reaction.
• Cells can speed up reactions in at least two ways:
  ➢ Compartmentalization: Rapid removal of products into a separate compartment.
  ➢ Use the products of one reaction as the reactants of an immediately following reaction.
• In the second case above, cells can cause the energetically unfavorable transition X→Y to occur if an enzyme catalyzing the X→Y reaction is supplemented by a second enzyme that catalyzes the energetically favorable reaction Y→Z.
• If the overall free energy change for the reaction X→Y→Z is negative, then the reaction X→Y will proceed rapidly, although by itself it would have been thermodynamically impossible.
Activated Carrier Molecules and their Role in Biosynthesis

- The energy released by the oxidation of food molecules is used in the construction of other organic molecules.
- The energy is temporarily stored as chemical bond energy in activated “carrier molecules” or coenzymes.
- These molecules transfer energy from one site to another by rapid diffusion.
- The coenzymes store their energy either in
  - high energy bonds, e.g. phosphoanhydride bonds in ATP.
  - as high energy electrons in another class of molecules.
Activated Carrier Molecules and their Role in Biosynthesis

- **ATP** is the most widely used activated carrier molecule in cells.
- The hydrolysis of ATP releases energy for doing cellular work and biosynthesis.
  \[
  \text{ATP} \rightarrow \text{ADP} + \text{P}_i.
  \]
- The reverse reaction forming ATP is driven by energy from sunlight or oxidation of glucose.
- The above reactions together constitute the mechanism by which cells are powered.
Activated Carrier Molecules and their Role in Biosynthesis

- *NADH* and *NADPH* are important electron carriers: supply reducing power needed for certain reactions.
- NADP+ is the oxidized form of NADPH.
- NADP+ contains a nicotine amide ring and is called nicotinamide adenine dinucleotide phosphate.
- By adding a proton (H+) and two high energy electrons to NADP+, one obtains NADPH.
- NADPH provides strong reducing power.
- NADH and NAD+ differ from NADPH and NADP+ respectively in that the bottom phosphate group is missing in the former.
Structure of NADP+ and NADPH

NADP+ oxidized form

NADPH reduced form

nicotinamide ring

RIBOSE

ADENINE

this phosphate group is missing in NAD+ and NADH
Role of NADH and NADPH

- NADH has a role as an intermediate in the *catabolic* system of reactions that generate ATP.
- NADPH operates with enzymes that catalyze *anabolic* reactions, supplying the high-energy electrons needed to synthesize energy-rich biological molecules.
- There are two different sets of activated carrier molecules for two different pathways.
- Thus, the cell has some amount of independent control over these pathways.
Acetyl CoA

- *Acetyl coenzyme A* (acetyl CoA) is another important activated carrier molecule.
- The bond between the acetyl group and the CoA molecule is a high energy bond.
- Acetyl CoA can readily transfer the acetyl group to other molecules making certain reactions possible.
Schematic Structure of Acetyl CoA

Acetyl group

CoA

High energy bond

nucleotide

adenine

ribose

O

O

O

O
Conclusion

• All the activated carrier molecules discussed contain a nucleotide.

• ATP is the only one that is involved in actual nucleic acid synthesis.

• The important role played by activated carrier molecules in biology:
  
  ➢ The condensation reactions involved in the synthesis of polysaccharides, proteins and nucleic acids are all powered by the energy derived from ATP hydrolysis.
Proteins
Proteins : An Introduction

Proteins are:

• the most versatile functional molecules found in a cell.
• responsible for carrying out nearly all cellular functions.
• constitute most of the dry mass of a cell.
Proteins : An Introduction

For Example :

• proteins called enzymes promote many chemical reactions.
• proteins serve as signaling molecules from one cell to another.
• proteins serve as moving parts of tiny molecular machines.
• antibodies, toxins, hormones, etc are all proteins.
Protein Structure and Function

• Proteins adopt a large number of different three dimensional shapes.
• Thus, they can perform versatile functions.
• The shape of a protein is specified by its amino acid sequence.
• A protein molecule is made from a long chain of amino acids.
Protein Structure and Function

• Each amino acid in the protein molecule is linked to its neighbor by a peptide bond.
• The repeated sequence of atoms along the chain is referred to as the polypeptide backbone.
• The side chains are unique to each protein.
• They are crucial in determining its distinctive properties.
Protein Structure and Function

• Proteins fold up into different shapes because of different sets of weak noncovalent bonds such as:
  ➢ hydrogen bonds
  ➢ ionic bonds
  ➢ Van der Waals attractions
  ➢ hydrophobic forces

• Consider the hypothetical protein in the following figure which is made up of 4 amino acids (2 polar, 2 non-polar) and its conformation in aqueous environment.
A simple protein with polar and non-polar side chains

Three dimensional conformation in an aqueous environment
Protein Structure and Function

- A given protein will fold into a conformation of lowest energy.
- A protein can be unfolded or *denatured* by treatment with certain solvents.
- When the solvent is removed, the protein *renatures* into its original conformation.

*Chaperone* proteins assist protein folding inside cells by speeding up the folding process.
Protein Structure and Function

- Proteins range in size from about 30 amino acids to more than 10,000 amino acids.
- Majority of proteins are between 50 and 2000 amino acids long.
- The conformation of a given protein is very complex.
- There are a few common structural motifs that underlie the various conformations.
- This facilitates a general study of protein structure.
The $\alpha$-helix and the $\beta$-sheet

- These structures result from hydrogen bonding between the $\text{N} - \text{H}$ and $\text{C} = \text{O}$ groups in the polypeptide backbone.
- Hence, these two structures are quite common.
- The side chains are not involved in hydrogen bonding.
- Thus, they can be formed by many different amino acid sequences.
Protein alpha helix
The $\alpha$-helix

- **$\alpha$-helix** generation: when a single polypeptide chain turns around itself to make a rigid cylinder.
  - A hydrogen bond is formed between every fourth peptide bond.
  - This gives rise to a regular helix with a complete turn every 3.6 amino acids.
  - Sometimes, $\alpha$-helices wrap around each other to form a particularly stable structure: a *coiled coil*.

- The following figure shows a transmembrane protein crossing the lipid bi-layer.
• hydrophobic side chains are exposed to the lipid bilayer.

• hydrophilic side chains face the inside of the helix.
The $\beta$-sheet

- $\beta$-sheets can form either from:
  - neighboring polypeptide chains that run in the same orientation. (i.e. parallel chains)
  - a polypeptide chain that folds back and forth upon itself. (i.e., antiparallel chains)

- These sheets are held together by hydrogen bonds that connect the peptide bonds in the neighboring chains.
- Both these types produce a very rigid structure.

- The following figure shows both parallel and anti-parallel $\beta$-sheets:
Schematic diagram of parallel and anti-parallel $\beta$ sheets
Levels of Organization in Proteins

The structure of a protein can be studied at four different levels:

• *primary structure* which corresponds to the amino acid sequence.

• *secondary structure* which corresponds to the existence of \( \alpha \)-helices and \( \beta \)-sheets.

• *tertiary structure* which corresponds to the 3-dimensional conformation.

• *quarternary structure* which corresponds to studying the complete structure for proteins made up of *more than one* polypeptide chain.
Protein Domain

• A protein domain is produced by any part of a polypeptide chain that can fold independently into a compact, stable structure.

• A domain usually contains between 50 and 350 amino acids.

• It is the modular unit from which many larger proteins are constructed.
Evolution of Proteins

• In theory, a vast number of different polypeptide chains could be made.
• For a polypeptide chain 250 amino acids long, $20^{250}$ different polypeptide chains are possible.
• Only a small fraction of these would adopt a stable three-dimensional conformation.
• Proteins that do not have a single stable conformation are not biologically useful.
• Hence, they have been eliminated during the process of evolution.
Evolution of Proteins

• A present day protein has a single stable conformation with properties that enable it to perform its function.

• Consider the protein hemoglobin found in blood.

• This protein has the right 3-D shape that enables it to bind oxygen atoms and carry them from the lungs to the tissues.
Protein Families

- Once a protein evolves into a stable conformation, its structure could be modified slightly to enable it to perform new functions.

- Such proteins can be grouped into protein families.

- Members of each such family display many similarities, including their three dimensional conformation.
Protein Subunits

- Large protein molecules often contain more than one polypeptide chain.

- Each polypeptide chain is called a protein subunit.

- The subunits bind to each other by weak noncovalent bonds.

- When the polypeptide chains are identical, we have:
  - dimers - 2 polypeptide chains
  - tetramers - 4 polypeptide chains
Other proteins contain two or more different types of polypeptide chains.

For instance, the protein hemoglobin contains two identical α-globin subunits and two identical β-globin subunits.

Proteins can assemble into filaments, sheets or spheres.

The protein units can bind to form a very long helix or a circular ring.
Identical protein units forming a dimer, a helix and a ring
• Structures such as viruses and ribosomes (protein making machines) are built from:
  - a mixture of one or more types of proteins.
  - RNA or DNA molecules.

• All these structures can be isolated in pure form and dissociated into their constituent molecules.

• It is possible to mix the isolated components together and watch the structures reassemble spontaneously.

• A helix is a common motif in biological structures and results from similar units connected end to end.
• Some proteins have elongated fibrous shapes; such as:
  ➢ α-keratin (found in hair)
  ➢ collagen
  ➢ elastin

• Extracellular proteins are often stabilized by covalent cross
  linkages such as the disulphide bonds between cysteine side
  chains.
• These serve as “atomic staples” to stabilize the three
  dimensional conformation of the protein outside the cell.
• Inside the cell, such staples are not needed and hydrogen
  ions are readily available.
Proteins

Disulphide bonding between cysteine side chains
Protein-Ligand Interactions

- Proteins bind to other molecules (called their ligands) with great specificity.

- The ligand must fit precisely into a groove on the protein surface. This allows for the protein and its ligand to be linked together strongly by large numbers of weak covalent bonds.

- This perfect kind of fitting enables a protein to pick out its ligand in a crowded environment.

- The region of a protein that associates with a ligand is called a binding site.
Antibodies

- **Antibodies**, or *immunoglobulins*, are a class of proteins that have a highly developed binding capacity.
- They are produced by the immune system in animals in response to foreign molecules.
- Each antibody binds to a particular target molecule tightly and either inactivates the target directly or marks it for destruction.
- Antibodies are Y-shaped molecules with two identical binding sites.
- Each site is complementary to a small portion of the surface of the target (antigen) molecules.
A typical antibody molecule
Antibodies

• An individual animal can make billions of different antibody molecules.

• Each antibody molecule has a distinct antigen binding site.

• Antibodies defend us against infection in at least two different ways:
  ➢ Antibody and antigen aggregates are ingested by phagocytic cells.
  ➢ Antibody coated bacteria or viruses are killed by special proteins in the blood.
**B-cells**

- Antibodies are made by a class of white blood cells, called *B cell lymphocytes* or *B cells*.
- Each B cell carries a different membrane-bound antibody molecule on its surface that serves as a receptor for recognizing a specific antigen.
- When the appropriate antigen binds to the receptor, the B cell is stimulated to divide and to secrete large amounts of the same antibody.
- This phenomenon can be used to raise the level of antibodies in animals and forms the basis of *immunization* and *immunotherapy*. 
• Antibodies can also be used to purify molecules.
• In *immunoaffinity column chromatography*, a mixture of A and other molecules is passed through a column containing beads coated with anti-A antibodies.
• The flow through the column is discarded and the A molecules can be recovered by appropriately treating the beads.

• In *immunoprecipitation*, a mixture of molecules containing A is treated with anti-A antibodies.
• The aggregate of A molecules plus anti-A antibodies can be collected by centrifugation.
• *Monoclonal antibodies* are large quantities of a single type of antibody molecule.

• They can be obtained by fusing a B cell (taken from an animal injected with antigen A) with a cell of a B cell tumor (divides indefinitely but does not make the antibody).

• The resulting cell divides indefinitely and also produces the antibody.
Antibodies as Molecular Tags

- Antibodies can also be used as molecular markers or molecular tags.
- One can couple a set of antibody molecules to a fluorescent dye, a colloidal gold particle or some other special tag.
Isolating Proteins from Cells

• To study proteins, we first need to isolate them from cells and tissues.

• The first step is to break open the cell by disrupting its plasma membrane.
  
  Four different approaches commonly used:
  
  ➢ break cells with high frequency sound.
  ➢ use a mild detergent to make holes in the plasma membrane
  ➢ force cells through a small hole using high pressure.
  ➢ shear cells between a close-fitting rotating plunger and the thick walls of a glass vessel.
Isolating Proteins from Cells

• The resulting *homogenate* (or *abstract*) contains molecules from the cytosol as well as the membrane bounded organelles.

• Careful homogenization can leave the membrane bounded organelles intact.

• The next step is to separate out the different components.

• There are a number of procedures used for separation such as *centrifugation*, *velocity sedimentation* and *equilibrium sedimentation*. 
Centrifugation

- The homogenate is placed in test tubes and rotated at high speeds in a centrifuge.

- Components separate out mainly on the basis of their size and density.

- Repeated centrifugation at progressively higher speeds will fractionate the cell homogenate into its different components.
Separating out a mixture via centrifugation

Drain the supernatant and subject to additional centrifugation
Velocity sedimentation

• Subcellular components are carefully layered over a dilute salt solution.

• These components sediment at different speeds according to their size.

• When sedimented through such a dilute solution, different cell components separate into distinct bands.

• After an appropriate time, the components can be collected individually.
Equilibrium sedimentation

- This procedure can be used to separate cellular components on the basis of their *buoyant density*, independently of their shape or size. It makes use of a centrifuge.

- The sample is usually layered on top of a steep density gradient that contains a very high concentration of sucrose or cesium chloride.

- Each subcellular component will move up or down when centrifuged until it reaches a position where its density matches that of its surroundings and then it will move no more.

- A series of distinct bands will form inside the test tubes and these bands can be separated out by collecting from the base of the test tube.
Separating a mixture of proteins

• Proteins are very diverse molecules; they differ from each other by size, shape, charge, hydrophobicity, etc.

• All of these properties can be used to separate proteins from each other so that they can be studied individually.

• *Column chromatography* and *Gel electrophoresis* are experimental procedures that can be used to separate out a mixture of proteins.
Column Chromatography

- The mixture of proteins is forced through a vertical column containing a matrix.

- The sample is positioned on top of the matrix and solvent from a reservoir is used to force the sample down the matrix.

- Different choices for the matrix lead to different kinds of chromatography.
Column Chromatography

• In *Ion Exchange Chromatography*, the matrix consists of positively or negatively charged beads that retard proteins of the opposite charge. Thus separation is achieved on the basis of charge.

• In *Gel-filtration Chromatography*, the matrix consists of tiny porous beads. Protein molecules that are small enough to enter the holes in the beads are delayed and travel more slowly through the column. In this case, the separation is achieved on the basis of size.

• In *Affinity Chromatography*, the matrix consists of molecules that interact specifically with the protein of interest, for instance, an antibody.
Proteins that bind specifically to such a column can finally be released by a pH change or by concentrated salt solutions, and they emerge highly purified.
Gel Electrophoresis

- This procedure is used to separate proteins on the basis of their size and net charge.

- In *SDS PAGE* (*sodium dodecyl sulphate polyacrylamide gel electrophoresis*), protein separation is achieved on the basis of size.

- *Isoelectric focusing* allows us to separate proteins on the basis of their charge.
SDS PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis)

- The individual polypeptide chains form a complex with negatively charged molecules of SDS.

- When subjected to an electric field, they migrate as a negatively charged SDS-protein complex through a slab of porous polyacrylamide gel.

- The proteins migrate at a rate that is proportional to their molecular weight; thus protein separation is on the basis of their size.
Isoelectric Focusing

• For any protein, there is a characteristic pH, called the isoelectric point, at which the protein has no net charge and therefore will not move in an electric field.

• In isoelectric focusing, proteins are electrophoresed in a narrow tube of polyacrylamide gel in which a pH gradient is established by a mixture of special buffers.

• Each protein moves to a point in the gradient that corresponds to its isoelectric point and stays there; thus allowing protein separation on the basis of charge.
Two dimensional polyacrylamide gel electrophoresis

- This method combines the two procedures just discussed to separate proteins on the basis of both size and electric charge.

- In this method, first isoelectric focusing is used to separate proteins on the basis of electric charge.

- Then SDS PAGE is applied in a direction perpendicular to the earlier one to achieve further separation on the basis of size.
Two dimensional polyacrylamide gel electrophoresis
Protein Structure Determination

• There are a number of different approaches that can be used to determine protein structure.

• In the first approach, a reagent is used to isolate the N-terminus amino acid from the rest of the polypeptide chain.

• The isolated amino acid can then be identified using a full set of known amino acid derivative standards e.g., the position to which a particular amino acid migrates under SDS PAGE.

• This process can be repeatedly used to determine the entire amino acid sequence of the protein.

• However, it is often found that the amino-terminus amino acid is chemically blocked; in that case, this method cannot be used.
Protein Structure Determination

• The second approach makes use of the fact that certain chemicals selectively cleave proteins at particular amino acid locations. For instance, the enzyme trypsin cleaves on the carboxyl side of lysine or arginine residues while the chemical cyanogen bromide cuts peptide bonds on the carboxyl side of methionine.

• Thus these reagents can produce a few relatively large peptides.

• These can be separated (using, for instance, gel electrophoresis) to create a peptide map that is diagnostic of the protein from which the peptides are generated.

• The sequencing of overlapping peptides can be used to laboriously piece together the sequence of the original protein.
Protein Structure Determination

- The third approach is based on recent advances in DNA technology.

- We note that knowing the sequence of as few as 20 amino acids of a protein is often enough to allow a DNA probe to be designed, so that the gene encoding the protein can be cloned.

- Once the gene, or the corresponding cDNA (complementary DNA) has been sequenced, the rest of the protein's amino acid sequence can be deduced by reference to the genetic code.

- In addition to the above approaches, the three dimensional structure of protein molecules has been studied using *X-ray crystallography* and *nuclear magnetic resonance spectroscopy*. 
Proteins that are Enzymes

- The binding strength between an enzyme and its substrate (or an antibody and its antigen) is measured by the *equilibrium constant* $K$.
- Consider the reaction:

$$A + B \xrightarrow{\text{association}} AB$$

- For this reaction, the reaction rate, called the *association rate* is given by:

$$\text{association rate} = \text{association rate constant} (k_{on}) \times [A] \times [B]$$
Proteins that are Enzymes

- Consider the reverse reaction:

\[ AB \xrightarrow{\text{dissociation}} A + B \]

- For this reaction, the *dissociation rate*, is given by:

\[ \text{dissociation rate} = \text{dissociation rate constant (} k_{\text{off}}\text{)} \times [AB] \]

- At equilibrium, the association rate equals the dissociation rate, therefore:

\[ k_{\text{on}} [A_e] [B_e] = k_{\text{off}} [AB_e] \]
Proteins that are Enzymes

• The quantity

\[ K = \frac{[AB_e]}{[A_e][B_e]} = \frac{k_{on}}{k_{off}} \]

is called the *equilibrium constant*.

• If $K$ is large then $[AB_e] >> [A_e]$ and $[B_e]$, the binding between A and B is very strong, the free energy of AB minus the free energy of A+B is very negative, and the reaction is dominant in the forward direction.
Proteins that are Enzymes

- Enzymes are powerful and highly specific catalysts.
- Each enzyme catalyzes only one very specific chemical reaction.
- Enzymes catalyzing similar reactions can be grouped together and have special names:
  - *hydrolases* are enzymes that catalyze a hydrolytic cleavage reaction;
  - *nucleases* are enzymes that break down nucleic acids by hydrolyzing phosphodiester bonds between nucleotides;
Proteins that are Enzymes

• *proteases* are enzymes that break down proteins by hydrolyzing peptide bonds between amino acids;

• *synthases* is the general name for enzymes that synthesize molecules in the anabolic reactions by condensing two smaller molecules together;

• *isomerases* are enzymes that catalyze the rearrangement of bonds within a single molecule;

• *polymerases* are enzymes that catalyze polymerization reactions such as the synthesis of DNA and RNA;
Proteins that are Enzymes

• *kinases* are enzymes that catalyze the addition of phosphate groups to molecules;

• *phosphatases* are enzymes that catalyze the hydrolytic removal of a phosphate group from a molecule;

• *oxido-reductases* is the general name for enzymes that catalyze reactions in which one molecule is oxidized while the other is reduced. Enzymes of this type are often called oxidases, reductases and dehydrogenases;
Proteins that are Enzymes

- *ATPases* are enzymes that hydrolyze ATP.

- Many proteins have an energy harnessing ATPase activity as part of their function.

For e.g.:
- motor proteins such as *myosin* (a muscle protein that is responsible for generating movement)
- membrane transport proteins such as the *sodium-potassium pump* (a protein that pumps sodium and potassium ions across cell membranes.)
Reaction Rate vs. Substrate Concentration

• As the substrate concentration is increased, the reaction rate initially goes up as more substrate molecules are available to interact with the enzyme.

• Beyond a certain substrate concentration, the reaction rate saturates since the enzyme is already functioning at its maximal speed and cannot process substrate molecules any faster.

• This limiting velocity of the enzyme is called $V_{\text{max}}$. 
Reaction Rate vs. Substrate Concentration

- **Rate of reaction**
  - $V_{max}$
  - $0.5V_{max}$

- **Substrate concentration**
  - $K_M$
Reaction Rate vs. Substrate Concentration

- The substrate concentration at which a given enzyme can function at half its maximal velocity is called the $K_m$ for that enzyme.
- It measures the affinity of the particular enzyme for the substrate.

- If two enzymes compete for the same substrate then the values of their $V_{\text{max}}$'s and $K_m$'s determine which reaction gets preference.
- If the substrate is limited, then the enzyme with the lower $K_m$ gets preference while if the substrate is not a limiting factor, then the enzyme with the higher $V_{\text{max}}$ dominates.
• Despite their versatility, there are instances where a protein needs the help of a non-protein molecule for performing functions that would have otherwise been impossible.

For example:

• A molecule of hemoglobin needs the help of four *heme* groups (ring shaped molecules) in order to be able to perform its function of picking up oxygen in the lungs and releasing it in the tissues;

• The signal receptor protein *rhodopsin* in the retina detects light by means of a small molecule called *retinal* that is embedded in the protein.
Regulation of Catalytic Activities of Enzymes

• The cell can regulate how much of each enzyme to produce by regulating *gene expression*.

• The cell can control enzymatic activities by confining sets of enzymes to particular subcellular compartments.

• An enzyme can change its activity in response to other molecules that it encounters in the cell.

• An example of an enzyme changing its activity in response to the presence of other molecules is *feedback inhibition*. 
Feedback Inhibition

- In feedback inhibition, X produces Y which produces Z. However, Z inhibits the production of X. This is an example of negative feedback (single loop).
- In general, there will be multi-loop feedbacks.
Negative Feedback: Another example

- When there is a rise in ADP in the cell (which is indicative of a low energy state) several enzymes involved in the oxidation of sugar molecules are activated.

- This stimulates the cell to convert ADP to ATP.

- This lowers the level of ADP in the cell and restores the desired energy level.
Allosteric Proteins

- In feedback inhibition, often times, it is found that the regulatory molecule has a shape totally different from the shape of the substrate of the enzyme.
- In this case, the feedback inhibition is not caused by the regulatory molecule binding preferentially to the same binding site on the enzyme.
- Instead, the enzyme has two binding sites, one for the substrate and the other for the regulatory molecule.
- Such proteins are called *allosteric proteins*.
- When the regulatory molecule binds the protein at one site, the conformation of the protein is slightly altered such that the substrate is no longer able to bind at the other binding site.
Regulation of protein function: change in conformation

- Eucaryotic cells commonly regulate a protein's function by adding a phosphate group covalently to one of its amino acid side chains.
- Because each phosphate group carries two negative charges, the enzyme-catalyzed addition of a phosphate group to a protein can cause a major conformational change.
- Removal of the phosphate group by a second enzyme returns the protein to its original conformation and restores its initial activity.
Regulation of protein function: change in conformation

- So, in eucaryotic cells, protein kinases and protein phosphotases play an important role in switching proteins from inactive to active conformations and vice versa.
- For instance, proteins undergo dramatic conformational changes when they bind GTP (guanosine triphosphate). The protein bound to GTP is active.
- The hydrolysis of the bound GTP returns the protein to an inactive configuration where it still binds GDP.
- After a while, the bound GDP is released but the protein remains inactive until it binds a GTP again.
Directed Movement in Cells

- Motor proteins can produce *directed movement* in cells by coupling the hydrolysis of ATP to conformational changes in the protein.
- In the absence of ATP hydrolysis, the protein would move back and forth in different directions generating random movement.
- By coupling the movement to ATP hydrolysis, which is an energetically favorable reaction, the movement steps are made irreversible thereby producing directed movement.
- This is the mechanism by which DNA and RNA polymerases move along a template DNA strand.
A motor protein that can produce directed movement
DNA
DeoxyriboNucleic Acid
the molecule used in cells to store genetic information.
• Genetic information must be passed on from a cell to its daughter cells at cell division.

• This information is passed on from generation to generation of organisms through the reproductive cells.
Genes

• *Genes* are the information-containing elements that determine the characteristics of a species as a whole and of the individuals within it.

• Natural questions that come up are:
  – the chemical composition of the genes.
  – the mechanism by which they store information which can be replicated in an almost unlimited fashion.
Chromosomes

- By the early twentieth century, biologists knew that genes were carried on chromosomes.
- Biochemical analysis of chromosomes revealed that they are made up of DNA and protein.
Two closely related strains of the bacterium *streptococcus pneumonia* were used:

1. The **R strain** which appears **Rough** under the microscope, is **not pathogenic**, and does not cause pneumonia when injected into animals.

2. The **S strain** which appears **Smooth** under the microscope, is **pathogenic**, and causes pneumonia when injected into animals, resulting in death.
• Live R strain cells were grown in the presence of *heat killed* S strain cells.

• Animals injected with these cultured R strain cells died leading to the conclusion that the *R strain cells had been transformed to S strain cells.*

  molecules that carry heritable information are present in the heat killed S strain cells.
• Cell free extract from these transformed R-strain cells was fractionated into classes of molecules.

• The resulting RNA, proteins, lipids and carbohydrates were all found to be of the R strain type.

• Only the DNA was found to have been transformed to the S strain type.

DNA carries the heritable information.
Structure of DNA was discovered by James Watson and Francis Crick in 1953.
DNA Structure

• A DNA molecule consists of two complementary chains of nucleotides.

• Four different nucleotide bases make up DNA
  Adenine (A), Guanine (G), Cytosine (C) and Thymine (T)

• A pairs with T through two hydrogen bonds while G pairs with C through three hydrogen bonds.

  $A \leftarrow T \quad G \rightarrow C$
DNA Structure

- The pairing of A with T and G with C ensures that each pair is of similar width, thus holding the sugar phosphate backbones an equal distance apart along the DNA molecule.

- In addition, the two sugar phosphate backbones wind around each other to form a double helix.
Encoding Genetic Information

- DNA encodes information in the order, or sequence of the nucleotides along each strand.
- Each base A, G, C or T can be considered as a letter in a four letter alphabet that is used to spell out biological messages in the chemical structure of the DNA.
- Organisms differ from one another because their respective DNA molecules have different nucleotide sequences.
Encoding Genetic Information

• The linear sequence of nucleotides in a gene spells out the linear sequence of amino acids in a protein.

• The exact correspondence between the four-letter nucleotide alphabet of DNA and the twenty-letter amino acid alphabet of protein, called the genetic code, is discussed in the next chapter.

• Here, we will focus on the mechanisms by which DNA is replicated, possible errors that occur during DNA replication and the mechanisms used to correct them.
Genome

- Most genes are short stretches of DNA encoding a single protein.
- However, not all of the DNA in a gene is used to encode the protein that it specifies.
- *regulatory regions*: specify when and in what amounts the protein encoded by a gene is made.
- The complete set of information in an organism's DNA is called the *genome*. 
DNA Replication

- At each cell division, the cell must copy its genome in order to pass it to both the daughter cells.
- Each strand of DNA contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand.
- Therefore, each strand can act as a *template*, or mold, for the synthesis of a new complementary strand.
Complementary DNA Strands

The strand S can serve as a template for synthesizing a new strand S' while S' can serve as a template for synthesizing a new strand S.
DNA Replication Scheme

(i) the two strands separate out.
(ii) each strand can be used as a template for synthesizing its complementary strand.
(iii) the newly synthesized strand and the template used to create it pair together to form a new double helix.

Replication “semi-conservative”: In each round of replication, one of the strands in each double helix is “conserved” from the previous round of replication.
Replication Origins

- Replication begins at replication origins where the two DNA strands separate.

- This separation of the two DNA strands is usually carried out by *initiator proteins* that bind to the DNA and pull the two strands apart.

- Since DNA that is rich in AT base pairs has fewer hydrogen bonds it is usually the site of replication origins.

- A bacterial genome: usually has only one replication origin.

- Human genomes: there are about 10,000 replication origins.
The core of the DNA replication machine is formed by an enzyme called *DNA polymerase*. This enzyme synthesizes the new DNA using one of the old strands as a template. This enzyme catalyzes the addition of nucleotides to the 3' end of a growing DNA strand by the formation of a phosphodiester bond between this end and the 5' phosphate group of the incoming nucleotide.
The nucleotides enter the reaction initially as energy rich *nucleoside triphosphates*, which provide the energy for the polymerization reaction.

The hydrolysis of one phosphoanhydride bond in the *nucleoside triphosphate* provides the energy for the condensation reaction that links the nucleotide monomer to the chain and releases pyrophosphate. (two inorganic phosphate ions linked together by a phosphoanhydride bond)
DNA Replication

Original Double stranded DNA

Opened up Single stranded DNA templates ready for DNA synthesis
DNA Replication Fork

- DNA polymerase can synthesize DNA only in the 5' to 3' direction.
- This causes a problem since the replication fork is asymmetrical.
- The DNA strand, whose 3' end has to grow as the fork moves (outward from the replication origin) is synthesized continuously and this strand is called the leading strand.
- The DNA strand whose 5' end must grow as the fork moves is synthesized discontinuously. This strand is called the lagging strand.
leading strand template

lagging strand template

Replication Fork

Okazaki fragments
Proofreading

• DNA polymerase is a self-correcting enzyme.
• *proofreading*: error-correcting activity.
• Before the enzyme adds a nucleotide to a growing DNA chain, it checks whether the previous nucleotide added has been correctly base-paired to the template strand.
• If the previous nucleotide added has been correctly base-paired to the template strand, the polymerase adds the next nucleotide; if not, the polymerase removes the mispaired nucleotide and tries again.
• Thus DNA polymerase possesses both a 5'-to-3' polymerization activity and a 3'-to-5' nuclease activity.
Primer

- Because of the proofreading activity, DNA polymerase cannot start a completely new DNA strand.
- A different enzyme is needed to begin a new DNA strand.
- This enzyme does not synthesize DNA but synthesizes an RNA strand (about 10 nucleotides long) using the DNA strand as a template.
- This RNA strand is called a primer and the enzyme that synthesizes it is called a primase.
• For the leading strand, an RNA primer is needed only to start replication at a replication origin.

• However, for the lagging strand, a new primer is needed every time a new DNA fragment has to be synthesized.

• To produce a continuous new DNA strand from the many Okazaki fragments synthesized on the lagging strand, three additional enzymes are used.
Enzymes to produce continuous DNA strand from many separate pieces of RNA and DNA

(i) a *nuclease* which disintegrates the RNA primer.

(ii) a *repair polymerase* which replaces the disintegrated RNA with DNA.

(iii) a *DNA ligase* which joins the 5'-phosphate end of one DNA fragment to the 3'-hydroxyl end of the next one.
Proteins used in Replication

1. DNA polymerase
2. Primase
3. *Helicase* which is a protein that uses the energy of ATP hydrolysis to move along the DNA, opening the double helix as it moves.
4. *Single strand binding protein* which clings to the single-stranded DNA exposed by the helicase and prevents it from immediately reforming base pairs.
5. *Sliding clamp* which keeps the DNA polymerase firmly positioned on the DNA template as the polymerase slides along.

On the lagging strand, for obvious reasons, the sliding clamp releases the polymerase from the DNA each time that an Okazaki fragment has been completed.
• Because all of the information specifying a living organism is contained in its DNA, cells go to great lengths to maintain the integrity of the DNA.
• One aspect of maintaining this integrity is that DNA replication must be carried out very accurately.
• The proofreading activity of DNA polymerase ensures that in DNA replication, only about one error is made for every $10^7$ nucleotides copied. (error rate: $10^{-7}$)
• Elaborate DNA repair mechanisms ensure that this accuracy is enhanced even further so that only one error is made for every $10^9$ nucleotides copied. (error rate: $10^{-9}$)
Mutation

- Despite the elaborate repair mechanisms that are in place, sometimes a permanent change in the DNA, called a *mutation*, can occur.
- Such changes over the long run (billions of years) have led to the wide variety of living species that we see on earth today.
- However, in the short run, they can spell disaster for the organism. For instance, *cancer is a disease that can result from the accumulation of DNA mutations.*
Terminology related to Genetics and Heredity

- A *gene* is generally a stretch of DNA that codes for a protein.
- An *allele* is a variant of a gene.
- Although different individuals may possess the same gene, the exact DNA sequences are not identical, i.e., they possess different alleles of the same gene.
- A *wild-type allele* is the kind normally present in the population.
- A *mutant allele* is one that differs from the wild type.
Terminology related to Genetics and Heredity

• In many situations, altering parts of the genome of an organism will not result in a change at the macroscopic observational level. Such a change is called a *genotypic* change or characteristic.

• On the other hand, a characteristic that manifests itself at the observational level is called a *phenotypic* change or characteristic.
Terminology related to Genetics and Heredity

- In humans and many other organisms, each cell except the reproductive ones, contains two copies of each gene.
- Consequently, in such organisms, mutating only one copy of a gene may not result in the manifestation of the associated phenotype.
- Such a mutation would be called a *recessive mutation*.
- On the other hand, a *dominant mutation* is one where mutating one copy of a gene is sufficient for the phenotype to manifest itself.
Terminology related to Genetics and Heredity

- When the two alleles for a particular gene in a genome are of the same type, the genome is said to be *homozygous* for that gene.

- On the other hand, when the two alleles are of different types, the genome is said to be *heterozygous* for that gene.
Diseases as a result of DNA mutations

- *Sickle cell anemia* is an inherited disease that is caused by a single DNA base mutation in the β-globin gene (which is the gene that codes for one of the subunits of hemoglobin).

- Humans have two copies of this gene, one inherited from each parent.

- To have full blown sickle cell anemia, one must have two defective copies of this gene.
Diseases as a result of DNA mutations

• If an individual has one defective copy, then he or she does not show symptoms of the disease although they can pass it on to their offspring (with another person having a defective copy of the same gene).

• Hence, a person with one defective β-globin gene is a carrier of sickle cell anemia.

• Clearly, the mutation involved in sickle cell anemia is a recessive one and the carriers of this disease have a genome which is heterozygous for the β-globin gene.
Diseases as a result of DNA mutations

• Colon cancer is caused by the failure to repair DNA damage in the colonic cells.

• Humans inherit two copies of a DNA repair gene, one from each parent.

• If one defective copy is inherited then it is very likely that at some point in time colon cancer will result when the other copy gets accidentally mutated.
Diseases as a result of DNA mutations

• The incidence of colon and other types of cancers goes up dramatically with age since, unfortunately, the DNA repair mechanisms slow down as one ages.

• Also enough mutations have to accumulate before cancer results and so time is required for the initiation and progression of the disease.
DNA mismatch repair

- Without DNA mismatch repair, mutations would develop much faster and the incidences of cancer would go up catastrophically.
- The DNA mismatch repair system recognizes the errors in DNA replication and corrects them and thereby prevents them from propagating.
- It is not yet known how the DNA repair machine preferentially acts on the newly synthesized DNA strand and leaves the original strand unchanged.
DNA damage

- In addition to replication errors, the DNA in our cells is *continually suffering damage* either spontaneously or due to agents such as sunlight, chemicals, etc.

- Examples are:
  - *depurination* and *deamination*
  - *ultraviolet radiation* in sunlight
Depurination

- In *Depurination*, which occurs spontaneously and in large numbers, a purine base (A or G) is removed resulting in a DNA strand that has missing bases (as shown in the figure).

- When such a strand with missing bases is used as a template for DNA replication, replication errors are bound to occur.
Deamination

- In *Deamination*, an amino group is lost, e.g., a Cytosine is changed to a Uracil (as shown in the figure).

- Since C pairs with G but U pairs with A, if the DNA strand after deamination is replicated then *base substitution* will occur.
The ultraviolet radiation in sunlight can damage DNA by causing the formation of covalent linkages between two adjacent thymine bases; thus forming the thymine dimer.
Common characteristics of DNA Repair Mechanisms

- Nearly all of the repair mechanisms depend on the existence of two copies of the genetic information, one in each strand of the DNA double helix.

- Most DNA damage creates structures that are never encountered in an undamaged DNA strand.

- This is thought to facilitate the identification of the damaged DNA strand.

- The basic pathway for repairing damage to DNA involves three steps.
1. The DNA damage is recognized and removed by one of a variety of different nucleases.

2. A repair DNA polymerase binds to the 3'- hydroxyl end of the cut DNA strand and fills in the gap by making a complementary copy of the information stored in the undamaged strand.

3. When the repair DNA polymerase has filled in the gap, a break remains in the sugar-phosphate backbone of the repaired strand. This nick in the helix is sealed by DNA ligase.
• Failure to repair damaged DNA can result in serious diseases.

• For instance, failure to repair thymine dimers can result in the disease called *xeroderma pigmentosum*.

• People with this inherited disease are prone to severe sunburn and skin cancer.
Conclusion

• The high fidelity with which the integrity of the DNA is maintained means that closely related species have very similar DNA sequences.

• For instance, humans and chimpanzees have DNA sequences that are about 98% identical.

• Even apparently unrelated species such as humans and whales share a lot of similarity in their DNA sequences, strengthening the hypothesis that they must have originated from a common ancestor at some point in evolutionary time.
Transcription and Translation
DNA Language

• The DNA contained in the genome of an organism primarily codes instructions for making proteins.

• Recall that there are only four DNA bases A, G, C, T.

• So an instruction written out in the language of DNA is basically an instruction written out in a language that has a 4-letter alphabet.
• Proteins are made up of amino acids linked together by peptide bonds.

• Since there are 20 possible amino acids that occur in nature, the amino acid sequence for a protein can be thought of as a biological message written out using letters from a 20-letter alphabet.

• A natural question: How does the DNA message coded in the 4-letter alphabet map into the amino acid sequence of the appropriate protein?

• In this chapter, we will provide a detailed answer to this question.
RNA – Ribonucleic Acid

• The DNA does not direct protein synthesis by itself but acts through certain intermediaries.

• When a particular protein is needed by the cell, the nucleotide sequence of the appropriate portion of the DNA molecule is first copied into another type of nucleic acid — RNA (ribonucleic acid).
Transcription and Translation

- RNA copies of short segments of the DNA are used as templates to direct the protein synthesis.
- The process of copying the DNA into the appropriate RNA strands is referred to as transcription.
- The process of producing the protein from the information in the RNA is referred to as translation.
• When a gene is being transcribed, it is said to be *expressed* or *turned ON*.

• The usual flow of genetic information is from *DNA to RNA to protein* - this fundamental principle is referred to as the *central dogma of molecular biology*.

• All cells, from the simplest bacterium to complex organisms such as humans, express their genetic information in this way.
• The use of an RNA intermediate makes it possible to more rapidly produce large amounts of a particular protein.

• This is because many identical RNA copies can be made from the same gene.

• Each RNA copy can be used to simultaneously produce many identical protein molecules.
• By controlling the efficiency of transcription and translation of the different genes, a cell can produce small amounts of some proteins and large amounts of others.

• In addition, a cell can change gene expression in response to the temporally changing needs for a particular protein.
DNA and RNA

RNA differs chemically from DNA in two aspects:

- The nucleotides in RNA are *ribonucleotides*, that is, they contain the sugar *ribose* instead of the sugar deoxyribose that is present in DNA.

Complementary Base Pairs

• Since U pairs with A, the complementary base-pairing properties described in the last chapter for DNA apply also to RNA.

• Therefore, in RNA, the base pairs are:
  • A – U (as compared to A – T in DNA)
  • G – C (similar in DNA)
DNA and RNA Structure

- DNA and RNA differ quite dramatically from each other in their overall structure.
- DNA always occurs in nature as a double-stranded helix while RNA is single-stranded.
- Unlike DNA, a single-stranded RNA chain can fold up into a variety of shapes, just as a polypeptide chain folds up to form the final shape of a protein.
• RNA can not only store information in the sequence of its nucleotides like DNA, but can also serve as a catalyst (in certain situations) or as a part of a larger molecular structure.

• The following figure shows a single-stranded RNA molecule assuming a particular shape resulting from the formation of hydrogen bonds between complementary base pairs:
A single stranded RNA molecule assuming a three dimensional structure.
Transcription

• During transcription, the cell copies the required part of the DNA sequence into a nucleotide sequence of RNA.

• To initiate transcription, a small portion of the DNA double helix is unwound.

• One of the two strands of the DNA double helix then acts as a template for RNA synthesis.
RNA Polymerase

- The RNA synthesis is carried out by an enzyme called *RNA polymerase* which catalyzes the formation of an RNA strand complementary to the DNA template.
- The process is similar to DNA replication by DNA polymerase and uses energy rich nucleoside triphosphates which are hydrolyzed to drive the formation of the RNA strand.
Transcription and DNA Replication

However, transcription and DNA replication have some major differences:

• Since RNA polymerase synthesizes RNA, it catalyzes the linkage of ribonucleotides and not deoxyribonucleotides.
• Unlike DNA polymerase, RNA polymerases do not possess nucleolytic proof reading activity. RNA polymerases can start an RNA chain without the need for a primer.
Transcription and DNA Replication

• Also, transcription is less accurate than DNA replication. (1 error for every $10^4$ nucleotides copied; as compared to 1 error in every $10^7$ nucleotides copied in the case of DNA replication.

• Since the DNA serves as a permanent storage of information while the RNA is only a temporary intermediate, it is only imperative that DNA replication be much more accurate than transcription.
Transcription and DNA Replication

- In transcription, the newly formed RNA strand does not remain hydrogen bonded to the DNA template strand.

- Instead, the DNA helix reforms and displaces the RNA chain.

- Thus, RNA molecules produced by transcription are single stranded.
Transcription and DNA Replication

• RNA molecules tend to be much shorter than DNA molecules since they are copied from only a limited region of DNA.

• Also, the immediate release of the RNA strand from the DNA as it is synthesized means that many RNA copies can be made from the same gene in a relatively short time.
- Not all RNAs produced in a cell are meant to be translated into protein.

- *Messenger RNAs* (*m*RNAs) are the RNAs that direct the synthesis of proteins.

- The *ribosomal RNAs* (*r*RNAs) form the core of the ribosomes, on which mRNAs are translated into proteins.

- The *transfer RNAs* (*t*RNAs) form the adaptors that select amino acids and hold them in place on a ribosome for their incorporation into proteins.
• In eucaryotes, each gene coding for a single protein is typically transcribed into a separate mRNA molecule.

• In bacteria, a single mRNA is often transcribed from several adjacent genes and, therefore, contains information for several different proteins.

• The way in which transcription is initiated differs somewhat between procaryotes and eucaryotes.
• In procaryotes, the RNA polymerase molecules slide rapidly along the DNA.

• The molecule tightly latches on to the DNA when it encounters a region called a *promoter* which contains a sequence of nucleotides indicating the starting point for RNA synthesis.

• Then, it opens up the double helix immediately in front of it to expose the nucleotides on a short stretch of DNA on each strand.
• One of the two exposed DNA strands then acts as a template for synthesis of the complementary RNA.

• The RNA chain elongation continues until the RNA polymerase encounters a second signal in the DNA, the *terminator* (or stop site).

• Here, the polymerase halts and releases both the DNA template and the newly made single-stranded RNA chain.
• A subunit of bacterial polymerase, called *sigma* (σ) *factor*, is primarily responsible for recognizing the promoter sequence on DNA.

• Once the RNA polymerase has latched onto the promoter and has synthesized approximately 10 nucleotides of RNA, the sigma factor is released.

• The RNA polymerase moves forward and continues the transcription process.
After the polymerase reaches the terminator sequence and is released, it reassociates with a free sigma factor and is ready to start transcription again from a promoter region.

The figure is a schematic diagram showing an RNA polymerase molecule transcribing a bacterial gene.

Here, the lower DNA strand serves as the template for RNA synthesis.
• Transcription can only proceed in the 5' to 3' direction for the synthesized strand.

• Moreover, a portion of DNA can be transcribed only if it is preceded by a promoter sequence.
• This ensures that unnecessary transcription does not take place.

• Since procaryotes do not have a nucleus, procaryotic transcription takes place in the cytoplasm.
• As mRNA molecules are transcribed, ribosomes immediately get on to the 5' end of the RNA transcript and protein synthesis starts.
• In eucaryotes, on the other hand, the DNA is enclosed within the nucleus.

• Transcription takes place in the nucleus but protein synthesis takes place in the cytoplasm.

• So, before a eucaryotic mRNA can be translated into protein, it must be transported out of the nucleus and into the cytoplasm.
RNA Processing

• Before the RNA is exported from the nucleus, it goes through several different RNA processing steps.

• Depending on which type of RNA is being produced, these transcripts are processed in different ways before leaving the nucleus.

• In eucaryotes, the RNA produced by transcription, but not yet processed, is often called the primary transcript.
• Primary transcripts destined to become mRNA molecules undergo the following two processing steps:

(1) the addition of a methyl guanine (methyl G) cap at the 5' end which usually occurs just after the RNA polymerase has synthesized the 5' end of the primary transcript and before it has completed transcribing the entire gene.

(2) the 3' end of eucaryotic RNAs are trimmed at a particular sequence of nucleotides and then a series of repeated adenine (A) nucleotides are added at the cut end.

• This processing step is referred to as polyadenylation.
• The trimming and nucleotide addition steps are carried out by two different enzymes.

• The sequence of adenine nucleotides is referred to as a poly A tail and is generally a few hundred nucleotides long.

• The two modifications, capping and polyadenylation, are thought to increase the stability of the mRNA molecule and to aid its export from the nucleus to the cytoplasm.

• They later serve to indicate to the protein synthesis machinery that both ends of the message are present and, therefore, the message is complete.
• Unlike bacterial genes, eucaryotic genes contain both coding and non-coding regions. The *coding regions*, called *exons* are interrupted by *non-coding regions*, called *introns*.

• The exons are usually shorter than the introns, and the coding portion of a gene is often only a small fraction of the total length of the gene.

• For RNAs destined to become mRNAs, the introns are removed after capping and polyadenylation. This is carried out in the nucleus before the mRNA is sent out to the cytoplasm. The process is known as *RNA splicing* and results in a functional mRNA molecule with uninterrupted coding sequences.
• A natural question: how does the cell determine which parts of the primary transcript should be removed?

• Each intron contains a few short nucleotide sequences that act as cues for its removal.

• These sequences are found near each end of the intron and are the same or very similar in all introns.

• Introns are removed from RNA by enzymes that are composed of a complex of protein and RNA.

• These splicing enzymes are called small nuclear ribonucleoprotein particles (snRNPs).
• The presence of numerous introns in eucaryotic DNA makes genetic recombination between exons of different genes more likely.

• Here, we note that eucaryotic genes for new proteins could have evolved quite rapidly by the combination of parts of pre-existing eucaryotic genes.

• Moreover, the primary transcripts of many eucaryotic genes, containing three or more exons, can be spliced in various ways to produce different mRNAs, depending on the cell type in which the gene is being expressed. This is referred to as alternative splicing.
• At each intron, a group of snRNPs assembles on the RNA, cuts out the intron, and rejoins the RNA chain — releasing the excised intron as a lariat.
• The length of time that an mRNA molecule persists in the cell affects the amount of protein produced from it. Different mRNA molecules have different lifetimes.

• Although the genomes of current day procaryotes do not contain any introns, it is thought that their earlier ancestors may have had introns in their genes.

• Procaryotic cells may have evolved and gotten rid of their introns in order to have a smaller genome which would allow faster DNA replication, thus facilitating very rapid reproduction (which is a characteristic that current day procaryotes do possess).
Translation

• In translation, the sequence of nucleotides in an mRNA molecule have to be decoded to produce the appropriate proteins.

• The code used by nature for this purpose is called the *genetic code* : the map from the nucleotide based language of RNA to the amino acid based language of protein.

• The genetic code was completely worked out in the early 1960s.

• Since there are only 4 different RNA bases and 20 amino acids that occur in nature, it is clear that, at a minimum, the code has to be a *triplet code* (confirmed by experiments in the 1960’s).
Codon

- The sequence of nucleotides in the mRNA molecule is read consecutively in groups of 3, each such group specifying either the start of a protein, the end of a protein, or an amino acid.
- Each nucleotide triplet is called a **codon**.

- There are 20 possible amino acids and 64 possible codons.
- The codons **UAA, UAG, UGA** are **stop codons** and signify the end of translation. The codon **AUG** specifies the amino acid **methionine** and also indicates the **start of a protein**.

- All the other codons code for some amino acid.
Universal Genetic Code

• Since there are 20 different amino acids and 61 codons coding for amino acids, there must be multiple codons coding for the same amino acid.

• In other words, the genetic code is degenerate.

• The genetic code is also universal in the sense that all known life on this planet makes use of the same genetic code.

• We have not included the entire genetic code here, since it is not crucial for our purposes.
Reading Frames

- Since each triplet of nucleotides codes for an amino acid, an RNA sequence can be translated in any one of three reading frames, depending on where the decoding process begins.

- The choice of the reading frame is important since advancing the reading frame by even one nucleotide can alter all the amino acids being coded by the RNA sequence leading to a protein that is completely different from the intended one.

- We will see later how a punctuation signal at the beginning of each message sets the correct reading frame.
tRNA : Adaptors

• The triplets in an mRNA molecule specify the sequence of amino acids that have to be linked together to produce the particular protein.

• To actually produce the protein, some kind of a decoder or adaptor molecule is required.

• tRNAs serve as adaptor molecules that can recognize and bind both to the codon on the mRNA and, at another site on their surface, to the amino acid.

• A tRNA molecule consists of a chain of about 80 ribonucleotides.
A typical tRNA molecule
• Parts of this chain assume particular shapes by complementary base pairing as shown in the figure.

• However, two regions of unpaired nucleotides are crucial to the function of tRNA in protein synthesis:

  1. the *anticodon* which is a set of three consecutive nucleotides that pairs with the complementary codon in an mRNA molecule;

  2. a short single stranded region at the 3' end of the molecule, which is the site where the amino acid that matches the codon is attached to the tRNA.
• The degeneracy of the genetic code implies that either there is more than one tRNA for many of the amino acids or that some tRNA molecules can base-pair with more than one codon.
• In fact, both situations occur.
• Some amino acids have more than one tRNA and some tRNAs are constructed so that they require accurate base-pairing only at the first two positions of the codon and can tolerate a mismatch (or wobble) at the third position.
• This wobble base pairing explains why many of the alternative codons for an amino acid differ only in their third nucleotide.
• For instance, the codons CCA, CCC, CCG and CCU all code for the amino acid proline.
• Specific enzymes called *aminoacyl-tRNA synthetases* covalently couple each amino acid to its appropriate set of tRNA molecules.

• Specific nucleotides in both the anticodon and the amino acid accepting arm allow the correct tRNA to be recognized by the synthetase enzyme.

• The synthetase-catalyzed reaction that attaches the amino acid to the 3' end of the tRNA is powered by ATP hydrolysis, and it produces a high-energy bond between the tRNA and the amino acid.

• The energy of this bond is used at a later stage in protein synthesis to covalently link the amino acid to the growing polypeptide chain.
Ribosomes

• The RNA message is decoded on *ribosomes*.

• These are large molecular machines that facilitate accurate and rapid translation of mRNA into protein.

• The ribosome travels along the mRNA chain, capturing complementary tRNA molecules, holding them in position, and bonding together the amino acids that they carry so as to form a protein chain.

• A ribosome is made from more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules called *ribosomal RNAs* (rRNAs).
• A typical living cell contains millions of ribosomes in its cytoplasm.
• In a eucaryote, the ribosomal subunits are made in the nucleus, by the association of newly transcribed rRNAs with ribosomal proteins, which have been transported into the nucleus after their synthesis in the cytoplasm.
• The individual ribosomal units are then exported to the cytoplasm to take part in protein synthesis.
• A ribosome is made up of two units — a small subunit and a large subunit that fit together to form a complete ribosome with a mass of several million daltons.
• The structure of a typical ribosome is shown in the figure.
• The small subunit matches the tRNAs to the codons of the mRNA, while the large subunit catalyzes the formation of the peptide bonds that link the amino acids together into a polypeptide chain.

• The two subunits come together on an mRNA molecule usually near its beginning (5' end) to begin the synthesis of a protein.
Ribosomes

• A ribosome contains four binding sites for RNA molecules: one is for the mRNA and three (called the A-site, the P-site and the E-site) are for tRNAs.

• A tRNA molecule is held tightly at the A- and P-sites only if its anticodon forms base pairs (allowing for wobble) with a complementary codon on the mRNA molecule that is bound to the ribosome.

• The A- and P- sites are sufficiently close together that their two tRNA molecules are forced to form base pairs with adjacent codons on the mRNA molecule.
Having introduced tRNAs and ribosomes, we are now in a position to take a detailed look at the steps involved in protein synthesis.

Let us first focus on examining the mechanism by which the ribosome adds a new amino acid to a growing polypeptide chain.

The three steps involved are shown in the figure.
• In the first step, the polypeptide chain containing the amino acids 1, 2 and 3 has just been synthesized, and the tRNA corresponding to amino acid 3 is still attached to amino acid 3 while being located in the P-site of the ribosome. The codon for amino acid 4 being located on the mRNA in the A-site of the ribosome, the corresponding tRNA with its bound amino acid comes and occupies the A-site of the ribosome.

• In the second step, the small ribosomal subunit (with the bound mRNA) moves relative to the large ribosomal subunit in such a way that the tRNAs 3 and 4 get shifted to the E and P sites respectively on the ribosome. Simultaneously, the amino acid 3 detaches from its tRNA and links up to amino acid 4 (still attached to its tRNA).
• In the third step, the tRNA for amino acid 3 is expelled from the E-site of the ribosome, and the small ribosomal subunit moves relative to the large ribosomal subunit so that they line up as before.

• This sets the stage for the tRNA corresponding to amino acid 5 to get into the A-site of the ribosome so that we are essentially back to step 1, but with a polypeptide chain one unit longer, and the entire cycle can repeat.
Peptidyl Tranferase Enzyme

• The central reaction of protein synthesis, in which an amino acid bound to its tRNA is detached from the tRNA and linked to the growing polypeptide chain, is catalyzed by *peptidyl transferase enzyme activity*, which is part of the ribosome.

• The catalytic part of the ribosome in this case is thought to be not one of the proteins but rather one of the rRNAs in the large ribosomal subunit.

• Let us now examine how protein synthesis starts and how it ends.
• Codons in the mRNA signal where to start and where to stop protein synthesis.

• The translation of an mRNA begins with the codon AUG, and a special tRNA is required to initiate translation.

• The *initiator tRNA* always carries the amino acid methionine (in bacteria, a modified form of methionine — formylmethionine — is used) so that newly made proteins all have methionine as the first amino acid at their amino-terminal end. This methionine is usually removed later by a specific protease.

• The initiator tRNA is distinct from the tRNA that normally carries methionine.
• In eucaryotes, the initiator tRNA (which is coupled to methionine) is first loaded into the small ribosomal subunit along with additional proteins called *initiation factors*.

• Of all the charged tRNAs in the cell, only the charged initiator tRNA is capable of *tightly binding the small ribosomal subunit* as shown in the figure.

• The loaded ribosomal subunit binds to the 5' end of an mRNA molecule, which is recognized in part by the methyl G cap present in eucaryotic mRNA.

• The small ribosomal subunit then moves forward along the mRNA searching for the first AUG.

• When this AUG is encountered, several initiation factors dissociate from the small ribosomal subunit to make way for the large ribosomal subunit to assemble and complete the ribosome.
initiator tRNA

small ribosomal subunit with translation initiation factors bound (not shown)

mRNA BINDING

mRNA

5' AUG 3'

RNA cap

INITIATOR tRNA MOVES ALONG RNA SEARCHING FOR FIRST AUG

5' AUG 3'

INITIATION FACTORS DISSOCIATE

E PA
LARGE RIBOSOMAL SUBUNIT BINDS

AMINOACYL-tRNA BINDS (step 1)

FIRST PEPTIDE BOND FORMS (step 2)

etc.
Once the ribosome assembly is completed, the next tRNA with attached amino acid can bind the codon in the A-site and protein synthesis can continue as described earlier.

The mechanism for selecting a start codon in bacteria is different.

Bacterial mRNA have no 5' methyl G caps to tell the ribosome where to begin searching for the start of translation.

Instead, they contain specific ribosome-binding sequences, up to six nucleotides long, that are located a few nucleotides upstream of the AUGs at which translation is to begin.

Unlike a eucaryotic ribosome, a procaryotic ribosome can readily bind directly to a start codon that lies in the interior of an mRNA, as long as a ribosome-binding site precedes it by several nucleotides.
• Consequently, procaryotic mRNAs are often *polycistronic*, i.e., they encode several different proteins.

• In contrast, a eucaryotic mRNA usually carries the information for a single protein.

• The end of the protein-coding message is signaled by one of several codons (UAA, UAG, or UGA) called *stop codons*.

• These are not recognized by a tRNA and do not specify an amino acid, but instead signal to the ribosome to stop translation.

• Proteins known as *release factors* bind to any stop codon that reaches the A-site on the ribosome.
• This binding alters the activity of the peptidyl transferase in the ribosome, causing it to catalyze the addition of a water molecule instead of an amino acid to the peptidyl-tRNA.

• This reaction frees the carboxyl end of the growing polypeptide, and since only this attachment normally holds the growing polypeptide to the ribosome, the completed protein chain is immediately released into the cytoplasm.

• Even before a ribosome has finished translating an mRNA, other ribosomes can get on the mRNA at more upstream locations and initiate translation.

• Thus, several ribosomes could be simultaneously working on the same mRNA molecule forming what are called polyribosomes.
• The amount of protein in each cell is regulated by carefully controlled protein breakdown.
• Proteins vary enormously in their life span.
• Some proteins may last for months or even years; others last for days, hours or even seconds.

• A natural question that arises is how the cell controls these lifetimes.

• Cells have specialized pathways that enzymatically break proteins down into their constituent amino acids (a process termed *proteolysis*).
• enzymes that degrade proteins are known collectively as proteases.
• One function of proteolytic pathways is to rapidly degrade those proteins whose lifetimes must be short.

• Another is to recognize and eliminate proteins that are damaged or misfolded.

• Most proteins degraded in the cytosol of eucaryotic cells are broken down by large complexes of proteolytic enzymes, called proteasomes.

• The proteins imported into proteasomes for degradation have usually been marked out for destruction by the covalent attachment of a small protein called ubiquitin.
• In all of life that we see around us today, DNA, RNA and proteins have very specialized individual roles.

• It is speculated that life initially originated in an RNA world. Since RNA is capable of both information storage as well as carrying out catalytic functions, this is a very plausible conjecture.

• It is thought that the initial RNA formed during the violent conditions present on the ancient earth.

• Later on, DNA and proteins took over some tasks from the RNA since DNA was better suited for information storage and proteins could perform a whole variety of other functions including serving as catalysts, resulting in the state of affairs we see today.
Chromosomes and Gene Regulation
Chromosomes and Gene Regulation

• In this chapter, we study the organization of DNA inside cells and the various factors that play a role in determining whether and to what extent a particular gene is expressed in a cell.
• The genome of an organism encodes all of the RNA and protein molecules that are needed to make its cells.
• However, not every gene needs to be expressed all the time.
• Even the simplest single-celled bacterium can use its genes selectively, switching genes ON and OFF so that it makes different metabolic enzymes depending on the food sources that are available to it.
• In multicellular organisms, *gene expression* is under even more elaborate control.
• Almost all of the cells of a multicellular organism (with the exception of the reproductive cells) contain the same genome and the apparent differences are caused by the fact that different cell types express different genes.
Organization of DNA into Chromosomes

- In *eucaryotes*, the DNA in the nucleus is distributed among a set of different chromosomes.
- Each chromosome consists of an enormously long DNA molecule that is folded and compacted by certain proteins.
- The complex of DNA and protein is called *chromatin*.
- In addition to the DNA packaging proteins, chromosomes are also associated with proteins involved in DNA replication, DNA repair and gene expression.
- In *procaryotes*, on the other hand, the DNA is organized into *one circular chromosome*. This is carried out by some proteins but not too much is known about the details.
Chromosomes contd.

• Human cells, with the exception of the germ cells (egg and sperm) each contain two copies of each chromosome, one inherited from the mother and one from the father.

• The two copies are called *homologous chromosomes*, the one inherited from the father is called the *paternal homolog* while the one inherited from the mother is called the *maternal homolog*.

• The only *nonhomologous chromosomes* are the *sex chromosomes* in males, where a *Y* chromosome is inherited from the father and an *X* chromosome is inherited from the mother.

• Females, on the other hand, have homologous sex chromosomes since they inherit *X* chromosomes from both parents.
Distinguishing Chromosomes

• The standard way of distinguishing one chromosome from another is to stain them with dyes that bind to certain types of DNA sequences.
• These dyes mainly distinguish DNA that is rich in A-T nucleotide pairs from DNA that is rich in G-C nucleotide pairs.
• This produces a characteristic pattern of bands along each chromosome and such a pattern is called a karyotype.
• Since the pattern for each chromosome is unique, these bands can be used to distinguish one chromosome from another.
• They can also be used to search for chromosomal abnormalities which characterize certain inherited birth defects at the prenatal stage and predispose individuals to certain types of cancers.
Chromosome States

- *Cell cycle* can be broadly divided into *interphase* and *mitosis*.

- During interphase, transcription, translation and DNA replication take place and so the *chromosome is in an extended state*, and cannot be easily distinguished under a light microscope: *interphase chromosomes*.

- During mitosis (or nuclear division), the chromosomes have already replicated and one copy needs to be delivered to each daughter cell. Consequently, during this phase, *the chromosomes are highly compacted* and are visible under a light microscope: *mitotic chromosomes*. 
Specialized DNA Sequences in Chromosomes

• to ensure that chromosomes replicate efficiently and are correctly apportioned between the two daughter cells during cell division.
• *Replication origins* : the DNA replication begins at a replication origin. Most eucaryotic chromosomes contain many replication origins to ensure that the entire chromosome can be replicated rapidly.
• *Centromere* : presence of this sequence allows one copy of each duplicated chromosome to be pulled into each daughter cell when a cell divides. During mitosis, a protein complex called a *kinetochore* forms at the centromere and attaches the duplicated chromosomes to the mitotic spindle, allowing them to be pulled apart towards the opposite ends of the dividing cell.
• *Telomere* : this specialized DNA sequence is found at each of the two ends of a chromosome. Telomeres contain repeated nucleotide sequences that enable the ends of chromosomes to be replicated.
Nucleosomes

- An individual nucleosome core particle consists of a complex of 8 histone proteins — two molecules each of histones H2A, H2B, H3 and H4 — and a double-stranded DNA around 146 nucleotide pairs. The histone octamer forms a protein core around which the double-stranded DNA helix winds.
Nucleosome contd.

- The term *nucleosome* refers to a *nucleosome core particle* plus an *adjacent DNA linker* (about 50 nucleotide base pairs long).
- The formation of nucleosomes converts a DNA molecule into a chromatin thread approximately *one-third of its initial length*.
- Histones are small proteins with a high proportion of *positively charged amino acids*.
- These positive charges allow the histones to bind tightly to the negatively charged sugar phosphate backbone of DNA, regardless of the precise nucleotide sequence.
- There are additional levels of chromatin packing, one of which is facilitated by a fifth histone H1 which is thought to pull the nucleosomes closer into a regular repeating array.
Heterochromatin

• The chromatin in an interphase chromosome is not in the same packing state throughout the chromosome.

• *Heterochromatin:* most highly condensed form of interphase chromatin.

• Heterochromatin typically makes up about 10% of an interphase chromosome, and in mammalian chromosomes, it is typically concentrated around the centromere and telomere regions.

• Heterochromatin is *transcriptionally inactive.*
Heterochromatin in the interphase X chromosomes of female mammals

- Recall that female mammals have two X chromosomes.
- Because a double dose of the corresponding protein in females could possibly be disastrous, female mammals permanently inactivate one of the two X chromosomes by condensing it into heterochromatin early in embryonic development.
- Thereafter, in all of the many progeny of the cell, the condensed and inactive state of that X chromosome is propagated.
- The rest of the interphase chromatin, which is in a variety of extended states, is called euchromatin.
- In a typical differentiated eucaryotic cell, some 10% of the euchromatin is in a state in which it is either actively being transcribed or is easily available for transcription; this is known as active chromatin and is the least condensed form of chromatin in the interphase chromosome.
Position effects on gene expression

• If a particular gene is moved from a region of heterochromatin to a region of active chromatin or vice versa, its transcriptional activity is likely to be altered.
Gene Regulation

• Central question in the study of gene regulation: how a cell specifies which of its many thousands of genes should be expressed as proteins.
• This question is especially important for multicellular organisms because as the organism develops, many different cell types are created starting essentially from the same precursor cells.
• This *differentiation* arises because cells make and accumulate different sets of RNA and protein molecules.
Experiment to demonstrate that different cell types from the same organism contain the same genome.
Housekeeping Proteins

- Many of the proteins in cells are found in all cell types.
- For instance, the proteins needed to make DNA polymerase, RNA polymerase, ribosomes, etc. have generic functions and are found in all cell types.
- These proteins are called *housekeeping proteins* and the genes that encode them are called *housekeeping genes*.
Specialized Proteins

- Ex: hemoglobin is made in reticulocytes, the cells that develop into red blood cells, but it cannot be detected in any other cell type.
- Insulin is made exclusively by the $\beta$ cells in the pancreas.
Control of Protein production

A cell can control the proteins it makes by:

(1) controlling when and how often a given gene is transcribed,
(2) controlling how the primary transcript is spliced or otherwise processed
(3) selecting which mRNAs are translated by ribosomes or
(4) selectively activating or inactiviting proteins after they have been made.
Transcription Control

- Transcription is controlled by proteins binding to regulatory DNA sequences.
- The promoter region of a gene attracts the enzyme RNA polymerase and correctly orients it to begin its task of making an RNA copy of the gene.
- The promoters include the *initiation site* and a sequence of about 50 nucleotides that extends “upstream” from the initiation site. This region contains sites that are required for the RNA polymerase to bind to the promoter.
- *Regulatory DNA sequences*: needed in order to switch the gene ON or OFF.
- Some regulatory DNA sequences are as short as 10 nucleotide pairs and act as simple gene switches that respond to a single signal. This occurs primarily in bacteria.
- Other regulatory sequences can be very long (as many as 10,000 base pairs) and act as molecular microprocessors, taking in several inputs to determine the transcription rate. This occurs in eucaryotic cells.
• The regulatory DNA sequences do not act by themselves. They must be bound by gene regulatory proteins that recognize them. Gene regulatory proteins can either suppress transcription or enhance it.

• *Repressors* turn genes OFF while *activators* turn them ON.

Ex:

![Diagram of operon](image)

*Operon*: several genes that are transcribed as a single mRNA.

Inside the promoter region, there is also a DNA regulatory sequence. When a repressor molecule binds to this sequence, RNA polymerase cannot get on to the promoter and so the transcription of this operon stops. However, the repressor molecule can bind this sequence only if it also binds several molecules of tryptophan.
Constitutive Gene Expression

• So when the level of tryptophan in the cell is high, the transcription of the operon stops.
• On the other hand, when the level of tryptophan in the cell falls, the repressor is no longer able to bind the regulatory sequence and so RNA polymerase can start transcription of the operon so that more tryptophan can be produced.
• The tryptophan repressor protein is *always present in the cell* though at a low level.
• Such unregulated gene expression is known as *constitutive gene expression* as opposed to *induced gene expression*, which occurs in response to some stimulus.
Role of Activator Proteins
RNA Polymerase

- The initiation of gene transcription in eucaryotic cells is a complex process.
- We next discuss differences between initiation of transcription in procaryotes and eucaryotes.
- Procaryotic cells contain a single type of RNA polymerase while eucaryotic cells have three different RNA polymerases called RNA polymerase I, RNA polymerase II and RNA polymerase III.
- RNA polymerases I and III transcribe the genes encoding transfer RNA, ribosomal RNA, and small RNAs that play a structural role in the cell.
- RNA polymerase II, on the other hand, transcribes the vast majority of eucaryotic genes, including all those that encode proteins.
Transcription Factors

- Eucaryotic RNA polymerase II cannot initiate transcription on its own.
- To initiate transcription, eucaryotic RNA polymerase requires the help of a large set of proteins called transcription factors.
- These are thought to position the RNA polymerase correctly at the promoter, to aid in pulling apart the two strands of DNA to allow transcription to begin, and to release RNA polymerase from the promoter once transcription begins. The assembly of transcription factors and their role in initiating transcription at the appropriate location are shown in the next slide.
Role of Transcription Factors
• Once the initiation complex has assembled, the general transcription factor TFIIH (which contains a protein kinase enzyme as one of its subunits) phosphorylates the RNA polymerase.

• This phosphorylation is thought to help the polymerase disengage from the cluster of transcription factors, allowing transcription to begin.
In eucaryotic cells, the gene regulatory proteins can influence the initiation of transcription even when they are bound to DNA thousands of nucleotide pairs away from the promoter.
• Finally, eucaryotic transcription initiation must take into account the packing of DNA into nucleosomes and more compact forms of chromatin structure.
• It is believed that the nucleosomes prevent the general transcription factors of RNA polymerase from assembling on the DNA and can, therefore, hinder the initiation of eucaryotic transcription.
Combinatorial Control

- Most bacterial genes are controlled by a single activator or repressor protein.
- On the other hand, most eucaryotic gene regulatory proteins work as part of a “committee” of regulatory proteins, all of which are necessary to express the gene in the right cell, in response to the right conditions, at the right time, and at the required level.
- This is referred to as *combinatorial control* and is similar to Boolean logic familiar to most engineers.
- Indeed, using Boolean functions to model relationships between expressed genes is a theme that we will encounter in a later part of the course.
• In eucaryotic cells, combinatorial control of gene expression can create different cell types.
• For instance, fibroblasts can be converted into myoblasts by expressing the gene Myosin D (MyoD).
• It appears that fibroblasts, which are derived from the same broad class of embryonic cells as muscle cells, have already accumulated all the other necessary gene regulatory proteins required for the combinatorial control of the muscle-specific genes and addition of MyoD completes the unique combination that directs the cells to become muscle.
Consider the hypothetical example. Here there are three genes of interest, each of which could be in one of two states, ON denoted by 1 and OFF denoted by 0. Clearly, there are $8 (= 2^3)$ possible gene expression patterns in this case and, by selectively expressing one or more genes in the progeny of this cell, we can end up with up to eight different possible cell types after three rounds of cell division.
• Some highly specialized cells never divide again once they have differentiated, eg. nerve cells or neurons
• But there are many other differentiated cells, such as liver cells, that will divide many times in the life of an individual.
• All these cell types give rise only to cells like themselves when they divide.
• This means that the changes in gene expression that give rise to a differentiated cell must be stored in memory and passed on to its progeny through all the subsequent cell divisions.
• Cells have several ways of accomplishing this.
• One possible mechanism involves a positive feedback loop where a key gene regulatory protein activates transcription of its own gene in addition to that of other cell type specific genes. The transcription of this gene regulatory protein ensures that this particular protein is turned ON in the new daughter cells so that they inherit the same cell type as that of their parent.
• In another mechanism, the condensed chromatin structure is faithfully propagated from parent to daughter cell even though DNA replication intervenes. An instance of this is the inactivation of one X chromosome in female mammals, something that we have already discussed.
Concluding Example

• Experiment to demonstrate that alteration of the expression status of a single gene can trigger the development of an entire organ and, that too, at an abnormal location.

• Now, a gene called ‘Ey’ in flies and ‘Pax-6’ in vertebrates is crucial for eye development.

• In this experiment involving fruit flies the gene ‘Ey’ is expressed early in development using artificial means in cells that normally go on to form legs. The result is that in the corresponding fruit flies, eyes develop in the middle of the legs.

• Thus this experiment clearly shows that the action of just one gene regulatory protein can turn on a cascade of gene regulatory proteins whose actions can result in the formation of an organized group of many different types of cells.
Genetic Variation
Genetic Variation

• In this chapter, we discuss mechanisms by which the DNA in organisms can undergo changes over a long period of time.
• Although cells go to great lengths to maintain the integrity of the DNA, permanent changes in the DNA, called mutations, do accumulate over time.
• The vast diversity of life that we see around us today has arisen through changes in the DNA that have accumulated over evolutionary time.
• This genetic variation or diversity refers to the genomic differences between different species as well as between members of the same species.
• We note that even members of the same species have genomes that are quite different from each other.
• In fact, no two human beings have identical genomes unless they happen to be identical twins.
**Genetic Variation**

- It is generally believed that the conditions on this earth have undergone dramatic changes over billions of years.
- Thus, in order for life to propagate, it is essential that the “survivors” be able to **adapt to changing conditions**.
- It is believed that genetic variation, though not always beneficial, is responsible for conferring **survivability** in a changing environment.

- There are three main mechanisms by which genetic variation can arise:
  (i) rare **mistakes in DNA replication** and repair;
  (ii) **DNA recombination** and the activity of viruses and **mobile genetic elements** that can move into and out of the DNA;
  (iii) the **reassortment of the gene pool** of the species into new combinations during sexual reproduction.
Genetic Variation in Bacteria

• Among procaryotes, E. coli is a model organism for genetic studies.
• E. coli is said to have a haploid genome since it has only one copy of each gene in its genome.
• Consequently, the effect of a mutation at the gene level will manifest itself in the phenotype.
• In contrast, organisms such as ourselves are called diploid organisms, since our genomes contain two copies of each gene.
• Consequently, a DNA mutation in one chromosome will not necessarily result in an observable phenotype.
• In this context, recall how an individual with only one defective β-globin gene does not exhibit the full blown symptoms of sickle-cell anemia.
Genetic Variation in Bacteria

- E. coli, like other procaryotes, reproduces by fission-type cell division.
- The DNA replicates and the two identical strands of DNA move to the two ends of the growing bacterium.
- The bacterium then splits in two, producing two daughter cells, each containing a genome identical to that of the parent cell.
- In the presence of sufficient nutrients, a population of E. coli doubles in number every 20 to 25 minutes.
- Thus, in less than a day, a single E. coli can produce more than 5 billion descendants.
- Every time that a cell divides, the DNA has to be replicated which means that there is a possibility of some replication errors.
- The rapid rate of division of E. coli means that very large populations of E. coli cells in which DNA mutations have occurred can be produced quite rapidly.
Genetic Variation in Bacteria

- E. coli bacteria can be quite easily used to demonstrate how DNA mutations confer survivability in a changing environment.
- For instance, consider the antibiotic rifampicin. Like many other antibiotics, this drug binds tightly to RNA polymerase inside the E. coli cell and prevents it from transcribing DNA to RNA.
- This inhibition eventually blocks the synthesis of new proteins, and the E. coli bacterium dies.
- However, in a large population of E. coli, say $10^9$ cells, there will be some cells that are rifampicin resistant.
- If such a population is treated with rifampicin, most of the cells will die.
- However, the rifampicin-resistant cells will survive and ultimately take over the population.
- The rifampicin resistance here comes from mutations in the DNA which allow RNA polymerase to transcribe even in the presence of rifampicin.
Bacteria and Virus

• The principle that we have encountered here, namely the ability of antibiotics to block bacterial transcription, is the basis for current day treatment of bacterial infections using antibiotics.
• The reason that this treatment is ineffective against viral infections is that a virus does not have its own transcription and translation machinery which the antibiotic could have targeted.
• Instead, as we shall see, a virus relies on hijacking the replication machinery of the host cell in order to propagate itself.
• Bacterial cells can acquire genes from other bacteria.
• If we mix a laboratory strain of E. coli that lacks one of the enzymes for making an essential amino acid with another strain that lacks one of the enzymes for making another essential amino acid, and if the mixture is allowed to grow together for a few hours and then transferred to a medium that lacks both amino acids, many rapidly growing bacteria can be found in the new medium. The genome of the new bacterial strain is composed of normal genes for the synthesis of both the essential amino acids.
Bacterial Mating

• Genes can be transferred from one bacterium to another by a process called **bacterial mating**.
• The ability to initiate mating and gene transfer, seen in some bacteria, is conferred by genes contained in bacterial **plasmids**.
• Plasmids are small, circular double stranded DNA molecules that are separate from the larger bacterial chromosome.
• A plasmid that commonly initiates mating in E. coli is the F plasmid, or **fertility plasmid**.
• When a bacterium carrying the F plasmid (the donor) encounters a bacterium lacking the plasmid (the recipient), a cytoplasmic bridge is formed between the two cells.
• The F plasmid DNA is replicated and transferred from the donor through the bridge to the recipient.
• Finally, the bridge breaks down and the two bacteria, both of which now contain the F plasmid, can act as donors in subsequent encounters with recipient bacteria.
Gene transfer by bacterial mating
Bacterial Transformation

- The F plasmid is necessary for mating because it carries genes that encode some of the proteins required to make the cytoplasmic bridge and to transfer the DNA.
- Bacterial mating via the F plasmid does not generate much genetic variation since the F plasmid that is transferred contains only a small number of genes.
- However, occasionally the F plasmid can get integrated into the bacterial chromosome with the result that when it now initiates mating and gene transfer, it can now take parts of the bacterial chromosome with it.
- Bacteria can take up DNA from their surroundings.
- For instance, DNA can be taken up from other dead bacteria in the surroundings. This process is called transformation since it can transform one strain of bacteria into another (recall the experiment from Chapter 5 where one strain of bacteria got transformed into another).
Homologous Recombination

- The most important route by which DNA becomes incorporated into a bacterial genome is called *homologous recombination*.
- Homologous recombination can take place between two DNA molecules of *similar nucleotide sequence*.
- Two double-stranded DNA molecules that have regions of very similar (homologous) DNA sequence align so that their homologous sequences are in register.
- This is shown in the figure. The same figure also shows how they can then cross over. The net result is homologous recombination.

- Cells utilize specialized proteins to facilitate homologous recombination. Such recombination enzymes are well characterized in bacteria, but are only beginning to be understood in eucaryotic cells.
Homologous recombination
• Two such exchanges can neatly replace a long stretch of bacterial DNA with a homologous but not identical DNA fragment from another source.
• The same mechanism can result in the F plasmid DNA getting integrated into the bacterial chromosome.
Gene Transfer by Bacteriophages

- Bacterial viruses, or *bacteriophages*, are viruses that invade bacterial cells.
- A virus is usually composed of DNA enclosed inside a protein coat.
- A virus enters a bacterial cell and uses the cell’s DNA replication, transcription and translation machinery to produce:
  (i) more copies of its own DNA
  (ii) its coat protein.

- The replicated DNA is packaged into additional protein coats to produce progeny viruses which can leave the current bacterial cell and invade other cells.
- Viral reproduction is generally lethal for the infected cell and the cell bursts open (*lyses*) as a result of the infection.
Gene Transfer by Bacteriophages

- When a virus invades a bacterial cell one of two things can happen: (i) the cell lyses and a large number of copies of the virus are released (ii) the viral genome gets integrated into the bacterial chromosome and at each subsequent bacterial cell division, the viral genome is also replicated along with the bacterial genome and passed on.
- An environmental insult such as exposure to ultraviolet light can induce the viral genome to leave the host chromosome and begin a lytic phase of viral replication.
- Most of the time, route (i) above is followed but on rare occasions, route (ii) is also followed.
- The viral genome usually integrates with the bacterial genome by site specific recombination that is carried out by an enzyme called integrase.
Transduction

- On leaving a host chromosome, the viral DNA will occasionally remove itself inaccurately and bring along a neighboring piece of host DNA in place of part of its own DNA.
- This host DNA will be packaged into a virus particle along with the viral DNA.
- So when the new virus infects a new host, it introduces, along with the viral DNA, DNA derived from the previous host.
- This bacterial DNA can become part of the new host chromosome in at least a couple of ways:
  (i) the incoming virus integrates into the new host chromosome
  (ii) if the incoming virus does not destroy the host, the passenger bacterial DNA can become a permanent part of the host’s genome by homologous recombination.
- This process is called *transduction*. 
Transposons

• Many bacterial and eucaryotic genomes contain stretches of DNA called transposable elements (or transposons), which can move from place to place within the chromosome by a process called transposition.

• Transposons move within the DNA of their host by means of special recombination enzymes called transposases, encoded by the transposable element, to create great genetic diversity.

• In non-replicative transposition, the stretch of DNA from the donor is removed and incorporated into the recipient while in replicative transposition, the donor DNA remains intact while a copy is made for incorporation into the recipient DNA.

• We note that most alterations in the genome are harmful to the individual bacterium and these are quickly eliminated from the population. However, alterations also confer survivability in a changing environment, as was demonstrated by the example involving the antibiotic rifampicin.
Nonreplicative and replicative transposition brought about by transposons
Sources of Genetic Change in Eucaryotic Genomes

• Unlike bacterial DNA, eucaryotic DNA has both coding and non-coding regions.
• In procaryotes, the rate of cell division is very high and hence there is a strong selective pressure to minimize the amount of superfluous DNA in the genome.
• This is probably the reason that procaryotic genomes have gotten rid of most of the spacer DNA.
• Eucaryotic genomes are also characterized by a large amount of gene duplication that has occurred over evolutionary time.
• As a result, there can be several genes belonging to the same family.
• The most well-documented example is the β-globin gene family and the human genome has a total of 5 β-globin genes.
• These genes encode the β subunits of the various hemoglobins produced at different times during embryonic, fetal and adult life.
Gene Duplication

- **Gene duplication** is thought to occur from a rare recombination event between two homologous chromosomes.
- Consider the case of the β-globin gene duplication. Instead of aligning properly for a crossover, the two homologous chromosomes align in an improper fashion as shown in the figure.
- After the crossover, the long chromosome has two copies of the globin gene, while the short chromosome lacks the original globin gene.
- Consequently, the individuals that inherit the short chromosome would be expected to be eliminated from the population while the individuals that inherit the long chromosome would be expected to have two β-globin genes instead of one.
- Genes encoding new proteins can also be created by the recombination of exons.
- The general scheme, which is shown in the figure, is the same as before except that an exon within a gene, rather than the entire gene is duplicated.
Globin gene

Homologous chromosomes (aligned properly for a crossover)

Short repeated homologous DNA sequence

misalignment

"Unequal" crossing-over

Globin gene

Long chromosome

Short chromosome
Duplication of exons

Diagram showing the process of duplication of exons through unequal crossing-over. The diagram illustrates how exons A and B are duplicated, resulting in a long chromosome on one side and a short chromosome on the other.
Exon shuffling

• Without introns there would be very few sites on the original gene at which a recombinational exchange between homologous chromosomes could duplicate the domain without damaging it.
• Therefore, introns greatly increase the probability that DNA duplications will give rise to functional genes encoding functional proteins.
• Moreover, the presence of introns greatly increases the probability that a chance recombination event can generate a functional hybrid gene by joining together two initially separate exons coding for quite different protein domains.
• This is referred to as exon shuffling.
Transposable Elements and Viruses

- About 10% of the human genome consists of two families of *transposable* sequences.
- Transposable DNA elements move from place to place by the mechanisms discussed earlier for procaryotic transposons.
- However for *eucaryotes*, there are also *retrotransposons* for which an RNA copy is first made using RNA polymerase following which DNA copies are made using *reverse transcriptase*.
- It is these DNA copies that are then inserted into the target.
- Transposition by reverse transcription is shown in the figure.
- Examples of human retrotransposons are the so-called *L1 transposable element* and the *Alu sequence*.
- The evolution of genomes has been greatly accelerated by transposable elements.
- The insertion of a transposable element in a regulatory region for a gene will often, by disrupting or adding short regulatory sequences, have a dramatic effect on gene expression.
Transposition by reverse transcription
Another source of genetic variation in eucaryotes is the activity of viruses. Like bacteriophages, viruses that infect eucaryotic cells are fully mobile elements that can move into or out of cells. Viral genomes can be made of DNA or RNA and can be single-stranded or double-stranded. An important class of viruses is retroviruses, which reverse the normal flow of genetic information. These viruses have a genome that is made of RNA, and have a protein coat that encapsulates the genome and the enzyme reverse transcriptase. The schematic diagram for a retrovirus hijacking a host cell is shown in the figure.
Retroviruses

• Here, the enzyme reverse transcriptase is first used to make a (single-stranded) complementary DNA copy of the viral genome.
• The enzyme DNA polymerase (present in the host cell) is then used to create a double stranded DNA copy of the viral genome, which is then integrated into the host DNA.
• Transcription and translation from the integrated genome produces copies of the viral RNA, reverse transcriptase and the coat protein and all of these can be packaged together to produce additional retroviruses.

• Retroviruses that have picked up host genes can make cells cancerous.
• Very few human cancers are caused by retrovirus infection, but these viruses are a prominent cause of cancers in some animals.
Cancer causing retrovirus

• The most well known example of a cancer-causing virus is the Rous Sarcoma virus.
• This virus can pick up a gene from its chicken host.
• This gene called src is unnecessary baggage from the point of view of the virus, but it has profound consequences for cells that are infected by the virus.
• The normal src gene in the chicken genome encodes a protein kinase that is involved in the control of cell division and is called a proto-oncogene.
• However, the src gene carried by the virus is not quite identical to the normal cellular gene and the difference gives rise to its ability to cause a cancer.
• This mutated src gene is called an oncogene and it causes the infected cell to divide uncontrollably.
Sexual Reproduction and the Reassortment of Genes

- Bacteria reproduce asexually and this gives rise to offspring that are identical to the parent.
- Sexual reproduction, on the other hand, involves the mixing of genomes from two individuals to produce offspring that are genetically distinct from one another and from both their parents.
- Sexual reproduction occurs in diploid organisms, in which each cell contains two sets of chromosomes, one inherited from each parent.
- The actual cells that carry out sexual reproduction in diploid organisms are called the germ cells or gametes.
- These gametes, which are haploid cells, are of two types — the egg and the sperm in animals.
- The haploid germ cells are generated from their diploid precursors by a special type of cell division called meiosis.
Meiosis

• Our main focus is on showing the staggering amount of genetic variation that results from meiosis.
• A schematic diagram showing how meiosis produces haploid gametes from a diploid precursor cell is shown in the figure.
• For clarity of presentation, we focus attention on only one pair of homologous chromosomes.
• During meiosis, the chromosomes first duplicate and then the homologous paternal and maternal chromosomes line up next to each other.
• Next, crossovers take place by homologous recombination and two rounds of cell splitting occur to produce four different haploid gametes.
Diploid germ cell precursor

DNA replication

Pairing of homologous duplicated chromosome

Chromosome crossing-over (recombination)
Sexual Reproduction and the Reassortment of Genes

- Clearly, in the absence of a crossover, each pair of homologous chromosomes will give rise to two different types of gametes depending on whether it contains the paternal homolog or the maternal homolog.
- With 23 chromosome pairs in the human genome, each individual can produce $2^{23}$ possible gametes.
- The actual number of possible gametes is much higher because of crossovers. (On an average, there is at least one crossover on each chromosome.)
- To develop a new organism, two such gametes of the opposite kind (egg and sperm) have to fuse to form the single diploid cell that will undergo repeated mitosis to develop into the new organism.
- Thus, the number of possible ways in which the genetic material gets reshuffled during sexual reproduction is staggering.
- This is the reason that it is not at all surprising that no two individuals (except identical twins) have the same genome.
DNA Technology
In this chapter, we will look at some of the modern technologies that enable us to create genetic variation in the laboratory setting.

Humans unknowingly experimenting with DNA: Ex: the cross breeding between plants and that between animals are really instances of genetic reshuffling.

In the early 1970s, it became possible for the first time to isolate a given piece of DNA out of the many millions of nucleotide pairs in a typical chromosome.

This in turn made it possible to create new DNA molecules in the test tube and introduce them back into living organisms.

This is referred to as “Recombinant DNA Technology” or “Genetic Engineering.”
Techniques for Analyzing DNA Molecules

- The development of current day DNA technology has been made possible by certain technological advances.
- The first among these is the discovery of *restriction nuclease*, which are enzymes that cleave DNA at certain specific nucleotide sequences.
- The figure presents a few examples of restriction nucleases cleaving DNA at specific locations.
More Examples of Specific Cleaving of DNA by Restriction Nucleases

Restriction nucleases cleave DNA in a very predictable and replicable manner, unlike mechanical shear, which fragments DNA, but there is no way to predict where the fragmentation will occur. Furthermore, the fragmentation locations due to mechanical shear will change from one run of the experiment to another.
Gel Electrophoresis

Double stranded DNA

Cut with restriction nuclease I

Cut with restriction nuclease II

Small DNA pieces

Load onto slab of agarose gel

Direction of migration
Gel Electrophoresis

- In gel electrophoresis, fragments of DNA are loaded on to a slab of agarose gel and subjected to an electric field as shown in the figure.
- Since DNA is negatively charged, the fragments will migrate towards the bottom and the distance that each fragment travels will be inversely proportional to its size.
- If one incorporates the radioactive $^{32}$P isotope of phosphorous in DNA before electrophoresis, then the DNA fragment positions after electrophoresis can be detected by using the technique of autoradiography.
- By comparing the sizes of the DNA fragments produced from a particular region of DNA after treatment with different combinations of restriction nucleases, a physical map of the region can be constructed showing the location of each cutting site.
- Such a map is known as a restriction map.
Sequencing DNA using the \textit{dideoxymethod}

- A restriction map, however, does not provide the complete sequence information for the entire DNA sequence.
- To determine the complete DNA sequence of a given DNA strand, one can run a \textit{four-lane gel electrophoresis} following four separate DNA synthesis reactions as shown in the figure.
- The key idea behind this method is that if during DNA synthesis, a \textit{dideoxyribonucleoside triphosphate} is incorporated into the growing DNA strand instead of a deoxyribonucleoside triphosphate, then the 3' end of the DNA chain is \textit{chemically blocked} and, therefore, the chain cannot elongate any further.
- Thus, if we add dATP, dCTP, dGTP, dTTP in excess and a small amount of ddATP, then some DNA strands complementary to the given strand and terminating at the various A locations will be produced.
- Based on this, we outline the following procedure to determine the complete \textit{nucleotide sequence} of a given DNA strand.
Sequencing DNA using the dideoxymethod

- Take the double-stranded DNA strand. Pick one of the two strands as the DNA to be sequenced and use its complementary strand as the template.
- Four different chain-terminating dideoxyribonucleoside triphosphates (ddATP, dd-CTP, ddGTP, ddTTP) are used in four separate DNA synthesis reactions on copies of the same single-stranded DNA template.
- Each reaction, which is primed using an oligonucleotide (synthetic) primer, produces a set of DNA copies that terminate at different points in the sequence.
- The products of these four reactions are separated by electrophoresis in four parallel lanes of an agarose gel.
- The positions of the DNA fragments in each of the lanes can be used to piece together the sequence of the DNA strand that we were originally interested in.
Sequencing DNA using the dideoxymethod

### Deoxyribonucleoside Triphosphates

- **Deoxyribonucleoside triphosphate (dNTPs)**
  - Used in DNA synthesis to add nucleotides to the growing DNA strand.
  - **Example:** dATP, dCTP, dGTP, and dTTP.

- **Dideoxyribonucleoside triphosphate (ddNTPs)**
  - Used as a terminator in DNA sequencing to stop further elongation of the DNA strand.
  - **Example:** ddATP.

### DNA Sequencing Process

1. **Oligonucleotide Primer**
   - A short sequence of nucleotides used as a starting point for DNA synthesis.

2. **DNA Polymerase**
   - Enzyme that catalyzes the addition of nucleotides to the growing DNA strand.

3. **DNA Molecule to be Sequenced**
   - The single-stranded DNA molecule to be analyzed for its sequence.

4. **Termination of Chain Growth**
   - The incorporation of ddNTPs prevents further elongation of the DNA strand, allowing for the determination of the sequence at that point.

- **Rare Incorporation**
  - The incorporation of ddNTPs is rare, occurring with a frequency of around 1 in 10,000 nucleotides.

---

This diagram illustrates the principles behind the dideoxy method developed by Sanger, which is a cornerstone of modern DNA sequencing technology.
Sequencing DNA using the dideoxymethod
Nucleic Acid Hybridization and Associated Techniques

• By subjecting double-stranded DNA to high temperatures or high pH, the two strands of DNA can be made to separate. This is referred to as *denaturation* of the DNA.
• By slowly cooling the DNA or lowering the pH, the DNA can be made to *renature*, i.e. reform the double helix.
• The nucleic acid hybridization based techniques make use of the complementary base pairing properties of nucleotides: A pairs with T and G pairs with C.
• One application of DNA hybridization is in the *pre-natal diagnosis* of genetic diseases, for instance, sickle cell anemia.
• Examining a single gene in the human genome requires searching through a total genome of over three billion nucleotides, which is an incredibly daunting task.
Denaturation and Renaturation of DNA molecules

DNA double helix

High temp/high pH Denatures DNA

Slow cooling / low pH Renatures DNA

Denatured DNA
Nucleic Acid Hybridization and Associated Techniques

However, the tremendous specificity of DNA hybridization makes it possible to do so in a fairly tractable fashion.

For instance, for sickle cell anemia, the exact nucleotide change in the mutant gene is known.

For prenatal diagnosis of sickle cell anemia, DNA is extracted from fetal cells.

Two DNA probes are used to test fetal DNA - one corresponding to the normal gene sequence in the region of the mutation and the other corresponding to the mutant gene sequence.

A DNA probe is a short single-stranded DNA, an oligonucleotide that is used in hybridization reactions to detect nucleic acid molecules containing a complementary sequence.
Nucleic Acid Hybridization and Associated Techniques

- DNA samples from the fetus are first treated with restriction nucleases and all the resulting DNA fragments are electrophoresed through a gel.
- The gel is then treated with a DNA probe that detects only the restriction fragment that carries the β-globin gene.
- Using the two DNA probes, it is possible to distinguish whether the fetus contains one, two or no defective β-globin genes.
- The laboratory procedure used to visualize the hybridization is known as **Southern blotting** and it involves the following six steps:
  - Cleave double-stranded DNA using restriction nucleases.
  - Electrophorese the fragments to separate the fragments by length.
  - A sheet of nitrocellulose paper is laid over the gel and separated DNA fragments are transferred to the sheet by blotting.
Nucleic Acid Hybridization and Associated Techniques

As this occurs, the DNA is denatured and the single-stranded DNA fragments adhere firmly to the surface of the nitrocellulose sheet.

– The nitrocellulose sheet is carefully peeled off from the gel.
– The sheet containing the bound single-stranded DNA fragments is placed in a sealed plastic bag together with buffer containing a radioactively labeled DNA probe specific for the required DNA sequence. This gives the probe a chance to hybridize with its complement, if the latter is present on the sheet.
– The sheet is removed from the bag and washed thoroughly so that only probe molecules that have hybridized to the DNA on the paper remain attached. After autoradiography, the DNA that has hybridized to the labeled probe will show up as bands on the autoradiograph.
**DNA cloning**

- **DNA cloning** has two possible meanings in biology:
  - The act of making many *identical copies* of a DNA molecule.
  - The isolation of a particular stretch of DNA from the rest of the cell’s DNA. This isolation is facilitated by making many identical copies of the DNA of interest, thereby amplifying it relative to the rest of the DNA.

- Just as restriction nucleases can be used to break DNA into smaller fragments, an enzyme called **DNA ligase**, which we encountered earlier in the context of DNA replication and DNA repair, can be used to join DNA fragments together to produce a recombinant DNA molecule.

- Since DNA has the same chemical structure in all organisms, the use of this enzyme allows DNAs from any source to be joined together.

- Restriction nucleases and DNA ligase play an important role in cloning DNA. This can be carried out using a **bacterial plasmid**.
DNA cloning

- The purified plasmid DNA is exposed to a restriction nuclease that cleaves it in just one place, and the DNA fragment to be cloned is covalently inserted into it using DNA ligase.
- This recombinant DNA molecule is then introduced into a bacterium (usually E. coli) by transformation and the bacterium is allowed to grow in the presence of nutrients, where it doubles in number about every 20 minutes.
- After just a day, more than a billion copies of the plasmid would have been produced.
- The bacteria are then lysed, and the much smaller plasmid DNA is purified away from the rest of the cell contents, including the large bacterial chromosome.
- The DNA fragment can be recovered by cutting it cleanly out of the plasmid DNA using the appropriate restriction nuclease and separating it from the plasmid DNA by gel electrophoresis.
Construction of Human Genomic and cDNA Libraries

- Human genes can be isolated by DNA cloning.
- Dealing with the unfragmented 3 billion nucleotides of the complete human genome is a daunting task.
- This can be avoided by breaking up the total genomic DNA into smaller, more manageable pieces to make it easier to work with.
- To do so, the total DNA extracted from a tissue sample or a culture of human cells is cut up into a set of DNA fragments by restriction nuclease treatment, and each fragment is cloned using bacterial plasmids as described earlier.
- The collection of cloned DNA fragments thus obtained is known as a DNA library. In this case, it is called a genomic library, as the DNA fragments are derived directly from the chromosomal DNA.
Construction of Human Genomic and cDNA Libraries

• While producing a genomic library, one must make sure that each colony of E. coli produces clones of only one DNA fragment.
• This can be done by carrying out the entire procedure under the following favorable conditions:
  – DNA fragments are inserted into plasmid vectors under conditions that favor the insertion of one DNA fragment for each plasmid molecule.
  – These recombinant plasmids are mixed with a culture of E. coli at a concentration that ensures that no more than one plasmid molecule is taken up by each bacterium.
• A natural question that arises is how one can find a particular gene in this huge genomic library.
Construction of Human Genomic and cDNA Libraries

• If the sequence of the complementary DNA is known, one could make a probe and use it to identify the particular gene by exploiting the properties of nucleic acid hybridization.

• If, on the other hand, the sequence of the gene is not known, one can use protein sequencing to identify a few of the amino acids coded by the gene.

• By using the genetic code in reverse, the DNA sequences that code for these amino acid sequences can be deduced and a suitable DNA probe can be synthetically prepared.

• Using this probe, those rare bacterial clones containing human DNA complementary to the probe can be identified by DNA hybridization.

• It is possible that for a given gene several clones may be identified, as no single clone might contain the entire gene. This is especially true for long DNA sequences having lots of introns.
Construction of Human Genomic and cDNA Libraries

• For many applications of DNA technology, it is advantageous to obtain a clone that contains only the coding sequence of a gene, that is, a clone that lacks the intron DNA.

• It is a relatively simple matter to isolate a gene free of all its introns.

• For this purpose, a different type of library, called a complementary DNA (cDNA) library is used.

• The creation of a cDNA library involves the following steps:
  – First, starting from the mRNA that is expressed in a particular tissue or cell culture, construct the complementary DNA (cDNA) using reverse transcriptase.
  – Then, degrade the mRNA copy using an alkali (or a base) and use the single-stranded cDNA left over to produce a double-stranded cDNA copy of the original mRNA.

• The cDNA molecules can then be cloned, just like the genomic DNA fragments described earlier, to produce the cDNA library.
DNA and cDNA clones

Based on our discussion so far, we can highlight the following differences between DNA clones and cDNA clones:

Genomic clones represent a random sample of all of the DNA sequences found in an organism’s genome while cDNA clones contain fragments of only those genes that have been transcribed into mRNA in the tissue from which the RNA came. Since cells of different tissues produce distinct sets of RNA molecules, a different cDNA library will be obtained for each type of tissue.

Genomic clones from eucaryotes contain large amounts of repetitive DNA sequences, introns, gene regulatory regions, and spacer DNA, in addition to protein coding sequences, while cDNA clones contain only coding sequences.
Thus cDNA clones are particularly well suited for:
(a) deducing the amino acid sequence of a protein from the DNA;
(b) producing the protein in bulk by expressing the cloned gene in a bacterial or yeast cell

We note that hybridization allows even distantly related genes to be identified.

Hybridization can be carried out under conditions that allow even an imperfect match between a DNA probe and its corresponding DNA to form a stable double helix.

This can be used to identify closely related genes using a single DNA probe or even use a DNA probe for one species to identify the corresponding gene for another species.
Polymerase Chain Reaction (PCR)

• Polymerase chain reaction, or PCR, is a synthetic procedure that can be used to selectively replicate a given nucleotide sequence quite rapidly and in large amounts from any DNA that contains it.  
• Each cycle of the PCR reaction consists of three steps as shown in the figure.  
  • In the first step, heat is applied to separate out the two DNA strands of a double-stranded DNA molecule.  
  • In the second step, primers are hybridized on the two strands to mark the beginning of the regions of DNA to be amplified.  
  • In the third step, DNA polymerase and deoxyribonucleoside triphosphates are added so that DNA complementary to each of the two strands and originating at the two primers can be synthesized.  
• Thus, in one cycle of PCR, the quantity of DNA is amplified by a factor of 2.
Steps in PCR

1. **Heat to Separate Strands**
   - Double-stranded DNA is heated to separate the strands.

2. **Hybridization of Primers**
   - Primers hybridize to the separated strands.

3. **DNA Synthesis from Primers**
   - DNA polymerase, with dATP, dGTP, dCTP, and dTTP, synthesizes new DNA strands.

**First Cycle**
Polymerase Chain Reaction (PCR)

- If additional amplification is necessary, then the procedure can be repeated over and over again.
- In $n$ cycles of PCR, the DNA would have been amplified by a factor of $2^n$.
- In each cycle of the PCR reaction, the DNA strands are separated by heating.
- Consequently, the DNA polymerase used for PCR is a special kind of heat-resistant polymerase which is usually isolated from thermophilic (heat loving) bacteria.
- Otherwise, the DNA polymerase would have had to be replaced in each cycle of the PCR reaction.
- We next describe some common applications of PCR.
Applications of PCR

- PCR can be used to clone directly a particular DNA fragment. The main advantage of this procedure is that it does not require any cell culturing.

- PCR can be used to detect viral infection at very early stages.
- Here short sequences complementary to the viral genome are used as primers, and following many cycles of amplification, the presence or absence of even a single copy of a viral genome in a sample of blood can be ascertained.
- Once the amplification has been carried out using PCR, the virus detection can be done using gel electrophoresis.
- The schematic diagram for the entire procedure in the case of the HIV virus is shown in the figure.
Use of PCR in detecting viral (HIV) infection
Applications of PCR

- PCR has great potential in forensic science used to track down the perpetrator of a crime.
- The DNA sequences that create the variability used in this type of analysis contain runs of short repeated sequences, such as GTGTGT...., which are found in various positions (loci) in the human genome.
- The number of repeats in each run is highly variable in the population, ranging from 4 to 40 in different individuals.
- A run of repeated nucleotides of this type is commonly referred to as a VNTR (variable number of tandem repeats) sequence.
- Because of the variability in these sequences, each individual will usually inherit a different variant of each VNTR locus from their mother and from their father; two unrelated individuals will therefore not usually contain the same pair of sequences.
Applications of PCR

- A PCR reaction using primers that bracket the VNTR locus produces a pair of bands of amplified DNA from each individual, one band representing the maternal variant and the other representing the paternal variant.
- The length of the amplified DNA will depend on the exact number of repeats at the locus.
- The position of the amplified DNA after gel electrophoresis will depend on the number of repeats that are present at a particular locus.
- This fact can be used to narrow down the list of suspects in a criminal investigation.
Applications of PCR

The steps are as follows:

• Select a number of VNTR loci.
• Use primers to bracket each locus.
• Amplify the DNA using PCR and then run gel electrophoresis.

• This process should be carried out on DNA obtained from the forensic sample and also DNA samples from each of the suspected individuals.
• If the bands obtained during gel electrophoresis on the forensic sample do not match the bands obtained from a suspect, then this particular suspect can reasonably be eliminated from the list of suspects.
Genetic Engineering
Engineering DNA Molecules, Proteins and RNAs

- Advances made in DNA technology over the last three decades have opened up vast avenues for **genetic manipulation**.
- For instance, using DNA technology, completely novel DNA molecules can be constructed.
- To do so, one can cut a **plasmid vector** using a restriction nuclease and then insert a DNA fragment into the plasmid vector using DNA ligase.
- The procedure can be repeated to insert additional DNA fragments.
- Another application of DNA technology is the production of large quantities of **rare cellular proteins** using cloned DNA.
- To do so, one can make use of a plasmid that already has a promoter sequence inserted into it.
- Such a plasmid is called a **double stranded expression vector**.
Engineering DNA Molecules, Proteins and RNAs

• The protein coding sequence is inserted into the region immediately following the promoter by using a restriction nuclease and DNA ligase.
• By introducing the recombinant plasmid into bacterial cells, these cells will be overexpressing the mRNA and the corresponding protein.
• Yet another application of DNA technology is the artificial production of RNAs by in vitro transcription.
• Most RNAs found in a cell are present only in small quantities and are hard to isolate from the other cellular components.
• Once the gene coding for an RNA has been isolated, that RNA can be produced artificially in a test tube by letting RNA polymerase operate on the gene.
The function of a gene is best revealed by an organism that has had that gene mutated in some way.

Neither the complete nucleotide sequence of a gene nor the three-dimensional structure of a protein is sufficient to deduce a protein’s function.

Many proteins, such as those that have a structural role in the cell or normally form part of a large multienzyme complex, will have no obvious activity when viewed in isolation.

Mutants that lack a particular protein may clearly reveal what the normal function or functions of that protein are.

Prior to the advent of recombinant DNA technology, such mutations would have had to arise naturally and so would necessarily arise by chance and be unpredictable.

Recombinant DNA technology makes it possible to introduce precise mutations by the technique of site-directed mutagenesis.
Site directed mutagenesis

1. **Strand Separation**
   - Synthetic oligonucleotide primer containing desired mutated sequence
2. **Strand Completion by DNA Polymerase and DNA Ligase**
3. **Introduction into Cells, Replication and Segregation into Daughter Cells**
4. **Transcription**
   - Normal protein made by half the progeny cells
   - mRNA
5. **Translation**
   - Asp

6. **Transcription**
   - Protein with the single desired amino acid change made by half the progeny cells
   - Ala
Engineering Mutant Haploid Organisms

- In recombinant DNA technology, one can start with a cloned gene and proceed to make mutations in it *in vitro*.
- Then, by reintroducing the altered gene back into the organism from which it originally came, one can produce a mutant organism in which the gene’s function may be revealed.
- The gene to be mutated is first inserted into a plasmid and a sequence GAC (say) in one of the DNA strands in the gene is replaced by a mutated sequence GCC.
- The plasmid is then introduced into a cell. Due to the semi-conservative nature of DNA replication, every time the cell replicates, one half of the progeny end up inheriting the normal gene while the other half end up inheriting the mutated gene.
- The result is that half of the progeny produce the corresponding normal protein while the other half produce the mutated protein.
Transgenic Organisms

- Recombinant DNA technology can be used to carry out:
  - gene replacement
  - gene knockout
  - gene addition.

- Organisms into which a new gene has been introduced, or those whose genomes have been altered in other ways using recombinant DNA techniques, are known as transgenic organisms.
Engineering Transgenic Animals

• It is clear that for haploid organisms such as bacteria or yeasts, transgenic progeny can be produced quite easily.
• For diploids organisms (such as mice) it is more difficult, but still possible, to achieve gene replacements.

• This involves the following four steps:

Step 1
• An altered version of the gene is introduced into cultured embryonic stem cells.
• A few rare embryonic stem cells will have their corresponding normal genes replaced by the altered gene through homologous recombination.
• Although often laborious, these rare cells can be identified and cultured to produce many descendants, each of which carries an altered gene in place of one of its two normal corresponding genes.
Step 2

• These altered embryonic stem cells are injected into a very early mouse embryo.
• As a result, the cells are incorporated into the growing embryo, and a mouse produced by such an embryo will contain some somatic cells (cells other than the reproductive ones) that carry the altered gene.
• Some of these mice will also contain germ-line (precursors of reproductive) cells that contain the altered gene.

Step 3

• When bred with a normal mouse, some of the progeny of these mice will contain the altered gene in all of their cells.
• This is because two haploid reproductive cells of different kinds (egg and sperm) have to fuse together to produce the diploid cell that will undergo repeated cell division to produce the entire organism.
Step 4

- If two such mice are in turn bred, some of the progeny will contain two altered genes (one on each chromosome) in all of their cells.

- This again is a consequence of the meiotic process by which the haploid gametes are produced from their diploid precursors, and the fact that when the egg with an altered gene fuses with a sperm with the same gene altered, then the resulting diploid cell will have both copies of the gene altered.

- If the original gene alteration completely inactivates the function of the gene, then we will have what are called “knockout” mice.
Cell Division
Cell Division

• In this chapter, we discuss cell division, which plays a crucial role in the propagation of all life.
• The only way to make more cells is by division of those that already exist.
• A cell reproduces by carrying out a tightly controlled sequence of events in which it duplicates its contents and divides in two.
• This cycle of duplication and division, known as the cell cycle, is the essential mechanism by which all living things reproduce.
• In unicellular organisms such as bacteria or yeast, each cell division leads to a complete new organism, while in multicellular organisms, many rounds of cell division are required to make a new individual from the single-celled egg.
• In multicellular organisms, quite often cell division has to be carried out to sustain the organism in the steady-state.
• The details of the cell cycle vary from organism to organism and at different stages in an organism’s life.
Cell Division

- Certain characteristics, however, are universal.
- For instance, the cell that is undergoing division must replicate its DNA and pass on identical copies of the DNA to its two daughter cells.
- Procaryotes (or bacteria) do not have a nucleus and reproduce by a fission type of cell division.
- Here, the circular bacterial chromosome is replicated and one copy moves towards each end of the dividing bacterium.
- The cell wall and plasma membrane in the center of the dividing bacterium pinch inwards resulting in two daughter cells.
- Cell division in a eucaryotic cell is much more complicated.
- This is because most of the genetic information of the cell, namely its nuclear genome, is distributed between multiple chromosomes contained in the nucleus.
- In addition, the cytoplasm contains many organelles which must be duplicated and apportioned out equally between the two daughter cells.
Cell Division

There are three major questions that arise in the context of cell division:

• the mechanisms by which cells duplicate their contents;
• the mechanisms by which cells partition their duplicated contents and split in two;
• the mechanisms employed by cells to ensure that the different steps involved in cell division take place in the proper sequence.

• Partial answers to the first question have already been provided, for instance, in earlier chapters, we discussed DNA replication and protein synthesis.
• Here, we will focus on providing an answer to the second question.
• The answer to the third question will be taken up in the next chapter.
The duration of the cell cycle varies greatly from one cell type to another.

For instance, a single-celled yeast can divide every 90-120 minutes in ideal conditions, while a mammalian liver cell divides, on average, less than once a year.

In this chapter, we will focus on the sequence of events in a fairly rapidly dividing mammalian cell, with a cell cycle time of about 24 hours.

The eucaryotic cell cycle is broadly divided into two phases:

- (i) *M phase*, which is composed of mitosis (nuclear division) and cytokinesis (splitting of the cell in two);
- (ii) *interphase*, which is the period between one M phase and the next. The interphase is further divided into the remaining three phases of the cell cycle, as shown in the figure.
Eucaryotic cell cycle

- **M**: Mitosis and Cytokinesis
- **G2**: Growth, preparation for mitosis
- **G1**: Growth, increase in cell size
- **S**: Replication of DNA

**INTERPHASE**
Interphase

- During *S phase* (S=synthesis), DNA replication takes place.
- The S phase is flanked by two phases where the cell continues to grow.
- The *G1 phase* (G=gap) is the interval between the completion of the M phase and the beginning of the S phase.
- The *G2 phase* is the interval between the end of the S phase and the beginning of the M phase.
- There are particular times in the *G1* and *G2* phases when the cell makes a decision whether to proceed to the next phase or pause to allow more time to prepare.
- These times are called the *G1* and *G2* checkpoints, respectively.
- During all of interphase, a cell continues to transcribe genes, synthesize proteins, and grow in mass.
- Together *G1* and *G2* phases provide additional time for the cell to grow and duplicate its cytoplasmic organelles.
• Without the $G1$ and $G2$ phases, the cell would get progressively smaller at each cell division.

• This is, in fact, what happens to a fertilized egg during the first few cell divisions of its life, referred to as cleavage cell divisions.

• The first readily visible sign under a microscope that a cell is about to enter M phase is the progressive condensation of its chromosomes.

• This condensation makes the chromosomes less likely to get entangled and therefore physically easier to separate during mitosis.

• The separation of the duplicated condensed chromosomes is carried out by a transient structure called the mitotic spindle.

• For animal cells, the actual division of the cell is carried out by a contractile ring made up of two kinds of protein filaments (actin and myosin).

• Like the mitotic spindle, the contractile ring is also a transient structure that assembles at the appropriate time.
In plant cells, the **cytoplasmonic division** has to be carried out by a different mechanism since each cell is enclosed within a hard cell wall.

Small organelles such as **mitochondria** and **chloroplasts** are usually present in large numbers in each cell and will be safely inherited by the daughter cells if, on the average, their numbers simply double once every cell cycle.

Other larger organelles, such as the **Golgi apparatus** and the **endoplasmic reticulum** disintegrate into small fragments during mitosis, which increases the chance that the latter will be more or less evenly distributed among the daughter cells when the cell divides.

Subsequently, the organelles are reconstructed from the inherited fragments present in each daughter cell.

We next take a detailed look at **mitosis**, which is the type of **nuclear division** that most eucaryotic cells undergo.
Mitosis and Cytokinesis

- Before mitosis begins, each chromosome has been replicated and consists of two identical *chromatids* (called *sister chromatids*) which are joined together along their length by interactions between proteins on the surface of the two chromatids.
- A typical pair of sister chromatids is shown in the figure.

Although mitosis proceeds as a continuous sequence of events, it is traditionally divided into five stages:

- *prophase*
- *prometaphase*
- *metaphase*
- *anaphase*
- *telophase*
A typical pair of sister chromatids
Stages in Mitosis

• During **prophase**, the replicated chromosomes condense and the mitotic spindle begins to assemble outside the nucleus.
• During **prometaphase**, the nuclear envelope breaks down, allowing the spindle microtubules to contact the chromosomes and bind to them.
• During **metaphase** the mitotic spindle gathers all of the chromosomes to the center (equator) of the spindle.
• During **anaphase** the two sister chromatids in each replicated chromosome synchronously split apart, and the spindle draws them to opposite poles of the cells.
• During **telophase** a nuclear envelope reassembles around each of the two sets of separated chromosomes to form two nuclei.
• In three of the stages described above, we have referred to the mitotic spindle.
• A natural question that comes up is how is the mitotic spindle formed and what exactly is its role.
Prophase

- The mitotic spindle starts to assemble in prophase.
- Towards the end of the S phase, the cell duplicates a structure called the *centrosome* to produce two daughter centrosomes, which initially remain together at one side of the nucleus.
- As *prophase* begins, the two daughter centrosomes separate and move to the opposite poles of the cell, driven by centrosome-associated motor proteins that use the energy of ATP hydrolysis to move along microtubules.
- Each centrosome serves to organize its own array of microtubules and the two sets of microtubules then interact to form the *mitotic spindle*.
- The figure shows the two centrosomes with their own array of microtubules.
Three classes of microtubules in a mitotic cell
• The microtubules that radiate from the centrosome in an interphase cell continuously polymerize and depolymerize by the addition and loss of units of a protein called *tubulin*.

• Individual microtubules, therefore, alternate between growing and shrinking and this process is called *dynamic instability*.

• The rapidly growing and shrinking microtubules extend in all directions from the centrosomes, exploring the interior of the cell.

• During prophase, while the nuclear envelope is still intact, some of these microtubules become stabilized against disassembly to form the highly organized mitotic spindle.

• This happens when some of the microtubules growing from opposite centrosomes interact, binding the two sets of microtubules together to form the basic framework of the mitotic spindle.

• The interacting microtubules are called *polar microtubules* (since they originate from the two poles of the spindle), and the two centrosomes that give rise to them are called the *spindle poles*. 
Prometaphase

- Chromosomes get attached to the mitotic spindle during prometaphase.
- **Prometaphase** starts abruptly with the disintegration of the nuclear envelope.
- Following this, the spindle microtubules, which have been lying in wait outside the nucleus, suddenly gain access to the replicated chromosomes and bind to them.
- The spindle microtubules bind the chromosomes through specialized protein complexes called *kinetochores*, which are formed on the chromosomes during late prophase.
- As we already mentioned, each replicated chromosome consists of two sister chromatids joined together along their length, and each chromatid is constricted at a region of specialized DNA sequence called the *centromere*.
- Just before prometaphase, kinetochore proteins assemble into a large complex on each centromere.
Prometaphase to Metaphase

- Once the nuclear envelope has disintegrated, a randomly probing microtubule encountering a kinetochore will bind to it, thereby capturing the chromosome.
- Such a microtubule is called a *kinetochore microtubule*, and it links the chromosome to a spindle pole.
- Not every microtubule emanating from the two centrosomes ends up forming a kinetochore microtubule or a polar microtubule.
- In fact, several of the microtubules remain unattached and are called *unattached microtubules*.
- During prometaphase, the chromosomes, which are now attached to the mitotic spindle, appear to move as if jerked around randomly in different directions.
- Finally they align at the equator of the spindle, halfway between the two spindle poles, thereby forming the *metaphase plate*. 
Metaphase

- The formation of the metaphase plate defines the beginning of \textit{metaphase}.
- The precise forces that act to bring the chromosomes to the equator are not well understood.
- However, the continual growth and shrinkage of the microtubules and the action of microtubule motor proteins are thought to play a role.
- In this context, we note that a continuous balanced addition and loss of tubulin subunits is required to maintain the mitotic spindle.
- Thus when tubulin addition is blocked in a mitotic cell using say the drug \textit{colchicine}, tubulin loss continues until the spindle disappears.
- This prevents the cell from dividing and so colchicine can be used in cancer therapy to prevent a cancer cell from proliferating.
Anaphase A

• At the start of anaphase, the connections between the two sister chromatids are cut by proteolytic enzymes, allowing each chromatid (now called a daughter chromosome) to be gradually pulled to the spindle pole to which it is attached.
• This movement is the result of two independent processes brought about by different parts of the mitotic spindle.
• These processes are called anaphase A and anaphase B, and their occurrence is more or less simultaneous.
• In anaphase A, the kinetochore microtubules shorten by depolymerization, and the two daughter chromosomes move towards their respective spindle poles.
Anaphase B

- In anaphase B, the spindle poles themselves move away from each other, further contributing to the segregation of the two groups of daughter chromosomes.
- The driving force for the movements of anaphase A is thought to be provided partly by the action of microtubule motor (movement generating) proteins operating at the kinetochore and partly by the loss of tubulin subunits that occurs mainly at the kinetochore end of the kinetochore microtubules.
- The driving force for the moving apart of the spindle poles in anaphase B is thought to be provided by the polymerization of the polar microtubules by the addition of tubulin units at their free ends.
The last step in mitosis is *telophase*.

During telophase a nuclear envelope reforms around each group of chromosomes to form the two daughter nuclei.

With the creation of the two daughter nuclei, the process of nuclear division (or mitosis) is complete.

M phase involves more than just the segregation of the daughter chromosomes and the formation of new nuclei.

It is also the time during which other components of the cell — membranes, organelles, and proteins — are distributed more or less evenly between the two daughter cells.
Cytokinesis

• This is achieved by *cytokinesis* and usually begins in anaphase but is not completed until after the two daughter nuclei have formed.

• Regarding cytokinesis, we note the following:
  (i) the mitotic spindle determines the plane of cytoplasmic cleavage, which is usually orthogonal to the orientation of the mitotic spindle;
  (ii) the *contractile ring* of animal cells, which is responsible for carrying out cytokinesis in such cells, is made up of the proteins *actin* and *myosin*;
  (iii) cytokinesis in plant cells involves the formation of a new *cell wall* (presumably because the plant cells in general are surrounded by a tough cell wall).
Meiosis

• In this section, we take a look at meiosis, which is another kind of eucaryotic cell division employed by sexually reproducing diploid organisms to produce the reproductive cells.
• In a diploid organism, all the cells of the body including the germ-line cells that give rise to the gametes are diploid; however, the gametes themselves are haploid.
• Thus the haploid gametes must be produced from their diploid precursors by a special type of cell division.
• This special type of cell division is called meiosis.
Homologues

- With the exception of the chromosomes that determine sex (the sex chromosomes), a diploid nucleus contains two very similar versions of each chromosome, one from the father called the paternal chromosome and one from the mother called the maternal chromosome.
- The two versions of each chromosome, however, are not genetically identical, as they carry different variants of many of the genes.
- They are, therefore, called homologous chromosomes, or homologues, meaning that they are similar but not identical.
- A diploid cell in a sexually reproducing organism consequently carries two similar sets of genetic information.
- In most cells, the paternal and maternal homologues maintain a completely separate existence as independent chromosomes.
Mitosis and Meiosis

- Mitosis and meiosis are similar in certain respects but there are important differences that we will be discussing.
- The following figure illustrates the key differences between mitosis and meiosis using a fictitious diploid cell with only one pair of homologous chromosomes.
- In mitosis, each chromosome replicates and the replicated chromosomes line up in random order at the metaphase plate; the two sister chromatids then separate from each other to become individual chromosomes, and the two daughter cells produced by cytokinesis inherit a copy of each paternal chromosome and a copy of each maternal chromosome.
- Thus both sets of genetic information are transmitted intact to the two daughter cells, which are, therefore, each diploid and genetically identical.
Duplicated chromosomes line up individually on the spindle.

Completion of cell division I.

Meiotic division II.

Cell division II.

Haploid gametes.

Cell division.
Mitosis and Meiosis

• In contrast, when diploid cells divide by meiosis they form haploid gametes with only one half the original number of chromosomes i.e. only one chromosome of each type instead of a pair of homologues of each type.
• Thus, each gamete acquires either the maternal copy or the paternal copy of a chromosome but not both.
• This reduction is needed so that when two gametes of opposite types (an egg and a sperm in animals) fuse at fertilization, the chromosome number is restored in the embryo to the original diploid number for that species.
• Since the assignment of maternal and paternal chromosomes to the gametes during meiosis occurs at random, the original maternal and paternal chromosomes are reshuffled into an incredibly large number of different combinations (as discussed earlier).
Mitosis and Meiosis

- As noted in that chapter, and as shown in the figure, in a meiotic cell division, the replicated homologous paternal and maternal chromosomes (including the two replicated sex chromosomes) pair up alongside each other before they line up on the spindle.
- This makes it possible for crossovers by homologous recombination to occur at this stage.
- In principle, since meiosis involves a halving of the number of chromosomes, it could have occurred by a simple modification of a normal mitotic cell division, such that the DNA replication (S phase) is omitted.
- For unknown reasons, the actual meiotic process is more complicated and involves DNA replication followed by two cell divisions instead of one (as shown in the previous figure).
Down syndrome

- Occasionally, the meiotic process occurs abnormally and homologues fail to separate — a phenomenon known as *nondisjunction*.
- In this case some of the gametes that are produced lack a particular chromosome, while others have more than one copy of it.
- Such gametes, when combined with a normal gamete of the opposite type, form abnormal embryos, most of which do not survive.
- Some, however, do and *Down syndrome* in humans is an example of a disease that is caused by an extra copy of Chromosome 21.
- This condition is, therefore, referred to in the scientific literature as *Trisomy 21*, meaning that such a person’s genome has three copies of Chromosome 21 instead of the two copies that normal humans have.
Cell Cycle Control, Cell Death and Cancer
Cell Cycle Control

• We introduce the phenomena of *cell cycle control* and *programmed cell death* in multicellular organisms and discuss how disruption in either of these can lead to the disease called *cancer*.

• Our discussion so far has focussed on two aspects of cell division: (i) replication of the contents of the cell; (ii) the actual partitioning of these replicated contents between the two daughter cells when the cell divides.

• There is yet another aspect of cell division that is crucially important, namely, the mechanism by which the cell controls the different chronological steps that are involved in cell division.
Cell Cycle Control System

- The events of the cell cycle occur in a fixed sequence, namely M phase followed by G1 phase, G1 phase followed by S phase, S phase followed by G2 phase and finally G2 phase followed by M phase again.
- This is ensured by a cell cycle control system which has to perform a number of functions.
  
  1. It has to activate the enzymes and other proteins responsible for carrying out each process in the cell cycle at the appropriate time, and then it has to deactivate them once the process is completed.
  2. It must also ensure that each stage of the cycle is completed before the next one is begun. (For instance, it has to make sure that DNA replication has been completed before mitosis begins, that mitosis has been completed before cytokinesis begins, and so on)
Cell Cycle Control System

(3) The control system must also take into account if the conditions outside the cell are conducive for division.

• For instance, in a multicellular organism the control system must be responsive to signals from other cells, such as those that stimulate cell division when more cells are needed.

• It is clear that the cell cycle control system plays a major role in the regulation of cell numbers in the tissues of the body.

• When this system malfunctions, it can result in cancer.
Cell Cycle Control System : An Analogy

• The operation of the cell cycle control system is very similar to that of the control system for any cyclic process.
• We will illustrate this by drawing the analogy with an automatic washing machine.
• The duty cycle of an automatic washing machine consists of the following five steps:
  (i) take in water, (ii) mix with detergent, (iii) wash the clothes, (iv) rinse them, and (v) spin them dry.
• In the eucaryotic cell cycle, these steps are analogous to (a) DNA replication, (b) mitosis, etc.
• Furthermore, the washing machine controller is itself regulated at certain critical points of the cycle by feedback from the processes that are being performed.
Checkpoints

- For instance, sensors monitor the water levels in a washing machine and send signals back to the controller to prevent the start of the next process before the current one has been completed.
- Similarly, the events of the cell cycle have to occur in a particular sequence, and this sequence must be preserved even if one of the steps takes longer than usual.
- For instance, all of the nuclear DNA must be replicated before the nucleus begins to divide and it is crucial for most cells to double in size before dividing in two as otherwise the cells would get progressively smaller at each cell division.
- The cell-cycle control system achieves all this by means of molecular brakes that can stop the cycle at various checkpoints.
• The control system in most cells has **checkpoints for cell size**, where the cell-cycle is halted until the cell has grown to an appropriate size.

• In *G1*, a **size checkpoint** allows the system to halt and the cell to grow further, if necessary, before a new round of DNA replication is triggered.

• Cell growth depends on an **adequate supply of nutrients** and other factors in the extracellular environment, and the *G1* checkpoint also allows the cell to check that the environment is favorable for cell proliferation before committing itself to the DNA replication (*S*) phase.

• A second **size checkpoint occurs in *G2***, allowing the system to halt before it triggers mitosis.

• The *G2* checkpoint also allows the cell to check that DNA replication is complete before proceeding to mitosis.
Cyclin-Dependent Kinases

- The cell cycle control system governs the cell-cycle machinery through the *phosphorylation of key proteins* that initiate or regulate DNA replication, mitosis, and cytokinesis.
- Recall that the phosphorylation reactions are carried out by enzymes called *kinases*.
- The protein kinases of the cell-cycle control system are present in dividing cells throughout the cell cycle.
- They are *activated*, however, only *at appropriate times* in the cycle, after which they quickly become deactivated again.
- This is made possible by a second set of protein components of the control system, which are called the *cyclins*.
- Cyclins have no enzymatic activity by themselves, but they have to *bind to the cell-cycle kinases* before the kinases can become enzymatically active.
- As a result, the kinases of the cell cycle control system are referred to as *cyclin-dependent protein kinases*, or *Cdks*. 


**Cyclin-Cdk**

- Cyclins derive their name from the fact that their concentrations vary in a cyclical fashion during the cell cycle.
- We will focus on the **cyclin-Cdk complex** that is responsible for driving cells into mitosis.
- This cyclin-Cdk complex is known as the *M-phase promoting factor* (MPF).
- The cyclin-Cdk complex that drives cells into M phase was first discovered through studies of cell division in frog eggs.
- The fertilized eggs of many animals are especially well-suited for biochemical studies of the cell cycle because (i) they are very large cells and (ii) they divide very rapidly.
- This is due to the fact that they undergo *cleavage divisions*, i.e. M phase followed by S phase and then M phase again with little or no G1 or G2 phases in between.
Xenopus oocyte

• By taking frog eggs at a particular stage of the cell cycle, an extract can be prepared that is representative of that cell-cycle stage.
• The biological activity of such an extract can then be tested by injecting it into a *Xenopus oocyte* (the immature diploid precursor of the unfertilized frog egg) and observing its effects on cell cycle behavior.
• The *Xenopus* oocyte is a convenient test system for detecting an activity that drives cells into M phase, as it has completed DNA replication and is arrested just before M phase of the first meiotic division.
• The oocyte is therefore at a stage in the cell cycle that is equivalent to the G2 phase of a mitotic cell cycle.
**M-phase promoting factor: MPF**

- In such experiments it was found that an abstract from an M-phase fertilized egg instantly drives the oocyte into M phase, whereas cytoplasm from a cleaving egg at other phases of the cycle does not.
- When initially discovered, the chemical composition and the mechanism of action of the factor responsible for this activity were unknown, and consequently the factor was simply called the *M-phase promoting factor* or MPF.
- MPF activity was found to oscillate dramatically during the course of each cell cycle as shown in the figure; it increased rapidly just before the start of mitosis and fell rapidly to zero towards the end of mitosis.
- Subsequent studies revealed that MPF contains a single protein kinase, which is required for its activity.
Variation of MPF activity and Cyclin concentration during different stages of the cell cycle
Cyclin

• By phosphorylating key proteins, the kinase causes several phenomena associated with mitosis to occur: the condensation of the chromosomes, the disintegration of the nuclear envelope and the formation of the mitotic spindle.

• However, the MPF kinase is not capable of acting by itself and has to have a specific cyclin bound to it in order to function.

• Biochemical experiments using cleaving clam eggs led to the discovery of cyclin.

• Cyclin was initially identified as a protein whose concentration rose gradually during interphase and then fell rapidly to zero as the cells went through M phase, repeating this performance in each cell cycle. (as shown in the previous figure)
The **MPF** is a protein complex containing two subunits: a regulatory subunit that is a cyclin and a catalytic subunit that is the mitotic Cdk.

Many of the cell-cycle control genes have been remarkably conserved during biological evolution. The human version of these genes will function perfectly well when introduced into a yeast cell.

The manufacture of the cyclin component of MPF starts immediately after cell division and continues steadily through interphase.

The cyclin accumulates, so that its concentration rises gradually and helps time the *onset of mitosis*; its subsequent rapid decrease helps initiate the exit from mitosis.

The sudden fall in the cyclin concentration during mitosis is the result of the *rapid degradation of the cyclin by the ubiquitin-dependent proteolytic system*.

The MPF activation initiates a process that, after some time delay, leads to the ubiquination and degradation of the cyclin, thereby turning the kinase off.
Cyclin-Cdk : Positive Feedback Mechanism

• From the figure we observe that the cyclin concentration increases gradually throughout interphase, whereas the MPF kinase activity switches on abruptly at the end of interphase.
• Thus, the cyclic variations in the cyclin concentration alone cannot completely explain MPF kinase activity.
• This is due to the fact that the kinase itself also has to be phosphorylated at one or more sites and dephosphorylated at others before it can become enzymatically active.
• The removal of the inhibitory phosphate groups by a specific protein phosphatase is the step that activates the kinase at the end of interphase.
• Once activated, a cyclin-Cdk complex can activate more cyclin-Cdk complexes by a positive feedback type of mechanism. This positive feedback type of mechanism is what causes the sudden explosive increase in MPF kinase activity that drives the cell abruptly into M-phase.
Other Cyclin-Cdk Complexes

- Here, we have focused attention on the cyclin-Cdk complex that constitutes the MPF.
- However, there are many varieties of cyclin and, in most eucaryotes, many varieties of Cdk that are involved in cell-cycle control.
- For instance, S phase cyclins trigger entry into S phase while G1 cyclins act earlier in G1.
- The latter bind to Cdk molecules to help initiate the formation and activation of the S-phase cyclin-Cdk complexes and thereby drive the cell toward S-phase.
- Note that the concentration of each type of cyclin rises and then falls sharply at a specific time in the cell cycle to control the timing of a particular stage.
- As before, the sudden fall in the cyclin concentration is the result of cyclin degradation by the ubiquitin pathway.
Cdk inhibitor proteins

• As we have already seen, the cell cycle control system triggers the events of the cell cycle in a specific order.
• If one of the steps is delayed, the control system of necessity must delay the activation of the following steps so that the sequence is maintained.
• This is accomplished by the action of molecular brakes that can stop the cell cycle at specific checkpoints.
• For most cases, the detailed molecular mechanisms involved are not well understood.
• However, in some cases, it is known that specific Cdk inhibitor proteins come into play.
Checkpoint: DNA Damage

- One of the best understood checkpoints stops the cell cycle in the G1 phase if the DNA is damaged, helping to ensure that a cell does not replicate damaged DNA.
- DNA damage causes an increase in both the concentration and activity of a gene regulatory protein called $p53$ (protein with a molecular weight of 53,000 units).
- When activated, $p53$ stimulates the transcription of a gene encoding a Cdk inhibitor protein called $p21$ (protein with a molecular weight of 21,000 units).
- This increases the concentration of the p21 protein, which binds to the S phase cyclin-Cdk complexes responsible for driving the cell into S phase and blocks their action.
• The arrest of the cell cycle in G1 allows the cell time to repair the damaged DNA before replicating it.

• If p53 is missing or defective, the unrestrained replication of damaged DNA leads to a high rate of mutation and the production of cells that tend to become cancerous.

• In fact, mutations in the p53 gene that permit cells with damaged DNA to proliferate play an important part in the development of many human cancers.
**G0 : Modified G1 State**

- Cells can **dismantle** their control system and withdraw from the cell cycle altogether.
- In the human body, for instance, nerve cells have to persist for a lifetime without dividing.
- Consequently, they enter a **modified G1 state called G0**, in which the cell-cycle control system is partly dismantled in that many of the Cdk's and cyclins disappear.
- As a result, the **nerve cell** remains permanently in the **G0 state**.
- Mammalian cells proliferate only if they are stimulated to do so by **signals** from other cells.
- If **deprived** of such signals, the cell cycle arrests at a **G1 checkpoint** and enters the **G0 state**.
Control of Cell Numbers in Multicellular Organisms

• **Unicellular organisms** such as bacteria and yeasts tend to grow and divide as fast as they can, and their rate of proliferation depends largely on the availability of nutrients in the environment.
• For an animal cell to proliferate, nutrients are not enough.
• It must also receive **stimulating signals** from other cells, usually its neighbors.
• In other words, in multicellular organisms, cell division is under very tight control.
• An important example of a brake that normally holds cell proliferation in check is the **retinoblastoma (Rb) protein**.
• The Rb protein, which is abundant in the nucleus of all **vertebrate cells**, binds to particular gene regulatory proteins, preventing them from stimulating the **transcription of genes** required for cell proliferation.
**Growth Factors**

- Extracellular signals such as *growth factors* that stimulate cell proliferation lead to the activation of the G1 cyclin-Cdk complexes mentioned earlier.

- These phosphorylate the Rb protein, altering its conformation so that it releases its bound gene regulatory proteins, which are then free to activate the genes required for cell proliferation to proceed.

- The stimulating signals that act to override the brakes on cell proliferation are mostly *protein growth factors*.

- One example of a protein growth factor is the so called *platelet-derived growth factor* (PDGF).

- When *blood clots* (in a wound, for example), blood platelets incorporated in the clot are triggered to release PDGF, which binds to receptor tyrosine kinases in surviving cells at the wound site, thereby stimulating them to proliferate and heal the wound.
**Hepatocyte growth factor**

- Another example of a protein growth factor is the *hepatocyte growth factor*.
- In this case, if part of the liver is lost through surgery or acute injury, cells in the liver and elsewhere produce this protein called hepatocyte growth factor which helps to stimulate the surviving liver cells to proliferate.
- Even in the presence of growth factors, normal animal cells do not keep on dividing indefinitely in culture.
- Even cell types that maintain the ability to divide throughout the lifetime of the animal stop dividing after a limited number of divisions.
Cell Senescence

- For instance, fibroblasts (a class of cells from which connective tissue is derived) taken from a human fetus stop dividing after 80 rounds of cell division, while fibroblasts taken from a 40-year-old adult stop after 40 rounds of cell division.
- This phenomenon is known as cell senescence.
- Fibroblasts from a mouse embryo, on the other hand, halt their proliferation after only about 30 divisions in culture.
- This may partly explain the difference in size between mice and humans.
- The mechanisms that halt the cell cycle in either developing or aging cells are not clearly understood at the present time, although the accumulation of Cdk inhibitor proteins and the loss of Cdks are likely to be involved.
Programmed Cell Death

- Animal cells require signals from other cells to avoid programmed cell death.
- These signals are called *survival factors*.
- If deprived of such survival factors, the cells activate an intracellular suicide program and die by a process called *programmed cell death* or *apoptosis*.
- This helps ensure that cells survive only when and where they are needed.
- The amount of programmed cell death that occurs in both developing and adult tissues is astonishing.
- In the developing vertebrate nervous system, for example, more than half of the nerve cells normally die soon after they are formed.
Programmed Cell Death

- In a healthy adult human, billions of cells die in the intestine every hour.
- A natural question that arises is what purpose is served by this huge amount of cell death. In some cases, the answers are clear.
- For instance, the sculpting of hands and feet in mammals or the developmental loss of the tail of a tadpole are brought about by apoptosis.
- In the case of the developing vertebrate nervous system, apoptosis is used to match the number of nerve cells to the number of target cells that require innervation.
- In yet another case an enlarged liver can be returned to normal size via apoptosis.
Apoptosis and Cell Necrosis

• We next examine how apoptosis uniquely differs from other kinds of cell death.
• Cells that die as a result of acute injury typically swell and burst and spill their contents all over their neighbors (a process called cell necrosis), causing a potentially damaging inflammatory response.
• By contrast, a cell that undergoes apoptosis dies very neatly, without damaging its neighbors.
• The cell shrinks and condenses, the cytoskeleton collapses, the nuclear envelope disassembles, and the nuclear DNA breaks up into fragments.
• Most important, the cell surface is altered, displaying properties that cause the dying cell to be cleaned up immediately, either by its neighbors or by a macrophage before there is any leakage of its contents.
Apoptotic Cascade

- The machinery that is responsible for this kind of controlled cell suicide seems to be similar in all animal cells.
- It involves a family of proteases which are themselves activated by proteolytic cleavage in response to signals that induce programmed cell death.
- The activated suicide proteases cleave, and thereby activate, other members of the family, resulting in an amplifying proteolytic cascade which is often referred to as an apoptotic cascade.
- The activated proteases then cleave other key proteins in the cell, killing it quickly and neatly.
Cancer: Breakdown of Cell Cycle Control

- *Cancers* are the products of *mutations* that set cells free from the usual controls on cell proliferation and survival.
- A cell in the body mutates through a series of chance events and acquires the *ability to proliferate* without the normal restraints.
- Its progeny inherit the mutations and give rise to a *tumor* that can grow without limit.
- *Oncogenes* (that turn on cell division) and *tumor-suppressor genes* (that function as brakes on cell division) play an important role in causing cancer.
- An example of an oncogene is a gene produced from the gene for the platelet derived growth factor (*PDGF*).
Proto-oncogene and Oncogene

• Here the normal PDGF gene is called a proto-oncogene and the corresponding cancer causing gene is called an oncogene.
• While the proto-oncogene stimulates cell proliferation when more cells are needed, the oncogene causes cell proliferation even when no new cells are needed.
• An example of a tumor suppressor gene is the retinoblastoma gene whose product normally acts as a brake on cell division.
• When this gene is mutated, the braking action may no longer be present leading to excessive cell proliferation.
**Proto-oncogene and Oncogene**

- In a normal diploid cell, there are two copies of each gene in the genome.
- Since a proto-oncogene accelerates cell division when called upon to do so, only one copy of a proto-oncogene needs to be mutated to an oncogene before excessive cell proliferation, and possibly cancer, can result.
- On the other hand, a tumor suppressor gene is associated with braking action on cell division and so, both copies have to be mutated before a loss of function, possibly leading to cancer, can occur.
- In almost all adult tissues, cells are continually dying and being replaced; through all the hurly-burly of cell replacement and tissue renewal, the architecture of the tissue must be maintained.
Cell proliferation & Apoptosis: an equilibrium

• We have some kind of a dynamic equilibrium between cell proliferation and apoptosis.

• Three main factors contribute to make this possible:

  (i) **Cell communication**: this ensures that new cells are produced only when and where they are required;
  (ii) **Selective cell-cell adhesion**: The selectivity of adhesion prevents the different cell types in a tissue from becoming chaotically mixed;
  (iii) **Cell memory**: Cells autonomously preserve their distinctive character and pass it on to their progeny. This preserves the diversity of cell types in the tissue.
• Different tissues in the body are renewed at different rates.
• For instance, nerve cells in the body never divide while liver cells divide about once a year.
• The cells lining the inside of the intestine, on the other hand, are turned over about once every three days.
• Many of the differentiated cells that need continual replacement are themselves unable to divide.
• More of such cells are generated from a stock of precursor cells, called stem cells, that are retained in the corresponding tissues along with the differentiated cells.
• Stem cells can divide without limit: some of the progeny of the stem cells will remain stem cells, while others will embark on a course leading irreversibly to terminal differentiation.
Metastases

- Cancer cells are defined by two heritable properties: they and their progeny
  (1) reproduce in defiance of the normal social constraints
  (2) invade and colonize territories normally reserved for other cells.
- Cells that have only property (1) will result in a benign tumor while cells that have properties (1) and (2) will result in a malignant tumor (cancer).
- Malignant tumor cells can break loose from the primary tumor, enter the blood stream or lymphatic vessels, and form secondary tumors, or metastases, at other sites in the body.
- This is referred to as the spreading of cancer.
Cancer: Consequence of Mutation

• It is important to note that cancer is a consequence of mutation and natural selection within the population of cells that form the body.
• Indeed, the mutations in cancer cells give them a competitive advantage over their normal neighbors.
• Unfortunately, it is this very fact that leads to disaster for the multicellular organism, since the controls on cell proliferation and survival, which are at the heart of maintaining tissue architecture, are disrupted.
• About $10^{16}$ cell divisions take place in the human body over the lifetime of an individual.
• Based on this, it can be shown that every single gene is likely to have undergone mutations on about $10^{10}$ separate occasions in any individual human being.
Cancer: Consequence of Mutation

- Typically, at least 5 or 6 independent mutations must occur in one cell to make it cancerous.
- This is because the body of a multicellular organism has different levels of protection against cancer.
- For instance, to become cancerous, an epithelial stem cell in the skin or the lining of the gut must undergo changes that not only enable it to divide more frequently than it should, but also let its progeny escape being sloughed off in the normal way from the exposed surface of the epithelium, enable them to displace their normal neighbors, and let them attract a blood supply sufficient to nourish tumor growth.
- Thus the mutations that give rise to cancer accumulate over a long time and that is why cancer is typically a disease of old age.
• Occasionally, however, individuals are encountered who have inherited a germ-line mutation in a tumor suppressor gene or an oncogene.
• For these people, unfortunately the number of mutations required is less and the disease occurs more frequently and at an earlier age.
• The families that carry such mutations are, therefore, prone to cancer.
Expression Microarrays
Focus at the RNA Level

- Cellular control results from multivariate activity among cohorts of genes and their products.
- Since all three levels in the central dogma: DNA, RNA, and protein interact, it is not possible to fully separate them; information from all realms must be combined for full understanding.
- The high level of interactivity between levels insures that a significant amount of the system information is available in each of the levels, so that focused studies provide useful insights.
- Much current effort is focused at the RNA level owing to measurement considerations.
- High-throughput technologies make it possible to simultaneously measure the RNA abundances of tens of thousands of mRNAs.
Expression Microarrays

- Expression microarrays result from a complex biochemical-optical system incorporating robotic spotting and computer image formation.
- These arrays are grids of thousands of different single-stranded DNA molecules attached to a surface to serve as probes.
- Two major kinds are those using synthesized oligonucleotides and those using spotted cDNAs.
- The basic procedure is to extract RNA from cells, convert the RNA to single-stranded cDNA, attach fluorescent labels to the different cDNAs, allow the single-stranded cDNAs to hybridize to their complementary probes on the microarray, and then detect the resulting fluor-tagged hybrids via excitation of the attached fluors and image formation using a scanning confocal microscope.
Expression Microarrays

- Relative RNA abundance is measured via measurement of signal intensity from the attached fluors.
- This intensity is obtained by image processing and statistical analysis.
- Particular attention is paid to the detection of high- or low-expressing genes, and beyond that to expression-based phenotype classification and the discovery of multivariate inter-gene predictive relationships.
- Here, we briefly discuss microarrays, their hybridization-based foundation, normalization, and ratio analysis.
cDNA Microarrays

- The principle behind cDNA spotted arrays is that an mRNA transcript to be detected is copied into a complementary DNA (cDNA) and this copied form of the transcript is **immobilized** on a glass microscope slide.
- The slides are usually coated with poly-lysine or poly-amine to immobilize the DNA molecules on the surface.
- For **robotic spotting**, the robot dips its pins into solutions that contain the cDNA and then the tiny amounts of solution that adhere to the pins are transferred to the surface.
- Each pin produces a printed spot containing cDNA.
- Another method uses **ink-jet printing**, in which the cDNA is expelled from a small nozzle equipped with a piezoelectric fitting by applying electric current.
cDNA Microarrays

- A complete cDNA microarray is prepared by printing **thousands of cDNAs in an array format** on the glass slide, and these provide **gene-specific hybridization targets**.
- A schematic representation of the preparation, hybridization, image acquisition, and analysis for cDNA microarrays is shown in the figure.
- A digital image reflecting the abundance of different mRNAs in a sample is formed in a number of steps.
- RNA is extracted from the cells of interest, converted to cDNA, and then **amplified by a reverse transcriptase-polymerase chain reaction**.
- During the process, fluorescent molecules are attached to the DNA.
- If a specific mRNA molecule is produced by the cells, then the generated fluorescently labeled cDNA molecule will hybridize to its complementary single-stranded microarray probe.
Microarray flow chart

1. **Microarray Preparation**
   - Clones in 96-well plates
   - PCR Amplification
   - DNA Purification
   - Robotic printing
   - Poly-L-Lysine coated glass slide

2. **cDNA Probe Hybridization**
   - Reverse Transcription
   - Label with Fluor Dyes
   - UV-crosslink
   - Blocking Denature
   - Hybridize probe to microarray
   - Red probe
   - Green probe

3. **Confocal Microscope**
   - Photonmultiplier Tube
   - PMT
   - Red channel
   - Barrier filter
   - Dichroic mirror
   - Pinhole
   - Excitation Lasers
   - X-Y stage

4. **Expression Analysis**
   - Array Database
   - Computer Analysis
cDNA Microarrays

- The cDNA molecules that do not find their complementary single-stranded DNA sequences on the microarray are removed in a washing step.
- Since the fluorescent tags are attached to the cDNA strands that hybridize, the corresponding spots will fluoresce when provided fluorescence excitation energy and be detected at the level of emitted light.
- This yields a digital image whose intensities reflect levels of measured fluorescence, which in turn reflect mRNA abundances.
- In practice, it is commonplace to label mRNA molecules from distinct sources with different fluorescent tags and then co-hybridize them onto each arrayed gene.
- Two monochrome images are obtained from laser excitations at two different wavelengths.
cDNA Microarrays

- Monochrome images of the intensity for each fluorescent label are combined by placing each image in the appropriate color channel of an RGB image.
- In this composite image, one can visualize the differential expression of genes in the two cell types, the test sample typically being placed in the red channel, with the reference sample in the green channel.
- Intense red fluorescence at a spot indicates a high level of expression of that gene in the test sample relative to the reference sample.
- Conversely, intense green fluorescence at a spot indicates relatively low expression of that gene in the test sample compared to the reference.
cDNA Microarrays

- When both test and reference samples express a gene at similar levels, the observed array spot is yellow.
- Assuming that specific DNA products from two samples have an equal probability of hybridizing to the specific target, the fluorescent intensity measurement is a function of the amount of specific RNA available within each sample, provided samples are well-mixed and there is sufficiently abundant cDNA deposited at each target location.
- Ratios or direct intensity measurements of gene-expression levels between the samples can be used to detect meaningfully different expression levels between the samples for a given gene.
cDNA Microarrays

- When using cDNA microarrays, the signal must be extracted from the background.
- This requires image processing to extract signals, variability analysis, and measurement quality assessment.
- The objective of the microarray image analysis is to extract probe intensities or ratios at each cDNA target location and then cross-link printed clone information so that biologists can easily interpret the outcomes and high-level analysis can be performed.
- A microarray image is first segmented into individual cDNA targets, either by manual interaction or an automated algorithm.
- For each target, the surrounding background fluorescent intensity is estimated, along with the exact target location, fluorescent intensity and expression ratios.
cDNA Microarrays

- As with any random data, understanding variation is important to microarray data analysis.
- A major impediment to an effective understanding of variation is the large number of sources of variance inherent in microarray measurements.
- In many statistical analysis publications, the measured gene expression data are assumed to have multiple noise sources: noise due to sample preparation, labeling, hybridization, background fluorescence, different arrays, fluorescent dyes, and different printing locations.
- Various approaches have been taken to quantify and treat the noise levels in a set of experiments including log-transformed signal-plus-noise ANOVA models, mixture models, multiplicative models, ratio-distribution models, rank-based models less sensitive to noise distributions, replicates using mixed models, and quantitative noise analysis.
Normalization

• Besides variation due to random effects, such as biochemical and scanner noise, simultaneous measurement of mRNA expression levels via cDNA microarrays involves variation owing to system sources, including labeling bias, imperfections due to spot extraction, and cross hybridization.

• Even with the development of good extraction algorithms and the use of control probes at the array printing stage to aid in accounting for cross hybridization, we are left with labeling bias resulting from the fluorescent tags as systemic error.

• Although different experimental designs target different profiling objectives, be it global cancer tissue profiling or a single induction experiment with one gene perturbed, normalization to correct labeling bias is a common preliminary step before further statistical or computational analysis is applied.
Normalization

• The objective of this preliminary step is to reduce the variation between arrays.
• Normalization is usually implemented for an individual array and is then called *intra-array* normalization, which is what we consider here.
• Assessment of the effectiveness of normalization has mainly been confined to the ability to detect differentially expressed genes.
• It has also been shown that normalization can benefit *classification accuracy*; the significance of this benefit depending on the bias properties.
Normalization

• The simplest and most commonly used normalization is the *offset method*.
• To describe it, let the red and green channel intensities of the $k$th gene be $r_k$ and $g_k$, respectively.
• In many cases these are background-subtracted intensities.
• In an ideal case where two identical biological samples are labeled and co-hybridized to the array, we expect the sum of the log-transformed ratios to be 0.
• However, due to various reasons (dye efficiency, scanner PMT control, etc.), this assumption may not be true.
• If we assume that the two channels are equivalent, except for a signal amplification factor, then the ratio of the $k$th gene, $t_k$, can be calculated as follows:
\[
\log t_k = \log \left( \frac{r_k}{g_k} \right) - \frac{1}{N_q} \sum_{k=1}^{N_q} \log \left( \frac{r_k}{g_k} \right)
\]

- The second term in this equation is a constant offset that simply shifts the \( r_k \) vs. \( g_k \) scatter plot to a 45° diagonal line intersecting the origin and \( N_q \) is the number of probes that have measurement quality score of 1.
- In some cases the scatter plot may not be perfectly at a 45° diagonal line due to the difference when the scanner’s two channels may operate at different linear characteristic regions.
- In this case, full linear regression, instead of requiring the line to intersect at the origin, may be necessary.
- This is achieved by finding coefficients \( a \) and \( b \) to minimize the expectation \( E[(g_k-(a r_k-b))^2] \).
Normalization

- Some microarray expression levels may have large dynamic range that will cause systematic scanner deviations such as nonlinear response at lower intensity range and saturation at higher intensity.
- Although data falling into these ranges are commonly discarded for further analysis, the transition range, without proper handling, may still cause some significant error in differential expression gene detection.
- To account for this deviation, locally weighted linear regression (lowess) is regularly employed as a normalization method for such intensity-dependent effects.
- A newer method, loess, is essentially the same algorithm except that it uses a quadratic polynomial to fit the piecewise segments.
- The figure shows the effects of the three normalizations.
Effects of normalization: the left-hand scatter plot shows the regression lines and the red scatter plots show the normalized scatter plots, offset, linear, and lowess, from left to right.
Ratio Analysis

• It may be that the two channels in a cDNA microarray represent two sources of mRNA to be compared or that the red channel corresponds to the source of interest while the green channel serves as a reference channel.
• In either case, a basic question is whether the red intensity is significantly greater than or less than the green intensity.
• Such a question is naturally approached in the framework of hypothesis tests.
• For a microarray having \( n \) genes, with red and green fluorescent expression values labeled by \( R_1, R_2, \ldots, R_n \) and \( G_1, G_2, \ldots, G_n \), respectively, the issue is whether \( R_k \) is over- or under-expressed relative to \( G_k \).
Ratio Analysis

- Letting $\mu R_k$ and $\sigma R_k$ denote the mean and standard deviation of $R_k$ (similarly for $G_k$), the relevant hypothesis test is:

$H_0 : \mu R_k = \mu G_k$

$H_1 : \mu R_k \neq \mu G_k$

using the ratio test statistic $T_k = R_k/G_k$.

- There is biological plausibility for assuming that the coefficient of variation is approximately constant across the microarray.

- If we make the simplifying assumption of a constant coefficient of variation, meaning that

$\sigma R_k = c \mu R_k$

$\sigma G_k = c \mu G_k$

where $c$ denotes the common coefficient of variation (cv), then under the null hypothesis $H_0$, this equation implies that $\sigma R_k = \sigma G_k$. 

Ratio Analysis

• Assuming $R_k$ and $G_k$ to be normally and identically distributed, $T_k$ has the density function

$$f_{T_k}(t; c) = \frac{(1 + t)\sqrt{1 + t^2}}{c(1 + t^2)^2\sqrt{2\pi}} \exp\left[\frac{-(t - 1)^2}{2c^2(1 + t)}\right]$$

• Owing to the constant-$cv$ assumption, the subscript $k$ does not appear on the right-hand side of the equation.

• Hence, the density function holds for all genes and all ratios satisfying the null hypothesis can be pooled to estimate the parameter of the above equation.
Ratio Analysis

- The estimate is given by

\[
\hat{c} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{(t_i - 1)^2}{(t_i^2 + 1)}}
\]

where \( t_1, t_2, \ldots, t_n \) are ratio samples taken from a family of housekeeping genes on a microarray.
Ratio Analysis

• For a microarray derived from two identical mRNA samples co-hybridized on one slide (self-self experiment), the parameter $c$ (cv of the fluorescent intensity) provides the variation of assay.

• However, for any experiment, to guarantee the null hypothesis condition is not always possible.

• One alternative is to duplicate some or all clones where the same expression ratio is expected.

• For the ratio, $T$, of expression ratios, it can be shown that

$$\sigma_{\log T}^2 \approx 4c^2$$

when the measurement of the log-transformed expression level is approximately normally distributed.

• For any given experiment with some duplicated clones, $\sigma_{\log T}^2$ is easily calculated, and along with it the coefficient of variation of the assay.
Ratio Analysis

• In practical applications, a constant amplification gain $m$ may apply to one signal channel, in which case the null hypothesis may become $\mu R_k = m\mu G_k$.

• Under this uncalibrated signal detection setting, the ratio density can be modified by

$$f_T(t; c, m) = (1/m) f_T(t/m; c, 1)$$

• To decide whether expression levels of a gene from two samples are significantly different, we would like to find a confidence interval such that within the confidence interval, the null hypothesis cannot be rejected: the expression ratio, $T_k = R_k/G_k$, of the gene under consideration is not significantly deviated from 1.0 if the ratio is within the confidence interval.

• The confidence interval can be evaluated by integrating the ratio density function.
Ratio Analysis

- Since the confidence interval is determined by the parameter $c$, one can either use the parameter derived from preselected housekeeping genes, or a set of duplicated genes if they are available in the array.
- The former confidence intervals contain some levels of variation from the fluctuation of the biological system that also affect the housekeeping genes, while the latter contains no variation of the biological fluctuation, but contains possible spot-to-spot variation.
- Spot-to-spot variation is not avoidable if one wishes to repeat the experiment.
- The confidence interval derived from the duplicated genes is termed as the confidence interval of the assay.
- The constant-$cv$ condition permits all ratios satisfying the null hypothesis to be pooled to estimate the parameter.
- However, studies indicate that the $cv$ varies with the expression intensities.
Ratio Analysis

• The constant-$cv$ condition is made under the assumption that $R_k$ and $G_k$ are the expression levels detected from two fluorescent channels.

• The latter assumption is based on the assumption that image processing successfully extracts the true signal.

• This proves to be quite accurate for strong signals, but problems arise when the signal is weak in comparison to the background.

• The following figure illustrates how in many real-world situations the weaker expression signals (at the lower left corner of the scatter plot) produce a larger spread of gene placement.

• Even with image processing, the actual expression intensity measurement is of the form

$$ R_k = (SR_k + BR_k) - \mu BR_k $$

where $SR_k$ is the expression intensity measurement of gene $k$, $BR_k$ is the fluorescent background level, and $\mu BR_k$ is the mean background level.
Dispersion owing to weak signals
Ratio Analysis

- The measurable quantities are
  (1) signal with background, $SR_k + BR_k$
  (2) the surrounding background.
- The null hypothesis of interest is $\mu_{SR_k} = \mu_{SG_k}$.
- Taking the expectation in the previous equation yields
  $$\mu_{R_k} = E[R_k] = E[(SR_k + BR_k) - \mu_{BR_k}] = \mu_{SR_k}$$
- Since $\mu_{G_k} = \mu_{SG_k}$, the hypothesis test is still the one with which we are concerned, and we still apply the test statistic $T_k$.
- There is, however, a major difference.
- The assumption of a constant $cv$ applies to $SR_k$ and $SG_k$, not to $R_k$ and $G_k$, and the density function is not applicable.
- The ratio analysis performed under the constant-$cv$ condition does not apply in these new circumstances.
Synthetic Oligonucleotide Arrays

- A different approach than that of immobilizing cDNA strands copied from particular mRNA transcripts onto a surface is to fabricate synthetic polynucleotides onto the surface.
- For the popular Affymetrix chip, photolithography and solid-phase DNA synthesis are used to synthesize desired polynucleotide probes at specified coordinates on the slide.
- Since these synthetic probes are much shorter than the genes on a cDNA microarray, they must be designed in such a way that their collective action provides accurate measurement of transcript targets.
- Here we are discussing probes consisting of only 25 bases (25-mers) and a single 25-mer may pair with a large number of mRNAs.
Synthetic Oligonucleotide Arrays

• For each mRNA of interest, there is a set of non-overlapping (almost non-overlapping) 25-mer oligonucleotides to provide sequence-specific detection that can lead to accurate abundance measurement at the next stage of the process.
• Probes must be designed so that the final measurement for a particular gene is not unduly affected by the fact that the short probe will hybridize to transcripts from other genes and other kinds of RNA in the sample.
• The basic approach is to have, for each mRNA to be detected, a probe set for which different probes hybridize to different regions of the mRNA.
• In this way the measurement for a specific mRNA is obtained via averaging across a collection of probes, thereby mitigating the effects of noise, outliers, and cross-hybridization.
Synthetic Oligonucleotide Arrays

- To further mitigate the effects of cross-hybridization, two sets of probes can be used.
- *Perfect-match (PM)* probes are the ones designed to perfectly match selected nucleotide sequences in the target mRNA, whereas *mismatch (MM)* probes are derived from the PM probes by changing one base at a central location.
- If a PM probe is aimed at identifying a particular mRNA by matching a string of bases in the mRNA, then it is more probable that the mRNA will hybridize to the PM probe rather than to the MM probe derived from it.
- It is argued that subtraction of the intensity for the MM probe accomplishes a correction for both background noise and cross-hybridization.
Synthetic Oligonucleotide Arrays

• With this in mind, the expression level for a probe set on a microarray can be estimated by the *average difference*,

\[ AvDiff = \frac{1}{|A|} \sum_{k \in A} PM_k - MM_k \]

where \( A \) is the subset of all probes \( k \) such that \( PM_k - MM_k \) is within 3 standard deviations of the trimmed average of all differences \( PM_j - MM_j \), the smallest and largest difference being left out of the average, and where \( |A| \) is the number of probes in \( A \).

• Alterations of this basic average difference have been proposed.

• Various problems have been pointed out regarding the average difference, including the fact that MM often exceeds PM.
Classification
Classification

- **Pattern classification** plays an important role in genomic signal processing.
- For instance, cDNA microarrays can provide expression measurements for thousands of genes at once, and a key goal is to perform classification via different expression patterns — between different kinds of cancer, different stages of tumor development.
- This requires designing a classifier (decision function) that takes a vector of gene expression levels as input, and outputs a class label that predicts the class containing the input vector.
- Classifiers are designed from a sample of expression vectors.
- This involves assessing expression levels from RNA obtained from the different tissues with microarrays, determining genes whose expression levels can be used as classifier features (variables), and then applying some rule to design the classifier from the sample microarray data.
Classification

- Design, performance evaluation, and application must take this randomness due to both biological and experimental variability.
- Three critical issues arise:
  1. Given a set of features, how does one design a classifier from the sample data that provides good classification over the general population?
  2. How does one estimate the error of a designed classifier when data are limited?
  3. Given a large set of potential features, such as the large number of expression levels provided by each microarray, how does one select a set of features as the input to the classifier?
- Small samples (relative to the number of features) are ubiquitous in genomic signal processing and impact all three issues.
Classifier Design

• Classification involves a *feature vector* \( \mathbf{X} = (X_1, X_2, \ldots, X_d) \) on \( d \)-dimensional Euclidean space \( \mathbb{R}^d \) composed of random variables (*features*), a binary random variable \( Y \), and a *classifier* \( \psi : \mathbb{R}^d \rightarrow \{0, 1\} \) to serve as a predictor of \( Y \), which means that \( Y \) is to be predicted by \( \psi(\mathbf{X}) \).

• The values, 0 or 1, of \( Y \) are treated as class *labels*.

• The error, \( \varepsilon[\psi] \), of \( \psi \) is the probability that the classification is erroneous, namely, \( \varepsilon[\psi] = P(\psi(\mathbf{X}) \neq Y) \).

• It equals the expected (mean) absolute difference, \( E[|Y - \psi(\mathbf{X})|] \), between the label and the classification.

• Owing to the binary nature of \( \psi(\mathbf{X}) \) and \( Y \), it also equals the mean-square error.
Bayes Classifier

- An optimal classifier, $\psi_d$, is one having minimal error, $\varepsilon_d$, among all binary functions on $\mathbb{R}^d$, so that it is the minimal mean-absolute-error predictor of $Y$.
- $\psi_d$ and $\varepsilon_d$ are called the Bayes classifier and Bayes error, respectively.
- Classification accuracy, and thus the error, depends on the probability distribution of the feature-label pair $(X, Y)$.
- The posterior distribution for $X$ is defined by $\eta(x) = f_{X,Y}(x, 1)/f_X(x)$, where $f_{X,Y}(x, y)$ and $f_X(x)$ are the densities for $(X, Y)$ and $X$, respectively.
- $\eta(x)$ gives the probability that $Y = 1$, given $X = x$.
- In this binary setting, $\eta(x) = E[Y | x]$ is the conditional expectation of $Y$ given $x$. 


Bayes Classifier

- The error of an arbitrary classifier can be expressed as

\[
\varepsilon[\psi] = \int_{\{x : \psi(x) = 0\}} \eta(x) f_X(x) \, dx + \int_{\{x : \psi(x) = 1\}} (1 - \eta(x)) f_X(x) \, dx
\]

- The class conditional probability of \( Y \) given \( x \) is defined by \( P(Y = 1 | x) = \eta(x) \).

- As a function of \( X \), \( P(Y = 1 | X) \) is a random variable dependent on \( f_X(x) \).

- Since \( 0 \leq \eta(x) \leq 1 \), the right-hand side of the above equation is minimized by

\[
\psi_d(x) = \begin{cases} 
0, & \text{if } P(Y = 1 | x) \leq 1/2 \\
1, & \text{if } P(Y = 1 | x) > 1/2 
\end{cases}
\]

\( \psi_d(x) \) is defined to be 0 or 1 according to whether \( Y \) is less or more likely to be 1 given \( x \).
Bayes Classifier

- It follows from the above two equations that the Bayes error is given by

\[ \varepsilon_d = \int_{\{x: \eta(x) \leq 1/2\}} \eta(x) f_X(x) dx + \int_{\{x: \eta(x) > 1/2\}} (1 - \eta(x)) f_X(x) dx \]

- The problem with the Bayes classifier is that we typically do not know the class conditional probabilities, and therefore must design a classifier from sample data.
- An obvious approach would be to estimate the conditional probabilities from data, but often we do not have sufficient data to obtain good estimates.
- Moreover, good classifiers can be obtained even when we lack sufficient data for satisfactory density estimation.
Classification Rules

- Design of a classifier $\psi_n$ from a random sample $S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ of vector-label pairs drawn from the feature-label distribution requires a classification rule that operates on random samples to yield a classifier.

- A classification rule is a mapping $\Psi_n: [\mathbb{R}^d \times \{0, 1\}]^n \rightarrow F$, where $F$ is the family of $\{0, 1\}$-valued functions on $\mathbb{R}^d$.

- Given a sample $S_n$, we obtain a designed classifier $\psi_n = \Psi_n(S_n)$ according to the rule $\Psi_n$.

- The Bayes error $\varepsilon_d$ is estimated by the error $\varepsilon_n$ of $\psi_n$. There is a design cost $\Delta_n = \varepsilon_n - \varepsilon_d$, $\varepsilon_n$ and $\Delta_n$ being sample-dependent random variables.

- The expected design cost is $E[\Delta_n]$, the expectation being relative to all possible samples.

- The expected error of $\psi_n$ is decomposed according to

$$E[\varepsilon_n] = \varepsilon_d + E[\Delta_n]$$
Classification Rules

- Asymptotic properties of a classification rule concern large samples (as \( n \to \infty \)).
- A rule is said to be consistent for a distribution of \((X, Y)\) if \(\Delta_n \to 0\) in the mean, meaning \(E[\Delta_n] \to 0\) as \(n \to \infty\).
- For a consistent rule, the expected design cost can be made arbitrarily small for a sufficiently large amount of data.
- Since rarely is there a good estimate of the distribution in pattern recognition, rules for which convergence is independent of the distribution are desirable.
- A classification rule is universally consistent if \(\Delta_n \to 0\) in the mean for any distribution of \((X, Y)\).
- Universal consistency is useful for large samples, but has little consequence for small samples.
Classification Rules

• We will discuss a few well-known and well-studied classification rules, our intent being to use them to illustrate methodology and issues.
• For the basic nearest-neighbor (NN) rule, $\psi_n$ is defined for each $x \in \mathbb{R}^d$ by letting $\psi_n(x)$ take the label of the sample point closest to $x$.
• For the NN rule, no matter the distribution of $(X, Y)$,
  \[ \varepsilon_d \leq \lim_{n \to \infty} E[\varepsilon_n] \leq 2\varepsilon_d. \]
• It follows that $\lim_{n \to \infty} E[\Delta_n] \leq \varepsilon_d$.
• Hence, the asymptotic expected cost of design is small if the Bayes error is small; however, this result does not give consistency.
• More generally, for the $k$-nearest-neighbor ($k$NN) rule, $k$ odd, the $k$ points closest to $x$ are selected and $\psi_n(x)$ is defined to be 0 or 1 according to which is the majority among the labels of these points.
• If $k = 1$, this gives the NN rule. We will not consider even $k$. 
Classification Rules

• The limit of $E[\varepsilon_n]$ as $n \to \infty$ can be expressed analytically and various upper bounds exist.
• In particular, $\lim_{n \to \infty} E[\Delta_n] \leq (ke)^{-1/2}$.
• This does not give consistency, but it does show that the design cost gets arbitrarily small for sufficiently large $k$ as $n \to \infty$.
• The $k$NN rule is universally consistent if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.
• A classification rule can yield a classifier that makes very few, or no, errors on the sample data on which it is designed, but performs poorly on the distribution as a whole, and therefore on new data to which it is applied.
• This situation is exacerbated by complex classifiers and small samples.
• If the sample size is dictated by experimental conditions, such as cost or the availability of patient RNA for expression microarrays, then one only has control over classifier complexity.
Classification Rules

• The situation with which we are concerned is typically referred to as *overfitting*.
• The basic point is that a classification rule should not cut up the feature space in a manner too complex for the amount of sample data available.
• The problem is not necessarily mitigated by applying an error estimation rule to the designed classifier to see if it “actually” performs well since when there is only a small amount of data available, error-estimation rules are very imprecise, and the imprecision tends to be worse for complex classification rules.
• Hence, a low error estimate is not sufficient to fully overcome our expectation of a large design error when using a complex classifier with a small data set.
Classification Rules

- Overfitting is illustrated in the following figure, in which the 3NN rule is applied to two equal-variance circular Gaussian class conditional distributions.
- Parts (a) and (b) show the 3NN classifier for two 30-point samples and parts (c) and (d) show the 3NN classifier for two 90-point samples.
- Note the greater overfitting of the data for the 30-point samples, in particular, the greater difference in the two 30-point designed classifiers as compared to the difference between the two 90-point classifiers, and the closer the latter are to the Bayes classifier given by the vertical line.
3NN classification applied to two equal-variance circular Gaussian class conditional distributions: (a) for a 30-point sample; (b) for a second 30-point sample; (c) for a 90-point sample; (d) for a second 90-point sample.
Constrained Classifier Design

• To mitigate overfitting, we need to use constrained classification rules.
• Constraining classifier design means restricting the functions from which a classifier can be chosen to a class $C$.
• This leads to trying to find an optimal constrained classifier, $\psi_C \in C$, having error $\varepsilon_C$.
• Constraining the classifier can reduce the expected design error, but at the cost of increasing the error of the best possible classifier.
• Since optimization in $C$ is over a subclass of classifiers, the error, $\varepsilon_C$, of $\psi_C$ will typically exceed the Bayes error, unless the Bayes classifier happens to be in $C$.
• This cost of constraint (approximation) is

$$\Delta_C = \varepsilon_C - \varepsilon_d$$
Constrained Classifier Design

- A classification rule yields a classifier $\psi_{n,C} \in C$ with error $\varepsilon_{n,C}$, and $\varepsilon_{n,C} \geq \varepsilon_C \geq \varepsilon_d$.
- Design error for constrained classification is
  $$\Delta_{n,C} = \varepsilon_{n,C} - \varepsilon_C$$
- For small samples, this can be substantially less than $\Delta_n$, depending on $C$ and the rule.
- The error of the designed constrained classifier is decomposed as
  $$\varepsilon_{n,C} = \varepsilon_d + \Delta_C + \Delta_{n,C}$$
- The expected error of the designed classifier from $C$ can be decomposed as
  $$E[\varepsilon_{n,C}] = \varepsilon_d + \Delta_C + E[\Delta_{n,C}]$$
- The constraint is beneficial if and only if $E[\varepsilon_{n,C}] < E[\varepsilon_n]$, which means
  $$\Delta_C < E[\Delta_n] - E[\Delta_{n,C}]$$
Constrained Classifier Design

- If the cost of constraint is less than the decrease in expected design cost, then the expected error of $\psi_{n,C}$ is less than that of $\psi_n$.
- Choosing a complex classifier that can tightly fit the data reduces the constraint cost but increases the expected design error $E[\Delta_{n,C}]$, whereas choosing a simple classifier that does not overfit the data increases the constraint cost but decreases the expected design error.
- In most microarray experiments the sample size is very small and design error is such a problem that simple classifiers are almost always preferable.
- The matter can be graphically illustrated.
- For some rules, $E[\Delta_n]$ is nonincreasing, meaning that $E[\Delta_{n+1}] \leq E[\Delta_n]$.
- This means that the expected design error never increases as sample sizes increase, and it holds for any feature-label distribution. Such classification rules are called smart.
Constrained Classifier Design

- The nearest-neighbor rule is not smart because there exist distributions for which $E[\Delta_{n+1}] \leq E[\Delta_n]$ does not hold for all $n$.
- Now consider a consistent rule, constraint, and distribution for which $E[\Delta_{n+1}] \leq E[\Delta_n]$ and $E[\Delta_{n+1},c] \leq E[\Delta_n,c]$.
- The figure illustrates the design problem.
- The axes correspond to sample size and error. The horizontal dashed lines represent $\varepsilon_C$ and $\varepsilon_d$, the decreasing solid lines represent $E[\varepsilon_{n,C}]$ and $E[\varepsilon_n]$.
- If $n$ is sufficiently large, then $E[\varepsilon_n] < E[\varepsilon_{n,C}]$; however, if $n$ is sufficiently small, then $E[\varepsilon_n] > E[\varepsilon_{n,C}]$.
- The point $N_0$ at which the decreasing lines cross is the cut-off: for $n > N_0$, the constraint is detrimental; for $n < N_0$, it is beneficial.
- If $n < N_0$, then the advantage of the constraint is the difference between the decreasing solid lines.
Errors for unconstrained and constrained classification as a function of sample size.
Quadratic Discriminant Analysis

- A classical example of a constrained classification rule is quadratic discriminant analysis (QDA).
- To discuss it, let $R_k$ denote the region in $\mathbb{R}^d$ where the Bayes classifier has the value $k$, for $k = 0, 1$.
- As we have seen earlier, $x \in R_k$ if $f_{X,Y}(x, k) \geq f_{X,Y}(x, j)$ for $j \neq k$.
- Since $f(x, k) = f(x|k)f(k)$, upon taking the logarithm, this is equivalent to $x \in R_k$ if $d_k(x) \geq d_j(x)$, where the discriminant $d_k(x)$ is defined by
  \[d_k(x) = \log f(x|k) + \log f(k)\]
- If the conditional densities $f(x|0)$ and $f(x|1)$ are normally distributed, then
  \[f(x|k) = \frac{1}{\sqrt{(2\pi)^n \det[K_k]}} \exp \left[-\frac{1}{2}(x - u_k)'K_k^{-1}(x - u_k)\right]\]
  where $K_k$ and $u_k$ are the covariance matrix and mean vector, respectively.
Quadratic Discriminant Analysis

- The discriminant becomes

\[ d_k(x) = -(x - u_k)'K_k^{-1}(x - u_k) - \log(\det[K_k]) + 2 \log f(k) \]

- The form of this equation shows that the decision boundary \( d_k(x) = d_j(x) \) is quadratic — thus, QDA.

- If both conditional densities possess the same covariance matrix \( K \), then \( \log (\det [K_k]) \) can be dropped from \( d_k(x) \) and

\[ d_k(x) = -(x - u_k)'K_k^{-1}(x - u_k) + 2 \log f(k) \]

which is a linear function of \( x \) and produces hyperplane decision boundaries.

- If we assume that the classes possess equal prior probabilities \( f(k) \), then the logarithm term can be dropped.
Linear Discriminant Analysis

- The discriminant characterizes *linear discriminant analysis (LDA)*.
- This classifier is of the form

\[
\psi(x) = T\left[a_0 + \sum_{i=1}^{d} a_i x_i\right]
\]

where \(x = (x_1, x_2, \cdots, x_d)\) and \(T\) thresholds at 0 and yields \(-1\) or \(1\).
- It divides the space into two half-spaces determined by the hyperplane defined by the parameters \(a_0, a_1, \ldots, a_d\).
- The hyperplane is determined by the equation formed from setting the linear combination equal to 0.
- Equations for quadratic and linear discriminant analysis are derived under the *Gaussian* assumption, but in practice can perform well so long as the class conditional densities are approximately Gaussian.
QDA and LDA

• Owing to the greater number of parameters to be estimated for QDA as opposed to LDA, one can proceed with smaller samples with LDA than with QDA.

• Although QDA and LDA are derived under certain assumptions, they are often applied in situations where the assumptions are violated in order to achieve parametric quadratic or linear classifiers.

• In such circumstances, their performance depends on the actual feature-label distribution and the sample size. LDA appears to be more robust relative to the underlying assumptions than QDA.
Support Vector Machine (SVM)

- A popular classification rule to construct linear classifiers is the support vector machine (SVM).
- The figure shows a linearly separable data set and three hyperplanes.
- The outer lines pass through points in the sample data, and the third, called the maximal-margin hyperplane (MMH) is equidistant between the outer lines.
- It has the property that the distance from it to the nearest 1-labeled vector is equal to the distance from it to the nearest -1-labeled vector.
- The vectors closest to it are called support vectors.
- The distance from the MMH to any support vector is called the margin.
- The matter is formalized by recognizing that differently labeled sets are separable by the hyperplane $u \cdot x = c$, where $u$ is a unit vector and $c$ is a constant, if $u \cdot x_k > c$ for $y_k = 1$ and $u \cdot x_k < c$ for $y_k = -1$. 
Linear support vector machine
Support Vector Machine (SVM)

- For any unit vector $u$, the margin is defined by

$$\rho(u) = \frac{1}{2} \left( \min_{\{x_k : y_k = 1\}} u \cdot x_k - \max_{\{x_k : y_k = -1\}} u \cdot x_k \right)$$

- The MMH, which is unique, can be found by solving the following quadratic optimization problem: among the set of all vectors $v$ for which there exists a constant $b$ such that

$$v \cdot x_k + b \geq 1, \text{ if } y_k = 1$$
$$v \cdot x_k + b \leq -1, \text{ if } y_k = -1$$

find the vector of minimum norm $\|v\|$.  

- If $v_0$ satisfies this optimization problem, then the vector defining the MMH and the margin are given by $u_0 = v_0/\|v_0\|$ and $\rho(u_0) = \|v_0\| - 1$ respectively.
Support Vector Machine (SVM)

- If the sample is not linearly separable, then one has two choices: try to find a good linear classifier or to find a **nonlinear classifier**.
- In the first case, the preceding method can be modified.
- A second approach is to map the sample points into a higher dimensional space, find a hyperplane in that space, and then map back into the original space.
Regularization by Noise Injection

- We illustrate some results based on regularized linear classification in the case of breast tumors from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2, or from patients not expected to carry a hereditary predisposing mutation.

- Pathological and genetic differences appear to imply different but overlapping functions for BRCA1 and BRCA2, and in an early study involving expression-based classification, cDNA microarrays are used to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers.
Regularization by Noise Injection

• Using data from that study, the analytic noise-injection method has been applied to derive a linear classifier to discriminate BRCA1 tumors from BRCA2 and sporadic tumors.
• The figure shows the designed classifier based on genes KRT8 and DRPLA, along with its relation to the sample data pairs for these two genes.
• What can be inferred from this separation relative to population classification between the tumor classes depends on the method used to select the features KRT8 and DRPLA, the classification rule to construct the separating line, and the relationship between the estimated error based on the sample data and the true classifier error relative to the feature-label distribution.
Linear classification of BRCA1 versus BRCA2 and sporadic patients.
Feature Selection

• The **Bayes error is monotone**, meaning that if $A$ and $B$ are feature sets for which $A \subset B$, then $\varepsilon_B \leq \varepsilon_A$, where $\varepsilon_A$ and $\varepsilon_B$ are the Bayes errors corresponding to $A$ and $B$, respectively.

• Thus, relative to the Bayes error, if one has a large set of potential features, it is statistically safe to use all of the features when forming a classifier.

• However, if $\varepsilon_{A,n}$ and $\varepsilon_{B,n}$ are the corresponding errors resulting from designed classifiers on a sample of size $n$, then it cannot be asserted that $E[\varepsilon_{B,n}] \leq E[\varepsilon_{A,n}]$.

• Monotonicity does not apply to the expected errors of designed classifiers.

• It is commonplace for the expected error of designed classifiers to decrease and then increase for increasingly large feature sets. This is called the **peaking phenomenon**.
Peaking phenomenon
Feature Selection

• It is illustrated in the figure, where the horizontal axis corresponds to a sequence of features, $x_1, x_2, \ldots, x_d, \ldots$, and the vertical axis gives the error.

• For $d$ features, the Bayes error $\varepsilon_d$ continues to decline but the expected error, $E[\varepsilon_{d,n}]$, of the designed classifier decreases and then increases.

• One might hope that there is some way of avoiding checking all possible feature sets to find the best.

• Unfortunately, a classical result states that to be assured of finding the optimal $k$-element feature set from among $n$ features, one must check all $k$-element feature sets, unless one has some mitigating prior distributional knowledge, which is generally not the case in practical situations.

• A further problem is that, even were one able to check all feature sets, the need for error estimation can greatly impact finding the best one.
Feature Selection

• The most obvious approach to suboptimal feature selection is to consider each feature by itself and choose the $k$ features that perform individually the best.

• While easy, this method is subject to choosing a feature set with a large number of redundant features, thereby obtaining a feature set that is much worse than the optimal.

• Moreover, features that perform poorly individually may do well in combination with other features.

• A common approach to suboptimal feature selection is sequential selection, either forward or backward, and their variants.

• Sequential forward selection (SFS) begins with a small set of features, perhaps one, and iteratively builds the feature set.

• When there are $k$ features, $x_1, x_2, \ldots, x_k$, in the growing feature set, all feature sets of the form $\{x_1, x_2, \ldots, x_k, w\}$ are compared and the best one chosen to form the feature set of size $k + 1$. 
Sequential forward selection  
(SFS)

- A problem with SFS is that there is no way to delete a feature adjoined early in the iteration that may not perform as well in combination as other features.
- The SFS look-back algorithm aims to mitigate this problem by allowing deletion.
- For it, when there are \( k \) features, \( x_1, x_2, \ldots, x_k \), in the growing feature set, all feature sets of the form \( \{x_1, x_2, \ldots, x_k, w, z\} \) are compared and the best one chosen.
- All \((k + 1)\) element subsets are checked to allow the possibility of one of the earlier chosen features to be deleted, the result being the \( k + 1 \) features that will form the basis for the next stage of the algorithm.
- Flexibility can be added by considering sequential forward floating selection (SFFS), where the number of features to be adjoined and deleted is not fixed, but is allowed to “float”.
Feature Selection

• When selecting features via an algorithm like SFFS that employs error estimation within it, one should expect the choice of error estimator to impact feature selection, the degree depending on the classification rule and feature label distribution.

• In general, feature selection is an important and difficult problem, with many proposed algorithms but little theoretical support.

• For the most part, comparison of feature selection algorithms is via simulation in which algorithms are applied to particular feature-label distributions.

• Even if good feature sets exist, for small samples, the likelihood of finding one whose performance is close to optimal may be small.

• This is due to the difficulty of classifier design and inadequate error estimation. The problem is exacerbated for counting-based estimators like cross-validation on account of a multitude of ties among feature sets.
Error Estimation

- Error estimation is a key aspect of classification.
- If a classifier $\psi_n$ is designed from a random sample $S_n$, then the error of the classifier relative to a particular sample is given by
  \[ \varepsilon_n = E_F [ |Y - \psi_n(X)| ] \]
  where the expectation is taken relative to the feature-label distribution $F$.
- The expected error of the classifier over all samples of size $n$ is given by
  \[ E[\varepsilon_n] = E_{F_n} E_F [ |Y - \psi_n(X)| ] \]
  where the outer expectation is with respect to the joint distribution of the sample $S_n$.
- In practice the feature-label distribution is unknown and the expected error must be estimated.
Error Estimation

- If there is an abundance of sample data, then it can be split into *training* and *test* data.
- A classifier is designed on the training data, and its estimated error is the proportion of errors it makes on the test data.
- The problem with using both training and test data is that one would like to use all the data for design because increased data means decreased design cost.
- This is especially the case with small samples.
Error Estimation using the Training Data

- One approach is to use all sample data to design a classifier $\psi_n$, and estimate $\varepsilon_n$ by applying $\psi_n$ to the same data.
- The *resubstitution estimate*, $\tilde{\varepsilon}_n$, is the fraction of errors made by $\psi_n$ on the sample (training) data.
- Resubstitution is usually biased low, meaning $E[\varepsilon_n] \leq E[\tilde{\varepsilon}_n]$, and it is often severely low-biased when samples are small.
- As with all error estimators, performance depends on the classification rule.
- An extreme case is that the resubstitution estimate is always 0 for the nearest-neighbor classifier.
- A common way of performing error estimation using the same data on which the classifier is designed is to apply a re-sampling strategy.
**Cross-validation**

- *Cross-validation* is a re-sampling strategy in which classifiers are designed from parts of the sample, each is tested on the remaining data, and $\varepsilon_n$ is estimated by averaging the errors.
- In *k-fold cross-validation*, the sample $S_n$ is partitioned into $k$ folds $S_{(i)}$, for $i=1, 2, \ldots, k$.
- Each fold is left out of the design process and used as a test set, and the estimate, $\varepsilon_n^{cv(k)}$, is the average error committed on all folds.
- A $k$-fold cross-validation estimator is unbiased as an estimator of $E[\varepsilon_{n-n/k}]$, meaning $E[\varepsilon_n^{cv(k)}] = E[\varepsilon_{n-n/k}]$.
- The special case of $n$-fold cross-validation yields the *leave-one-out estimator*, $\hat{\varepsilon}_n^{loo}$, which is an unbiased estimator of $E[\varepsilon_{n-1}]$.
- While not suffering from severe low bias like resubstitution, cross-validation has large variance in small-sample settings and therefore its use is problematic.
Leave-one-out Estimation

- Focusing on the unbiasedness of leave-one-out estimation, $E[\hat{\varepsilon}_n^{\text{loo}}] = E[\varepsilon_{n-1}]$, so that $E[\hat{\varepsilon}_n^{\text{loo}} - \varepsilon_n] \approx 0$.
- Thus, the expected difference between the error estimator and the error is approximately 0.
- But we are not interested in the expected difference between the error estimator and the error.
- Rather, we are interested in the precision of the error estimator in estimating the error.
- Our concern is the expected deviation, $E[|\hat{\varepsilon}_n^{\text{loo}} - \varepsilon_n|]$, and unless the cross-validation variance is small, which it is not for small samples, this expected deviation will not be small.
Bootstrap Estimator

• *Bootstrap* is a general re-sampling strategy that can be applied to error estimation.
• A *bootstrap sample* consists of *n* equally-likely draws with replacement from the original sample *S*.
• Some points may appear multiple times, whereas others may not appear at all.
• For the basic bootstrap estimator, $\hat{\varepsilon}_n^b$, the classifier is designed on the bootstrap sample and tested on the points left out, this is done repeatedly, and the bootstrap estimate is the average error made on the left-out points.
• $\hat{\varepsilon}_n^b$ tends to be a high-biased estimator of $E[\varepsilon_n]$, since the number of points available for design is on average only $0.632n$.
• The *.632 bootstrap estimator* tries to correct this bias via a weighted average of $\hat{\varepsilon}_n^b$ and resubstitution

$$\hat{\varepsilon}_n^{b,632} = 0.368\hat{\varepsilon}_n^{res} + 0.632\hat{\varepsilon}_n^b$$
Bolstered-Resubstitution estimator

- In resubstitution there is no distinction between points near and far from the decision boundary;
- The *bolstered-resubstitution estimator* is based on the heuristic that, relative to making an error, more confidence should be attributed to points far from the decision boundary than points near it.
- This is achieved by placing a distribution, called a *bolstering kernel*, at each point and estimating the error by integrating the bolstering kernel for each point over the decision region that the point should not belong to and then *summing the integrals* (rather than simply counting the misclassified points as in resubstitution).
- The procedure is illustrated in the figure for a linear classifier.
- A key issue is the amount of bolstering (spread of the bolstering kernels).
Bolstering error for linear classification
Bolstered-Resubstitution estimator

• Since the purpose of bolstering is to improve error estimation in the small-sample setting, we need to be cautious about using bolstering kernels that require complicated inferences.
• Hence, zero-mean, spherical bolstering kernels with covariance matrices of the form $\zeta_i I$ are commonly employed.
• The choice of the parameters $\sigma_1, \sigma_2, \ldots, \sigma_n$ determines the variance and bias properties of the corresponding bolstered estimator.
• If $\sigma_1 = \sigma_2 = \ldots = \sigma_n = 0$, then there is no bolstering and the bolstered estimator reduces to the original estimator.
• As a general rule, larger $\sigma_i$ lead to lower-variance estimators, but after a certain point this advantage is offset by increasing bias.
• In situations where resubstitution is very low-biased, it may not be a good idea to spread the incorrectly classified points. This approach yields the semi-bolstered resubstitution estimator.
Clustering
Clustering

- A cluster operator takes a set of data points and partitions the points into clusters (subsets).
- Clustering has become a popular data-analysis technique in genomic studies using gene-expression microarrays.
- **Time-series clustering** groups together genes whose expression levels exhibit similar behavior through time.
- Similarity indicates possible co-regulation.
- Another way to use expression data is to take expression profiles over various tissue samples, and then cluster these samples based on the expression levels for each sample.
- This approach offers the potential to discriminate pathologies based on their differential patterns of gene expression.
Classification

- **Classification** exhibits two fundamental characteristics:
  1. classifier error can be estimated under the assumption that the sample data arise from an underlying feature-label distribution;
  2. given a family of classifiers, sample data can be used to learn the optimal classifier in the family.

- Once designed, the **classifier represents a mathematical model** that provides a decision mechanism relative to real-world measurements.
- The model represents scientific knowledge to the extent that it has predictive capability.
- The purpose of testing (error estimation) is to quantify the worth of the model.
Clustering has historically lacked both fundamental characteristics of classification.

Many validation techniques have been proposed for evaluating clustering results.

These are generally based on the degree to which clusters derived from a set of sample data satisfy certain heuristic criteria.

This is significantly different than classification, where the error of a classifier is given by the probability of an erroneous decision.

The problem is illustrated in the following figure.
Two “reasonable” cluster results
Clustering

• Both cluster results appear “reasonable,” but the performance of a clustering algorithm cannot be measured by observing the results of a single application; rather, as with classification, its performance must be measured relative to predictive results on a distribution.

• For clustering, error estimation must assume that clusters resulting from a cluster algorithm can be compared to the correct clusters for the data set in the context of a probability distribution, thereby providing an error measure.

• The key to a general probabilistic theory of clustering, including both error estimation and learning, is to recognize that classification theory is based on operators on random variables, and that the theory of clustering needs to be based on operators on random sets.
Examples of Clustering Algorithms
$k$-means

- If we envision clusters formed by points $x$ generated by a random sample $S$ from a mixture of $m$ circular Gaussian conditional distributions, then the points $a_0^S, a_1^S, \ldots, a_{m-1}^S$ that minimize

$$
\rho_S(a_0, a_1, \ldots, a_{m-1}) = \frac{1}{|S|} \sum_{x \in S} \min_{0 \leq j \leq m-1} \| x - a_j \|^2
$$

are a reasonable choice for the centroids of the $m$ classes arising from the conditional distributions.

- Let $V = \{ V_1, V_2, \ldots, V_m \}$ be the Voronoi partition of $\mathbb{R}^d$ induced by $a_0^S, a_1^S, \ldots, a_{m-1}^S$: a point lies in $V_k$ if its distance to $a_k^S$ is no more than its distance to any other of the points $a_0^S, a_1^S, \ldots, a_{m-1}^S$.

- For *Euclidean-distance clustering*, the sample points are clustered according to how they fall into the Voronoi partition.
**k-means**

- Direct implementation of Euclidean-distance clustering is computationally prohibitive.
- A classical iterative approximation is given by the *k-means algorithm*, where *k* refers to the number of clusters provided by the algorithm.
- Each sample point is placed into a unique cluster during each iteration and the means are updated based on the classified samples.
- Given a sample *S* with *n* points to be placed into *k* clusters, initialize the algorithm with *k* means, \( m_1, m_2, \ldots, m_k \), among the points; for each point \( x \in S \), calculate the distance \( ||x - m_i|| \), for \( i = 1, 2, \ldots, k \); form clusters \( C_1, C_1, \ldots, C_k \) by placing \( x \) into \( C_i \) if \( ||x - m_i|| \leq ||x - m_j|| \) for \( j = 1, 2, \ldots, k \); update \( m_1, m_2, \ldots, m_k \) as the means of \( C_1, C_1, \ldots, C_k \), respectively; and repeat until the means do not change.
**k-means**

- At each stage of the algorithm, the clusters are determined by the **Voronoi diagram** associated with $m_1, m_2, \cdots, m_k$.
- Two evident problems with the $k$-means algorithm are the prior assumption on the number of means and the choice of means to seed the algorithm.
Fuzzy $k$-means

- Eq. 1 can be rewritten as

$$\rho_S(a_0, a_1, \cdots, a_{m-1}) = \frac{1}{|S|} \sum_{i=1}^{n} \sum_{j=0}^{m-1} P(C_j|\mathbf{x}_i)^b \|\mathbf{x}_i - a_j\|^2$$

where $P(C_j|\mathbf{x}_i)$ is the probability that $\mathbf{x}_i \in C_j$, which is either 0 or 1, and is 1 only for the minimizing $j$ of Eq. 1.

- A fuzzy approach results from letting the conditional probabilities reflect uncertainty, so that cluster inclusion is not crisp, and letting $b > 0$ be a parameter affecting the degree to which a point can belong to more than a single cluster.

- The conditional probabilities are constrained by the requirement that their sum is 1 for any fixed $\mathbf{x}_i$,

$$\sum_{j=0}^{m-1} P(C_j|\mathbf{x}_i) = 1$$
Fuzzy $k$-means

- Let $p_j$ denote the prior probability of $C_j$. Since the conditional probabilities $P(C_j | x_i)$ are not estimable and are heuristically set, we view them as fuzzy membership functions.
- In this case, for the minimizing values of $\rho_S(a_0, a_1, \ldots, a_{m-1})$, the partial derivatives with respect to $a_j$ and $p_i$ satisfy $\partial \rho_S / \partial a_j = 0$ and $\partial \rho_S / \partial p_i = 0$.
- These partial-derivative identities yield

$$
\mathbf{m}_j = \frac{\sum_{i=1}^{n} P(C_j | x_i)^b x_i}{\sum_{i=1}^{n} P(C_j | x_i)^b}
$$

$$
P(C_j | x_i) = \frac{\| x_i - \mathbf{m}_j \|^{-1/(b-1)}}{\sum_{l=1}^{k} \| x_i - \mathbf{m}_l \|^{-1/(b-1)}}
$$
**Fuzzy k-means**

- These lead to the *fuzzy k-means* iterative algorithm.
- Initialize the algorithm with \( b, k \) means \( m_1, m_2, \ldots, m_k \), and the membership functions \( P(C_j | x_i) \) for \( j = 1, 2, \ldots, k \) and \( i = 1, 2, \ldots, n \), where the membership functions must be normalized so that their sum is 1 for any fixed \( x_i \).
- Re-compute \( m_j \) and \( P(C_j | x_i) \) by the previous two equations.
- Repeat until there are only small pre-specified changes in the means and membership functions.
- The intent of fuzzifying the \( k \)-means algorithm is to keep the means from getting “stuck” during the iterative procedure.
Self-Organizing Maps

- **Self-organizing maps** provide a different extension of the k-means concept.
- The idea is to map high-dimensional vectors in Euclidean space to a low-dimensional grid in a neighborhood-preserving manner, whereby we mean that vectors that are close in the high-dimensional space have close representations in the low-dimensional grid.
- To describe self-organizing maps, we begin with an ordered lattice, typically a one- or two-dimensional grid, that we index by $I = \{1, 2, \cdots, k\}$.
- Associated with $I$ is a neighborhood function $\eta_t$, defined on $I \times I$ and parameterized by $t = 0, 1, \cdots$, satisfying three properties:
  1. $\eta_t$ depends only on the distance between points in $I$, meaning that $\eta_t(i, j) = \eta_t(||i−j||)$
  2. $\eta_t$ is non-increasing relative to distance, meaning $\eta_t(i−j) \leq \eta_t(||u−v||)$ if $||u−v|| \leq ||i−j||$
  3. the domain of $\eta_t$ is non-increasing relative to $t$. 
Self-Organizing Maps

• The properties of the neighborhood function have been chosen so that points nearby a chosen point in $I$ will be updated in conjunction with the chosen point, with less adjustment for points further away from the chosen point, and this update neighborhood will decrease in size as the algorithm proceeds through its iterations, usually with the neighborhood reduced to a single point during the last steps of the algorithm.

• It is assumed that the input vectors to the algorithm are derived from a random sample and they lie in some bounded convex set in $\mathbb{R}^d$.

• Each iteration of a self-organizing-map algorithm is characterized by a vector-valued state vector

$$m(t) = (m_1(t), m_2(t), \cdots, m_k(t))$$

where $m_i(t) \in \mathbb{R}^d$ for $i = 1, 2, \cdots, k$.

• The algorithm is initialized at $m(0)$.

• The algorithm proceeds in a recursive fashion as described next.
Self-Organizing Maps

- Given an input vector $\mathbf{x}$ at time $t$, the index, $\iota_t$, of the component state closest to $\mathbf{x}$ is selected, namely,
  \[ \iota_t = \arg \min_{i \in I} \| \mathbf{x} - \mathbf{m}_i(t) \| \]
- The vector $\mathbf{m}(t)$ is then updated according to
  \[ \mathbf{m}_i(t + 1) = \mathbf{m}_i(t) + \beta_t \eta_t(\iota_t, i)[\mathbf{m}_i(t) - \mathbf{x}] \]
  for $i = 1, 2, \ldots, k$, where $\beta_t > 0$ is a parameter that can be lowered over time to lessen the adjustment of $\mathbf{m}(t)$.
- A critical role is played by the neighborhood function in achieving the preservation of neighborhood relations.
- The algorithm is very popular and used for various purposes; however, many theoretical questions remain regarding the organizational nature of the algorithm and convergence.
- Clustering is achieved in the same manner as the $k$-means algorithm, with the clusters determined by the Voronoi diagram associated with $\mathbf{m}_1(t), \mathbf{m}_2(t), \ldots, \mathbf{m}_k(t)$. 
Hierarchical Clustering

- Both $k$-means and fuzzy $k$-means are based on Euclidean-distance clustering.
- Another approach is to iteratively join clusters according to a similarity measure.
- The general hierarchical clustering algorithm is given by the following procedure:
  - Initialize the clusters by $C_i = \{x_i\}$ for $i = 1, 2, \ldots, n$ and by a desired final number $k$ of clusters.
  - Proceed to iteratively merge the nearest clusters according to the similarity measure until there are only $k$ clusters.
  - An alternative is to continue to merge until the similarity measure satisfies some criterion.
  - The merging process can be pictorially represented in the form of a dendrogram, where joined arcs represent merging.
  - As stated, the hierarchical clustering is agglomerative, in the sense that points are agglomerated into growing clusters.
Hierarchical Clustering

- One can also consider *divisive* clustering, in which, beginning with a single cluster, the algorithm proceeds to iteratively split clusters.
- Various similarity measures have been proposed. Three popular ones are the minimum, maximum, and average measures given by

\[
\begin{align*}
    d_{\text{min}}(C_i, C_j) &= \min_{x \in C_i, x' \in C_j} \| x - x' \| \\
    d_{\text{max}}(C_i, C_j) &= \max_{x \in C_i, x' \in C_j} \| x - x' \| \\
    d_{\text{av}}(C_i, C_j) &= \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i, x' \in C_j} \| x - x' \|.
\end{align*}
\]

- Hierarchical clustering using the minimum distance is called *nearest-neighbor* clustering.
- If it halts when the distance between nearest clusters exceeds a pre-specified threshold, then it is called *single-linkage clustering.*
Nearest-neighbor clustering

- Given a set of clusters at any stage of the algorithm, it merges the clusters possessing the nearest points.
- If we view the points as the nodes of a graph, then when two clusters are merged, an edge is placed between the nearest nodes in the two clusters.
- Hence, there are no closed loops created and the resulting graph is a tree.
- If the algorithm is not stopped until there is a single cluster, the result is a spanning tree, and the spanning tree is minimal.
- While the algorithm is intuitively pleasing owing to the manner in which it generates a hierarchical sub-cluster structure based on nearest neighbors, it is extremely sensitive to noise and can produce strange results, such as elongated clusters.
- It is also very sensitive to early mergings, since once joined, points cannot be separated.
Farthest-neighbor clustering

• *Farthest-neighbor* clustering results from using the maximum distance.
• If it halts when the distance between nearest clusters exceeds a pre-specified threshold, then it is called *complete-linkage clustering*.
• Given a set of clusters at any stage of the algorithm, it merges the clusters for which the greatest distance between points in the two clusters is minimized.
• This approach *counteracts the tendency toward elongation* from which nearest-neighbor clustering suffers.
• Finally, if the algorithm halts when the average distance between nearest clusters exceeds a pre-specified threshold, then it is called *average-linkage clustering*.
Hierarchical clustering for two types of lymphoma, DLBCL and MCL

- Figure shows a sample of 30 patients (columns) suffering from B-cell lymphoma and the expression profiles (rows) of 24 genes across the sample. The samples have been hierarchically clustered and the clusters correspond perfectly to two types of lymphoma, DLBCL and MCL.
Hierarchical Clustering

- The genes have also been hierarchically clustered and it appears from the figure that the red-labeled genes are up-regulated for MCL and down-regulated for DLBCL, whereas the green-labeled genes are up-regulated for DLBCL and down-regulated for MCL.
- Thus, the different gene clusters seem to “classify” the lymphomas.
- More precisely, the clustering might provide feature selection in the sense that actual classification might be accomplished by a two-gene feature set, one red-labeled and one green-labeled.
Clustering Accuracy

- A clustering algorithm operates on the point set as a whole by partitioning it.
- Its error is based on the accuracy of the partition.
- Essentially, if we assume the points come from different distributions, label them accordingly, and match the results from the clustering algorithm with the subsets of the partition in the best way possible, then the error is the number of mismatches.
- This idea can be formalized in the probabilistic theory of random sets.
Cluster Validation

- The error of a cluster operator can be defined in the context of random point sets in a manner analogous to classification error.
- One might loosely define a “valid” cluster operator as one possessing small error — or even more loosely as one that produces “good” clusters.
- Such a definition is vacuous unless it is supported by a definition of goodness.
- If one tries to evaluate a cluster operator in the absence of a probability distribution (labeled point process), then measures of validity (goodness) need to be defined.
- They will not apply to the cluster operator as an operator on random sets, but will depend on heuristic criteria.
- Their relation to future application of the operator will not be understood, for the whole notion of prediction depends on the existence of a random process.
Cluster Validation

- Nevertheless, a heuristic criterion can serve to give some indication of how the cluster operator is performing on the data at hand relative to some criterion.
- We will consider some methods that address the validity of a clustering output, or compare clustering outputs, based on heuristic criteria.
- These can be roughly divided into two categories.
- *Internal validation* methods evaluate the resulting clusters based solely on the data.
- *External validation* methods evaluate the resulting clusters based on pre-specified information.
- External validation involves a criterion based on the comparison of the clusters produced by the algorithm and a partition chosen by some other means, say one produced by the investigator’s understanding of the data.
- One criterion can simply be the number of mismatches, which provides an empirical error.
Cluster Validation

• If we were to randomly generate a set \( S \) of points and partition according to an underlying distributional model, then we would have a single-sample estimate of the error of the clustering algorithm relative to the distributional model.
• However, in the present situation no random process has been assumed as the generator of the point set \( S \).
• Rather than compute an empirical error directly, we can consider how pairs of points are commonly and uncommonly clustered by a cluster operator \( \zeta \) and a heuristic partitioning.
• Suppose that \( P_S \) and \( P_\zeta \) are the heuristic and \( \zeta \) partitions, respectively.
Cluster Validation

Define four quantities:

- **a** is the number of pairs of points in $S$ such that the pair belongs to the same class in $P_S$ and the same class in $P_\zeta$;
- **b** is the number of pairs such that the pair belongs to the same class in $P_S$ and different classes in $P_\zeta$;
- **c** is the number of pairs such that the pair belongs to different classes in $P_S$ and the same class in $P_\zeta$; and
- **d** is the number of pairs in $S$ such that the pair belongs to different classes in $P_S$ and different classes in $P_\zeta$.

- If the partitions match exactly, then all pairs are either in the **a** or **d** classes.
Cluster Validation

- The *Rand index* is defined by \( (a + d)/(a + b + c + d) \).
- The Rand index lies between 0 and 1. Numerous external validation procedures have been proposed.
- *Internal validation* methods evaluate the clusters based solely on the data, without external information.
- Typically, a heuristic measure is defined to indicate the goodness of the clustering.
- It is important to keep in mind that the measure applies only to the data at hand and therefore is not predictive of the worth of a clustering algorithm — even with respect to the measure itself.
- A common heuristic for spatial clustering is that, if the algorithm produces tight clusters and cleanly separated clusters, then it has done a good job clustering.
- We consider two indices based on this heuristic.
Cluster Validation

- Let $P = \{S_1, S_2, ..., S_m\}$ be a partition of $S$, $\delta(S_i, S_j)$ be a between-cluster distance, and $\sigma(S_i)$ be a measure of cluster dispersion.
- The **Davies-Bouldin index** is defined by
  \[
  \alpha(P) = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left\{ \frac{\sigma(S_i) + \sigma(S_j)}{\delta(S_i, S_j)} \right\}
  \]
- The **Dunn index** is defined by
  \[
  \beta(P) = \min_{i} \min_{j \neq i} \frac{\delta(S_i, S_j)}{\max_{l} \sigma(S_l)}
  \]
- Low and high values are favorable for $\alpha(P)$ and $\beta(P)$, respectively.
- As defined, the indices leave open the distance and dispersion measures, and different ones have been employed.
Cluster Validation

• If we want the classes far apart, an obvious choice for $\delta$ is

$$\delta(S_i, S_j) = \min_{x \in S_i, z \in S_j} ||x - z||$$

• For tight classes, an evident choice is the diameter of the class,

$$\sigma(S_i) = \max_{x, z \in S_i} ||x - z||$$

• Since these kinds of measures do not possess predictive capability, it appears difficult to assess their worth— even what it means to be “worthy.”

• But there have been simulation studies to observe how they behave.

• The danger of relying on heuristic validation indices has been demonstrated in a study that has shown weak correlation between many validation indices and clustering error across various clustering algorithms and random point processes.
Genetic Regulatory Networks
Cellular Gene Control

- A central focus of genomic research concerns understanding the manner in which cells execute and control the enormous number of operations required for normal function and the ways in which cellular systems fail in disease.
- In biological systems, decisions are reached by methods that are exceedingly parallel and extraordinarily integrated.
- Feedback and damping are routine even for the most common of activities.
Cellular Gene Control

• Traditional biochemical and genetic characterizations of genes do not facilitate rapid sifting of these possibilities to identify the genes involved in different processes or the control mechanisms employed.

• When methods do exist to focus genetic and biochemical characterization procedures on a smaller number of genes likely to be involved in a process, progress in finding the relevant interactions and controls can be substantial.

• The earliest understandings of the mechanics of cellular gene control were derived in large measure from studies of metabolism in cells.

• In metabolism, it is possible to use biochemistry to identify stepwise modifications of the metabolic intermediates and genetic complementation tests to identify the genes responsible for catalysis of these steps, and those genes and cis-regulator elements involved in control of their expression *(cis-regulator* is a DNA sequence that controls the transcription of a related gene).
Analysis of gene regulation

- Starting from the basic outline of the process, molecular biologists and biochemists have been able to build up a very detailed view of the processes and regulatory interactions operating within the metabolic domain.
- In contrast, for most cellular processes, general methods to implicate likely participants and to suggest control relationships have not emerged from classical approaches.
- The resulting inability to produce overall schemata for most cellular processes has meant that gene function has been, for the most part, determined in a piecemeal fashion.
Analysis of gene regulation

• Once a gene is suspected of involvement in a particular process, research focuses on the role of that gene in a very narrow context.
• This typically results in the full breadth of important roles for well-known, highly characterized genes being slowly discovered.
• A good example of this is the relatively recent appreciation that oncogenes such as Myc can stimulate apoptosis in addition to proliferation.
• Because transcriptional control is accomplished by a complex method that interprets a variety of inputs, the development of analytical tools that detect multivariate influences on decision-making present in complex genetic networks is essential.
• Modeling and analysis of gene regulation can substantially help to unravel the mechanisms underlying gene regulation and to understand gene function.
• This can have a profound effect on developing techniques for drug testing and therapeutic intervention for effective treatment of disease.
Nonlinear Dynamical Modeling of Gene Networks

Two salient aspects of a genetic regulatory system must be modeled and analyzed:

(1) the topology (connectivity structure)
(2) the set of interactions between the elements which determine the dynamical behavior of the system.

Exploration of the relationship between topology and dynamics may lead to valuable conclusions about the structure, behavior, and properties of genetic regulatory systems.

Numerous mathematical and computational methods have been proposed for construction of formal models of genetic interactions.
Nonlinear Dynamical Modeling of Gene Networks

These models share certain characteristics.

• they represent *systems* in that they characterize an *interacting group of components forming a whole*, can be viewed as a process that results in a transformation of signals, and generate outputs in response to input stimuli.

• they are *dynamical* in that they capture the *time-varying quality of the physical process* under study and can change their own behavior over time.

• they can be considered to be generally *nonlinear*, in that the interactions within the system yield behavior that is more complicated than the sum of the behaviors of the agents.

• These characteristics are representative of *nonlinear dynamical systems.*
Nonlinear Dynamical Modeling of Gene Networks

- Non-linear dynamic systems are composed of states, input and output signals, transition operators between states, and output operators.
- Most attempts to model gene regulatory networks fall within the scope of nonlinear dynamical systems, including probabilistic graphical models, such as Bayesian networks, neural networks and differential equations.
- Based on more recent evidence from genomics, nonlinear dynamical systems appear to provide the appropriate framework to support the modeling of genomic systems.
- To build a model for a specific application requires abstracting from the specifics of the problem, and the breadth of nonlinear dynamical systems facilitates modeling within their framework.
Nonlinear Dynamical Modeling of Gene Networks

• Many concepts relevant to genomic regulation have been characterized from the perspectives of mathematical theory, estimation of model parameters, and application paradigms.

• **Structural stability** concerns the persistent behavior of a system under perturbation. It captures the idea of behavior that is not destroyed by small changes to the system.
  
  This is certainly a property of real genetic networks, since the cell must be able to maintain homeostasis (ability of living systems to maintain internal equilibrium) in the face of external perturbations and stimuli.

• **Uncertainty** relative to model behavior and knowledge acquisition has been extensively explored. Information theory, traditionally used for communications technology applications, is well suited to study uncertainty measures, quantified through the use of probability theory.
Nonlinear Dynamical Modeling of Gene Networks

• *Distributed control* is common for complex systems, which have the property that no single agent is singularly in control of the system behavior; rather, control is dispersed among all agents, with varying levels of influence.

• This is the current view of genetic regulatory networks.

• To significantly change the global behavior of a system in a desired manner via external control, it is necessary to consider the effects holistically.

• This property is consistent with the inherent global stability of genetic networks in the presence of small changes to the system.
Nonlinear Dynamical Modeling of Gene Networks

• The preceding issue is addressed within control theory, where a central problem is *controllability*: how to select inputs so that the state of the system takes a desired value after some period of time.

• This is precisely the kind of issue that must be addressed for treatment of cancer and other genetically related diseases.

• In sum, nonlinear dynamical systems provide a framework for modeling and studying gene regulatory networks.
Model Selection

• A key question: which model one should use?

Model selection depends on the kind and amount of data available and the goals of the modeling and analysis.

The classical engineering trade-offs:
• Should a model be **fine**, with many parameters to capture detailed low-level phenomena, such as protein concentrations and kinetics of reactions, but thereby requiring a great deal of data for inference?
• Should it be **coarse**, with fewer parameters and lower complexity, thus being limited to capturing high-level phenomena, such as whether a gene is ON or OFF at a given time, but thereby having the advantage of requiring much smaller amounts of data?
Model Selection

• Model selection needs to obey the principle of Occam’s razor: model complexity should be sufficient to faithfully explain the data but not be greater.

• From a pragmatic engineering perspective, this is interpreted to mean that the model should be as simple as possible to sufficiently solve the problem at hand.

• In the context of a functional network, complexity is determined by the number of nodes, the connectivity between the nodes, the complexity of the functional relations, and the quantization.
Boolean Networks

• Here, we focus on the original deterministic version of the Boolean model.

• The more recently proposed stochastic extension will be presented later.

• The Boolean model is archetypical of logical functional models and many of the issues that arise with it arise in other regulatory network models.

• A key issue is intervention in gene regulatory networks and this has mainly been considered in the context of a probabilistic generalization of the Boolean model.
Boolean Model

- The regulatory model that has perhaps received the most attention is the **Boolean network model**. The model has been studied both in biology and physics.
- In the Boolean model, *gene expression is quantized to two levels: ON and OFF.*
- The expression level (state) of each gene is functionally related to the expression states of other genes using **logical rules**.
- Although binarization provides very coarse quantization, we note that it is commonplace to describe genetic behavior in **binary logical language**, such as *on* and *off*, *up-regulated* and *down-regulated*, and *responsive* and *non-responsive*.
Boolean Model

- In the context of expression microarrays, consideration of differential expression leads to the categories of low-expressed and high-expressed, thereby leading to binary networks, or to the categories of low-expressed, high-expressed, and invariant, thereby leading to ternary valued networks that are treated in much the same way as binary networks and often referred to as Boolean networks.
- Successful application of the Boolean model requires the inclusion of genes whose behavior is essentially binary (bi-modal).
- It has been demonstrated in the context of microarrays that there can be sufficiently many switch-like genes so that binary quantization can be successfully utilized for clustering and classification.
From the perspective of logical prediction, numerous Boolean relations have been observed in the NCI 60 Anti-Cancer Drug Screen cell lines such as:

\[ MRC1 = VSNL1 \lor HTR2C \]
\[ SCY A7 = CASR \land MU5SAC \]

Using classical methods there is ample evidence demonstrating inherent logical genomic decision making.

The following figure shows:
- a biologically studied regulatory pathway;
- its corresponding Boolean representation.
• for cells to move into the S phase, cdk2 and cyclin E work together to phosphorylate the Rb protein and inactivate it, thereby releasing cells into the S phase, and that misregulation can result in unregulated cell growth.
Boolean Networks

- A *Boolean network* is defined by a set of nodes, \( V = \{x_1, x_2, \ldots, x_n\} \) and a list of Boolean functions, \( F = \{f_1, f_2, \ldots, f_n\} \).
- Each \( x_k \) represents the state (expression) of a gene, \( g_k \), where \( x_k = 1 \) or \( x_k = 0 \), depending on whether the gene is expressed or not expressed.
- The Boolean functions represent the rules of regulatory interaction between the genes.
- Network dynamics result from a synchronous clock with times \( t = 0, 1, 2, \ldots \).
- The value of gene \( g_k \) at time \( t + 1 \) is determined by
  \[
  x_k(t + 1) = f_k(x_{k1}, x_{k2}, \ldots, x_{k,m(k)})
  \]
  where the nodes in the argument of \( f_k \) form the *regulatory set* for \( x_k \) (gene \( g_k \)).
Boolean Networks

• The numbers of genes in the regulatory sets define the *connectivity* of the network, with maximum connectivity often limited.
• At time point $t$, the state vector

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t))$$

is called the *gene activity profile (GAP)*.
• The functions together with the regulatory sets determine the network wiring.
• A Boolean network is a very coarse model; nonetheless, it facilitates understanding of the *generic properties of global network dynamics* and its *simplicity mitigates data requirements for inference*. 


Attractors in Boolean Networks

- **Attractors** play a key role in Boolean networks.
- Given a starting state, within a finite number of steps, the network will transition into a cycle of states, called an *attractor*, and absent perturbation will continue to cycle thereafter.
- Each attractor is a subset of a *basin* composed of those states that lead to the attractor if chosen as starting states.
- The basins form a partition of the state space for the network.
- Non-attractor states are transient. They are visited at most once on any network trajectory.
Attractors in Boolean Networks

- The following figure provides a transition-flow schematic for a Boolean network containing six genes, with states $0 = 000000$, $1 = 000001, \ldots$, $63 = 111111$.
- There are three singleton attractors, 32, 41, and 55.
- There are four transient levels, where a state in level $k$ transitions to an attractor in $k$ time points.
- The attractors of a Boolean network characterize the long-run behavior of the network and have been conjectured by Kauffman to be indicative of the cell type and phenotypic behavior of the cell.
• Real biological systems are typically assumed to have short attractor cycles, with singleton attractors being of special import.
• For instance, it has been suggested that apoptosis and cell differentiation correspond to some singleton attractors and their basins, while cell proliferation corresponds to a cyclic attractor along with its associated basin.
• Changes in the Boolean functions, via mutations or rearrangements, can lead to a rewiring in which attractors appear that are associated with tumorigenesis.
• This is likely to lead to a cancerous phenotype unless the corresponding basins are shrunk via new-rewiring, so that the cellular state is not driven to a tumorigenic phenotype, or, if already in a tumorigenic attractor, the cell is forced to a different state by flipping one or more genes.
• The objective of cancer therapy would be to use drugs to do one or both of the above.
Coefficient of Determination

• By viewing gene status across different conditions, say, via microarrays, it is possible to establish relationships between genes that show variable status across the conditions.
• Owing to limited replications, we assume that gene expression data are quantized based on some statistical analysis of the raw data.
• One way to establish multivariate relationships among genes is to quantify how the estimate for the expression status of a particular target gene can be improved by knowledge of the status of some other predictor genes.
• This is formalized via the coefficient of determination (CoD), which is defined by

$$ CoD = \frac{\varepsilon_0 - \varepsilon_{opt}}{\varepsilon_0} $$

where $\varepsilon_0$ is the error of the best numerical predictor of the target gene in the absence of observation & $\varepsilon_{opt}$ is the error of the optimal predictor of the target gene based on the predictor genes.
Coefficient of Determination

• This nonlinear form of the CoD is essentially a nonlinear, multivariate generalization of the familiar goodness of fit measure in linear regression.

• The CoD measures the degree to which the best estimate for the transcriptional activity of a target gene can be improved using the knowledge of the transcriptional activity of some predictor genes, relative to the best estimate in the absence of any knowledge of the transcriptional activity of the predictors.

• The CoD is a number between 0 and 1, a higher value indicating a tighter relationship.

• The following figure shows a CoD diagram for the target gene p53 and predictor genes p21 and MDM2, in which the CoDs have been estimated in the context of a study involving stress response.
CoD diagram for p21 and MDM2 predicting p53
Coefficient of Determination

• We see that the individual CoDs for p21 and MDM2 are 0.227 and 0.259, respectively, but when used jointly, the CoD for the predictor set \{p21, MDM2\} increases to 0.452.

• Biologically, it is known that p53 is influential but not determinative of the up regulation of both p21 and MDM2, and hence it is not surprising that some level of prediction of p53 should be possible by a combination of these two genes.

• Note that the prediction of p53 by p21 and MDM2 results from the regulation of p53 on them.

• Going the other way, the same study found the CoD for p53 predicting p21 to be 0.473.

• The increased predictability of p53 using both MDM2 and p21 is expected because increasing the size of the predictor set cannot result in a decrease in CoD.
Danger of Marginal Analysis

- In the figure, MDM2 and p21 have very similar CoDs relative to p53 and there is a significant increase when they are used in combination.
- On the other hand, it may be that very little, if any, predictability is gained by using predictors in combination.
- Moreover, it may be that the individual predictors have CoDs very close (or equal) to 0, but when used in combination the joint CoD is 1.
- This kind of situation shows that it is risky to assume that a predictor g1 and target g0 are unrelated because the CoD of g1 predicting g0 is very low.
- This situation is akin to that in classification where a feature may be poor if used alone but may be good if used in combination with other features.
- The issue in both settings is the danger of *marginal analysis* – drawing conclusions about variables from marginal relations instead of joint (multivariate) relations.
Probabilistic Boolean Networks

- Given a target gene, several predictor sets may provide equally good estimates of its transcriptional activity, as measured by the CoD.
- Moreover, one may rank several predictor sets via their CoDs.
- Such a ranking provides a quantitative measure to determine the relative ability of each predictor set to improve the estimate of the transcriptional activity of the particular target gene.
- While attempting to infer inter-gene relationships, it makes sense to not put all our faith in one predictor set; instead, for a particular target gene, a better approach is to consider a number of predictor sets with high CoDs.
- Considering each retained predictor set to be indicative of the transcriptional activity of the target gene with a probability proportional to its CoD represents feature selection for gene prediction.
Probabilistic Boolean Networks

- Having inferred inter-gene relationships in some manner, this information can be used to model the evolution of the gene activity profile over time.
- It is unlikely that the determinism of the Boolean-network model will be concordant with the data.
- One could pick the predictor set with the highest measure of predictability, but as remarked previously in the case of the CoD, there are usually a number of almost equally performing predictor sets, and for them we will have only estimates from the data.
- By associating several predictor sets with each target gene, it is not possible to obtain with certainty the transcriptional status of the target gene at the next time point; however, one can compute the probability that the target gene will be transcriptionally active at time $t + 1$ based on the gene activity profile at time $t$. 
Probabilistic Boolean Network: Markovian System

• The time evolution of the gene activity profile then defines a stochastic dynamical system.
• Since the gene activity profile at a particular time point depends only on the profile at the immediately preceding time point, the dynamical system is Markovian.
• Such systems can be studied in the established framework of Markov Chains and Markov Decision Processes.
• These ideas are mathematically formalized in probabilistic Boolean networks (PBNs).
• In a PBN, the transcriptional activity of each gene at a given time point is a Boolean function of the transcriptional activity of the elements of its predictor sets at the previous time point.
Instantaneously random PBN

• The choice of Boolean function and associated predictor set can vary randomly from one time point to another.
• For instance, when using the CoD, the choice of Boolean function and predictor set can depend on CoD based selection probabilities associated with the different predictor sets.
• This kind of probabilistic generalization of a Boolean network, in which the Boolean function is randomly selected at each time point, defines an instantiateously random PBN.
Context-sensitive PBN

• Instead of simply assigning Boolean functions at each time point, one can take the perspective that the data come from distinct sources, each representing a context of the cell.

• From this viewpoint, the data derive from a family of deterministic networks and, were we able to separate the samples according to the contexts from which they have been derived, then there would in fact be CoDs with value 1, indicating deterministic biochemical activity for the wiring of a particular constituent network.

• Under this perspective, the only reason that it is not possible to find predictor sets with CoD equal (or very close to) 1 is because they represent averages across the various cellular contexts.
Context-sensitive PBN

- This perspective results in the view that a PBN is a collection of Boolean networks in which one constituent network governs gene activity for a random period of time before another randomly chosen constituent network takes over, possibly in response to some random event, such as an external stimulus or genes not included in the model network.
- Since the latter is not part of the model, network switching is random.
- This model defines a *context-sensitive PBN*.
- The probabilistic nature of the constituent choice reflects the fact that the system is open, not closed.
- The idea is that network changes result from the genes responding to latent variables external to the model network.
- The context-sensitive model reduces to the instantaneously random model by having network switching at every time point.
PBNs : Generalization

• Much of the theory and application of PBNs applies directly to the more general case which need not possess binary quantization and which are also called PBNs, owing to the multi-valued logical nature of functional relations for finite quantization.

• A particularly important case is ternary quantization, where expression levels take on the values +1 (up-regulated), −1 (down-regulated), and 0 (invariant).

• PBN is composed of a set of \( n \) genes, \( x_1, x_2, \ldots, x_n \), each taking values in a finite set \( V \) (containing \( d \) values), and a set of vector-valued network functions, \( f_1, f_2, \ldots, f_r \), governing the state transitions of the genes.

• To every node \( x_i \), there corresponds a set \( F_i = \{ f_j^{(i)} \}_{j=1,\ldots,l(i)} \), where each \( f_j^{(i)} \) is a possible function, called a predictor, determining the value of gene \( x_i \) and \( l(i) \) is the number of possible functions assigned to gene \( x_i \).
• Each network function is of the form $f_k = (f_{k_1}^{(1)}, f_{k_2}^{(2)}, \ldots, f_{k_n}^{(n)})$, for $k = 1, \ldots, r$, $1 \leq k_i \leq l(i)$ and where $f_{k_i}^{(i)} \in F_i$ ($i = 1, 2, \ldots, n$).

• Each vector function $f_k : \{0, 1\}^n \rightarrow \{0, 1\}^n$ acts as a transition function (mapping) representing a possible realization of the entire PBN.

• Thus, given the value of all genes, $(x_1, x_2, \ldots, x_n)$, $f_k (x_1, x_2, \ldots, x_n) = (x_1', x_2', \ldots, x_n')$ gives us the state of the genes after one step of the network given by the realization $f_k$.

• The choice of which network function $f_j$ to apply is governed by a selection procedure.

• At each time point a random decision is made as to whether to switch the network function for the next transition, with a probability $q$ of a change being a system parameter.
PBNs

• If a decision is made to change the network function, then a new function is chosen from among $f_1, f_2, \ldots, f_r$, with the probability of choosing $f_k$ being the selection probability $c_k$.

• Now, let $\mathbf{F} = (f^{(1)}, f^{(2)}, \ldots, f^{(n)})$ be a random vector taking values in $F_1 \times F_2 \ldots \times F_n$.

• That is, $\mathbf{F}$ can take on all possible realizations of the PBN.

• Then, the probability that predictor $f^{(i)}_j$ is used to predict gene $i$ ($1 \leq j \leq l(i)$) is equal to

$$c^{(i)}_j = P\{f^{(i)} = f^{(i)}_j\} = \sum_{k: f^{(i)}_k = f^{(i)}_j} P\{\mathbf{F} = f_k\}.$$  

• Since $c^{(i)}_j$ are probabilities, they must satisfy

$$\sum_{j=1}^{l(i)} c^{(i)}_j = 1.$$
PBNs

• It is not necessary that the selection of Boolean functions composing a specific network be independent.
• This means that it is not necessarily the case that

\[ P\{f^{(i)} = f_j^{(i)}, f^{(l)} = f_k^{(l)}\} = P\{f^{(i)} = f_j^{(i)}\} \cdot P\{f^{(l)} = f_k^{(l)}\}. \]

• A PBN is said to be independent if the random variables \( f^{(1)}, f^{(2)}, \ldots, f^{(n)} \) are independent.
• In the dependent case, product expansions such as the one given in the above equation, as well as ones involving more functions, require conditional probabilities.
• If the PBN is independent, then there are \( L = \prod_{i=1}^{n} l(i) \) realizations (constituent Boolean networks).
• Moreover, for an independent PBN, if the \( k^{th} \) network is obtained by selecting \( f_{i_r}^{(i)} \) for gene \( i, i = 1, 2, \ldots, n, 1 \leq i_r \leq l(i) \), then the selection probability \( c_k \) is given by \( c_k = \prod_{i=1}^{n} c_{i_r}^{(i)} \).
Perturbation

- A PBN with perturbation can be defined by there being a probability $p$ of any gene changing its value uniformly randomly to another value in $V$ at any instant of time.
- A network switch corresponds to a change in a latent variable causing a structural change in the functions governing the network, for instance, in the case of a gene outside the network model that participates in the regulation of a gene in the model,
- A random perturbation corresponds to a transient value flip that leaves the network wiring unchanged, as in the case of activation or inactivation owing to external stimuli such as mutagens, heat stress, etc.
Markov Chain

• The state space $S$ of the network together with the set of network functions, in conjunction with transitions between the states and network functions, determine a Markov chain,

• The states of the Markov chain are of the form $(x^i, f_j)$.

• If there is random perturbation, then the Markov chain is ergodic, meaning that it has the possibility of reaching any state from another state and that its stationary distribution becomes a steady-state distribution.

• In the special case when $q = 1$, a network function is randomly chosen at each time point and the Markov chain consists only of the PBN states.

• For a PBN, characterization of its long-run behavior is described via the Markov chain it defines.
Attractor cycles of a PBN

- An instantaneously random PBN has equivalence classes of communicating states analogous to the basins of attraction for Boolean networks.
- If there is perturbation, then the Markov chain is ergodic, which then guarantees the existence of a global steady-state distribution.
- In general, whether the PBN is instantaneously random or context-sensitive, by definition its attractors consist of the attractors of its constituent Boolean networks.

Two events can remove a network from an attractor cycle $C$:
- a perturbation
- a network switch
Attractor cycles of a PBN

- A perturbation can send it to a different state, and assuming the constituent network remains unchanged and there are no further perturbations for a sufficient time, then it will return to C if the perturbation leaves it in the basin of C or it will transition to a different attractor cycle of the same constituent network if the perturbation sends it to a different basin.

- A network switch will put it in the basin of an attractor cycle for the new constituent network and it will transition to the attractor cycle for that basin so long as the constituent network remains unchanged and there are no further perturbations for a sufficient time.

- Whereas the attractor cycles of a Boolean network are mutually disjoint, the attractor cycles of a PBN can intersect because different cycles can correspond to different constituent Boolean networks.

- Assuming that the switching and perturbation probabilities are very small, a PBN spends most of its time in its attractors.
Network Inference

We confront three impediments in designing genetic regulatory networks from experimental data:

- model complexity
- limited data
- lack of appropriate time-course data to model dynamics

- Numerous approaches to the *network inference problem* have been proposed in the literature, many based on gene expression microarray data.
- Here, we briefly outline some of the proposed methods for PBNs and the rationale behind each of them.
As first proposed, the inference of the PBN is carried out using the CoD.

For each gene in the network, a number of high CoD predictor sets are found and these high-CoD predictor sets determine the evolution of the activity status of that particular gene.

Furthermore, the selection probability of each predictor set for a target gene is assumed to be the ratio of the CoD of that predictor set to the sum of the CoDs of all predictor sets used for that target gene.

This approach makes intuitive sense since it is reasonable to assign the selection probability of each predictor set in a PBN to be proportional to its predictive worth as quantified by the CoD.
A second approach to PBN construction uses mutual information clustering and reversible-jump Markov-chain-Monte-Carlo predictor design.

First, mutual-information-minimization clustering is used to determine the number of possible parent gene sets and the input sets of gene variables corresponding to each gene.

Thereafter, each (predictor) function from the possible parent gene sets to each target gene is modeled by a simple neural network consisting of a linear term and a nonlinear term, and a reversible-jump Markov-chain-Monte-Carlo (MCMC) technique is used to calculate the model order and the parameters.

Finally, the selection probability for each predictor set is calculated using the ratio of the CoDs.
Autonomous Subnetworks

• In most expression studies, there is some degree of previous knowledge regarding genes that play a role in the phenotypes of interest, for instance, p53 in unregulated proliferation.

• To take advantage of this knowledge, and to obtain networks relating to genes of interest, it has been proposed to construct networks in the context of directed graphs by starting with a seed consisting of one or more genes believed to participate in a meaningful subnetwork.

• Given the seed, a network is grown by iteratively adjoining new genes that are sufficiently interactive with genes in the growing network in a manner that enhances subnetwork autonomy.

• The proposed algorithm has been applied using both the CoD and the Boolean-function influence, which measures interaction between genes.
Autonomous Subnetworks

- The algorithm has the benefit of producing a collection of small tightly knit autonomous subnetworks as opposed to one massive network with a large number of genes.
- Such small subnetworks are more amenable to modeling and simulation studies, and when properly seeded are more likely to capture a small set of genes that may be maintaining a specific core regulatory mechanism.
- The following figure shows a melanoma network grown from four seed genes, WNT5A, RTN1, S100B and SNCA.
Melanoma network grown from four seed genes
WNT5A, RTN1, S100B and SNCA

• The CoD has been used as the measure of gene interaction and the boxes denote the seed genes while the ellipses are the genes added by the algorithm.
• The solid lines represent strong connections (connection strengths exceeding 0.3) while the dotted lines represent weak connections (connection strengths between 0.2 and 0.3).
• This network reveals some very interesting insights that are highly consistent with prior biological knowledge derived from earlier gene expression studies using melanoma cell lines.
• For instance, it is known that the WNT5A gene product has the capability to drive aspects of cell motility and invasiveness.
That being the case, it is to be expected that genes playing a part in either mediating extracellular matrix remodeling/interaction, such as MMP3 (matrix metalloproteinase 3), SERPINB2 (serine (or cysteine) proteinase inhibitor), and MCAM (melanoma adhesion molecule), or cellular movement, such as MYLK (myosin light polypeptide kinase), would share regulatory information with WNT5A.

On the other hand, it is not known how WNT5A regulation is coupled to other genes playing a part in melanoma cell proliferation, such as MAP2K1 (mitogen-activated protein kinase kinase 1), and the regulation of apoptosis, such as CASP3 (cysteine aspartate protease 3) and BIRC1 (baculoviral IAP repeat-containing 1).

Nevertheless, it is quite possible that high-level coordination of these activities exists through either specific circuitry, or as a consequence of differing extra-cellular interactions that arise from metastatic cell movement.
Steady-State Data

- A key issue in network design arises because much of the currently available gene-expression data comes to us from steady-state phenotypic behavior and does not capture any temporal history.
- Consequently, the process of inferring a PBN, which is a dynamical system, from steady-state data is a severely ill-posed inverse problem.
- Steady-state behavior constrains the dynamical behavior of the network but does not determine it.
- Therefore, building a dynamical model from steady-state data is a kind of overfitting.
- It is for this reason that a designed network should be viewed as providing a regulatory structure that is consistent with the observed steady-state behavior.
Steady-State Data

- Also, it is possible that several networks may emerge as candidates for explaining the steady-state data.
- Under the assumption that we are sampling from the steady-state, a key criterion for checking the validity of a designed network is that much of its steady state mass lies in the states observed in the sample data because it is expected that the data states consist mostly of attractor states.
- A number of recent papers have focused on network inference keeping in mind that most of the data states correspond to steady-state behavior.
- In one of these, a fully Bayesian approach has been proposed that emphasizes network topology.
Bayesian approach

• The method computes the possible **parent sets of each gene**, the corresponding **predictors** and the **associated probabilities** based on a neural-network model, using a **reversible jump MCMC technique**; and an MCMC method is employed to search the network configurations to find those with the highest Bayesian scores to construct the PBNs.

• This method has been applied to a melanoma cell line data set.

• The steady-state distribution of the resulting model contains attractors that are either identical or very similar to the states observed in the data, and many of the attractors are singletons, which mimics the biological propensity to stably occupy a given state.

• Furthermore, the **connectivity rules** for the most optimally generated networks constituting the PBN were found to be remarkably similar with strong interactions between the components.
Constraint and Operational Criteria

• If we consider network inference from the general perspective of an ill-posed inverse problem, then one can formalize inference by postulating criteria that constitute a solution space in which a designed network must lie.

For this we propose two kinds of criteria:

• *Constraint criteria* are composed of restrictions on the form of the network, such as biological and complexity constraints.

• *Operational criteria* are composed of relations that must be satisfied between the model and the data.
Constraint and Operational Criteria

• Examples of constraint criteria include limits on connectivity and attractor cycles.

• One example of an operational criterion is some degree of concordance between sample and model CoDs, and another is the requirement that data states are attractor states in the model.

• The inverse problem may still be ill-posed with such criteria, but all solutions in the resulting space can be considered satisfactory relative to the requirements imposed by the criteria.

• This kind of approach has been implemented by finding constituent Boolean networks satisfying constraints such as limited attractor structure, transient time, and connectivity.
Dynamic Bayesian Networks

• In addition to the ongoing effort to infer PBNs, there has been a continuing effort to infer Bayesian and dynamic Bayesian networks (DBNs).
• A Bayesian network is essentially a compact graphical representation of a joint probability distribution.
• This representation takes the form of a directed acyclic graph in which the nodes of the graph represent random variables and the directed edges, or lack thereof, represent conditional dependencies, or independencies.
• The network also includes conditional probability distributions for each of the random variables.
Dynamic Bayesian Networks

- In the case of genetic networks, the values of the nodes can correspond to gene-expression levels or other measurable events, including external conditions.
- There is a precisely characterized relation between certain DBNs and PBNs in the sense that they can represent the same joint distribution over their corresponding variables.
- PBNs are more specific in the sense that the mapping between PBNs and DBNs is many-to-one, so that a DBN does not specify a specific PBN.
Intervention
Intervention

- The ultimate objective of genetic regulatory network modeling is to use the network to design different approaches for affecting network dynamics in such a way as to avoid undesirable phenotypes, for instance, cancer.
- The pragmatic manifestation of this goal is the development of therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease.
- Mitigation would be accomplished by the use of drugs to act on the gene products.
- Engineering therapeutic tools involves synthesizing nonlinear dynamical networks, analyzing these networks to characterize gene regulation, and developing intervention strategies to modify dynamical behavior.
Intervention

- Changes in network connectivity or functional relationships among the genes in a network, via mutations or re-arrangements, can lead to steady-state behavior associated with tumorigenesis, and this is likely to lead to a cancerous phenotype unless corrective therapeutic intervention is applied.

Intervention studies using PBNs have used three different approaches:

(i) resetting the state of the PBN to a more desirable initial state and letting the network evolve from there.

(ii) changing the steady-state behavior of the network by minimally altering its rule-based structure.

(iii) manipulating external (control) variables that alter the transition probabilities of the network.
Intervention

• In this chapter, we discuss the intervention results obtained using the first two approaches.

• The results obtained using the third approach and their variants will be discussed in the next chapter.

• Given a PBN, the transition from one state to the next takes place in accordance with certain transition probabilities and their dynamics.

• Therefore, intervention can be studied in the context of homogeneous Markov chains with finite state spaces.
PBN Notation

- To characterize the Markov chain associated with an instantaneously random PBN, we first focus on **Boolean networks**.
- For BNs the state vector $x(k)$ at any time step $k$ is essentially an $n$-digit binary number whose decimal equivalent is given by
  \[ y(k) = \sum_{j=1}^{n} 2^{n-j} x_j(k). \]
- As $x(k)$ ranges from 00...0 to 11...1, $y(k)$ takes on all values from 0 to $2^n-1$.
- Here, we define: $z(k) = 1 + y(k)$.
- As $x(k)$ ranges from 00...0 to 11...1, $z(k)$ takes on all values from 1 to $2^n$. 
PBN Notation

• The mapping from $x(k)$ to $z(k)$ is one-to-one and onto, and hence invertible.
• Thus, instead of the binary representation $x(k)$ for the state vector, we can equivalently work with the decimal representation $z(k)$ for which the state space is $S = \{1, 2, 3, \ldots, 2^n\}$.
• Each $z(k)$ can be uniquely represented by a basis vector $w(k) \in \mathbb{R}^{2^n}$, where $w(k) = e_{z(k)}$.
• If $z(k) = 1$, then $w(k) = [1, 0, 0, \ldots]$. 
• The evolution of the vector $w(k)$ proceeds according to the difference equation:
  
  $$w(k + 1) = w(k)A$$

where $A$ is a $2^n \times 2^n$ matrix having only one non-zero entry (equal to one) in each row.
PBN Notation

- This equation derived for a deterministic Boolean network has a stochastic counterpart for a PBN.
- To arrive at this counterpart, let $w(k)$ denote the probability distribution vector for a PBN at time $k$, i.e. $w_i(k) = Pr\{z(k) = i\}$.
- Then it can be shown that $w(k)$ evolves according to the equation
  \[ w(k + 1) = w(k)A \]
  where $A$ is the stochastic matrix of transition probabilities.
- This completes our limited discussion of PBNs, as Markov chains.
- As with the majority of the literature, in this section we have focused on binary quantization. As seen already, most of the theory and application carry over to any finite quantization in a fairly obvious fashion.
- It is in the ternary setting that we will consider the application of external control in the next chapter.
Intervention by Flipping the Status of a Single Gene

• Recognizing that a key goal of PBN modeling is the discovery of possible intervention targets (genes) by which the network can be “persuaded” to transition into a desired state or set of states, in this section, we consider the effects of intervention by deliberately affecting a particular gene in an instantaneously random PBN.

• Whereas in Boolean networks attractors are hypothesized to correspond to functional cellular states, in PBNs this role is played by irreducible subchains.

• Absent the possibility of perturbation ($p = 0$), a PBN is unable to escape from an irreducible subchain, implying that the cellular state cannot be altered.

• If $p$ is positive, then the Markov chain is ergodic and there is a chance that the current cellular state may switch to another cellular state by means of a random gene perturbation.
Intervention by Flipping the Status of a Single Gene

- Flipping the values of certain genes is more likely to achieve the desired result than flipping the values of some other genes.
- Our goal is to discover which genes are the best potential “lever points”, in the sense of having the greatest possible impact on desired network behavior so that we can intervene with them by changing their value (1 or 0) as needed.
- In addition, we wish to be able to intervene with as few genes as possible in order to achieve our goals.
- To motivate the discussion, let us illustrate the idea with an example.
Example 1.1

- Suppose we are given a PBN consisting of three genes $x_1$, $x_2$, $x_3$.
- There are two functions $f_1^{(1)}$, $f_2^{(1)}$ associated with $x_1$, one function $f_1^{(2)}$ associated with $x_2$ and two functions $f_1^{(3)}$, $f_2^{(3)}$ associated with $x_3$ as shown in the truth table.
- This truth table results in four possible Boolean networks $N_1 = (f_1^{(1)} f_1^{(2)} f_1^{(3)})$, $N_2 = (f_1^{(1)} f_1^{(2)} f_2^{(3)})$, $N_3 = (f_2^{(1)} f_1^{(2)} f_1^{(3)})$ and $N_4 = (f_2^{(1)} f_1^{(2)} f_2^{(3)})$ possessing the probabilities $c_1 = 0.3$, $c_2 = 0.3$, $c_3 = 0.2$ and $c_4 = 0.2$, respectively.
- The state diagram of the Markov chain corresponding to this PBN is also shown.
- Suppose that we are currently in state (111) and wish to eventually transition to state (000).
- The question is, with which of the three genes, $x_1$, $x_2$, $x_3$, should we intervene such that the probability is greatest that we will end up in (000).
## Truth Table

<table>
<thead>
<tr>
<th>$x_1 x_2 x_3$</th>
<th>$f_1^{(1)}$</th>
<th>$f_2^{(1)}$</th>
<th>$f_1^{(2)}$</th>
<th>$f_1^{(3)}$</th>
<th>$f_2^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>001</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>010</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>011</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>111</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$c^{(i)}_{ji}$</td>
<td>0.6</td>
<td>0.4</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
State Transition Diagram
Example 1.1 Contd.

Objective is to reach state 000 starting from state 111
- Which gene should be flipped?

Example 1.1 Contd.

Objective is to reach state 000 starting from state 111
• Which gene should be flipped?

The states 110 and 101 will always return to state 111 with probability 1 (as there are no outward edges from the states in the shaded class C). Hence gene 1 should be flipped.
Random Gene Perturbations

• Assume there is independent random perturbation with $p > 0$, so that the Markov chain is ergodic and every state will eventually be visited.

• The question of intervention can be posed in the sense of reaching a desired state as soon as possible.

• For instance, in example 1, if $p$ is very small and we are in state (111), then it will be a long time until we reach (000) and setting $x_1 = 0$ is much more likely to get us there faster.

• Hence, we are interested in the probability $F_k(x, y)$ that, starting in state $x$, the first time the PBN will reach some given state $y$ will be at time $k$.

• This is known as the first passage time from state $x$ to state $y$.

• For $k = 1$, $F_k(x, y) = A(x, y)$, which is just the transition probability from $x$ to $y$. 
Random Gene Perturbations

• For $k \geq 2$, it can be shown that:  
  
  $$F_k(x, y) = \sum_{z \in S - \{y\}} A(x, z) F_{k-1}(z, y).$$

• We can examine our results by considering:
  
  $$H_{K_0}(x, y) = \sum_{k=1}^{K_0} F_k(x, y)$$

  which is the probability that the network, starting in state $x$, will visit state $y$ before time $K_0$.

• As a special case, when $K_0 = \infty$, $H_{K_0}(x, y)$ is the probability that the chain ever visits state $y$, starting at state $x$, which is equal to 1 since the Markov chain is ergodic.

• The mean first passage time from state $x$ to state $y$ is defined as
  
  $$M(x, y) = \sum_{k} k F_k(x, y).$$

• $M(x, y)$ is the average time it will take to get from state $x$ to state $y$. 
Example 1.2

• Continuing with example 1.1, it is possible to compute the entries of the matrix $A$.
• Supposing $p = 0.01$, the steady-state distribution is given by $[0.0752, 0.0028, 0.0371, 0.0076, 0.0367, 0.0424, 0.0672, 0.7310]$, where the leftmost element corresponds to (000) and the rightmost to (111).
• The PBN spends much more time in state (111) than in any other state.
• Let our starting state $x$ be (111) and the destination state $y$ be (000), as before.
• Should we intervene with gene $x_1$, $x_2$, or $x_3$?
• Using first passage time, we compute $F_k((011), (000))$, $F_k((101), (000))$, and $F_k((110), (000))$.
• The figure shows the plots of $H_{K_0}(x, y)$ for $K_0 = 1, 2, \ldots, 20$ and for the three states of interest, namely, (011), (101), and (110).
Example 1.2

• The plots indicate that starting at state (011), the network is much more likely to enter state (000) sooner than by starting at states (110) or (101).

• For instance, during the first 20 steps, there is almost a 0.25 probability of entering (000) starting at (011), whereas starting at (110) or (101), there is only a 0.05 probability.

• Thus, we should intervene with gene $x_1$ rather than with $x_2$ or $x_3$.

• Were we to base intervention on mean first passage time then the best gene for intervention would be the one possessing the smallest mean first passage time to the destination state.

• For this example, the mean first passage times corresponding to the perturbations of genes $x_1$, $x_2$, and $x_3$ are 337.51, 424.14, and 419.20, respectively.

• Since the first one is the smallest, this again supports the conclusion that gene $x_1$ is the best candidate for intervention.
To summarize the results of this section, given an initial state $x$, we generate different states $x^{(i)} = x \oplus e_i$, $i = 1, 2, \ldots, n$, where $e_i$ is the unit binary vector with a 1 in the $i^{th}$ coordinate, by perturbing each of the $n$ genes, and compute $H_{K_0}(x^{(i)}, y)$ for some desired destination state $y$ and constant $K_0$.

Then, the best gene for intervention is the one for which $H_{K_0}(x^{(i)}, y)$ is maximum; that is, given a fixed $K_0$, the optimal gene $x_{i_{opt}}$ satisfies

$$i_{opt} = \arg \max_i H_{k_0}(x^{(i)}, y).$$

Alternatively, by minimizing the mean first passage times, the optimal gene satisfies

$$i_{opt} = \arg \min_i M(x^{(i)}, y).$$
Intervention to Alter the Steady-State Behavior

- The type of intervention described in the last section can be useful for modulating the dynamics of the network but it does not alter the underlying network structure.
- Accordingly, the stationary distribution remains unchanged.
- However, an imbalance between certain sets of states can be caused by mutations of the “wiring” of certain genes, thereby permanently altering the state-transition structure and, consequently, the long-run behavior of the network.
- Develop a methodology for altering the steady-state probabilities of certain states or sets of states with minimal modifications to the rule-based structure.
- The motivation is that these states may represent different phenotypes or cellular functional states, such as cell invasion and quiescence, and we would like to decrease the probability that the whole network will end up in an undesirable set of states and increase the probability that it will end up in a desirable set of states.
Intervention to Alter the Steady-State Behavior

- One way to accomplish this is by altering some Boolean functions (predictors) in the PBN.
- An additional goal is to alter as few functions as possible.
- Formal methods and algorithms have been developed for addressing such a problem. Here we briefly discuss the results.
- Consider an instantaneously random PBN with perturbation and two sets of states $A, B \subseteq \{0, 1\}^n$.
- Since the Markov chain is ergodic, each state $x \subseteq \{0, 1\}^n$ has a positive stationary probability $\pi(x)$.
- Thus, we can define $\pi(A) = \sum_{x \in A} \pi(x)$, and $\pi(B)$ similarly.
- Suppose that we are interested in altering the stationary probabilities of these two sets of states in such a way that the stationary probability of $A$ is decreased and the stationary probability of $B$ is increased by $\lambda$, $0 < \lambda < 1$. 
Intervention to Alter the Steady-State Behavior

• As already mentioned above, these two states may represent two different cellular functional states or phenotypes.
• In order to achieve this, suppose we alter function $f_{j0}^{(i0)}$ by replacing it with a new function $g_{j0}^{(i0)}$.
• The probability $c_{j0}^{(i0)}$ corresponding to $g_{j0}^{(i0)}$ must remain the same as for $f_{j0}^{(i0)}$, since $c_1^{(i)} + c_2^{(i)} + \ldots + c_{l(i)}^{(i)} = 1$.
• Thus, we have a new PBN whose stationary distribution we can denote by $\mu$.
• Letting $\mu(A)$ and $\mu(B)$ be the stationary probabilities of $A$ and $B$ under the altered PBN model, we pose the following optimization problem.
Intervention to Alter the Steady-State Behavior

- Given sets A and B, predictor functions $f_{j}^{(i)}$ together with their selection probabilities $c_{j}^{(i)}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, l(i)$, and $\lambda \in (0, 1)$, select $i_{0}$ and $j_{0}$, and a function $g_{j_{0}}^{(i_{0})}$ to replace $f_{j_{0}}^{(i_{0})}$, such that $\varepsilon(\pi(A) - \lambda, \mu(A))$ and $\varepsilon(\pi(B) + \lambda, \mu(B))$ are minimum among all $i$, $j$, $g_{j}^{(i)}$, where $\varepsilon(a, b)$ is some error function, such as the absolute error $\varepsilon(a, b) = |a - b|$.

- An additional constraint can be that $g_{j_{0}}^{(i_{0})}$ has no more essential variables than $f_{j_{0}}^{(i_{0})}$.

- In this scenario, we are only allowing the alteration of one predictor function.

- More generally, we can pre-select a number of predictor functions that we are willing to alter.
Example 1.3

• For the PBN of Example 1.1, the figure shows the state transition diagram assuming no perturbation \((p = 0)\).
• From the figure we see that there are two absorbing states, (000) and (111).
• Suppose (111) corresponds to cell invasion (and rapid proliferation) and (000) corresponds to quiescence.
• Now assume perturbation probability \(p = 0.01\).
• A simple analysis based on the probability transition matrix shows that the stationary probabilities of states (000) and (111) are 0.0752 and 0.7310, respectively.
• Thus, in the long run, the network will be in quiescence only 7\% of the time and will be in proliferation 73\% of the time.
• Suppose we wish to alter this imbalance and require the stationary probabilities to be approximately 0.4 for both (000) and (111). The other six states will then be visited only 20\% of the time.
Example 1.3

- In the framework of the optimization problem already described, $A = \{(111)\}$, $B = \{(000)\}$, $\pi(A) = 0.7310$, $\pi(B) = 0.0752$, $\mu(A) = \mu(B) = 0.4$, and $\lambda = 0.3279$.

- Finally, suppose we are allowed to change only one predictor function.

- In the truth table of example 1.1, this corresponds to changing only one column, while keeping the selection probabilities $c_j^{(i)}$ unchanged.

- Thus, there are five possible columns (predictors) and 256 possibilities for each.

- The $5 \times 256 = 1280$ possible alterations have been generated and the stationary probabilities $\mu(000)$ and $\mu(111)$ have been computed for each. (see figure).

- The optimal values of $\mu(000)$ and $\mu(111)$ for the error function $\varepsilon(a, b) = |a - b|$ are indicated by an arrow in the figure.

- The objective function to be minimized is $|\mu(000) - 0.4| + |\mu(111) - 0.4|$. 
Example 1.3

- The colors of the circles represent which predictor is altered.
- For example, red denotes that predictor $f_1^{(1)}$ is altered.
- The optimal predictor is the one that alters $f_2^{(1)}$ for gene 1 (column 2 in the truth tables) and the truth table of the optimal predictor is $(00010101)^T$.
- This predictor achieves the stationary probabilities $\mu(000) = 0.4068$ and $\mu(111) = 0.4128$.

The structure of the plot in the figure reveals an interesting phenomenon:

- The two stationary probabilities exhibit regularities, forming clusters of points arranged in a linear fashion, with different directions.
- In fact, this phenomenon has been observed in numerous examples.
- It appears that the alterations of different predictors tend to occupy different parts of the space, implying that for a given predictor, there is a certain “range of action” that can be achieved by manipulating it.
Example 1.3

- This suggests that a brute-force search for the optimal predictor alteration may possibly be avoided by following a number of search directions simultaneously, with the more promising ones being explored further.
- This, in turn suggests the use of genetic algorithms for optimization.
- In fact, genetic algorithms have been used to solve the optimal structural intervention problem posed here and the resulting savings in computational effort have been remarkable.
- Nonetheless, this remains an essentially brute force procedure and better approaches need to be developed.
External Intervention Based on Optimal Control Theory
Review of BN’s and PBN’s

• Deterministic Boolean Networks

• n genes : \( g_1, g_2, \ldots, g_n \)

• \( x_i = \) expression status of \( g_i \)

• \( x_i = 0 \)  \( \Rightarrow \)  \( g_i \) not expressed

• \( x_i = 1 \)  \( \Rightarrow \)  \( g_i \) expressed

• Let $x_i(k)$ be the value of $x_i$ at time step $k$.
• Assume $x_i(k+1) = f_i(x_1(k), \ldots, x_n(k))$ where $f_i$ is a Boolean function.
• Example: $x_1 \textbf{ And } x_2 \quad x_1 \textbf{ Or } x_2$

<table>
<thead>
<tr>
<th>$x1$</th>
<th>$x2$</th>
<th>$f_1(x1,x2)$</th>
<th>$f_2(x1,x2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

[Diagram of AND and OR gates]
• Define $x(k) = [x_1(k), \ldots, x_n(k)]$
• Then $x(k+1) = [f_1(x(k)), \ldots, f_n(x(k))]$

**Boolean Network**

```
010  -->  111
  |     |
  v     v
001  -->  000

Truth Table

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Alternative Representation of BN

• Given $x(k)$ define

$$z(k) = 1 + \sum_{j=1}^{n} 2^{n-j} x_j(k)$$

$$w(k) = e_{z(k)}$$

Then

$$w(k + 1) = w(k) A$$

Where $A$ is a $2^n \times 2^n$ matrix
Generalization to PBN’s

Here $x_i(k+1) = \begin{cases} 
    f_i^1(x(k)) \text{ with probability } c_i^1 \\
    f_i^2(x(k)) \,, \,, \,, \, \, c_i^2 \\
    f_i^{l(i)}(x(k)) \text{ with probability } c_i^{l(i)} 
  \end{cases}$
Generalization to PBN’s

Here \( x_i(k+1) = \begin{cases} f_i^1(x(k)) \text{ with probability } c_i^1 \\ f_i^2(x(k)) \,, \,, \,, \text{ with probability } c_i^2 \\ f_i^{l(i)}(x(k)) \text{ with probability } c_i^{l(i)} \end{cases} \)
Alternative Representation of PBN

\[ w(k + 1) = w(k) \cdot A \]

- Here \[ w_i(k) = \Pr(z(k) = i) \]
- \( A \) is a stochastic matrix of transition probabilities

\[ a_{ij} = \Pr(z(k + 1) = j \mid z(k) = i) \]

- Standard Homogenous Markov chain

---

External Optimal Control in PBN’s

• Transition Probabilities depend on external control inputs e.g. chemotherapy, radiation, etc.

• Assume $m$ control inputs $u_1, u_2, \ldots, u_m$.

• Each input can take on the values 0 (not applied) or 1 (applied).

• The values of the control inputs can be changed with time.
• Control input vector at time $k = [u_1(k), \ldots, u_m(k)]$

• As before define

$$v(k) = 1 + \sum_{j=1}^{m} 2^{m-j} u_j(k)$$

• Then $v(k)$ can take on $2^m$ values corresponding to the different control activity profiles

• We now have $w(k+1) = w(k) A( v(k))$
• Transition probability matrix depends on control input.
• Controlled homogenous Markov chain (Markov Decision Process)
• Extensively studied in many areas like Queuing theory etc.
• Optimal control of Markov chains.
• Choose $v(0), v(1), \ldots$ to minimize a particular cost function.
• Choice of cost function?
• Suppose treatment horizon is finite $k=0, 1, \ldots, M-1$.
• Let $C_k(z(k), v(k))$ denote the cost of applying control $v(k)$ at state $z(k)$. (Input from Biologists).
• Then the cost of control over \( M-1 \) periods is

\[
E \left[ \sum_{k=0}^{M-1} C_k \left( z(k), v(k) \right) \bigg| z(0) \right]
\]

• Note: even if \( z(0) \) is deterministic subsequent states are random

\[ \Rightarrow \quad \text{Expectation is necessary} \]
• Net result of control action: state ends up in \( z(M) \).
  (again random)
• Must penalize \( z(M) \) in the cost to reduce chances of ending up in undesirable states
• Define \( C_M(z(M)) \) to be the terminal cost of ending up in state \( z(M) \).
- To determine $C_M(z(M))$ set all controls to zero.
- Partition states into Equivalence classes
- Assign higher penalties to states associated with rapid cell proliferation or reduced apoptosis and lower penalties for states associated with normal cell cycle. (Input from Biologists).
• Total cost

\[
E \left[ C_M(z(M)) + \sum_{k=0}^{M-1} C_k(z(k), v(k)) \middle| z(0) \right]
\]

• Assume \( v(k) = \mu_k(z(k)) \)

where \( \mu_k : \{1, 2, \ldots, 2^n\} \rightarrow \{1, 2, \ldots, 2^m\} \)
Optimal Control Problem

Minimize $E \left[ C_M(z(M)) + \sum_{k=0}^{M-1} C_k(z(k), \mu_k(z(k))) \mid z(0) \right]$

subject to

$\Pr(z(k+1) = j \mid z(k) = i, v(k)) = a_{ij}(v(k))$

• solved using dynamic programming
WNT5A Example

- WNT5A network
- Subset of experiment with 31 cell lines and 587 genes
- Closely knit 10 gene network reduced to 7.
- Uses ternary data (-1,0,1).
Multivariate Relationship Between the Genes of the 7-gene WNT5ANetwork.
Control: Perfect Information

• Objective: Down regulate WNT5A
• 5 step control action
• Terminal Penalties: 0 for down-regulated, 3 for unregulated, 6 for up-regulated
Case 1

- Use \textbf{WNT5A} itself as control
  
  control :- set WNT5A = -1 and let it evolve from there to next time step.

- Cost to use control =1.

- Get suitable decision set for all 2187 states at all 5 time steps.

Case 2

- Use \textbf{Pirin} as a control gene
A few observations

**WNT5A control**

- We could take WNT5A to -1 in the final state starting from any initial state.
- Probability of WNT5A=-1 is always higher with control than without control at all time points.
- In the absence of control action some initial states transition to bad absorbing states.
Pirin Control

- Using Pirin WNT5A cannot be down regulated to -1 for any initial state.
- In the intermediate steps, the probability of WNT5A = -1 is not always higher with control than without it.
- However for the final state, probability of WNT5A = -1 is always much higher with control than without it.
- Bad absorbing states can be avoided using control.
Significant Extensions

• Imperfect Information case
• Context-sensitive PBNs
• Controlling family of Networks
• Controlling Steady-State behavior.
Imperfect Information Case

- Perfect information about states not available.
- Sufficient statistic $P_{X_k|I_k}$ is used.
- $I_k$ is the information vector (total information available at time $k$).

Context-sensitive PBNs

Context-sensitive PBNs are much closer approximation to biological reality.

- $n$ genes, $N = 2^n$ states
- Composed of $k$ Boolean networks.
- Network Transition Probability $= q$
- Perturbation Probability $= p$

Example $n = 3$ genes, $N = 8$ states, $k = 2$ Boolean networks
**Mutually exclusive events**

**EVENT 1**
The current network function is applied, the PBN transitions accordingly, and the network function remains the same for the next transition.

Start state 000, starting network BN1, and event 1 occurs.
EVENT 2
The current network function is applied, the PBN transitions accordingly, and a new network function is selected for the next transition.

Start state 000, starting network BN1, and event 2 occurs
EVENT 3
There is a random perturbation and the network function remains the same for the next transition.

Start state 000, starting network BN1, and event 3 occurs
EVENT 4
There is a random perturbation and a new network function is selected for the next transition.

Start state 000, starting network BN1, and event 4 occurs
Expected Cost for a Finite Horizon Problem of Length 5 Originating from the Different Initial States

Controlling family of Networks

- Collection of Boolean Networks obtained from the steady-state data.
- Control algorithm can be derived to apply on the family of networks.
- Weighting of each network is adaptively estimated from the available data.

Controlling Steady-State Behavior

• Changing the steady-state using stationary control
  (independent of time, only dependent on state).
• Infinite horizon cost considered to design the optimal control.
Stationary Policy Applied on WNT5a network using pirin as control gene (WNT5A is MSB)

Original PBN Steady State Distribution

Steady State Distribution using Stationary Policy

Overview of steps in modeling and control of Probabilistic Boolean networks

1. Grid Alignment
2. Segmentation
3. Gene Expression Extraction
   - 1.72
   - 2.25
   - 0.94
   - 1.56
4. Hypothesis testing
   - (1) Housekeeping Genes
   - (2) Dual Genes
   - Discretization
5. Gene Selection
6. Seed Algorithm
7. Feature Selection
8. Finite horizon Control
   - Perfect State Information
9. Finite horizon Control
   - Imperfect State Information
10. Control of a family of Networks
11. Design of Optimal Control Policy
12. Formulation
13. Penalty Assignment
   - Dynamic Programming
14. PBN Generation
   - (1) CoD
   - (2) Prescribed Attractor Structure Algorithm
   - (3) Bayesian Connectivity Algo
15. Original Steady State
16. Steady State using Control
17. Infinite Horizon Control
Concluding Remarks

• Controllability in a gene regulatory network

State A → State B

Standard Control theory definition: There exists a control input that can take us from State A to state B in finite time.

Which gene to be used as a modulator?

Conclusions Contd.

• Robustness of the Intervention Strategies to Modeling Errors

  sources of error:
  
  – Noisy Microarray data
  – Discrete Quantization
  – Limited amount of data for Inference

Worst-case design?

Preliminary results on min-max robust control
Conclusions Contd.

• Stay in touch with Biology/Biologists.
• Necessary Mathematical theory or something pretty close is probably available in the literature.
• Should not get carried away by elegant mathematics.
• Critical need for experimental validation of network models.