Lecture I. Basics of Molecular Biology

Yu Xiaoxue, Yuan Ling, Yuan Xiang, Zen Fanfan

February 6, 2003

1 Brief history of Bioinformatics

In 1866, Gregor Mendel discovered genetics. Mendel's experiments on peas unveiled some biological elements called genes, which pass information from generation to generation. At that time, people thought genetic information was carried by some "chromosomal" protein, however.

Later in 1869, DNA was discovered. But it was not until 1944 that Avery and McCarty demonstrated DNA is the major carrier of genetic information, but not protein. Remarkable as it is, this discovery is often referred as the start of bioinformatics. In 1953, another historic discovery enabled great advances in both biology and bioinformatics: James Watson and Francis Crick deduced the three dimensional structure of DNA, which is a double helix.

Later in 1960, the genetic code, namely how the mapping from DNA to peptide (protein) is done, was elucidated. It is by the means of combining three nucleotides in the DNA as a codon, and mapping each of them to one amino acid in the peptide.

Starting from the 1970's, several important biotechnology techniques were developed. Firstly, DNA sequencing techniques, like sequence segmentation and electrophoresis were developed. These enabled the identification of DNAs given just a tissue found on a human-body. Moreover, in 1985, the groundbreaking technique, Polymerase-Chain-Reaction (PCR) was invented. By exploiting natural replication, DNA samples can be easily amplified using PCR, so that they are enough for doing experiment.

Starting from the 1980's, scientists began to sequence the genomes. From 1980-1990, complete sequencing of the genomes of various organisms, like that of the E.Coli, was done successfully. And probably the most remarkable event was the launch of Human Genome Project (HGP) in 1989. Originally, it was planned to be completed in 15 years; however, thanks to more and more emerging techniques, in the year of 2001, the first draft of the human genome was published.

2 Cell

Our body consists of a number of organs. Each organ composes of a number of tissues, and each tissue composes of cells of the same type.

Cell performs two types of functions: (1) It carries out various chemical reactions necessary to maintain our life; (2) It passes the information for maintaining life to the next generation.

Essentially, protein performs the chemical reactions, namely, it is responsible for the first function that we mentioned above. DNA stores and passes information, thus, it is responsible for the second function. And RNA is the intermediate between DNA and proteins; it has some functions of proteins, as well as some of DNA's.

3 Protein

Protein is a sequence composed of an alphabet of 20 amino acids. The length is in the range of 20 to more than 5000 amino acids. In average, each protein contains around 350 amino acids.

Due to several intra-molecular forces, like the hydrogen bonds, protein folds into three-dimensional shape, which forms the building blocks and performs most of the chemical reactions within a cell.

The building block of protein, amino acid, consists of one amino group, one carboxyl group and one R group (see Figure 1).

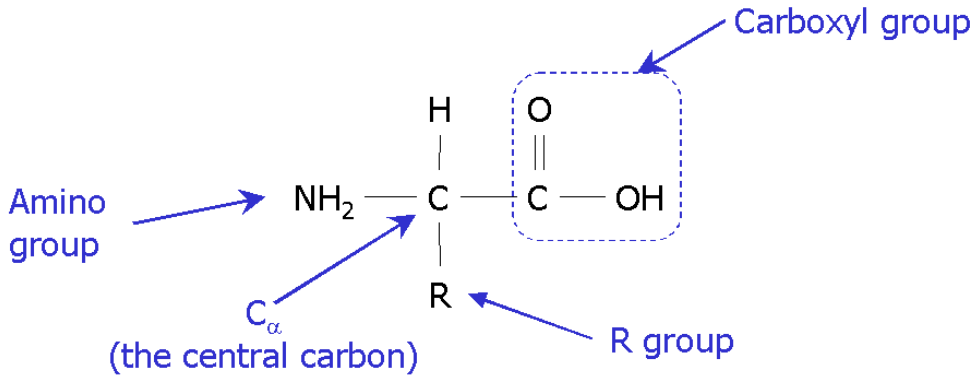

Figure 1: Structure of Amino Acid

There are several ways to classify the amino acids. One is according to the electronic charge of the amino acid. Positively charged (basic) amino acids include: Arginine (Arg, R), Histidine (His, H), and Lysine (Lys, K). Negatively charged (acidic) amino acids include: Aspartic acid (Asp, D) and Glutamic acid (Glu, E). Polar amino acids include: Asparagine (Asn, N), Cysteine (Cys, C), Glutamine (Gln, Q), Glycine (Gly, G), Serine (Ser, S), Threonine (Thr, T), and

Tyrosine (Tyr, Y). They are overall uncharged, but have uneven charge distribution. As a result, they can form hydrogen bonds with water and are called hydrophilic. Normally this group of amino acids is found on the outer surface of a folded protein. Nonpolar amino acids include: Alanine (Ala, A), Isoleucine (Ile, I), Leucine (Leu, L), Methionine (Met, M), Phenylalanine (Phe, F), Proline (Pro, P), Tryptophan (Trp, W), and Valine (Val, V). They are overall uncharged too, but have uniform charge distribution. Thus they cannot form hydrogen bonds with water. They are called hydrophobic, and tend to appear on the inside surface of a folded protein.

As mentioned, protein or polypeptide chain is formed by joining the amino acids together via a peptide bond. The formation and the structure of the peptide bond are shown in Figure 2 below.

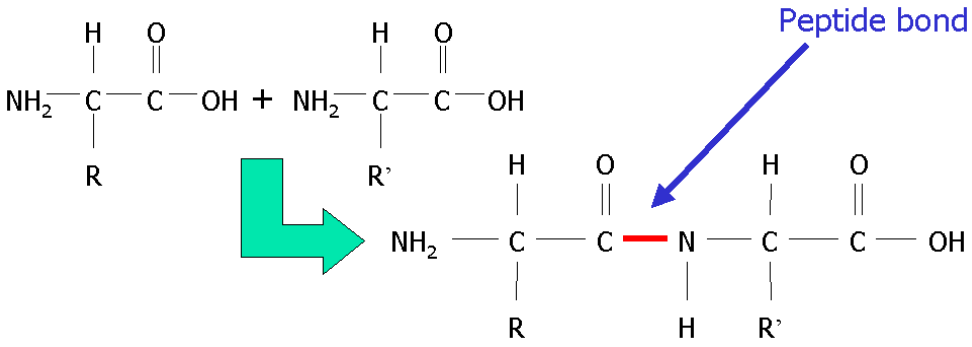

Figure 2: Formation and Struction of the Peptide Bond

One end of the polypeptide is the amino group, which is called the N-terminus. The other end of the polypeptide is the carboxyl group, which is called the C-terminus.

4 DNA

DNA stores the instruction needed by the cell to perform daily life function. DNA can be thought of as a large cookbook with recipes for making every protein in the cell. The information in DNA is used like a library. Library books can be read and reread many times, but they are not used up or given away. The library keeps them for others to use. Similarly, the information in the genes is read, perhaps millions of times in the life of an organism, but the DNA itself is never used up.

4.1 Components of DNA

DNA is double stranded- two strands line up antiparrallel to each other. The double strands are interwoven together and form a double helix. Show in figure 3

and figure 4. If you look at it in detail, you could observe that DNA has a ladder-like structure. The two uprights of the ladder are a structural backbone that supports the rungs of the ladder. Each rung is made of two chemicals called bases that are paired together. These bases are the letters of the genetics code, but the code has a small alphabet – only four letters. The different sequences of letters along the DNA ladder make up our genes. Biological machinery in the cell reads the genetic code of a gene in order to carry out gene's biological function. There are some greater detail to present.

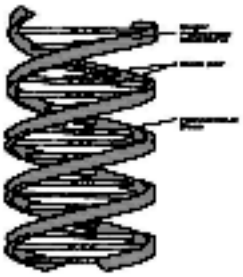

Figure 3: Double Stranded DNA

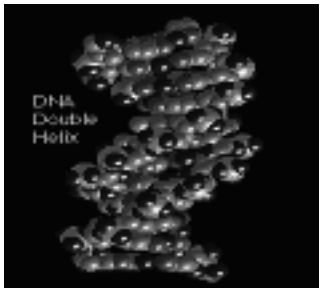

Figure 4: Double Stranded Helix

4.1.1 Structure of Nucleotide

DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide". Each nucleotide, As shown in figure 5 consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group.

The nucleotide structure can be broken down into 2 parts. The sugar-phosphate backbone and the base. All nucleotides share the sugar-phosphate backbone. Nucleotide polymers are formed by linking the monomer units together using an oxygen on the phosphate, and a hydroxyl group on the sugar.

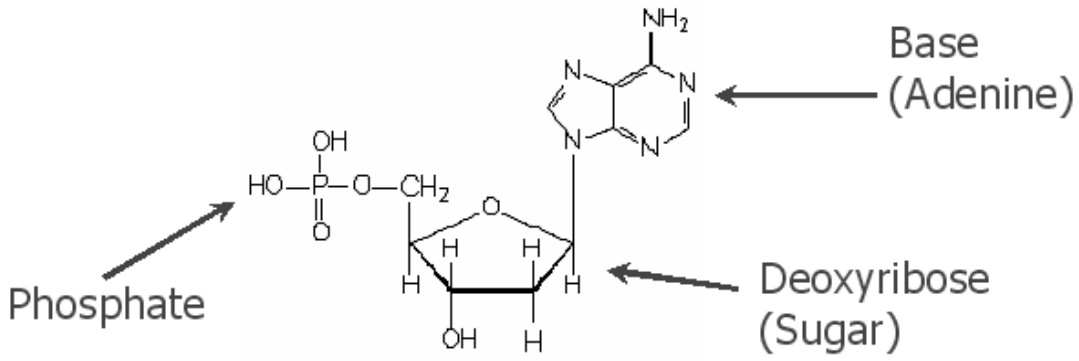

Figure 5: Structure of Nucleotide

4.1.2 Forms of nucleotides

Nucleotides can have 1, 2, or 3 phosphate groups. Monophosphate nucleotides have only 1 phosphate, which are building blocks for DNA. Diphosphate nucleotides have 2 phosphate groups and Triphosphate nucleotides have 3 phosphate groups, which are used to transport energy in the cell. Triphosphate nucleotides show in figure 6.

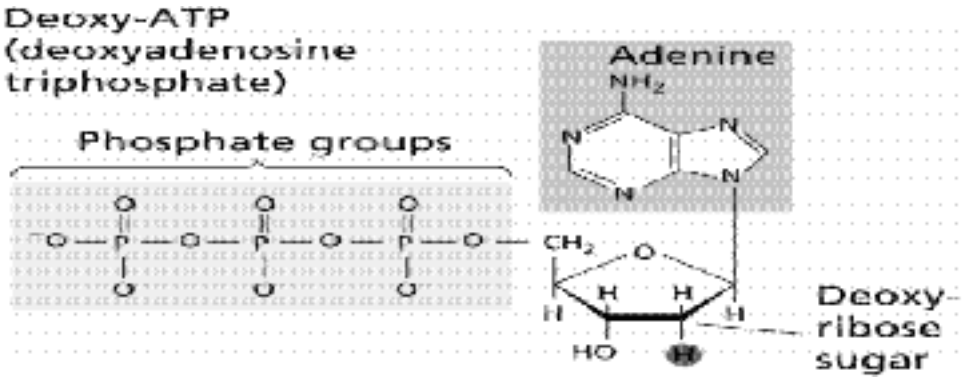

Figure 6: Triphosphate Nucleotides

There are five different types of nucleotides, differing only in the nitrogenous base. The five nucleotides are given one letter abbreviations as shorthand for the five bases. The structure show in figure 7.

A is for adenine
C is for cytosine
G is for guanine
T is for thymine
U is for uracil

The base on each nucleotide is different, but they still show similarities. *adenine* (A) and *guanine* (G) are purines, notice the two ring structure, with the

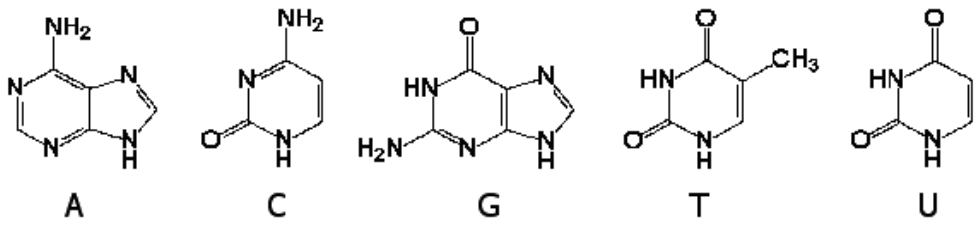

Figure 7: Five Types of Nucleotides: A C G T U

differences in the molecules coming in the groups attached to the ring. Likewise, *cytosine* (C), *thymine* (T) and *uracil* (U) are pyrimidines and share a similar structure. They have a 1-ring structure, but differ in their side groups.3-dimension figures show in figure 8 and figure 9.

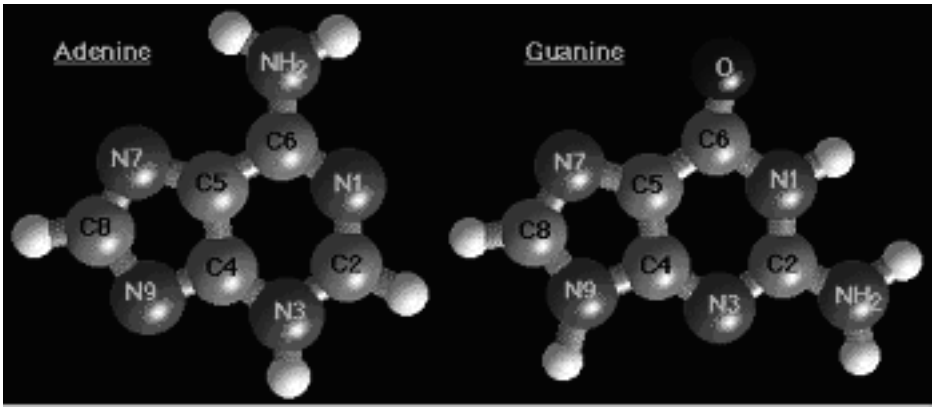

Figure 8: Purines: Adenine and Guanine

For purines, the 9 atoms that make up the fused rings (5 carbon, 4 nitrogen) are numbered 1-9. All ring atoms lie in the same plane.

For pyrimidines, the 6 atoms (4 carbon, 2 nitrogen) are numbered 1-6. Like purines, all pyrimidine ring atoms lie in the same plane.

DNA only uses A, C, G, T.RNA uses A, C, G, U.

4.1.3 Base Pairing

If two DNA are adjacent to one another, the bases along the polymer can interact with complementary bases in the other strand. Adenine(A) is capable of forming hydrogen bonds with thymine(T) and cytosine(C) can base pair with guanine(G). Adenine forms two hydrogen bonds with thymine, cytosine forms 3 with guanine. Show in figure 10.

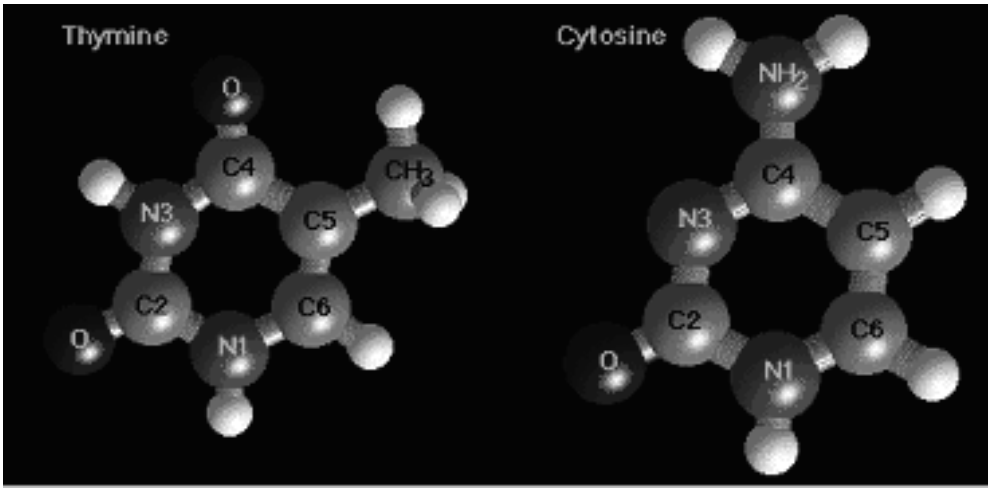

Figure 9: Pyrimidines: Thymine and Cytosine

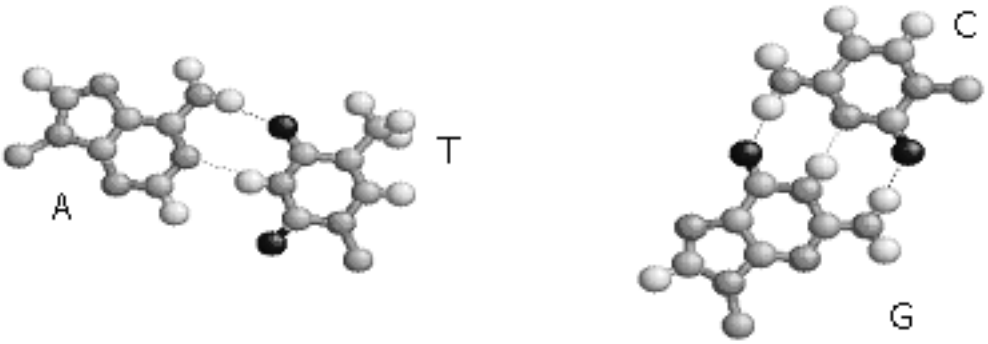

Figure 10: Watson-Crick Pairs

4.2 Orientation of DNA

One strand of DNA is generated by chaining together nucleotides. It forms a phosphate-sugar backbone. It has direction: from 5' to 3'. Because DNA always extends from 3' end. Another strand goes from 3' to 5'. It is the reverse complement. Show in figure 11.

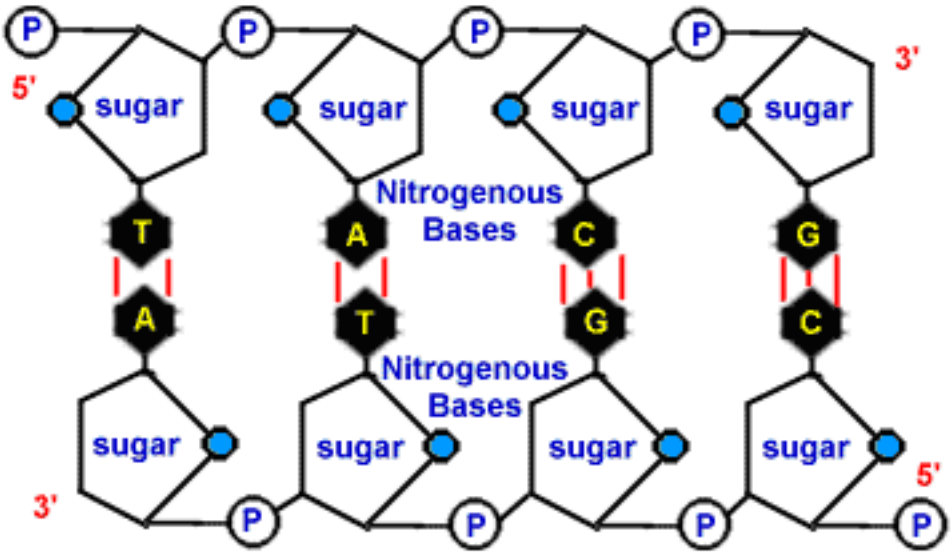

Figure 11: Orientation of a DNA

Cells contain two strands of DNA that are exact mirrors of each other. When correctly aligned, A can pair with T and G can pair with C. Because these strands are mirrors of each other, the amount of A is equal to the amount of T and the amount of C is equal to the amount of G in any double stranded DNA molecule. In solution, the two strands can find each other and form a double helix. This reaction is favorable because of the numerous hydrogen bonds that can be formed between the complementary bases. Show in figure 12.

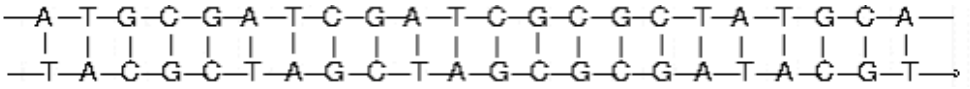

Figure 12: Example of DNA double delix

4.3 Replication of DNA

DNA must be able to replicate itself to pass on this genetic information. It must do this precisely. The overall process is surprisingly simple to understand, but involves the action of numerous proteins in a very complex molecular dance. Another reason why DNA is double stranded is that double strands eases DNA replicate. When the two strands separate, each one serves as a template to make another complementary strand.

This process is known as semi conservative replication. The process is :

1. When a cell split, the double strands of DNA split into two separate strands.
2. Each strands serve as a template to synthesize the reverse complement strand.
3. To start the replication process, an initial primer (short nucleotide chain) must exist. The primer hybridize to the template strand.
4. Replication is accomplished by the coordinated efforts of many cellular enzymes (about 20). One of the best understood enzymes is polymerase, which forms the phospodiester bond between the phosphate residue on the sugar of the incoming nucleotide and OH residue on the sugar of the growing DNA chain. Synthesis is always from 5' to 3'. This enzyme can also proof read and correct any mistakes made along the way. Show in figure 13.

4.4 Form and Location of DNA

DNA usually exists in linear form. E.g. in human, yeast, DNA exists in linear form.

In some simple organism, DNA exists in circular form, E.g. in E. coli, exists in circular form.

Before mentioning the location of DNA, there are some items to explain:

Two types of organisms: Prokaryotes and Eukaryotes.

Prokaryotes: Organisms with a single cell compartment bounded by one or more membranes. Their cells have no nuclei (e.g. bacteria, E.coli).

Eukaryotes: Organisms whose cells contain a nucleus (which contains the genetic material) surrounded by cytoplasm, contained within a plasma membrane. Their cells have nuclei. (e.g. plant and animal, HeLa).

Difference between prokaryotes and eukaryotes:

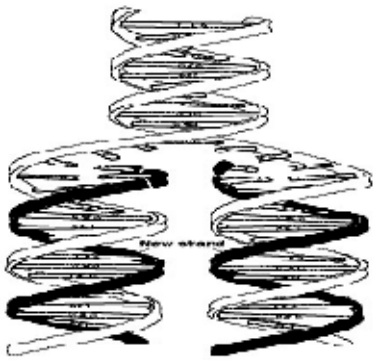

Figure 13: Replication of DNA

1. Eukaryotic cells are larger, more structured than prokaryotic cells.
2. Eukaryotic have extensive internal membrane structures, intracellular compartments (organelles).
3. Eukaryotes have more proteins, more DNA.

In Prokaryote, DNA swims within the cell.
In Eukaryotes, DNA locates within the nucleus

5 Genome, Chromosome, and Gene

5.1 Genome

The *genome* is an organism's complete set of DNA. All the genetic information in an organism is referred collectively as a 'genome'. Genomes vary widely in size: the smallest known genome for a free-living organism (a bacterium) contains about 600,000 DNA base pairs, while human and mouse genomes have some 3 billion. Except for mature red blood cells, all human cells contain a complete genome.

5.2 Chromosome

The 3 million bases of the human genome are not all in one continuous strand of DNA. Rather, the human genome is divided into 23 separate pairs of DNA, called *chromosomes*. Chromosomes are strands of DNA wound around histone proteins. Humans have 22 pairs of chromosomes numbered 1 to 22 (also called autosomes) and the X and Y sex chromosomes. A typical cell has 2 copies of each of the numbered chromosomes, one from the mother and one from the father, and two sex chromosomes. Females have two X chromosomes, while males have an X and a Y. This results in a total of 46 chromosomes in each cell.

The collection of chromosomes in an individual is called a karyotype. For example, the typical male karyotype has 22 pairs of autosomes, one X and one Y chromosome.

5.3 Gene

Each chromosome contains many *genes*, the basic physical and functional units of heredity. Genes are specific sequences of bases that encode a protein or an RNA molecule. Genes comprise of two of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

5.4 Complexity of the organism vs. genome size

Due to the huge amount of genes inside human genome, one might argue that the complexity of one organism is somewhat related to its genome size. But in fact, it's not the truth. People have known that the human genome has 3G base pairs, yet the Amoeba dubia (a single cell organism), even has up to 600G base pairs! Thus, genome size really has no relationship with the complexity of the organism.

5.5 Number of genes vs. genome size

We have just used the quantity of genes inside the genome as a kind of measurement of genome size, but it is totally wrong also. Let's take a look at the human genome and the genome of another Prokaryotic genome, E. coli. We have already known that there are 40,000 to 70,000 genes in human genome, as well as 3G base pairs. And the biologists have also made the estimation that the average length of a gene in human genome is around 1,000 to 2,000 base pairs. It's quite clear that many base pairs "disappeared" here, i.e., it seems that a great part of the human genome even doesn't contribute to the so-called "coding regions" of the genes. In fact, the biologists have discovered that less than 3call them "junk

DNA". Later when we discuss the gene structure in details, we will find out the importance of the coding regions of genes.

And how about E. coli genome? It has 5M base pairs and 4,000 genes. And the average length of a gene in E. coli genome is 1,000 base pairs. From these figures we can get that around 90genome is "useful", i.e., it consists of the coding regions. So, it seems that the Prokaryotic organism, E. coli, even has a better genome structure than human beings! In fact, the conclusion here is, for Eukaryotic genome, the genome size has nothing to do with the number of genes!

6 RNA

We have discussed both DNA and protein, and now we will go on to RNA. As we have mentioned before, RNA has the properties of both DNA and protein. First, similar to DNA, it can store and transfer information. Secondly, similar to protein, it can form complex 3D structure and perform some functions. So, it seems that we only need RNA to accomplish all the requirements for DNA and protein. Why we still need DNA and protein? The reason lies in a simple rule - when you want to do two different things at the same time, you can never do either one as perfectly as those people that only focus on only one thing. As the storage of information, RNA is not as stable as DNA, and that's why we still have DNA. And protein can perform more functions than RNA do, which is the reason that we still need protein.

6.1 Nucleotide for RNA

We have illustrated the nucleotide structure of DNA before. Now we will take a look at the nucleotides in RNA (Shown in figure 14. Similar to the nucleotide of DNA, the one in RNA also has Phosphate and Base. The only difference is that the nucleotide here has the third part of Ribose Sugar, instead of the Deoxyribose in the DNA nucleotide. The Ribose has an extra OH group at 2', which is different from the H group at the same place of Deoxyribose. And that's why we call these two different things "Ribonucleic Acid" and "Deoxyribonucleic Acid" - one is with the OH group, which contains the "O" molecule, yet the other one without.

6.2 RNA vs DNA

Besides the primitive difference that one OH group takes place of the H group, RNA has some other characteristics so that we can easily differentiate it from DNA. First of all, unlike the double helix structure of DNA, RNA is single stranded. One might doubt that with the simple single strand structure, RNA should perform even fewer functions than DNA. The hint here is just the extra

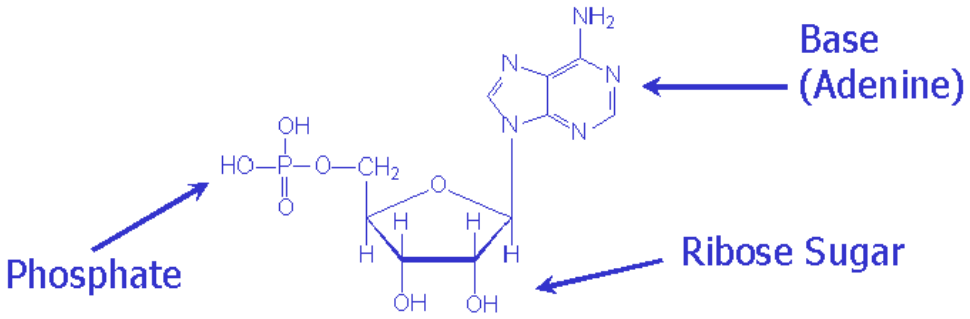

Figure 14: Diagram of nucleotide for RNA

OH group. Due to this extra OH, RNA can form more hydrogen bonds than DNA, so that it can form complex 3D structure to perform more functions. And finally, RNA uses Base U instead of the Base T that DNA uses. Base U is chemically similar to Base T. In particular, U is also complementary to A.

7 Mutation

We have talked about the main usage of DNA is to transfer information from one generation to another. If such a transfer is always absolutely correct, i.e., the new copy of information is exactly the same as the original one, we will lose both the evolution and many fatal disease. In fact, during the reproduction of DNA, RNA and protein, there exists something called "mutation". One can take mutation as a sudden change of genome.

For instance, when one segment of DNA is reproduced, a small sub-segment of it could be lost, duplicated or reversed. Furthermore, sometimes a total new segment could be inserted into the DNA segment.

It is the mutation that makes the new generation of cells or organisms might have something different from their ancestor. We can understand that this is just the basis of evolution. But it can also have some evil effect. For example, some "bad" mutation in human genome may cause fatal disease, such as cancer.

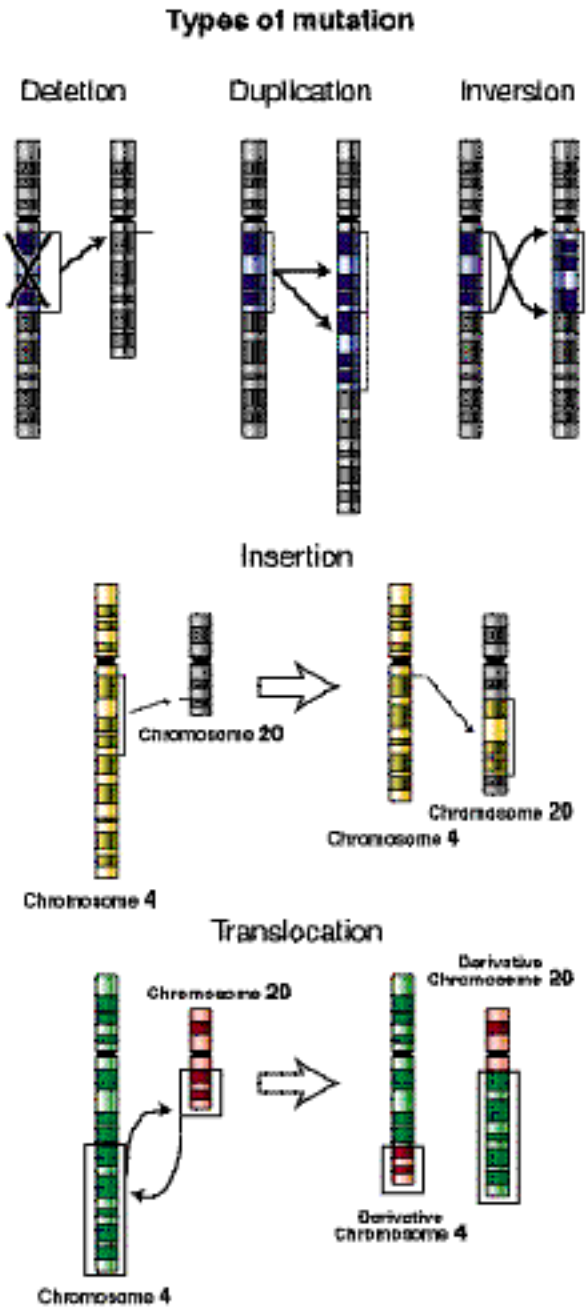

Figure 15: Types of mutation

8 Central Dogma

8.1 Definition

As we have mentioned before, RNA is an intermediate between DNA and protein. But how does it work? To be more specific, how can we get the protein from the gene? There is a rule called "Central Dogma" that defines the whole process. And we also call this process "Gene Expression". The expression of gene consists of two steps, Transcription and Translation. During the transcription period, a "messenger RNA" (mRNA) is synthesized from a DNA template resulting in the transfer of genetic information from the DNA molecule to the mRNA. And in the translation period, the mRNA directs the amino acid sequence of a growing polypeptide during protein synthesis, thus the information obtained from DNA is transferred to the protein.

Let's take at a look at the Central Dogma for Prokaryotes:

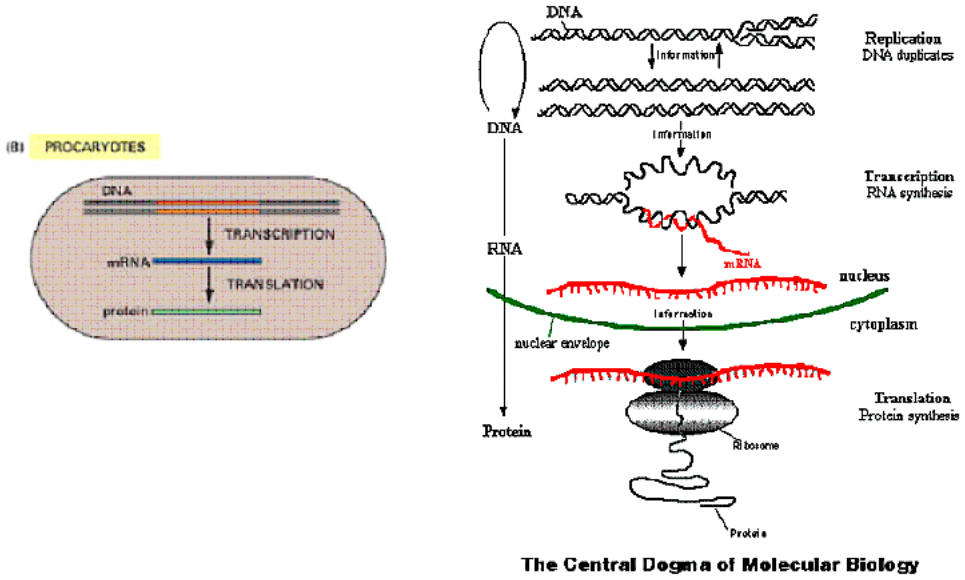

Figure 16: Central Dogma for Prokaryotes

8.2 Transcription (Procaryotes)

In general, during the transcription process of prokaryotes, the mRNA is synthesized from one strand of the DNA gene. First, an enzyme RNA polymerase temporarily separates the double- stranded DNA. Then the transcription begins at the "transcription start site", which is a kind of marker inside genome. The transcription follows such rules: The bases A, C and G are copied from DNA to

mRNA as exactly the same, but the T is replaced by U, as we have mentioned before. And the transcription will continue till another marker, the "transcription stop site".

8.3 Translation

And the translation process will synthesize a protein from an mRNA. In fact, each amino acid is encoded by a consecutive sequences of 3 nucleotides, called codon. And the decoding table from codon to amino acid is called "genetic code". Since each nucleotide could be one of the four types, A, C, G and U, there are totally $4^3 = 64$ different codons. But we have already known that there are altogether 20 different amino acids. Thus, the codons are not one-to-one correspondence to the 20 amino acids. From the diagram of the genetic code we could find out that several different codons all cause the same amino acid. Another important characteristic about the genetic code is that all organisms use the same decoding table, no matter it is from a butterfly, or human being.

8.4 Genetic code

The scientists have already got the complete genetic code for all organisms, just like below:

Again from this diagram we can see that multiple codons may correspond to the same amino acid. And some codons even don't correspond to any amino acid. In fact, they are just the "start signal" and "end signal".

8.5 More on Gene Structure

Now let's take a look at the structure of gene.

A gene consists of four regions, the regulatory region, the 5' untranslated region, the coding region and the 3' untranslated region. The coding region contains the codons for protein. It is also called "open reading frame". Its length is a multiple of 3 since each codon consists of three nucleotides. The coding region must begin with start codon, end with end codon, and the rest of its codons are not an end codon. The 5' untranslated region, coding region and 3' untranslated region together are also called the "mRNA transcript", because it is exactly what the mRNA copies from the DNA. And as we have mentioned before, there are some regions in the gene that are "useless", i.e., they will not be translated into protein. These "useless" regions are just the two "untranslated" regions here. Finally the regulatory region contains promoter, which regulate the transcription process. The promoter is in fact a DNA molecule to which RNA polymerase binds, initiating the transcription of mRNA.

First Position	Second Position of Codon: T	Second Position of Codon: C	Second Position of Codon: A	Second Position of Codon: G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Figure 17: Genetic Code

Figure 18: Gene Structure

8.6 The translation process

Now we can discuss the translation process in details. The translation process is handled by a molecular complex ribosome which consists of both proteins and ribosomal RNA (rRNA). First, the ribosome reads mRNA and the translation starts around the start codon (translation start site). Then, with the help of transfer RNA (tRNA, a class of RNA molecules that transport amino acids to ribosome for incorporation into a polypeptide undergoing synthesis.), each codon is translated to an amino acid. Finally the translation stops once ribosome read the stop codon (translation stop site).

8.7 More on tRNA

We have said that the translation from codon to amino acid is with the help of the transfer RNA, or "tRNA". Totally there are 61 different tRNAs, and each corresponds to a non-terminated codon in the genetic code table. Each tRNA folds to form a cloverleaf-shaped structure. On one side it holds an anticodon (A sequence of three adjacent nucleotides in tRNA designating a specific amino acid that binds to a corresponding codon in mRNA during protein synthesis), and on the other side it holds the appropriate amino acid. The structure of the tRNA is shown in figure 19:

8.8 Central Dogma for Eucaryotes

After discussing the Central Dogma for the Prokaryotes, we move to the Eukaryotes.

8.9 Introns and exons

For Eukaryotes, each gene contains Introns and Exons. Intron is a segment of gene situated between exons. It is not responsible for the coding of protein. So the Introns will be ultimately spliced out of the mRNA. And Exon is a nucleotide sequence in DNA that carries the code for the final mRNA molecule and thus defines the amino acid sequence during protein synthesis. Though the Introns seem "useless", it is quite amazing that in Eukaryotes, each gene can have many Introns, and each Intron may have thousands of bases.

8.10 Transcription (Eukaryotes)

Now we discuss the transcription process of Eukaryotes. Firstly a pre-mRNA is produced which contains both Introns and Exons. Then the 5' cap and poly-A tail are added to the pre-mRNA. After that, the RNA splicing removes the Introns and mRNA is produced. Finally the mRNAs are transported out of the

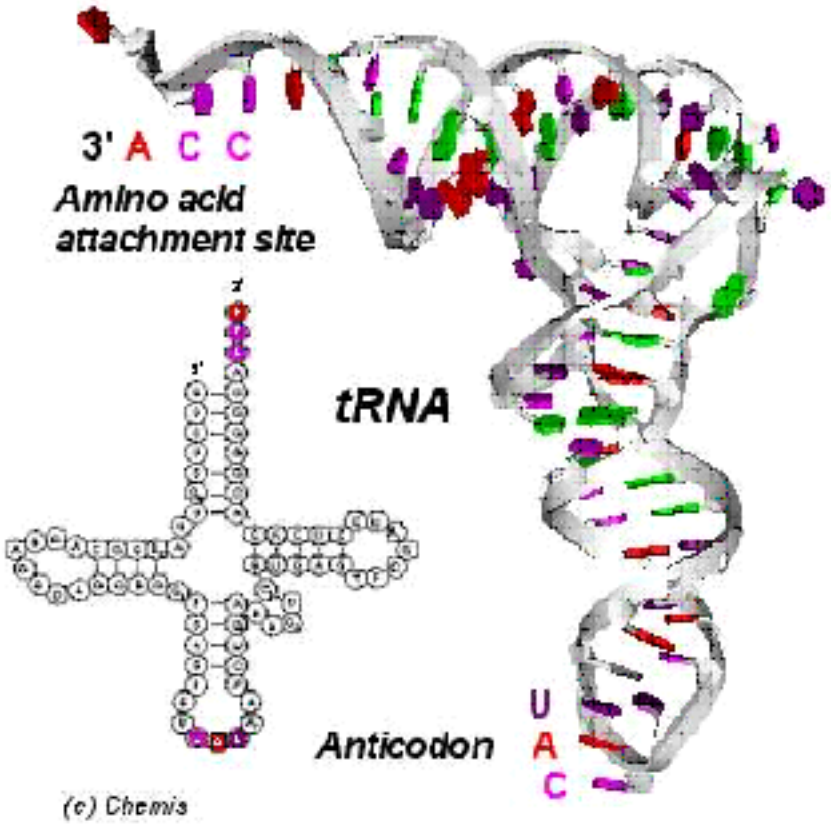

Figure 19: tRNA Structure

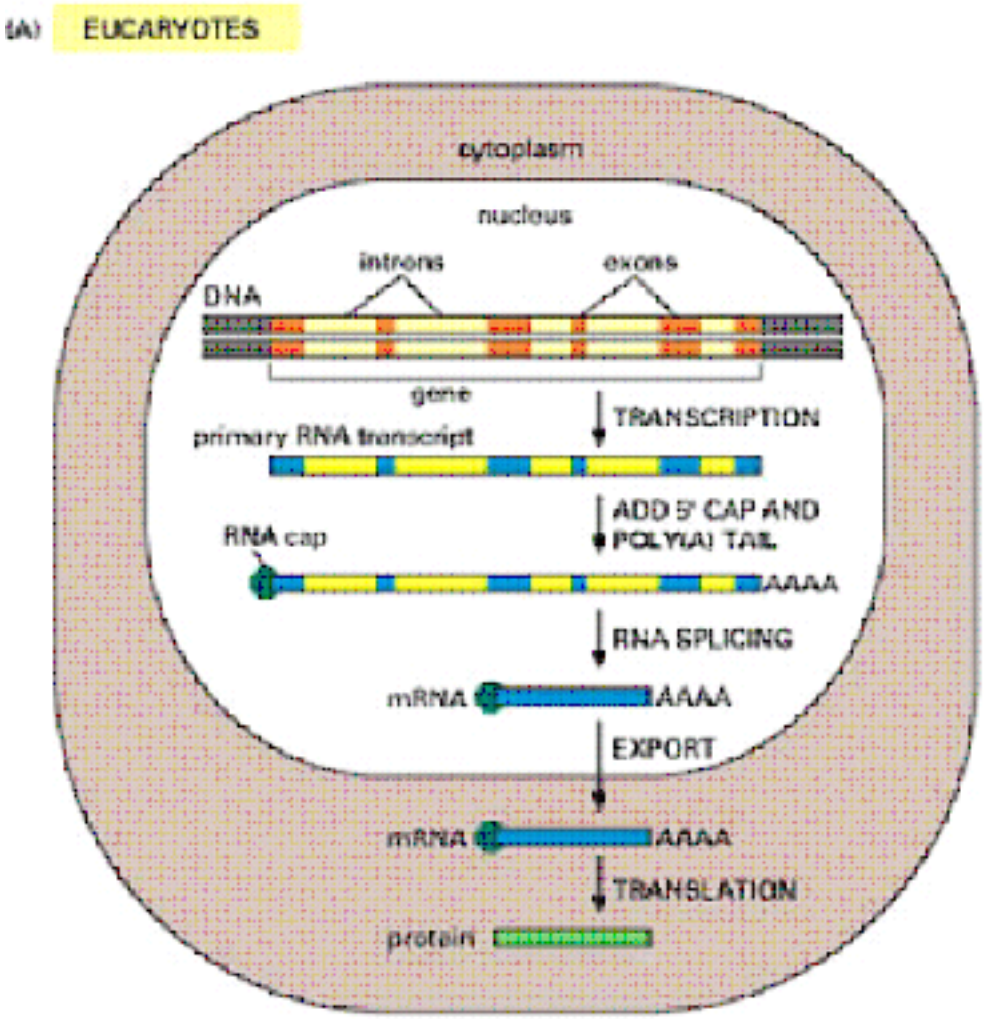

Figure 20: Central Dogma for Eukaryotes

nucleus. And it is translated to a protein using the process described in section 8.3.

9 Basic Biotechnological Tools

A vast range of technological tools have been developed to faciliate the scientists to study DNA in a more efficient manner. Basic tools help to cut and break DNA (using Restriction Enzymes, or Shotgun Method), to duplicate DNA fragments (using Cloning, or PCR), and to measure the length of DNA (using Gel Electrophoresis). Each of these tools is examined in this section.

9.1 Restriction enzymes

Restriction enzymes are DNA-cutting enzymes, which recognize certain point, called restriction site, in the DNA with a particular pattern and break it. Such process is called digestion. Naturally, restriction enzymes are used to break foreign DNA to avoid infection. Each type of restriction enzyme seeks out a single DNA sequence and precisely cuts it in one place. For instance, the enzyme shown here, EcoRI, cuts the sequence GAATTC, cutting between the G and the A. Similar to most of the other restriction enzymes, GAATTC is a palindrome, which means GAATTC is its own reverse complement.

9.2 Shotgun method

Sequencing method that involves randomly sequenced cloned pieces of the genome, with no foreknowledge of where the piece originally came from. Basic idea is to apply high vibration to a solution which has a large amount of purified DNA, so that each molecule is broken randomly into small fragments. This method contrasts with "directed" strategies, in which pieces of DNA from known chromosomal locations are sequenced. Because there are advantages to both strategies, researchers use both random (or shotgun) and directed strategies in combination to sequence the human genome.

9.3 Cloning

Given a piece of DNA X, the process of duplicating it into many pieces is called cloning. The basic steps involve:

1. Insert X into a plasmid vector with antibiotic-resistance gene and a recombinant DNA molecule is formed;

2. Insert the recombinant into the host cell (usually, E. coli).

3. Grow the host cells in the presence of antibiotic. Note that only cells with antibiotic-resistance gene can grow. What is more when we duplicate the host cell, X is also duplicated.

4. Select those cells with antibiotic-resistance genes.

5. Kill them and extract X.

Cloning is a time-consuming process normally requiring several days, whereas we will learn about a much quicker method known as PCR in the next section.

9.4 PCR

PCR is an acronym which stands for polymerase chain reaction. What is a polymerase? A polymerase is a naturally occurring enzyme, a biological macromolecule that catalyzes the formation and repair of DNA (and RNA). The accurate replication of all living matter depends on this activity. In the 1980s, Kary Mullis at Cetus Corporation conceived of a way to start and stop a polymerase's action at specific points along a single strand of DNA. What is the chain reaction? Mullis also realized that by harnessing this component of molecular reproduction technology, the target DNA could be exponentially amplified.

The PCR technique is basically a primer extension reaction for amplifying specific nucleic acids in vitro. The use of a thermostable polymerase allows the dissociation of newly formed complimentary DNA and subsequent annealing or hybridization of primers to the target sequence with minimal loss of enzymatic activity. Inputs for PCR includes: (1) Two oligonucleotides are synthesized, each complementary to the two ends of the region. They are used as primers. (2) Thermostable DNA polymerase TaqI.

PCR consists of repeating a cycle with three phases 25-30 times. Each cycle takes about 5 minutes.

Phase 1: separate double stranded DNA by heat;

Phase 2: cool; add synthesis primers;

Phase 3: Add DNA polymerase TaqI to catalyze 5' to 3' DNA synthesis.

Then, the selected region has been amplified exponentially.

PCR method is used to amplify DNA segments to the point where it can be readily isolated for use. When scientists succeeded in making the polymerase chain reaction perform as desired in a reliable fashion, they had a powerful technique for providing unlimited quantities of the precise genetic material molecular biologists and others required for their work. The example applications are: (1) Clone DNA fragments from mummies; (2) Detection of viral infections.

9.5 Gel electrophoresis

Gel electrophoresis, developed by Frederick Sanger in 1977, is a method that separates macromolecules-either nucleic acids or proteins-on the basis of size, electric charge, and other physical properties.

A gel is a colloid in a solid form. The term electrophoresis describes the migration of charged particle under the influence of an electric field. Electro refers to the energy of electricity. Phoresis, from the Greek verb phoros, means "to carry across." Thus, gel electrophoresis refers to the technique in which molecules are forced across a span of gel, motivated by an electrical current. Activated electrodes at either end of the gel provide the driving force. A molecule's properties determine how rapidly an electric field can move the molecule through a gelatinous medium.

Organic molecules such as DNA are charged. DNA is negatively charged. A gel is prepared which will act as a support for separation of the fragments of DNA. Holes are created in the gel. These will serve as a reservoir to hold the DNA solution. DNA solutions (mixtures of different sizes of DNA fragments) are loaded in a well in the gel. The gel matrix acts as a sieve for DNA molecules. Large molecules have difficulty getting through the holes in the matrix. Small molecules move easily through the holes. Because of this, large fragments will lag behind small fragments as DNAs migrate through the gel. As the process continues, the separation between the larger and smaller fragments increases. The mixture is separated into bands, each containing DNA molecules of the same length.

An application of gel electrophoresis is to reconstruct DNA sequence of length 500-800 within a few hours. The idea is to, firstly, generate all sequences end with A. Then we can use gel electrophoresis, the sequences end with A are separated into different bands. Such information tells us the positions of A's in the sequence. Similarly, we can process for C, G, and T accordingly.