Sequencing and Mapping

CAP 5937-01 Bioinformatics
Fall 2004
Amar Mukherjee

Biology Background

- The genome of an organism is the complete set of the DNA sequences that constitutes its total genetic information content in a cell. This information is wrapped up in a set of **chromosomes** in the cell.
- The eukaryotes also have an additional set of **extrachromosomal** genes. These are located outside the nucleus of the cell within the energy producing organelles called **mitochondria**.

Biology Background

- For plants and algae, there are genes located in the **chloroplasts**. By the word genome, we usually mean the **nuclear genome**.
- For prokaryotic cell, the genome is a circular DNA molecule.
- For eukaryotes, like human, the genome consists of a set of linear DNA molecules contained in different chromosomes.

Biology Background

- In most eukaryotes, there are two copies of each chromosome, and hence two copies of each gene. This is called the **diploid** complement.
- The nucleus of a **haploid** contains only one copy of each chromosome, found only in reproductive cells. The number of chromosomes in a genome is characteristic of a given species.

Example

Organism	Genome Size(kb)	No. of Chromosomes	Avg. no. of DNA/chromosome
Prokaryotes			
E.Coli	4 000	1	4000
Eukaroytes			
Yeast	20 000	16	1250
Fruit Fly	165 000	4	41 250
Human	3 200 000	23	130 000
Mouse	3 454 200		
Maize	15000 000	10	1 500 000
Salamander	90 000 000	12	7 500 000
Puffer Fish	375 000		

Comparative Genomics

Humans share many genes with mice, frogs, flies, and even bacteria and yeast. We all retain similar DNA sequences inherited from our shared ancestors who lived hundreds of millions of years ago.

Species	Chromosomes	Genes	Base Pairs
Human *(Homo sapiens)*	46 (23 pairs)	28-35,000	~3.1 billion
Mouse *(Mus musculus)*	40	22.5-30,000	~2.7 billion
Pufferfish *(Fugu rubripes)*	44	~31,000	~365 million
Malaria Mosquito *(Anopheles gambiae)*	6	~14,000	~289 million
Sea Squirt *(Ciona intestinalis)*	28	~16,000	~160 million
Fruit Fly *(Drosophila melanogaster)*	8	~14,000	~157 million
Roundworm *(C. elegans)*	12	19,000	~97 million
Bacterium *(E. coli)*	1*	~5,000	~4.1 million

*Bacterial chromosomes are chromoneres, not true chromosomes.

Courtesy of Joint Genome Institute

Gene Expressions

- Not all genes are used at all times
  - Out of 30,000 genes in the cell, 1000-5000 are expressed at a given time
- Gene expressed:
  - Transcription process initialized
  - Lots of things must come together for this
- Abundance of RNA and protein products determine gene expression level

Biology Background

- Obviously, genome size does not predict the complexity of the organism and also there is no direct correlation between the **genome size** and the **number of chromosomes**.
- It is generally true that it takes more genes to make the species more complex but there are also other factors.
- About 2-3% of the human nuclear genome actually takes part in the production of proteins.
- Even if we ignore the **introns**, apparently 70 to 80% of the genome is unused! This paradox may be due to the existence of highly repetitive DNAs.

Sequencing

- In order to understand the structure and functions of the genome, we need to first extract the **complete base-pair sequence** in the chromosomes.
- The goal of the Human Genome project was to obtain this complete DNA sequence information. The process of obtaining this information is called **sequencing**.
- Current available biotechnology does not allow sequencing a DNA molecule having more than a few hundred bp (less than 1000 bp).

Sequencing

- Before the genome project was started, biologists started sequencing thousands of **mRNAs** corresponding to coding genes.
- The process involved first purifying mRNA, then obtaining complementary DNA (cDNA) by **reverse transcriptase.**
- Sequencing the cDNA gives immediate information of the DNA of the original gene. However, the cDNA fragment containing a gene is considerably smaller than the genomic DNA.
- This difficulty has given rise to several challenging problems in computational biology.

Molecular Biology Laboratory Techniques : DNA Sequence

- **DNA Sequencing** : separating DNA segments according to size (Gel Electrophoresis)
- The DNA sequence can be read by a technique called ***gel electrophoresis*** which separates DNA molecules into groups depending on their lengths.
- Gel electrophoresis has high resolution; even fragments which differ by a single nucleotide can be separated.
- The sample molecules are placed in a gel under the influence of an electric field.

DNA Sequencing

- The DNA or RNA molecules (which are slightly negatively charged) can migrate towards the positive electric field.
- The speed of migration is inversely proportional to the length of the molecule; longer molecules move slow, shorter move faster.
- All molecules are initially placed at the top of the 'well' and after a few hours, the molecules move to different locations depending on its length.
- If the molecules are labeled with radioactive isotopes, their positions can be photographed on a film.

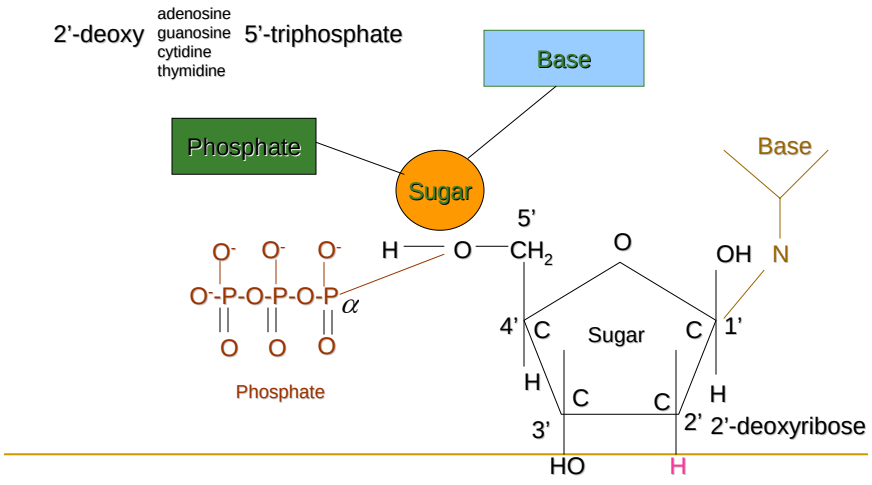
DNA Sequencing

- DNA or a RNA molecule can be sequenced using these techniques as follows.
  - Given a DNA molecule, obtain all fragments that end in a single letter *A.*
  - Similarly, obtain all sequences ending in *T, C* and *G.*
  - For example, if the sequence is GATTCGGATTTACT the fragments that end in *T* are *GAT,GATT*, *GATTCGGAT*, *GATTCGGATT*, *GATTCGGATTT* and the whole sequence *GATTCGGATTTACT*.

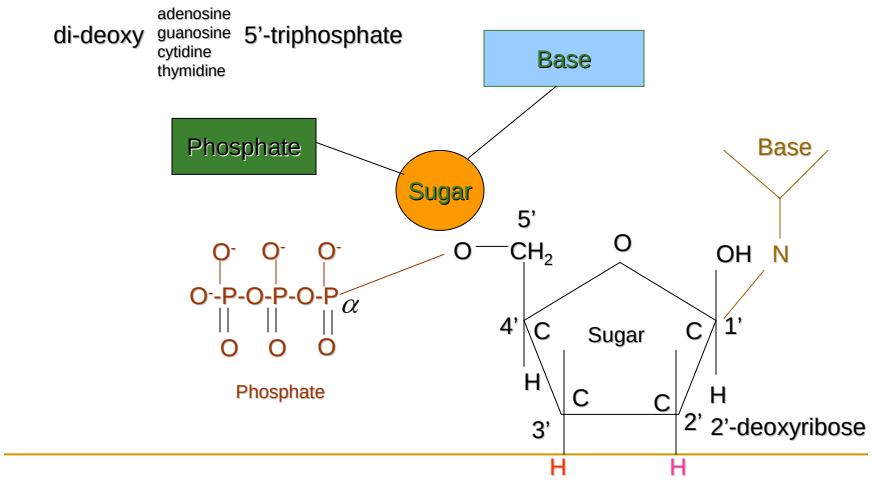
Dideoxy nucleoside triphosphates

- These subsequences are formed by special enzymatic chemical reactions in presence of DNA polymerase , ddATP, ddTTP, ddCTP and ddTGP which are used as ingredients to stop copying the DNA sequence.
- The replication of the DNA sequence is proceeds normally otherwise. WE will soon explain this point in detail.

Chemical Structure of a Nucleotide



Chemical Structure of a **ddATP, ddTTP, ddCTP and ddTGP**

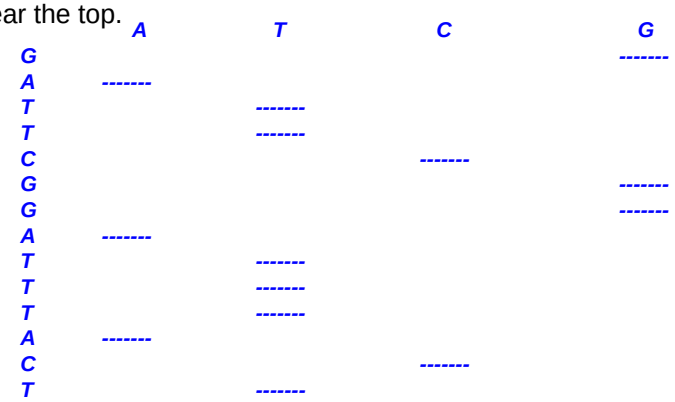


Wells

- In modern automated sequencing, the primer is replaced by a different fluorescent probes and the signals from the probes are detected by special detectors.
- After a period of incubation, these sequences are then placed in four wells, the *A*-well, the *T*-well, the *C*-well and the *G*-well and subjected to electric field simultaneously.
- We can conclude the precise sequence of the original fragment.

Separation in the well depending on length

- The figure below illustrates the principle. We assume here that the positive terminal is on the top and the shorter fragments leave their mark near the top.



Web Link

- If you now read the horizontal bars from top to bottom corresponding to the wells, you will get the entire sequence *GATTCGGATTTACT*
- For further details, see http://web.utk.edu/~khughes/main.htm
- The gel electrophoresis technique was developed in 1970 by Maxam and Gilbert and Sanger. Since the method obtained the DNA fragments by **chemical degradation of part of the sequence**, it was not very reliable. A more efficient and reliable method is to use PCR which we describe next.

Molecular Biology Laboratory Techniques : Cutting a DNA Sequence

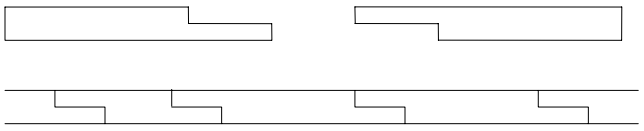
- **Cutting a DNA Sequence** -- Restriction Enzymes
- DNA is a long molecule. In order to be able to sequence it piece by piece, a biological pair of scissors is needed.
- Around 1973, Smith et. al. made a startling discovery in the course of their study of defense mechanism of bacterial cells from viral attack.
  - Observation: certain bacteria produced enzymes that can cut or break a double stranded DNA at specific points.
  - These proteins, called **restriction enzymes** can catalyze the **hydrolysis** of DNA (the process of breaking a molecule by adding water) at specific points called **restriction sites** that are determined by a specific sequence of base pair.

Palindromes

- The first such enzyme discovered called ***EcoRI*** could **cleave** or **digest** DNA molecules between *G* and *A* whenever it encountered the sequence 5'-*GAATTC-3'*.
  - Note that the sequence is its own reverse complement, i.e., if you read the single strand in the 3'-5' direction, you get the same sequence *GAATTC.* Such reverse complement sequences are called **palindromes.**
  - So, whenever such a sequence appears in one strand, it also appears in the other strand. Since the cuts are made in both strands between *G* and *A*, the remaining DNA pieces have **sticky ends.**

Example

5' ……GAATTC……3'
3'……CTTAAG……5'
5'……..G AATTC……3'
3'……..CTTAA G……5'



Chromosomal DNA and restriction enzyme cutting sites

Recombinant DNA

- The sticky ends themselves are naturally complementary to each other.
- This favors re-linking with another DNA piece cut with the same enzyme with the help of another glue enzyme called **ligase**.
- It is also possible to mix DNA from two different sources that have both been cut by using the same restriction enzyme.
- This allows combining fragments from two distinct DNA. Thus, restriction and ligase enzymes are nature's way of providing "cut and paste" editing facility for DNA sequences and have been used in **genetic engineering for recombinant DNA**.
- Even for the same DNA, the cut pieces may join together in different combinations generating overlapping DNA fragments. These are also recombinant DNA and can be cloned for further processing.

Blunt ends

- There are also restriction enzymes that do not create sticky ends, they create **blunt ends.**
- Such blunt cuts can also be ligated with other blunt-ended DNA molecules. In particular, small oligonucleotides can be legated at the blunt ends to have almost arbitrary combination of DNA ends.
- Since the discovery of *EcoR*I, more than 300 restriction enzymes have been found in other bacterial species and have been used in laboratories.
  - These are mostly 4-, 6- or 8-cutters;
  - it is rare to find an odd cutter since the palindromes must be of even length.
  - Finally, the restriction enzymes are sometimes called **endonucleases** because they cut the DNA in the middle of the sequence. There are enzymes called **exonulceases** that cut a DNA from only one end.

- Recently, PCR technology has replaced the restriction endonucleases in many applications.
- They are still used extensively in laboratories for routine subcloning and diagnostic purposes.
- Restriction enzymes cut the DNA into many different sizes of fragments ranging from 256 bp to 1 million bp.

- DNA molecules can also be broken down into *random* pieces by subjecting a solution of purified DNA to rapid mechanical vibrations. The fragments are then filtered;
- Multiple copies are made by cloning, and then sequenced by gel electrophoresis (or microarrays).
- Finally, the fragments are assembled to get the entire DNA sequence. We will describe each of these processes now.

Molecular Biology Laboratory Techniques : DNA Cloning

- In order to study a specific fragment of DNA sequence, we need to select the fragment and *amplify* it so that the solution contains a purified near-homogeneous population.
- The technique inserts the DNA piece in a ***vector***. A naturally occurring vector is a ***plasmid*** which is a circular DNA found in bacteria. Plasmids can infect bacteria such as *E.Coli*.

- Cutting plasmids with a restriction enzyme that has also been used to cut the DNA creating compatible sticky end.
- This allows formation of **recombinant plasmids**.
- The resulting molecule is then inserted into a suitable host (a bacteria or yeast cell) and the organism multiply under suitable conditions (temperature and nutrients), producing a colony of identical cell clones.
- The host is then killed and the resulting DNA pieces extracted and sequenced.

Explanation

- Vectors for cloning vary depending on the size of the DNA to be cloned.
- There are many types of cloning vectors available allowing varying sizes of DNA inserts to be amplified.
- The table below gives a partial list. This includes plasmids, viruses, yeast artificial chromosomes (YAC) and bacterial artificial chromosome (BAC) which were used to create overlapping clones for sequencing human genome.
- The details of the laboratory techniques to produce a purified clone set containing a specific DNA fragment as an insert in the vectors are not covered in this course.

Cloning Vector	Insert Size
Bacteriophage M13	1.5 kb
Plasmid	5 kb
Bacteriophage λ	25 kb
Cosmid	40 kb
BAC(bacterial artificial chromosome)	150 kb
YAC(yeast artificial chromosome)	500 kb

Molecular Biology Laboratory Techniques :Polymerase Chain Reaction (PCR)

- Restriction enzymes and plasmid cloning techniques are used routinely in many laboratory experiments.
- The discovery of PCR has replaced these techniques for large scale sequencing of genomes.
- Without PCR automated and fast sequencing technology would not have been possible.

- PCR is a cell-free method of amplifying a short (<15kb) fragment of a target DNA in large quantities.
- People have compared PCR with the Gutenberg printing press of DNA and Kary Mullis who invented PCR in 1983 got Nobel Prize.
- He thought it was a good idea because he “had been spending a lot of time writing computer programs”.
- PCR is a laboratory application of the concept of “recursion” in computer science.

- PCR technique depends on the existence of a ***primer sequence*** of 15-30 nucleotides long at the end of a target DNA .
- When added to a ***denatured*** DNA (single stranded DNA at temperature >91^0 C), the primers will bind to complementary sequences if the temperature is now cooled to =50^0 C.
- This process is called ***annealing***.

- Under the presence of a DNA polymerase at =720 C, the synthesis of new DNA strands complementary to both strands of the target DNA will start.
- PCR is called a “chain reaction” because both the newly synthesized DNA strands now act as templates for future iterations, doubling the number if DNA fragments at every cycle.
- This results in a huge quantity of the DNA fragments in a short time.
- For further details, see http://web.utk.edu/~khughes/main.htm.