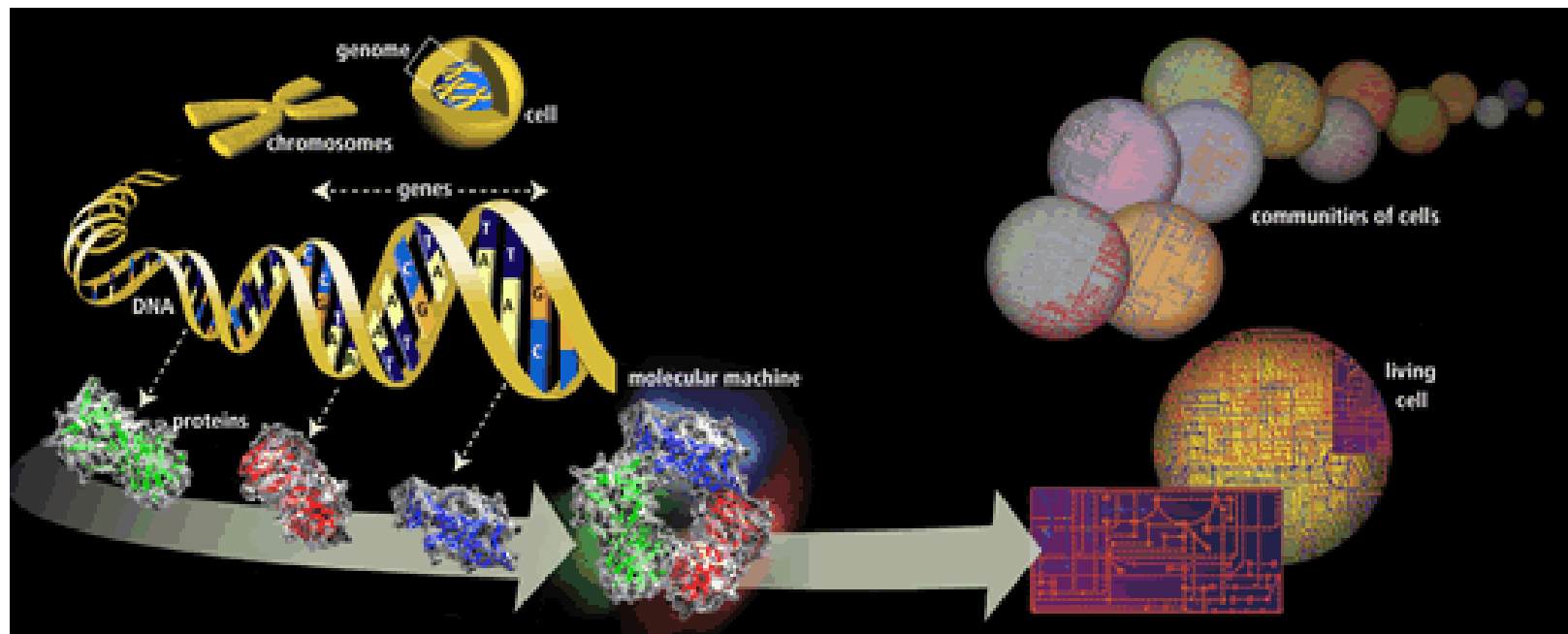


Molecular Biology Primer



Angela Brooks, Raymond Brown, Calvin Chen, Mike Daly,
Hoa Dinh, Erinn Hama, Robert Hinman, Julio Ng, Michael
Sneddon, Hoa Troung, Jerry Wang, Che Fung Yung

Outline:

1. History: Major Events in Molecular Biology
 2. What Is Life Made Of?
 3. What Is Genetic Material?
 4. What Do Genes Do?
 5. What Molecule Code For Genes?
 6. What Is the Structure Of DNA?
 7. What Carries Information between DNA and Proteins
 8. How are Proteins Made?
-

Outline Cont.

9. How Can We Analyze DNA

1. Copying DNA
2. Cutting and Pasting DNA
3. Measuring DNA Length
4. Probing DNA

10. How Do Individuals of a Species Differ

11. How Do Different Species Differ

1. Molecular Evolution
2. Comparative Genomics
3. Genome Rearrangement

12. Why Bioinformatics?

How Molecular Biology came about?

- Microscopic biology began in **1665**
 - *Robert Hooke* (1635-1703) discovered organisms are made up of cells
 - *Matthias Schleiden* (1804-1881) and *Theodor Schwann* (1810-1882) further expanded the study of cells in 1830s
 - **1865** *Gregor Mendel* discover the basic rules of heredity of garden pea.
 - An individual organism has two alternative heredity units for a given trait (**dominant trait** v.s. **recessive trait**)
 - **1869** *Johann Friedrich Miescher* discovered DNA and named it nuclein.
-

Major events in the history of Molecular Biology 1880-1902

- **1881** *Edward Zacharias* showed chromosomes are composed of **nuclein**.
- **1899** *Richard Altmann* renamed nuclein to **nucleic acid**.
- **By 1900**, chemical structures of all 20 amino acids had been identified
- **1902** – *Emil Hermann Fischer* wins Nobel prize: showed amino acids are linked and form proteins
 - **Postulated**: protein properties are defined by amino acid composition and arrangement, which we nowadays know as fact

Major events in the history of Molecular Biology 1900-1911

- **1911** – *Thomas Hunt Morgan* discovers genes on chromosomes are the discrete units of heredity
- **1911** – *Pheobus Aaron Theodore Levene* discovers RNA
- **1941** – *George Beadle* and *Edward Tatum* identify that genes make proteins
- **1950** – *Edwin Chargaff* find Cytosine complements Guanine and Adenine complements Thymine
- **1950s** – *Mahlon Bush Hoagland* first to isolate tRNA
- **1952** – *Alfred Hershey* and *Martha Chase* make genes from DNA

Major events in the history of Molecular Biology 1952 - 1960

- **1952-1953** *James D. Watson and Francis H. C. Crick* deduced the double helical structure of DNA



James Watson
and Francis Crick

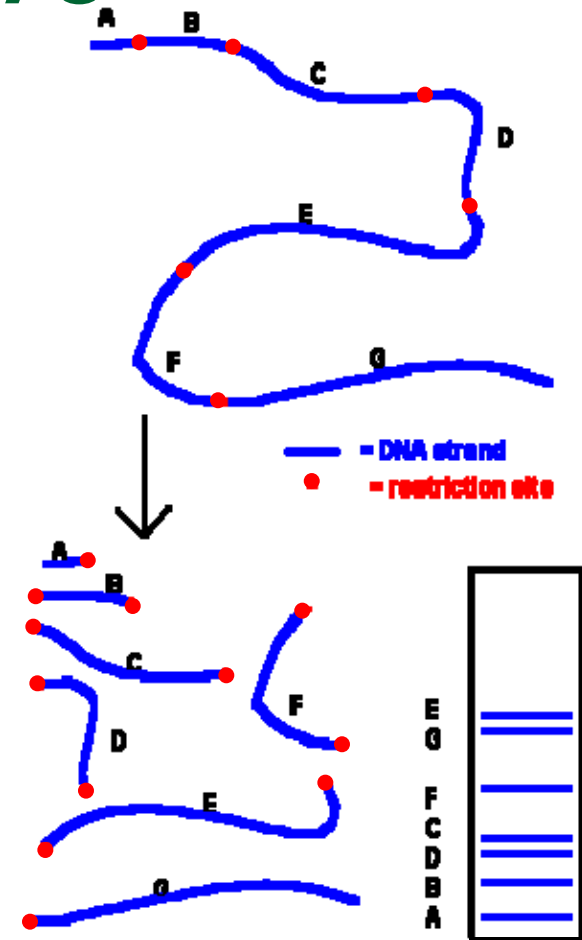
- **1956** *George Emil Palade* showed the site of enzymes manufacturing in the cytoplasm is made on RNA organelles called ribosomes.



George Emil Palade

Major events in the history of Molecular Biology 1970

- **1970** *Howard Temin* and *David Baltimore* independently isolate the first restriction enzyme
- DNA can be cut into reproducible pieces with site-specific endonuclease called restriction enzymes;
 - the pieces can be linked to bacterial vectors and introduced into bacterial hosts. (*gene cloning* or *recombinant DNA technology*)



Major events in the history of Molecular Biology 1970- 1977

- **1977** *Phillip Sharp* and *Richard Roberts* demonstrated that pre-mRNA is processed by the excision of introns and exons are spliced together.
- *Joan Steitz* determined that the 5' end of snRNA is partially complementary to the consensus sequence of 5' splice junctions.



Phillip Sharp



Richard Roberts



Joan Steitz

Major events in the history of Molecular Biology 1986 - 1995

- 1986 *Leroy Hood*: Developed automated sequencing mechanism
- 1986 Human Genome Initiative announced
- 1990 The 15 year Human Genome project is launched by congress of USA
- 1995 Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published (These maps provide the locations of "markers" on each chromosome to make locating genes easier)



Leroy Hood



Major events in the history of Molecular Biology 1995-1996

- **1995** *John Craig Venter*. First bacterial genomes sequenced
- **1995** Automated fluorescent sequencing instruments and robotic operations
- **1996** First eukaryotic genome-yeast-sequenced



John Craig Venter

Major events in the history of Molecular Biology 1997 - 1999

- **1997** E. Coli sequenced
- **1998** PerkinsElmer, Inc.. Developed 96-capillary sequencer
- **1998** Complete sequence of the *Caenorhabditis elegans* genome
- **1999** First human chromosome (number 22) sequenced



hádátko
Caenorhabditis elegans

Major events in the history of Molecular Biology 2000-2001

- **2000** Complete sequence of the euchromatic portion of the *Drosophila melanogaster* genome
- **2001** International Human Genome Sequencing: first draft of the sequence of the human genome published



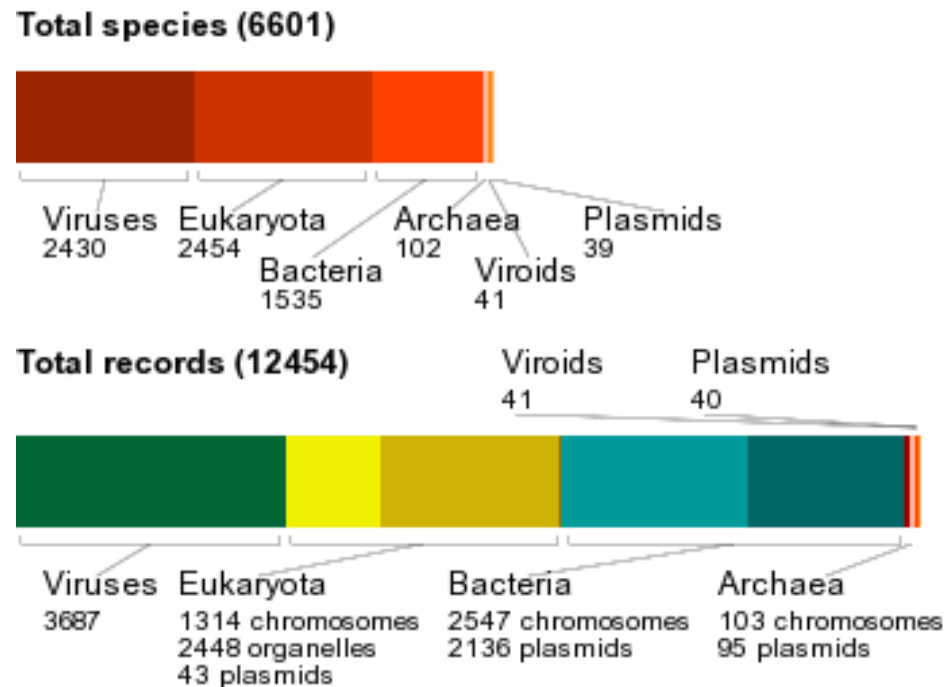
Major events in the history of Molecular Biology 2003- Present

- **April 2003** Human Genome Project Completed. Mouse genome is sequenced.
- **April 2004** Rat genome sequenced.
- Chimpanzee, ...

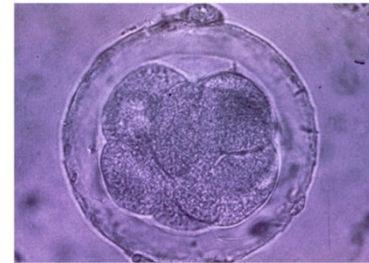
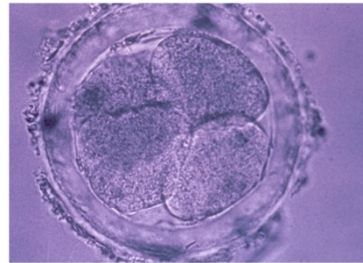
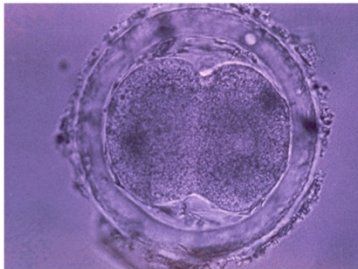
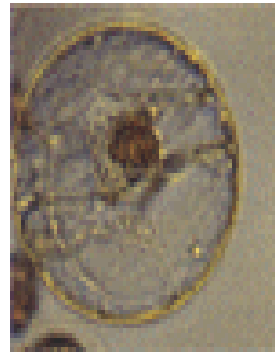
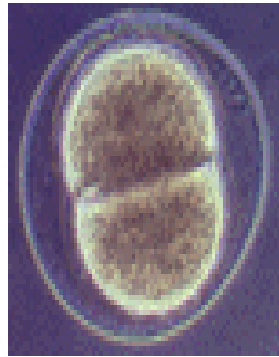
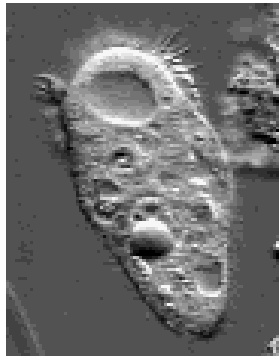


Major events in the history of Molecular Biology Present

- Whole genomes
- <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>
- 2010



Section 2: What is Life made of?



Outline For Section 2:

1. All living things are made of Cells
 - Prokaryote, Eukaryote
 2. Cell Signaling
 3. What is inside the cell: From DNA, to RNA, to Proteins
-

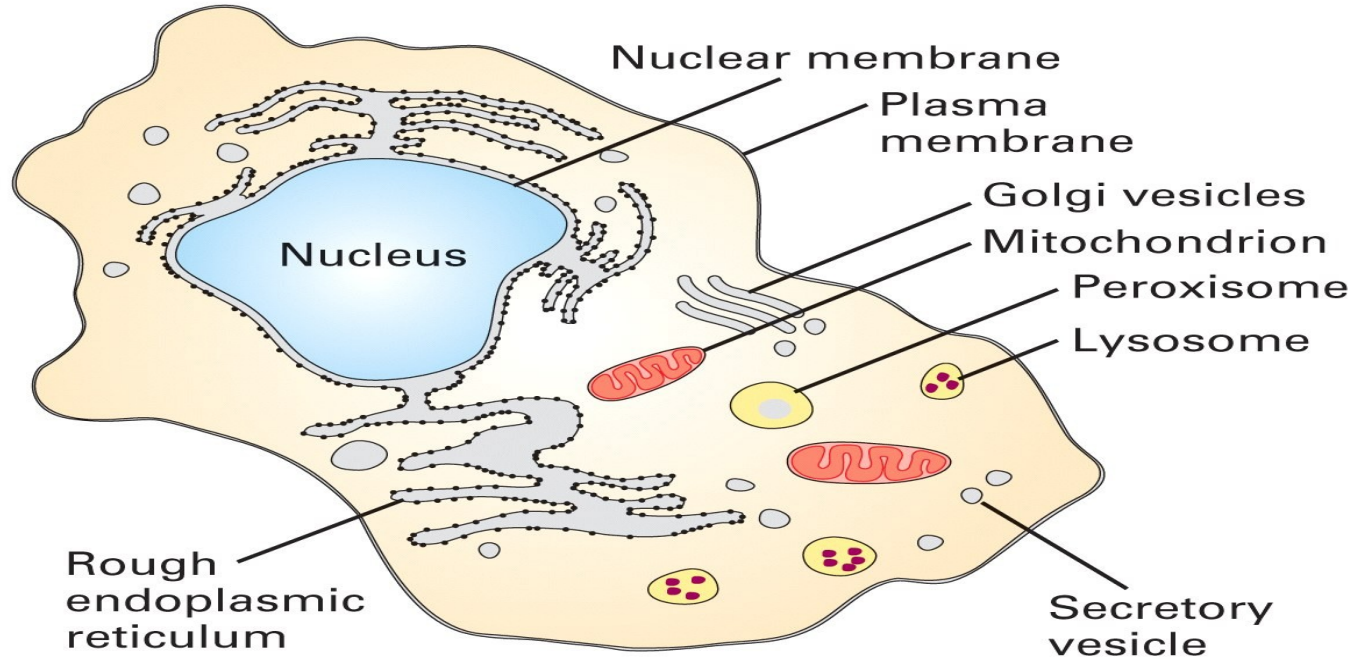
Cells

- **Fundamental working units** of every living system.
- Every organism is composed of one of two radically different types of cells:
 - **prokaryotic** cells or
 - **eukaryotic** cells.
- **Prokaryotes** and **Eukaryotes** are descended from the same primitive cell.
 - All existing prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

Cells

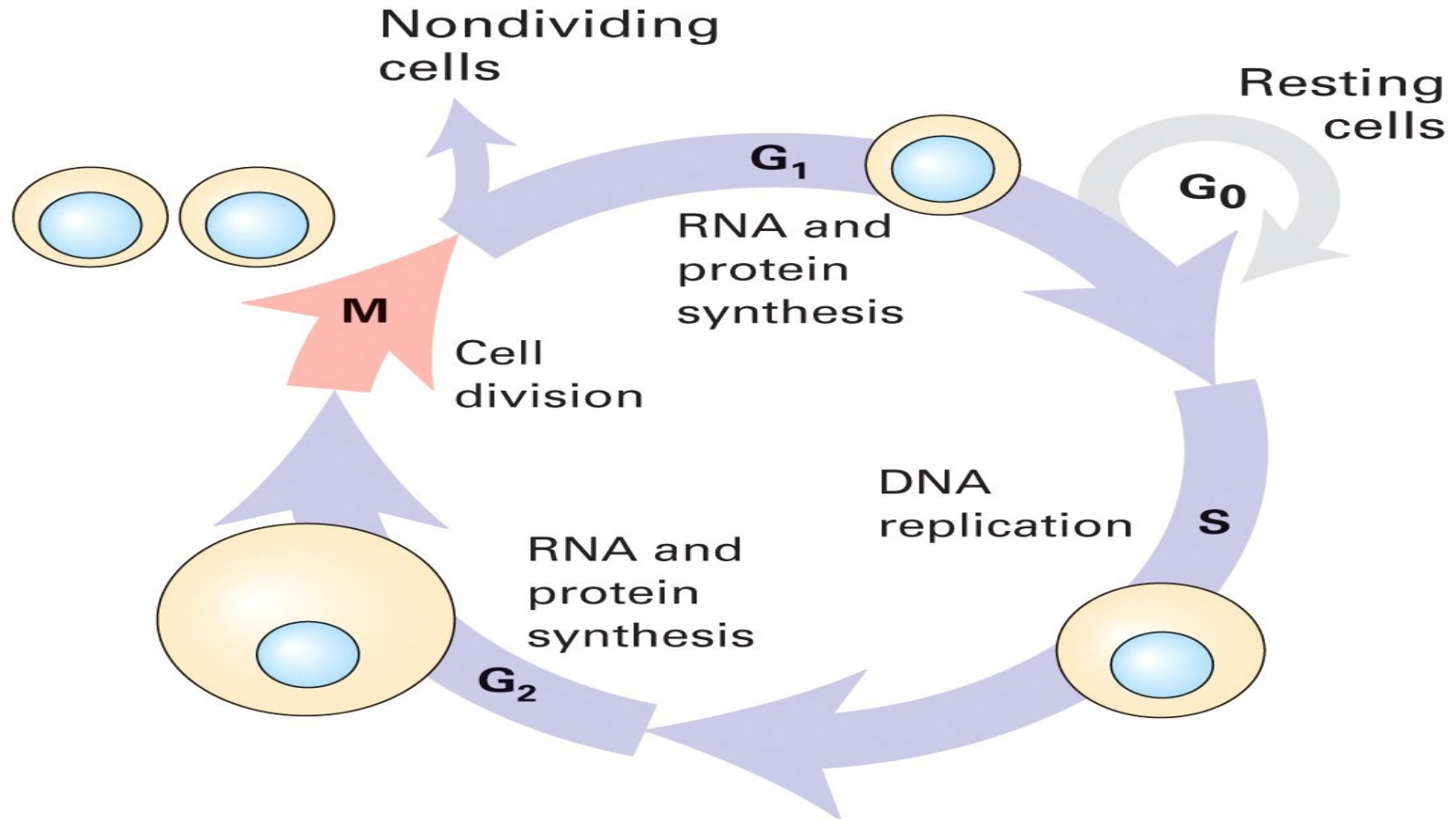
- Chemical composition – by weight
 - 70% water
 - 7% small molecules
 - salts
 - lipids
 - amino acids
 - nucleotides
 - 23% macromolecules
 - Proteins
 - Polysaccharides
 - lipids
 - biochemical (metabolic) pathways
 - translation of mRNA into proteins
-

Life begins with Cell



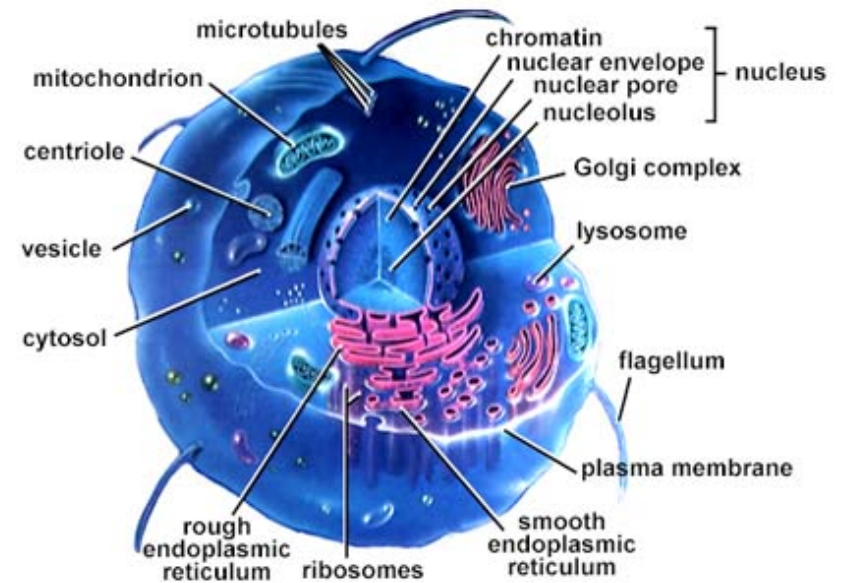
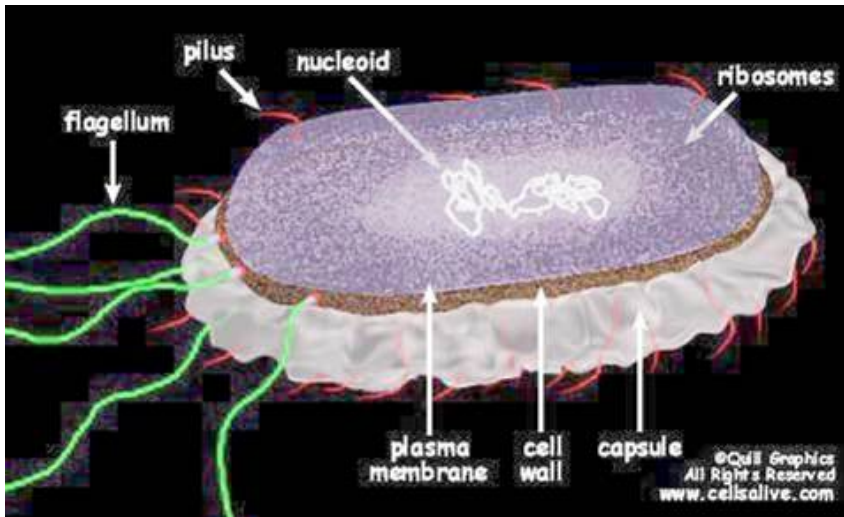
- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

All Cells have common Cycles

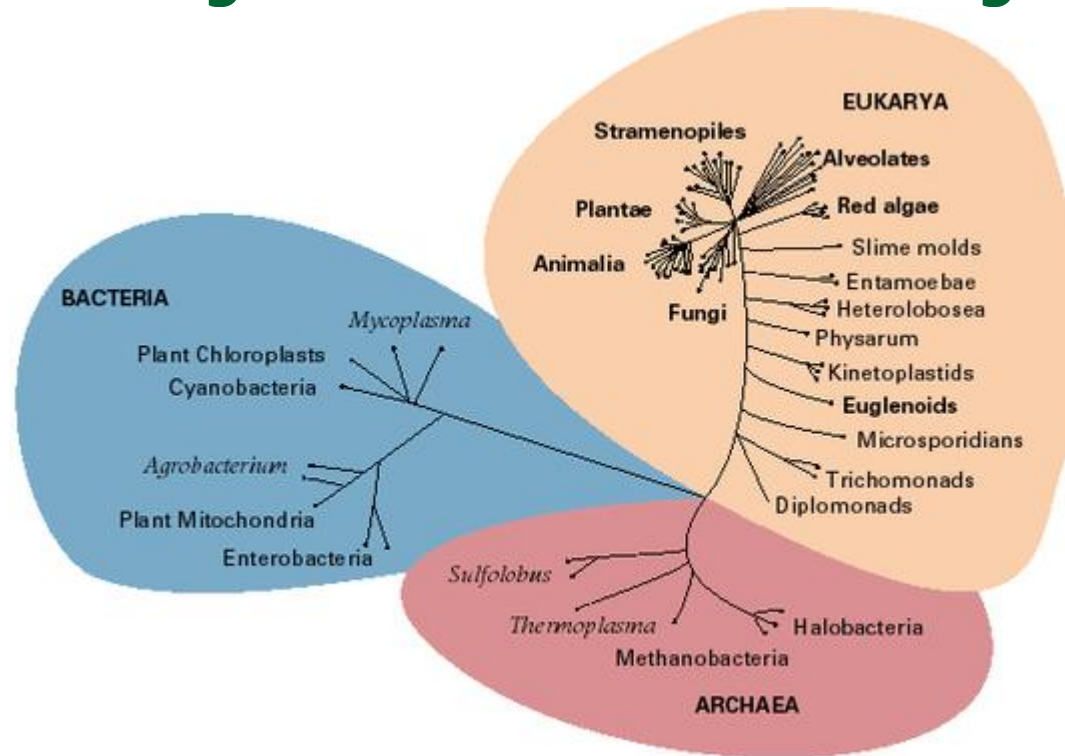


- Born, eat, replicate, and die

2 types of cells: Prokaryotes v.s. Eukaryotes



Prokaryotes and Eukaryotes



- According to the most recent evidence, there are three main branches to the tree of life.
- Prokaryotes include Archaea ("ancient ones") and bacteria.
- Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae.

Prokaryotes and Eukaryotes, continued

Prokaryotes	Eukaryotes
Single cell	Single or multi cell
No nucleus	Nucleus
No organelles	Organelles
One piece of circular DNA	Chromosomes
No mRNA post transcriptional modification	Exons/Introns splicing

Prokaryotes v.s. Eukaryotes

Structural differences

Prokaryotes

- Eubacterial (blue green algae) and archaeobacteria
- only one type of membrane – plasma membrane forms
 - the **boundary** of the cell proper
- The smallest cells known are bacteria
 - Ecoli cell
 - 3×10^6 protein molecules
 - 1000-2000 polypeptide species.

Eukaryotes

- plants, animals, Protista, and fungi
- complex systems of internal membranes forms
 - organelle and compartments
- The volume of the cell is several hundred times larger
 - Hela cell
 - 5×10^9 protein molecules
 - 5000-10,000 polypeptide species

Prokaryotic and Eukaryotic Cells

Chromosomal differences

Prokaryotes

- The genome of E.coli contains amount of 4×10^6 base pairs
- > 90% of DNA encode protein
- Lacks a membrane-bound nucleus.
 - Circular DNA and supercoiled domain
- Histones are unknown

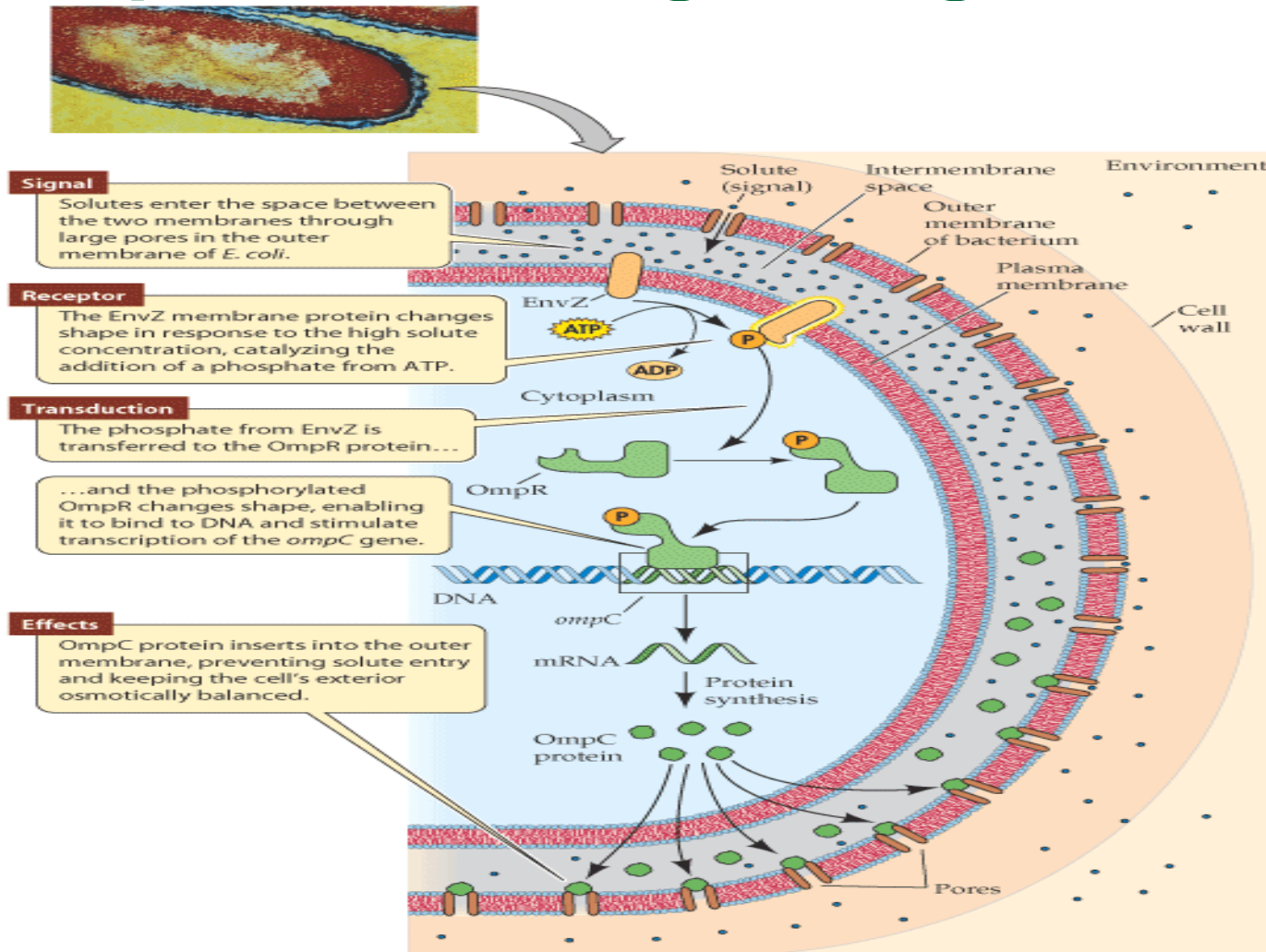
Eukaryotes

- The genome of yeast cells contains 1.35×10^7 base pairs
- A small fraction of the total DNA encodes protein.
 - Many repeats of non-coding sequences
- All chromosomes are contained in a membrane bound nucleus
 - DNA is divided between two or more chromosomes
- A set of five histones
 - DNA packaging and gene expression regulation

Signaling Pathways: Control Gene Activity

- Instead of having brains, cells make decision through complex networks of chemical reactions, called pathways
 - Synthesize new materials
 - Break other materials down for spare parts
 - Signal to eat or die
-

Example of cell signaling



Cells Information and Machinery

- Cells store all information to replicate itself
 - Human genome is around 3 billions base pair long
 - Almost every cell in human body contains same set of genes
 - But not all genes are used or expressed by those cells
 - Machinery:
 - Collect and manufacture components
 - Carry out replication
 - Kick-start its new offspring
- (A cell is like a car factory)
-

Overview of organizations of life

- **Nucleus = library**
 - **Chromosomes = bookshelves**
 - **Genes = books**
 - Almost every cell in an organism contains the same libraries and the same sets of books.
 - Books represent all the information (DNA) that every cell in the body needs so it can grow and carry out its various functions.
-

Some Terminology

- **Genome**: an organism's genetic material
 - **Gene**: a discrete units of hereditary information located on the chromosomes and consisting of DNA.
 - **Genotype**: The genetic makeup of an organism
 - **Phenotype**: the physical expressed traits of an organism
 - **Nucleic acids**: Biological molecules (RNA and DNA) that allow organisms to reproduce
-

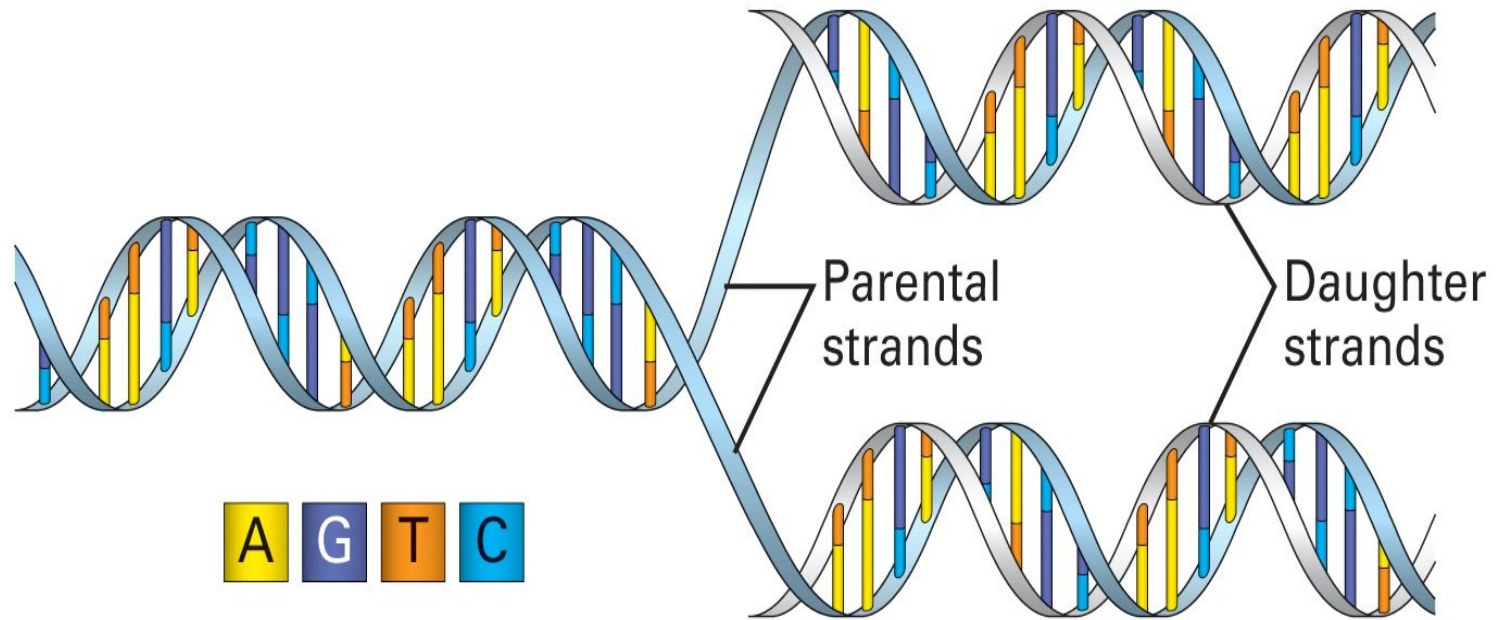
More Terminology

- The **genome** is an organism's complete set of DNA
 - a bacteria contains about 600,000 DNA base pairs
 - human and mouse genomes have some 3 billion
- human genome has 24 distinct chromosomes
 - Each chromosome contains many **genes**.
- **Gene**
 - basic physical and functional units of heredity
 - specific sequences of DNA bases that encode instructions on how to make **proteins**
- **Proteins**
 - Make up the cellular structure
 - large, complex molecules made up of smaller subunits called **amino acids**

All Life depends on 3 critical molecules

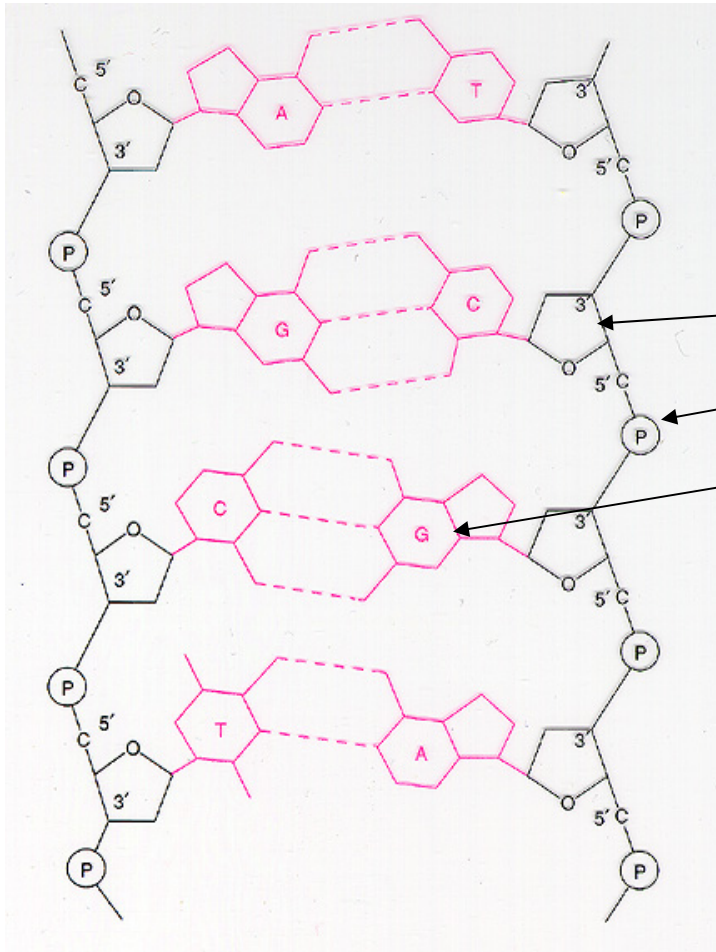
- **DNAs**
 - Hold information on how cell works
 - **RNAs**
 - Act to transfer short pieces of information to different parts of cell
 - Provide templates to synthesize into protein
 - **Proteins**
 - Form enzymes that send signals to other cells and regulate gene activity
 - Form body's major components (e.g. hair, skin, etc.)
-

DNA: The Code of Life



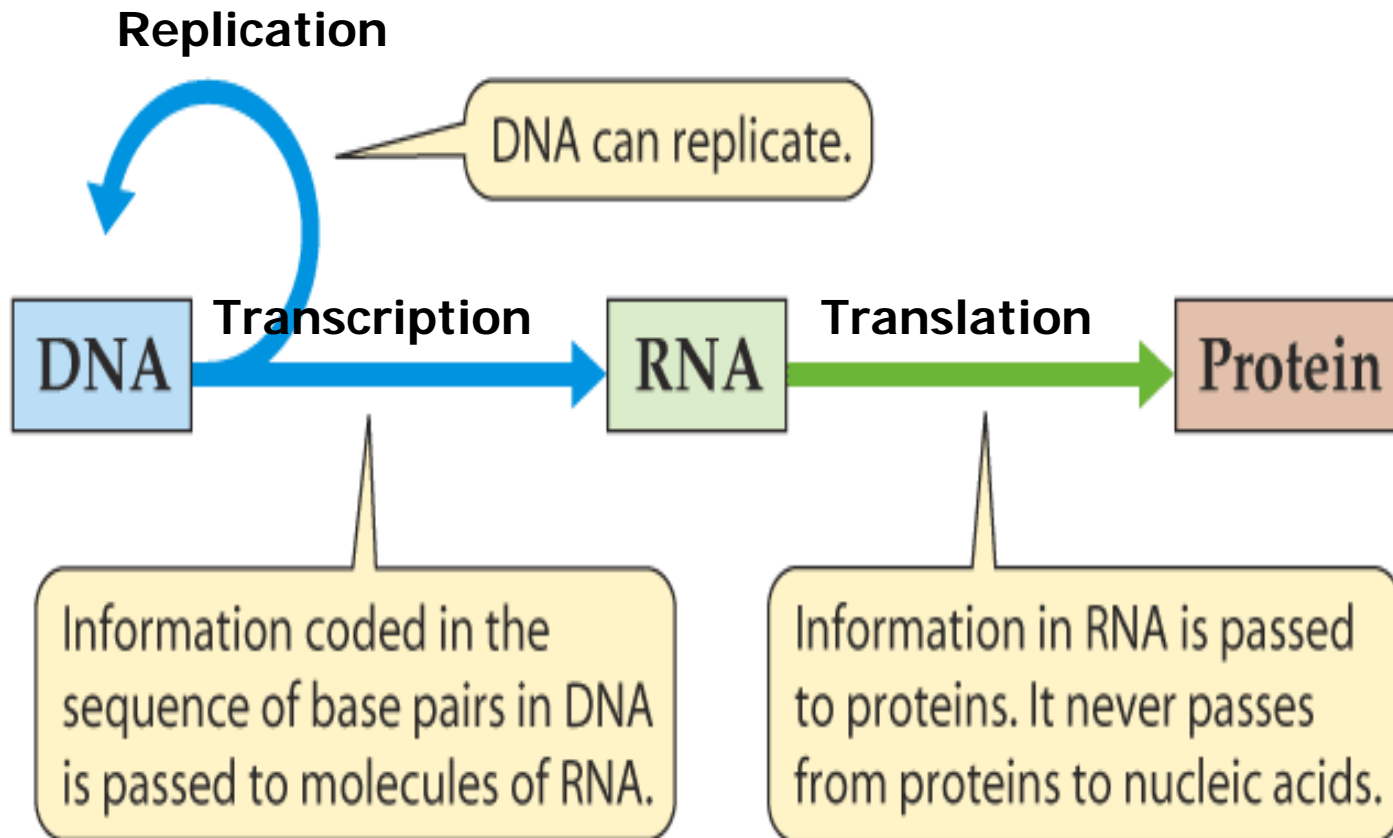
- The structure and the four genomic letters code for all living organisms.
- Adenine, guanine, thymine, and cytosine which pair A-T and C-G on complimentary strands.

DNA, continued

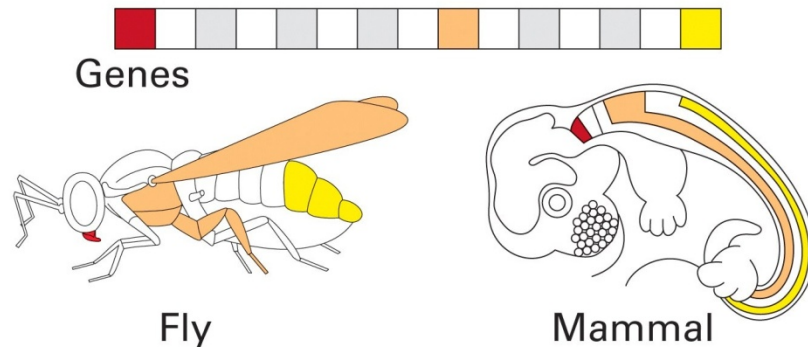


- DNA has a double helix structure which composed of
 - sugar molecule
 - phosphate group
 - and a base (A,C,G,T)
- DNA always reads from 5' end to 3' end for transcription replication
5' ATTTAGGCC 3'
3' TAAATCCGG 5'

DNA, RNA, and the Flow of Information



DNA the Genetics Makeup



- Genes are inherited and are expressed
 - **genotype** (genetic makeup)
 - **phenotype** (physical expression)

- On the left, is the eye's phenotypes of green and black eye genes.



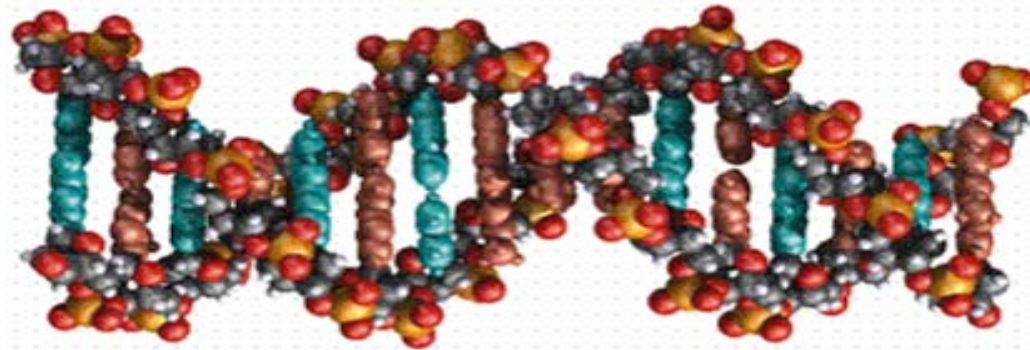
Cell Information: Instruction book of Life

- DNA, RNA, and Proteins are examples of strings written
 - in either the four-letter nucleotide of DNA and RNA (A C G T/U)
 - or the twenty-letter amino acid of proteins. Each amino acid is coded by 3 nucleotides called codon. (Leu, Arg, Met, etc.)

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

END of SECTION 2

Section 3: Genetic Material of Life



Outline For Section 3:

- What is genetic material?
 - ***Mendel's experiments***
 - *Pea plant experiments*
 - ***Mutations in DNA***
 - Good, bad, silent
 - ***Chromosomes***
 - Linked genes
 - Gene order
 - Genetic maps
 - Chromosomes and sexual reproduction
-

Mendel and his Genes

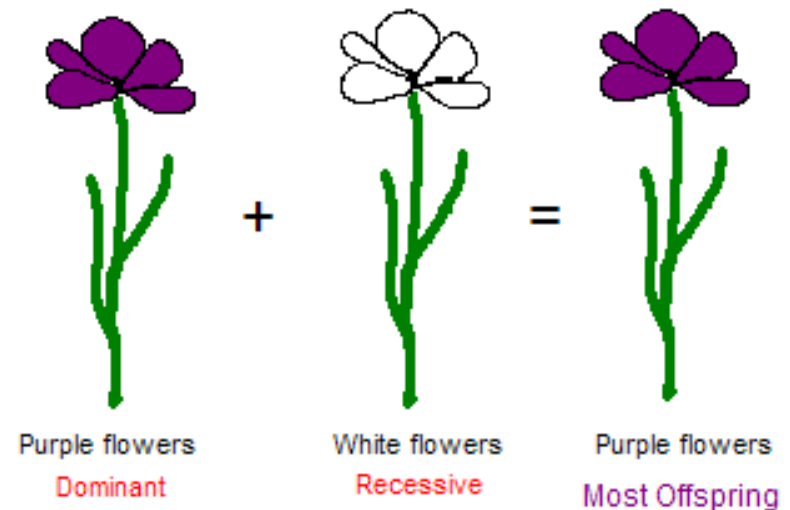
- What are genes?
 - physical and functional traits that are passed on from one generation to the next.
- Genes were discovered by Gregor Mendel in the 1860s while he was experimenting with the pea plant. He asked the question:

**Do traits come from a blend of both parent's traits
or from only one parent?**

The Pea Plant Experiments

- Mendel discovered that genes were passed on to offspring by both parents in two forms: dominant and recessive.

- The dominant form would be the phenotypic characteristic of the offspring



DNA: the building blocks of genetic material

- DNA was later discovered to be the molecule that makes up the inherited genetic material
- Experiments performed by Fredrick Griffith in 1928 and experiments with bacteriophages in 1952 led to this discovery
- DNA provides a code, consisting of 4 letters, for all cellular function.

Letters in DNA code: **C A G T**

MUtAsHONS

- The DNA can be thought of as a sequence of the nucleotides: C, A, G, or T
- What happens to genes when the DNA sequence is mutated?

Normal DNA sequence:

ATCTAG

Mutated DNA sequence:

ATCGAG



Genes are Organized into Chromosomes

- What are chromosomes?
 - It is a threadlike structure found in the nucleus of the cell which is made from a long strand of DNA. Different organisms have a different number of chromosomes in their cells.
- *Thomas Morgan* (1920s) - Evidence that genes are located on chromosomes was discovered by genetic experiments performed with flies.

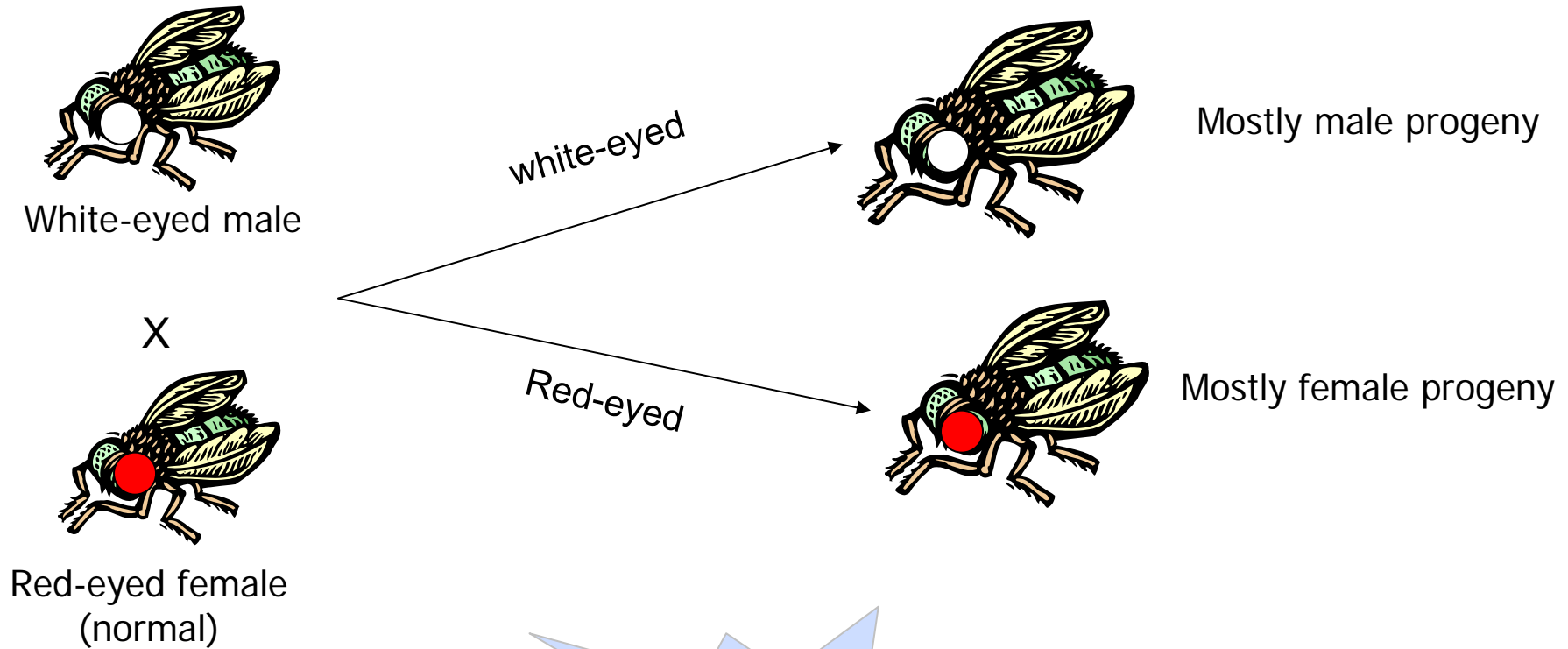


Portrait of Morgan

<http://www.nobel.se/medicine/laureates/1933/morgan-bio.html>



The White-Eyed Male



These experiments suggest that the gene for eye color must be linked or co-inherited with the genes that determine the sex of the fly. This means that the genes occur on the same chromosome; more specifically it was the X chromosome.

Linked Genes and Gene Order

- Along with eye color and sex, other genes, such as body color and wing size, had a higher probability of being co-inherited by the offspring → genes are linked.
- *Morgan* hypothesized that the closer the genes were located on the a chromosome, the more often the genes are co-inherited.

Linked Genes and Gene Order cont...

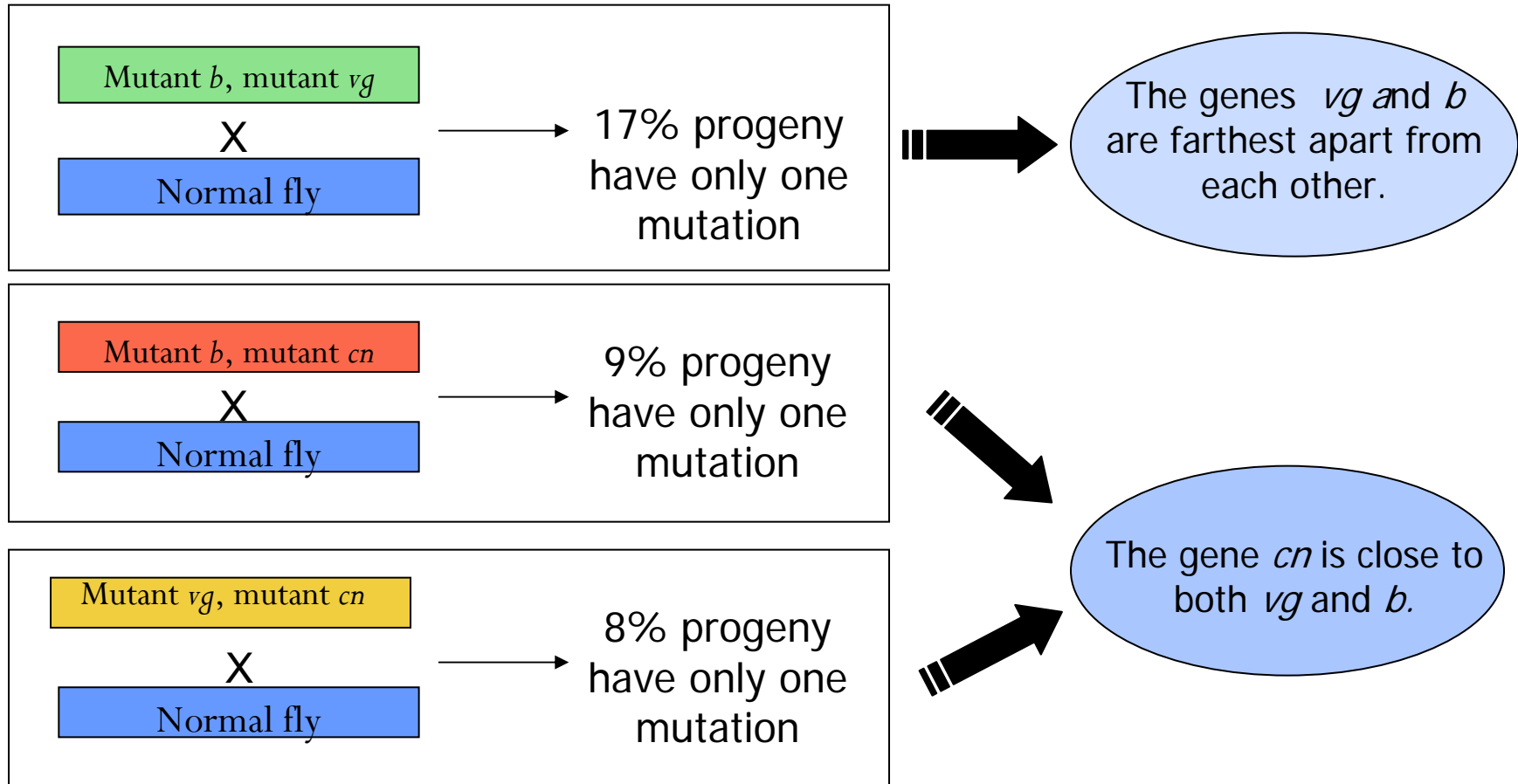
- By looking at the frequency that two genes are co-inherited, genetic maps can be constructed for the location of each gene on a chromosome.
- One of *Morgan's* students *Alfred Sturtevant* pursued this idea and studied 3 fly genes:



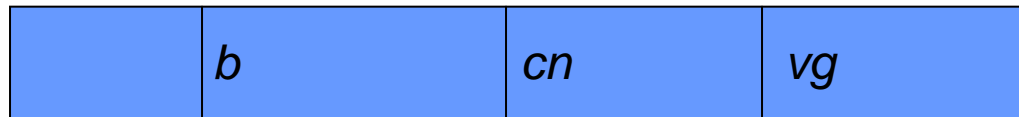
***en* - eye color**
***h* - body color**
***wg* - wing size**



What are the genes' order on the chromosome?

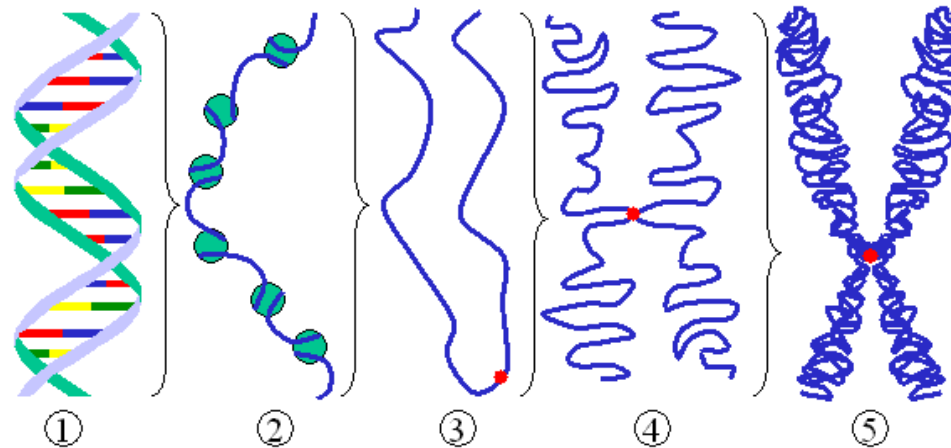


What are the genes' order on the chromosome?



*This is the order of the genes, on the chromosome,
determined by the experiment*

Genetic Information: Chromosomes



- 1) Double helix DNA strand.
- 2) Chromatin strand (**DNA** with **histones**)
- 3) Condensed chromatin during interphase with **centromere**.
- 4) Condensed chromatin during prophase
- 5) Chromosome during metaphase

Chromosomes

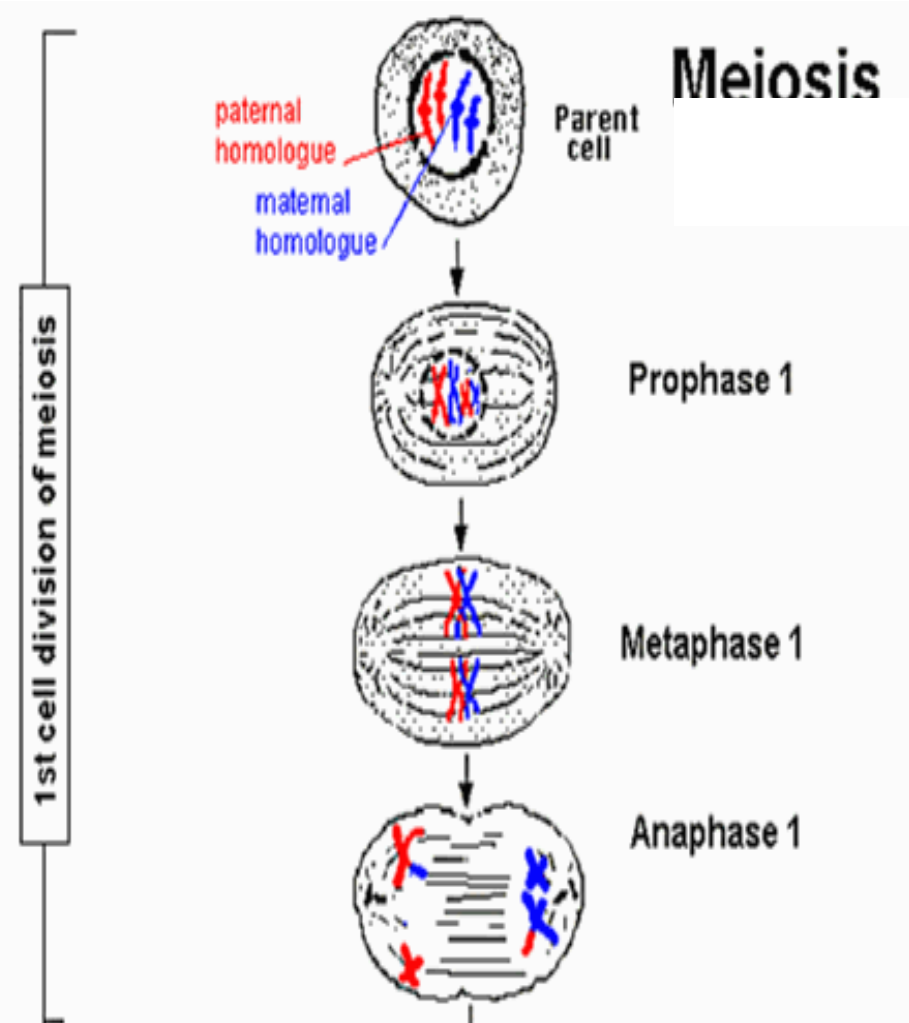
<u>Organism</u>	<u>Number of base pair</u>	<u>number of Chromosomes</u>
Prokaryotic		
Escherichia coli (bacterium)	4×10^6	1
Eukaryotic		
Saccharomyces cerevisiae (yeast)	1.35×10^7	17
Drosophila melanogaster (insect)	1.65×10^8	4
Homo sapiens (human)	2.9×10^9	23
Zea mays (corn)	5.0×10^9	10

Sexual Reproduction

- Formation of new individual by a combination of two haploid sex cells (gametes).
 - Fertilization – combination of genetic information from two separate cells that have one half the original genetic information
 - Gametes for fertilization usually come from separate parents
 1. Female produces an egg
 2. Male produces sperm
 - Both gametes are haploid, with a single set of chromosomes
 - The new individual is called a zygote, with two sets of chromosomes (diploid).
 - **Meiosis** is a process to convert a diploid cell to a haploid gamete, and cause a *change in the genetic information* to increase diversity in the offspring.
-

Meiosis

- Meiosis comprises two successive nuclear divisions with only one round of DNA replication.
- **First division of meiosis**
 - **Prophase 1:** Each chromosome duplicates and remains closely associated. These are called sister chromatids. Crossing-over can occur during the latter part of this stage.
 - **Metaphase 1:** Homologous chromosomes align at the equatorial plate.
 - **Anaphase 1:** Homologous pairs separate with sister chromatids remaining together.
 - **Telophase 1:** Two daughter cells are formed with each daughter containing only one chromosome of the homologous pair.

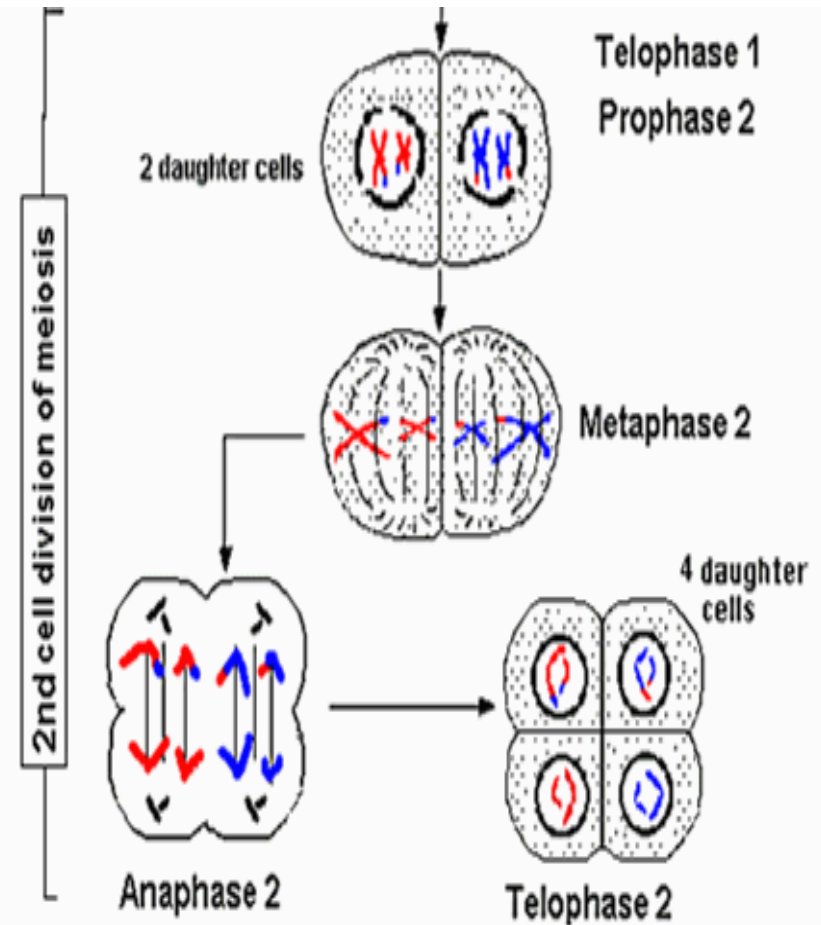


Meiosis

- **Second division of meiosis:** Gamete formation
 - **Prophase 2:** DNA does not replicate.
 - **Metaphase 2:** Chromosomes align at the equatorial plate.
 - **Anaphase 2:** Centromeres divide and sister chromatids migrate separately to each pole.
 - **Telophase 2:** Cell division is complete. Four haploid daughter cells are obtained
- One parent cell produces **four daughter cells**.

Daughter cells:

- half the number of chromosomes found in the original parent cell
- crossing over cause genetically difference.



END of SECTION 3

Section 4: What Do Genes Do?

Outline For Section 4:

- *Beadle and Tatum Experiment*
 - *Design of Life* (gene → protein)
 - protein synthesis
 - Central dogma of molecular biology
-

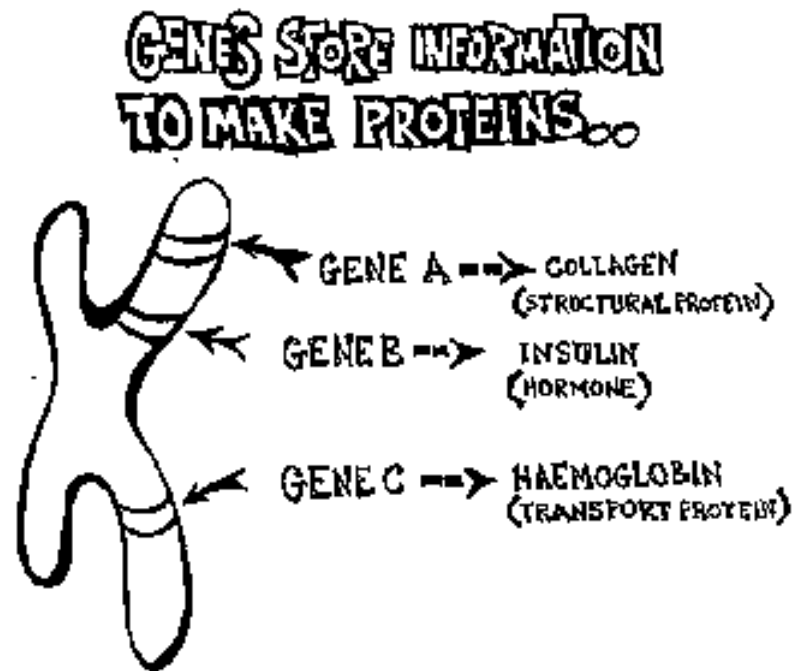
Beadle and Tatum Experiment

Conclusions

- Irradiated Neurospora survived when supplemented with Vitamin B6
- X-rays damaged genes that produces a protein responsible for the synthesis of Vitamin B6
- three mutant strains – substances unable to synthesize (Vitamin B6, Vitamin B1 and Para-aminobenzoic acid) essential growth factors
- crosses between normal and mutant strains showed differed by a single gene
- hypothesized that there was more than one step in the synthesis of Vitamin B6 and that mutation affects only one specific step
- Evidence: One gene specifies the production of one enzyme!

Genes Make Proteins

- Genome → genes → protein (forms cellular structural & life functional) → pathways & physiology



Proteins: Workhorses of the Cell

- 20 different **amino acids**
 - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.
 - Proteins do all essential work for the cell
 - build cellular structures
 - digest nutrients
 - execute metabolic functions
 - Mediate information flow within a cell and among cellular communities.
 - Proteins work together with other proteins or nucleic acids as "molecular machines"
 - structures that fit together and function in highly specific, lock-and-key ways.
-

END of SECTION 4

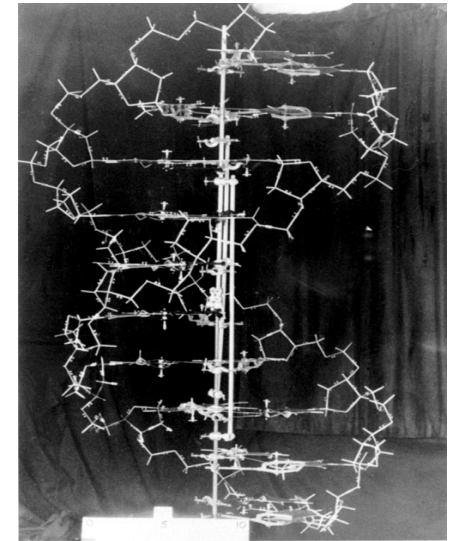
Section 5: What Molecule Codes For Genes?

Outline For Section 5:

- *Discovery of the Structure of DNA*
 - *Watson and Crick*
- *DNA Basics*

Discovery of DNA

- DNA Sequences
 - Chargaff and Vischer, 1949
 - DNA consisting of A, T, G, C
 - Adenine, Guanine, Cytosine, Thymine
 - Chargaff Rule
 - Noticing $\#A \approx \#T$ and $\#G \approx \#C$
 - A "strange but possibly meaningless" phenomenon.
- Wow!! A Double Helix
 - Watson and Crick, *Nature*, April 25, 1953
 - **1 Biologist (Watson)**
 - **1 Physics Ph.D. Student (Crick)**
 - + 900 words
 - = Nobel Prize
 - Rich, 1973
 - Structural biologist at MIT.
 - DNA's structure in atomic resolution.



Original DNA demonstration model (scale gives distance in Angstroms) Cold Spring Harbor Laboratory Archives



Watson and Crick walk along the Beach Cold Spring Harbor Laboratory Archives

Crick

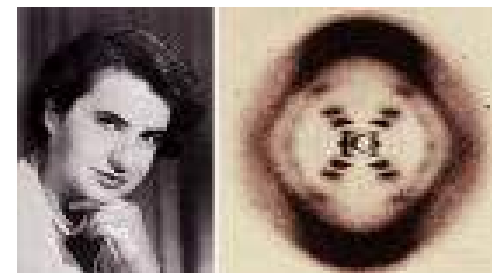
Watson

Watson & Crick – “...the secret of life”

- Watson: a zoologist, Crick: a physicist
- *“In 1947 Crick knew no biology and practically no organic chemistry or crystallography..”* – www.nobel.se
- Applying Chagraff’s rules and the X-ray image from Rosalind Franklin, they constructed a “tinkertoy” model showing the double helix
- Their 1953 *Nature* paper: *“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”*



Watson & Crick with DNA model



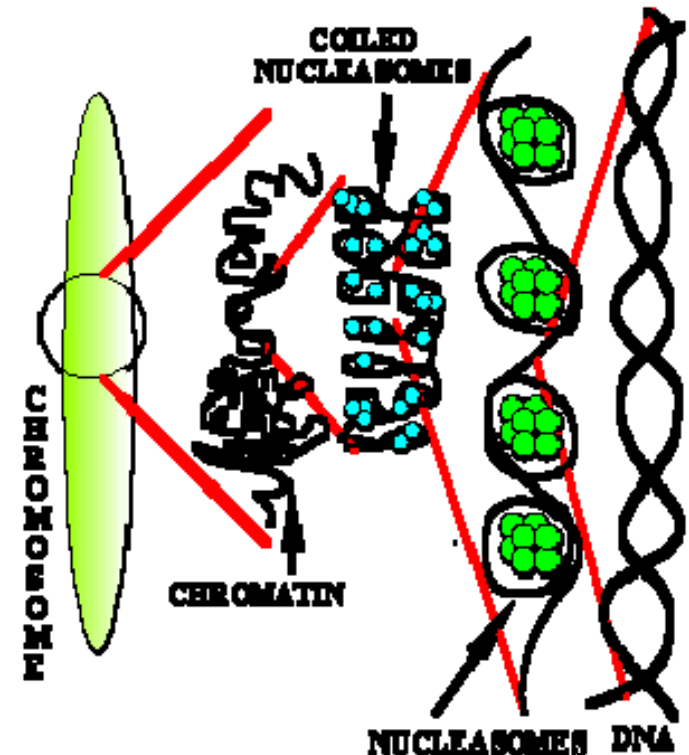
Rosalind Franklin with X-ray image of DNA

Double helix of DNA

- James Watson and Francis Crick proposed a model for the structure of DNA.
 - Utilizing X-ray diffraction data, obtained from crystals of DNA
 - This model predicted that DNA
 - as a helix of two complementary anti-parallel strands,
 - wound around each other in a rightward direction
 - stabilized by H-bonding between bases in adjacent strands.
 - The bases are in the interior of the helix
 - Purine bases form hydrogen bonds with pyrimidine.
-

DNA: The Basis of Life

- Humans have about 3 billion base pairs.
 - How do you package it into a cell?
 - How does the cell know where in the highly packed DNA to start transcription?
 - Special regulatory sequences
 - DNA size does not mean more complex
- Complexity of DNA
 - Eukaryotic genomes consist of variable amounts of DNA
 - Single Copy or Unique DNA
 - Highly Repetitive DNA



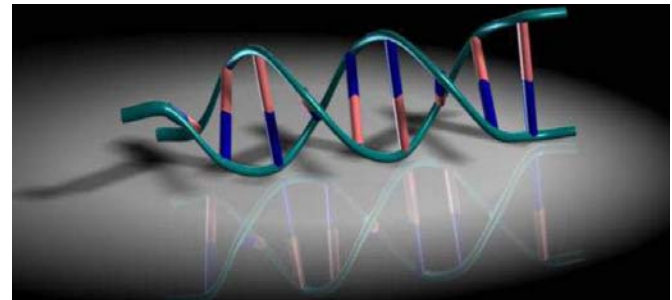
END of SECTION 5

Section 6: The Structure of DNA

CSE 181

Raymond Brown

May 12, 2004

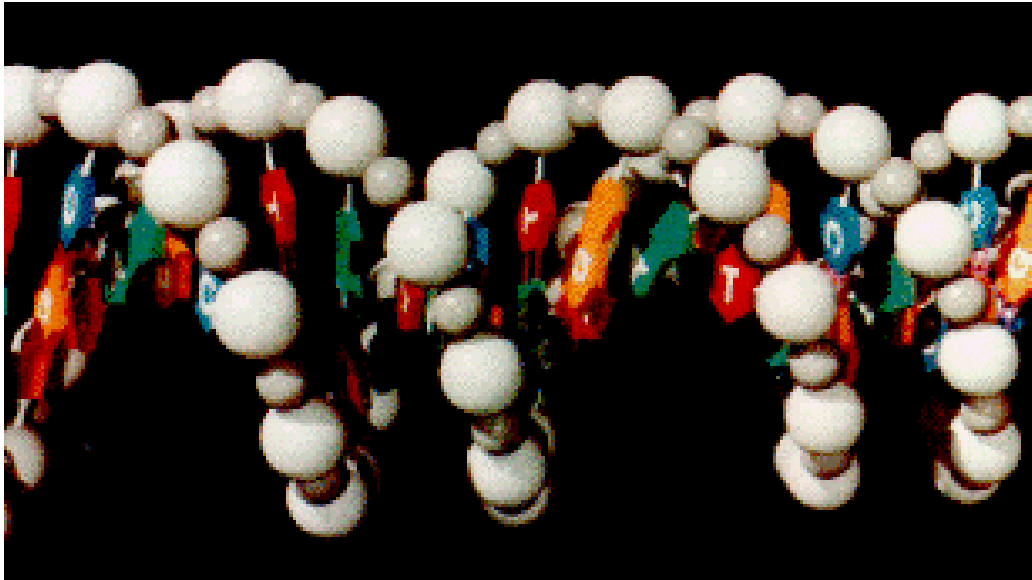


Outline For Section 6:

- *DNA Components*
 - Nitrogenous Base
 - Sugar
 - Phosphate
 - *Double Helix*
 - *DNA replication*
 - *Superstructure*
-

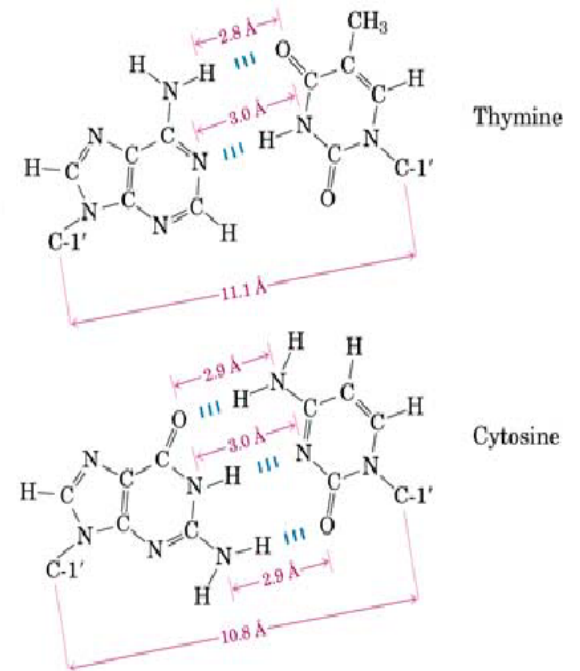
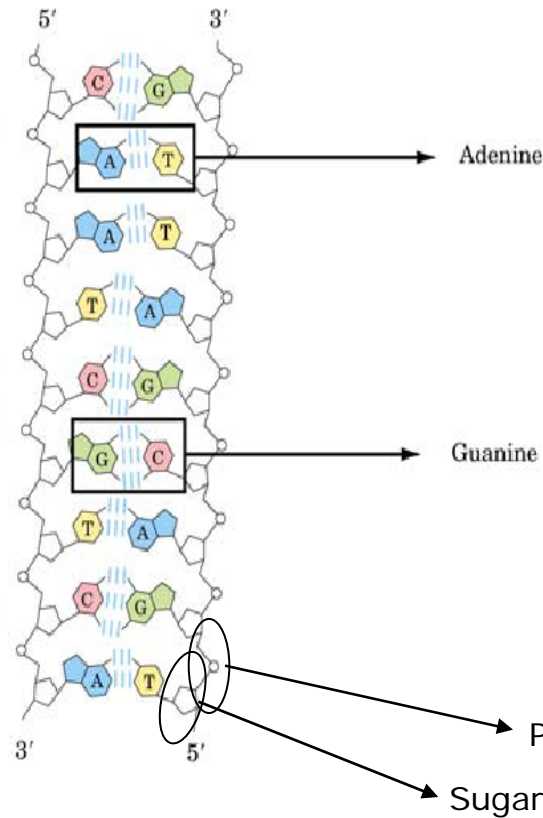
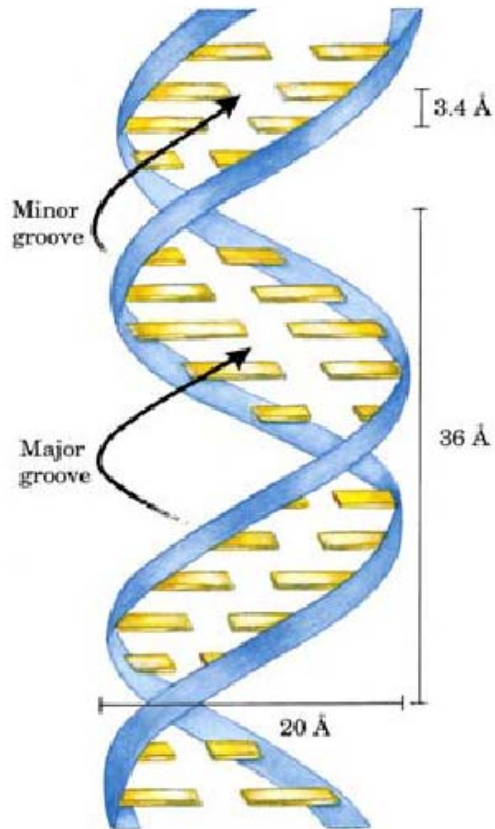
DNA

- Stores all information of life
- 4 “letters” base pairs. AGTC (adenine, guanine, thymine, cytosine) which pair A-T and C-G on complimentary strands.



Basic Structure

Watson-Crick base pair structures



Phosphate
Sugar

Basic Structure Implications

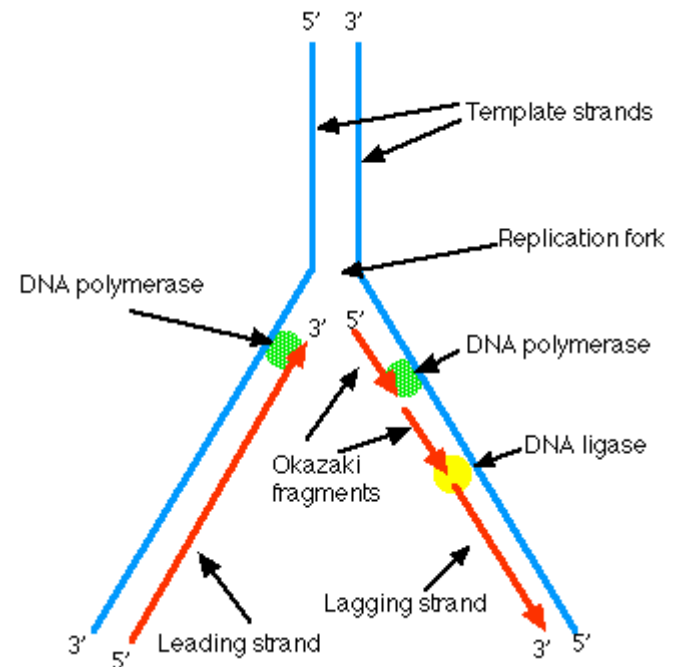
- **DNA is (-) charged due to phosphate:**
 - gel electrophoresis, DNA sequencing (Sanger method)
- **H-bonds form between specific bases:**
 - hybridization – replication, transcription, translation
 - DNA microarrays, hybridization blots, PCR
 - C-G bound tighter than A-T due to triple H-bond
- **DNA polymerization:**
 - 5' to 3' – phosphodiester bond formed between 5' phosphate and 3' OH

Double helix of DNA

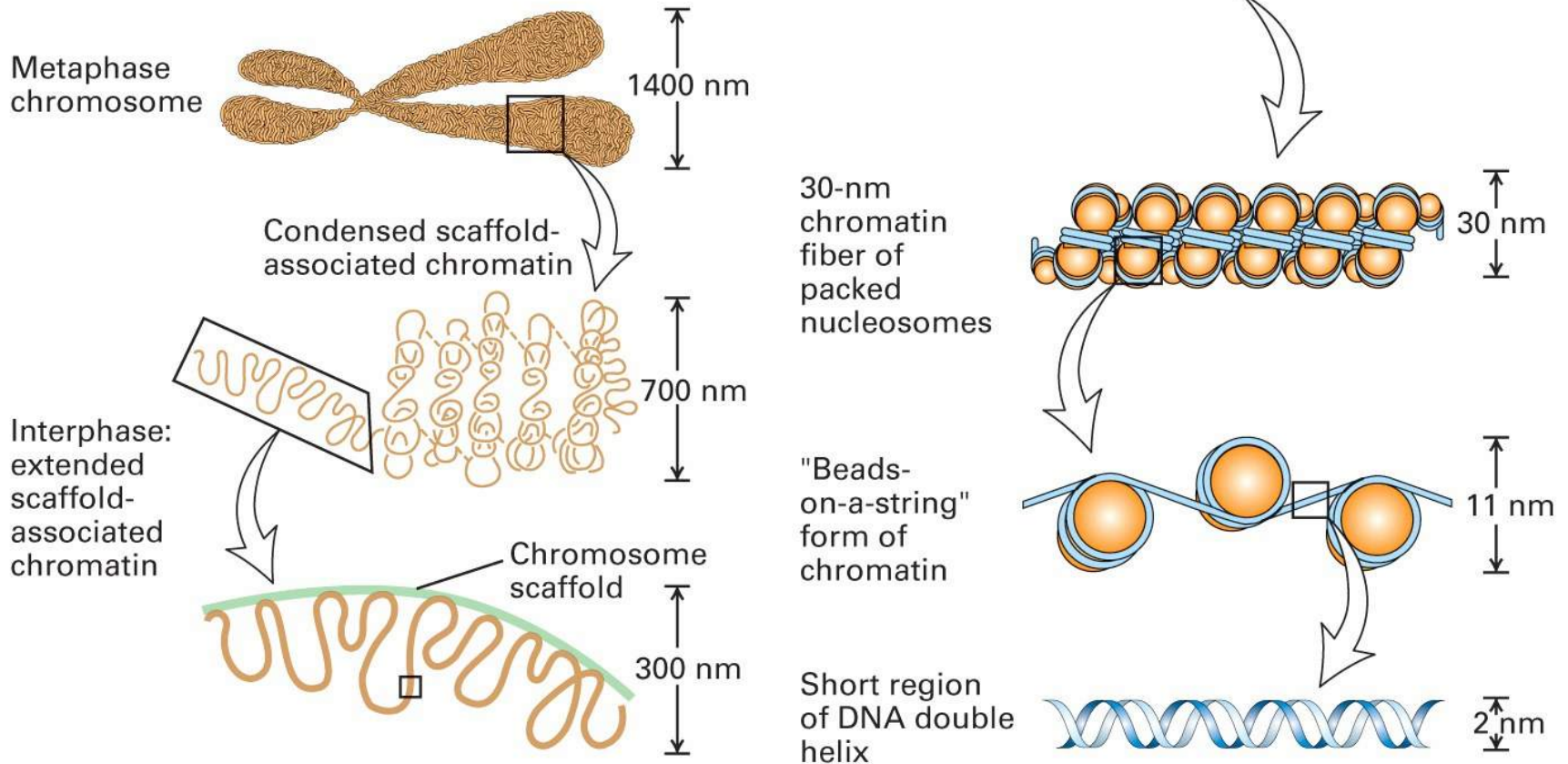
- The DNA strands are assembled in the 5' to 3' direction
 - by convention, we "read" them the same way.
 - The phosphate group bonded to the 5' carbon atom of one deoxyribose is covalently bonded to the 3' carbon of the next.
 - The purine or pyrimidine attached to each deoxyribose projects in toward the axis of the helix.
 - Each base forms hydrogen bonds with the one directly opposite it, forming base pairs (also called nucleotide pairs).
-

DNA - replication

- DNA can replicate by splitting, and rebuilding each strand.
- Note that the rebuilding of each strand uses slightly different mechanisms due to the 5' 3' asymmetry, but each daughter strand is an exact replica of the original strand.



Superstructure



Superstructure Implications

- DNA in a living cell is in a highly compacted and structured state
 - Transcription factors and RNA polymerase need ACCESS to do their work
 - Transcription is dependent on the structural state – SEQUENCE alone does not tell the whole story
-

END of SECTION 6

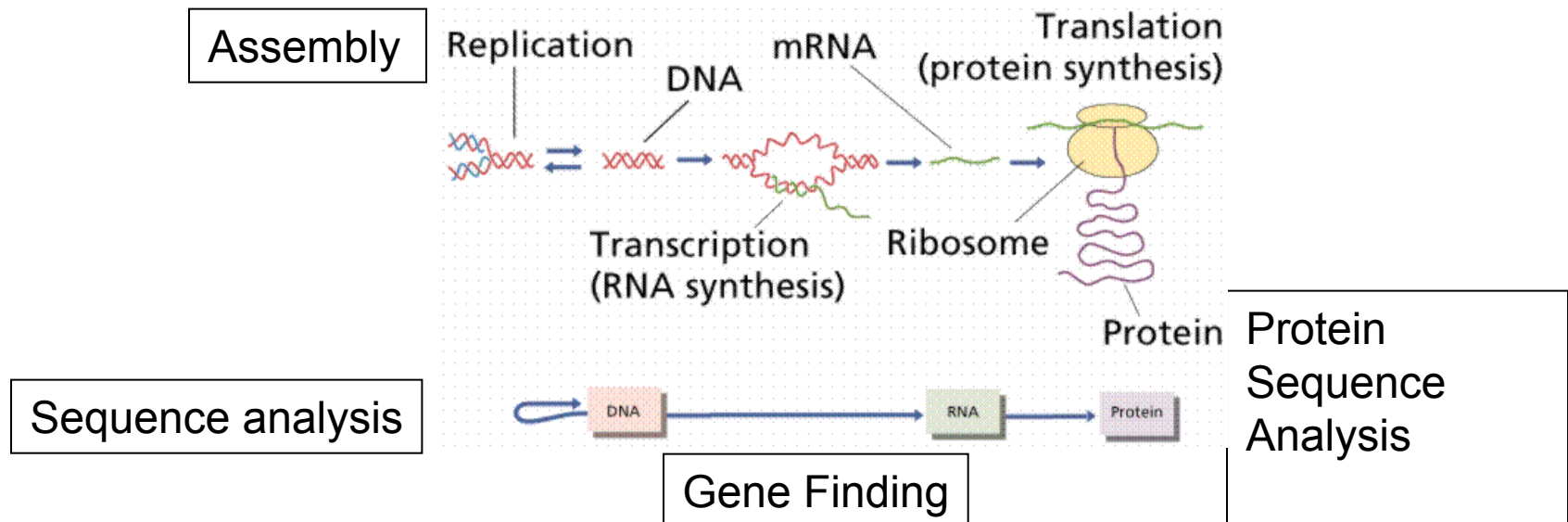
Section 7: What carries information between DNA to Proteins

Outline For Section 7:

- *Central Dogma Of Biology*
 - *RNA*
 - *Transcription*
 - *Splicing hnRNA -> mRNA*
-

Central Dogma of Biology

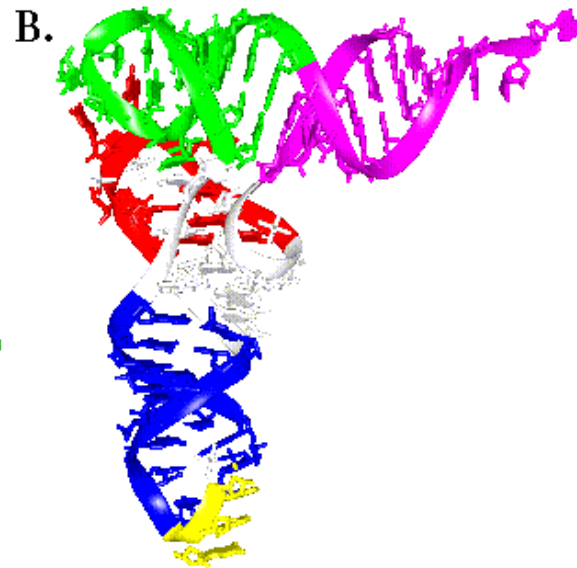
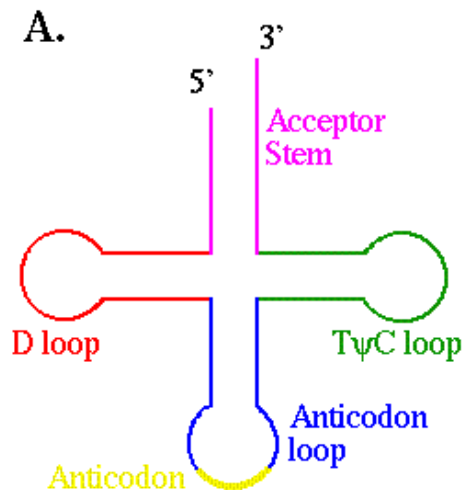
The information for making proteins is stored in DNA. There is a process (transcription and translation) by which DNA is converted to protein. By understanding this process and how it is regulated we can make predictions and models of cells.



RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Some forms of RNA can form secondary structures by “pairing up” with itself. This can have change its properties dramatically.

DNA and RNA
can pair with
each other.



RNA, continued

- Several types exist, classified by function
- **m**RNA – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s **m**essage out of the nucleus.
- **t**RNA – **t**ransfers genetic information from mRNA to an amino acid sequence
- **r**RNA – **r**ibosomal RNA. Part of the ribosome which is involved in translation.

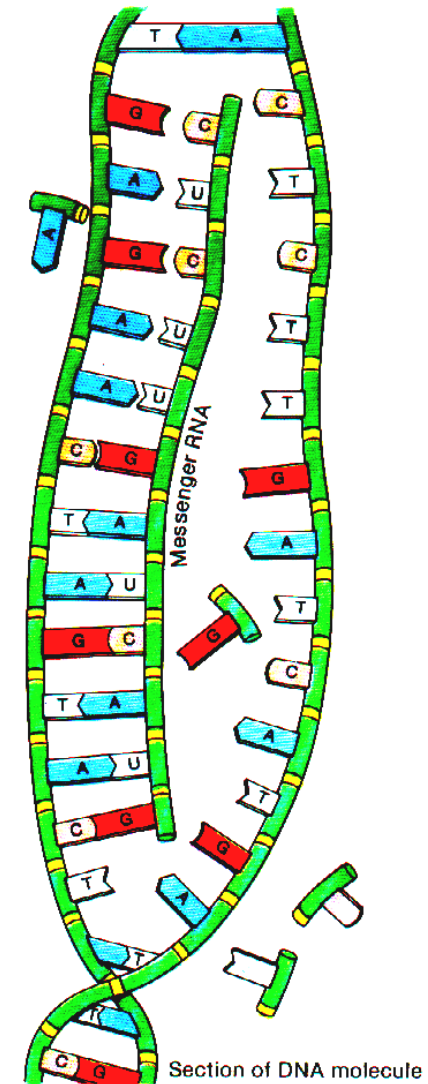
Terminology for Transcription

- **hnRNA (heterogeneous nuclear RNA)**: Eukaryotic mRNA primary transcripts whose introns have not yet been excised (**pre-mRNA**).
- **Phosphodiester Bond**: Esterification linkage between a phosphate group and two alcohol groups.
- **Promoter**: A special sequence of nucleotides indicating the starting point for RNA synthesis.
- **RNA (ribonucleotide)**: Nucleotides A, U, G, and C with ribose
- **RNA Polymerase II**: Multisubunit enzyme that catalyzes the synthesis of an RNA molecule on a DNA template from nucleoside triphosphate precursors.
- **Terminator**: Signal in DNA that halts transcription.

Transcription

- The process of making RNA from DNA
- Catalyzed by “transcriptase” enzyme - *RNA-polymerase*
- Needs a promoter region to begin transcription.
- ~50 base pairs/second in bacteria, but multiple transcriptions can occur simultaneously

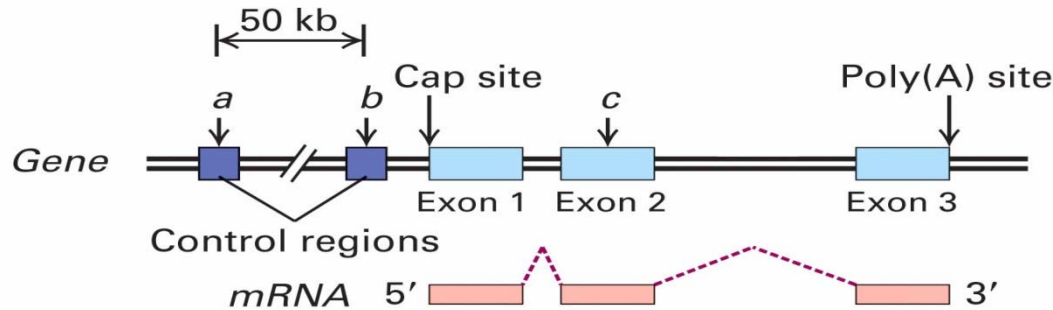
KEY
T = thymine
C = cytosine
A = adenine
G = guanine



Transcription, continued

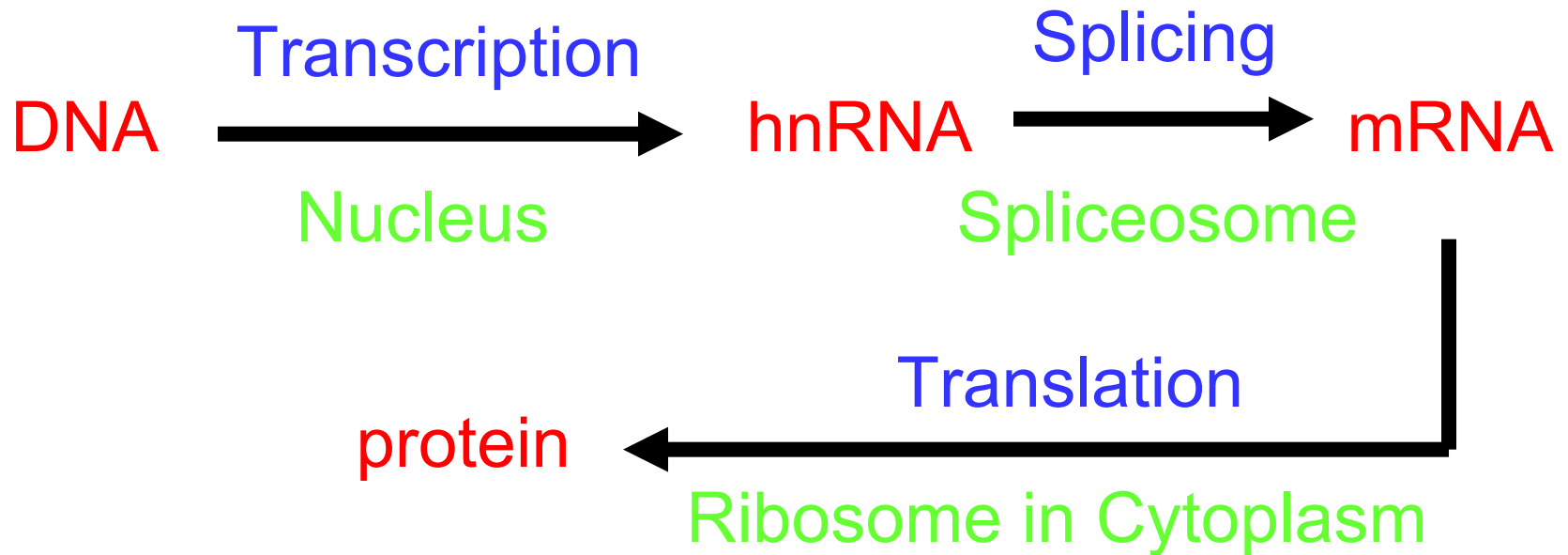
- Transcription is highly regulated. Most DNA is in a dense form where it cannot be transcribed.
 - To begin transcription requires a promoter, a small specific sequence of DNA to which polymerase can bind (~40 base pairs “upstream” of gene)
 - Finding these promoter regions is a partially solved problem that is related to motif finding.
 - There can also be repressors and inhibitors acting in various ways to stop transcription. This makes regulation of gene transcription complex to understand.
-

Definition of a Gene



- **Regulatory regions:** up to 50 kb upstream of +1 site
- **Exons:** protein coding and untranslated regions (UTR)
1 to 178 exons per gene (mean 8.8)
8 bp to 17 kb per exon (mean 145 bp)
- **Introns:** splice acceptor and donor sites, junk DNA
average 1 kb – 50 kb per intron
- **Gene size:** Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.

Central Dogma Revisited



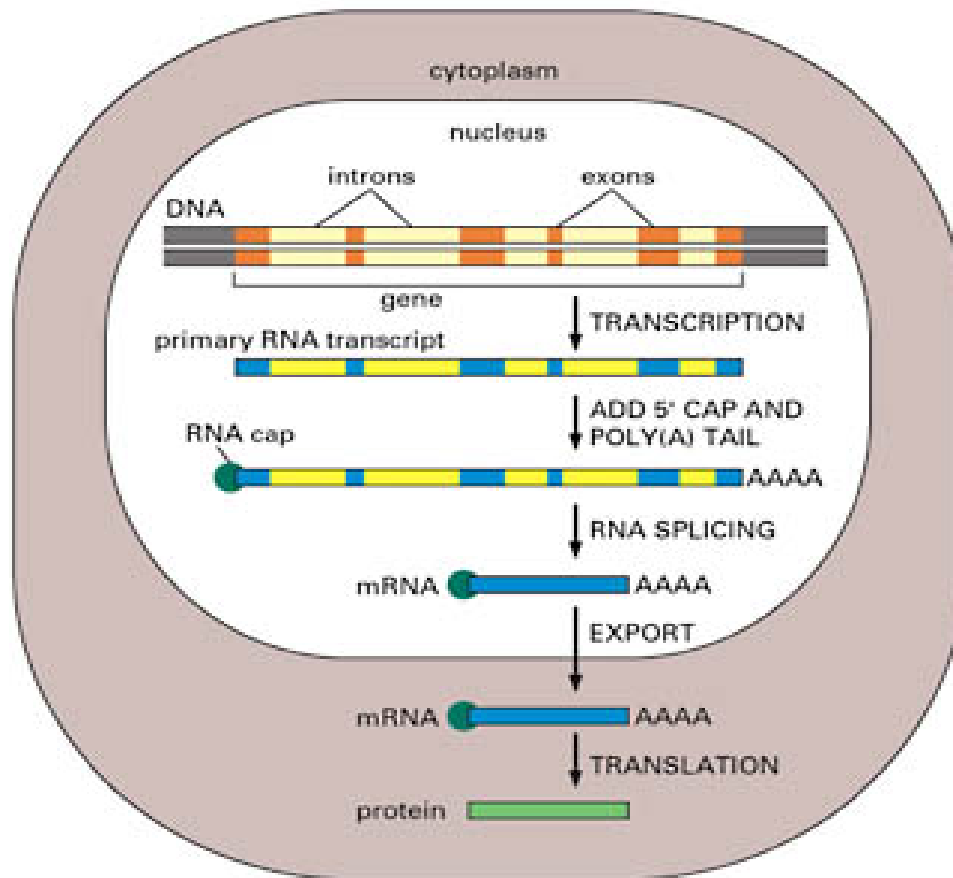
- **Base Pairing Rule:** A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- **Note:** Some mRNA stays as RNA (i.e. tRNA, rRNA).

Terminology for Splicing

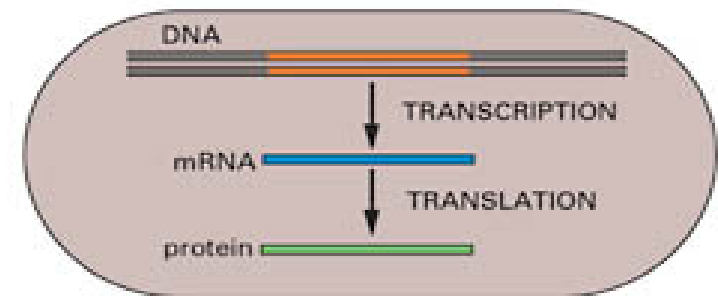
- **Exon**: A portion of the gene that appears in both the primary and the mature mRNA transcripts.
 - **Intron**: A portion of the gene that is transcribed but excised prior to translation.
 - **Lariat structure**: The structure that an intron in mRNA takes during excision/splicing.
 - **Spliceosome**: A organelle that carries out the splicing reactions whereby the pre-mRNA is converted to a mature mRNA.
-

Splicing

(A) EUCARYOTES



(B) PROCARYOTES



END of SECTION 7

Section 8: How Are Proteins Made? (Translation)

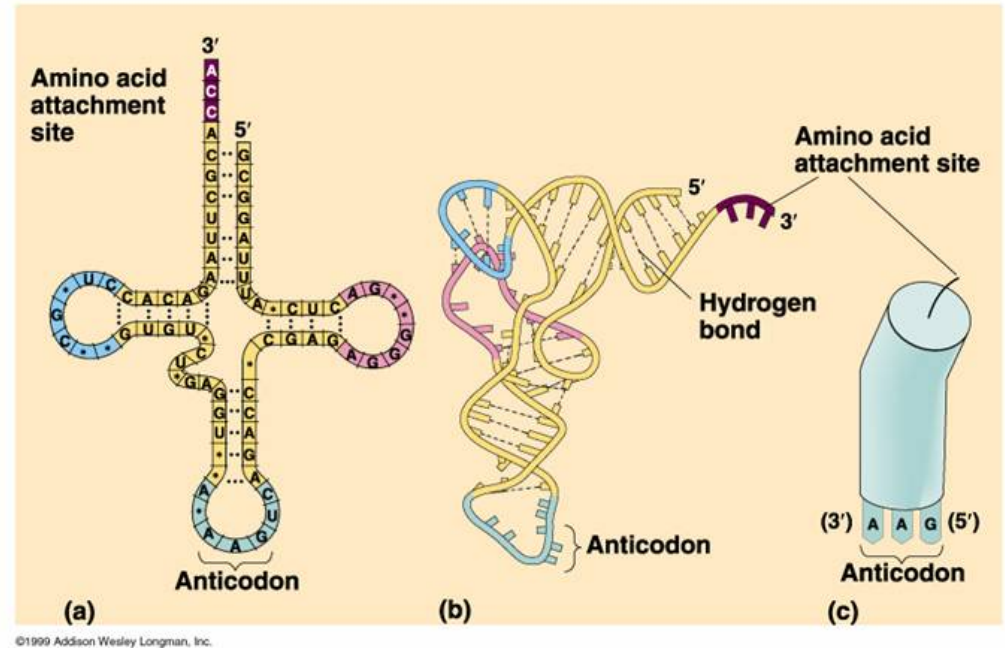
Outline For Section 8:

- *mRNA*
 - *tRNA*
 - *Translation*
 - *Protein Synthesis*
 - *Protein Folding*
-

Terminology for Ribosome

- **Codon**: The sequence of 3 nucleotides in DNA/RNA that encodes for a specific amino acid.
- **mRNA (messenger RNA)**: A ribonucleic acid whose sequence is complementary to that of a protein-coding gene in DNA.
- **Ribosome**: The organelle that synthesizes polypeptides under the direction of mRNA
- **rRNA (ribosomal RNA)**: The RNA molecules that constitute the bulk of the ribosome and provides structural scaffolding for the ribosome and catalyzes peptide bond formation.
- **tRNA (transfer RNA)**: The small L-shaped RNAs that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA.
- **Anticodon**: The sequence of 3 nucleotides in tRNA that recognizes an mRNA codon through complementary base pairing.

Purpose of tRNA



- The proper tRNA is chosen by having the corresponding anticodon for the mRNA's codon.
- The tRNA then transfers its aminoacyl group to the growing peptide chain.
- For example, the tRNA with the anticodon UAC corresponds with the codon AUG and attaches methionine amino acid onto the peptide chain.

Uncovering the code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there'd be 4^1 amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$
 - This triplet of bases is called a "codon"
 - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid

Translation

- The process of going from RNA to polypeptide.
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.
- Always starts with Methionine and ends with a stop codon

		SECOND POSITION					
		U	C	A	G		
U	phenyl-alanine	leucine	serine	tyrosine	cysteine	U	THIRD POSITION
				stop	stop	C	
				stop	tryptophan	A	
C	leucine	leucine	proline	histidine	arginine	U	THIRD POSITION
				glutamine		C	
						A	
A	isoleucine	methionine*	threonine	asparagine	serine	U	THIRD POSITION
				lysine	arginine	C	
						A	
G	valine	valine	alanine	aspartic acid	glycine	U	THIRD POSITION
				glutamic acid		C	
						A	
						G	

* end start

Proteins

- Complex organic molecules made up of amino acid subunits
- 20* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
- <http://www.bio.davidson.edu/biology/aatable.html> for complete list of chemical structures and abbreviations.
- Proteins are often enzymes that catalyze reactions.
- Also called “poly-peptides”

*Some other amino acids exist but not in humans.

Polypeptide v. Protein

- A protein is a polypeptide, however to understand the function of a protein given only the polypeptide sequence is a very difficult problem.
- Protein **folding** an open problem. The 3D structure depends on many variables.
- Current approaches often work by looking at the structure of homologous (similar) proteins.
- Improper folding of a protein is believed to be the cause of mad cow disease.

Protein Folding

- Proteins are not linear structures, though they are built that way
 - The amino acids have very different chemical properties; they interact with each other after the protein is built
 - This causes the protein to start fold and adopting it's functional structure
 - Proteins may fold in reaction to some ions, and several separate chains of peptides may join together through their hydrophobic and hydrophilic amino acids to form a polymer
-

END of SECTION 8

Section 9: How Can We Analyze DNA?

Outline For Section 9:

- ***9.1 Copying DNA***
 - *Polymerase Chain Reaction*
 - *Cloning*
 - ***9.2 Cutting and Pasting DNA***
 - *Restriction Enzymes*
 - ***9.3 Measuring DNA Length***
 - *Electrophoresis*
 - *DNA sequencing*
 - ***9.4 Probing DNA***
 - *DNA probes*
 - *DNA arrays*
-

Analyzing a Genome

- How to analyze a genome in four easy steps.
 - Cut it
 - Use enzymes to cut the DNA in to small fragments.
 - Copy it
 - Copy it many times to make it easier to see and detect.
 - Read it
 - Use special chemical techniques to read the small fragments.
 - Assemble it
 - Take all the fragments and put them back together. This is hard!!!
 - Bioinformatics takes over
 - What can we learn from the sequenced DNA.
 - Compare interspecies and intraspecies.
-

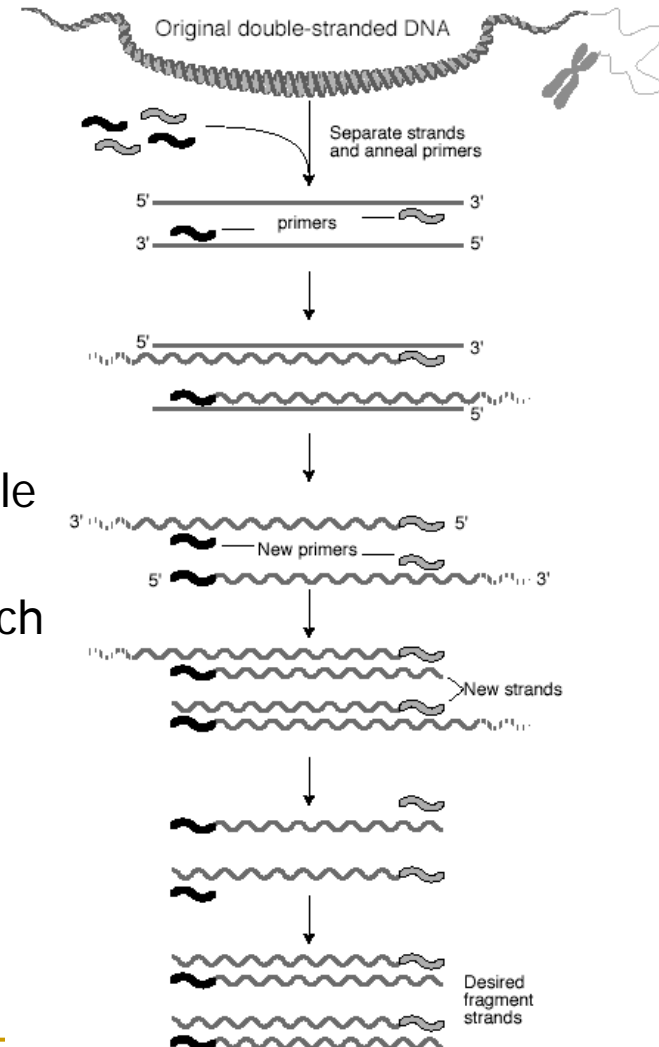
9.1 Copying DNA

Why we need so many copies

- Biologists needed to find a way to read DNA codes.
 - How do you read base pairs that are angstroms (equal to $0.1 \text{ nm} = 1 \times 10^{-10} \text{ m}$) in size?
 - It is not possible to directly look at it due to DNA's small size.
 - Need to use chemical techniques to detect what you are looking for.
 - To read something so small, you need a lot of it, so that you can actually detect the chemistry.
 - Need a way to make many copies of the base pairs, and a method for reading the pairs.
-

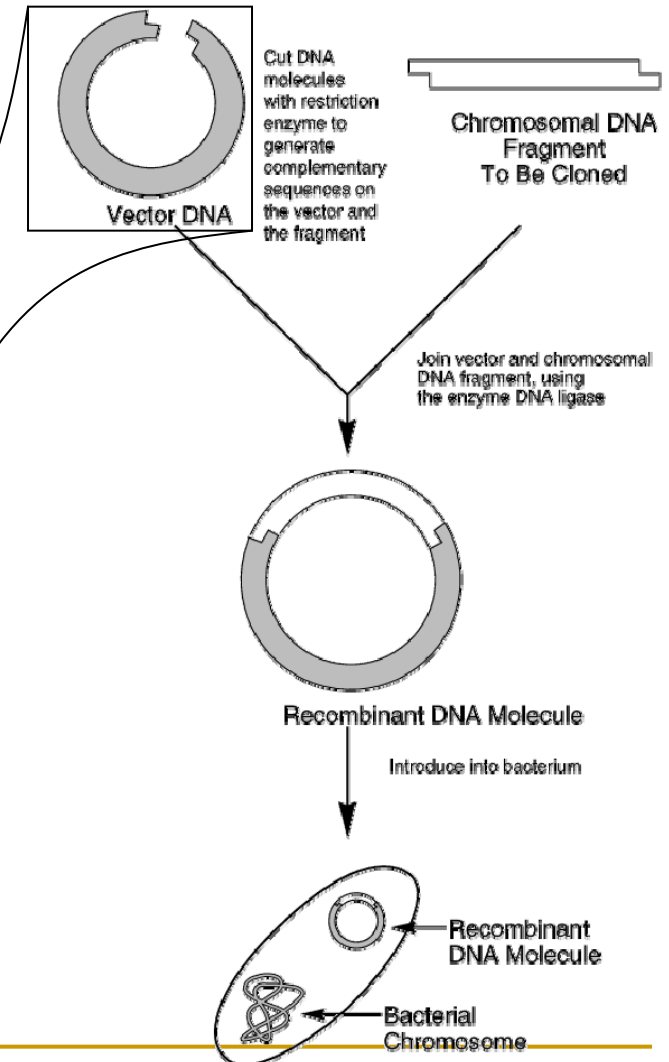
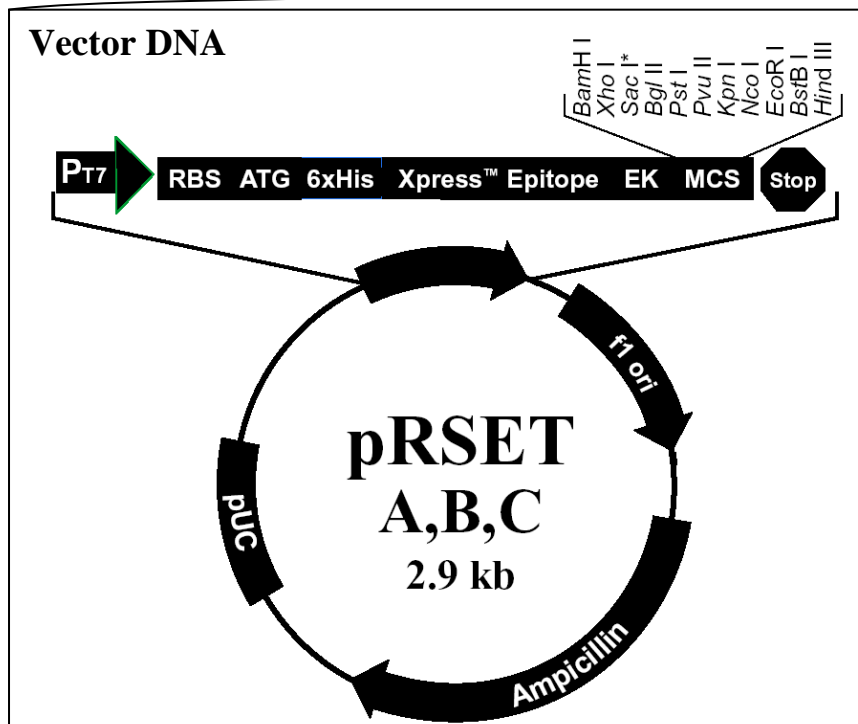
Polymerase Chain Reaction (PCR)

- Polymerase Chain Reaction (PCR)
 - Used to massively replicate DNA sequences.
- How it works:
 - Separate the two strands with low heat
 - Add some base pairs, primer sequences, and DNA Polymerase
 - Creates double stranded DNA from a single strand.
 - Primer sequences create a seed from which double stranded DNA grows.
 - Now you have two copies.
 - Repeat. Amount of DNA grows exponentially.
 - $1 \rightarrow 2 \rightarrow 4 \rightarrow 8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \dots$



Cloning DNA

- DNA Cloning
 - Insert the fragment into the genome of a living organism and watch it multiply.
 - Once you have enough, remove the organism, keep the DNA.
- Use Polymerase Chain Reaction (PCR)



9.2 Cutting and Pasting DNA

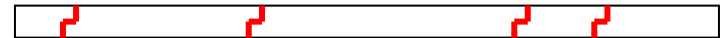
Restriction Enzymes

- Discovered in the early 1970's
 - Used as a defense mechanism by bacteria to break down the DNA of attacking viruses.
 - They cut the DNA into small fragments.
 - Can also be used to cut the DNA of organisms.
 - This allows the DNA sequence to be in a more manageable bite-size pieces.
 - It is then possible using standard purification techniques to single out certain fragments and duplicate them to macroscopic quantities.
-

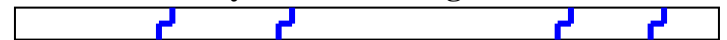
Cutting DNA

- Restriction Enzymes cut DNA
 - Only cut at special sequences
- DNA contains thousands of these sites.
- Applying different Restriction Enzymes creates fragments of varying size.

Restriction Enzyme “A” Cutting Sites



Restriction Enzyme “B” Cutting Sites



“A” and “B” fragments overlap

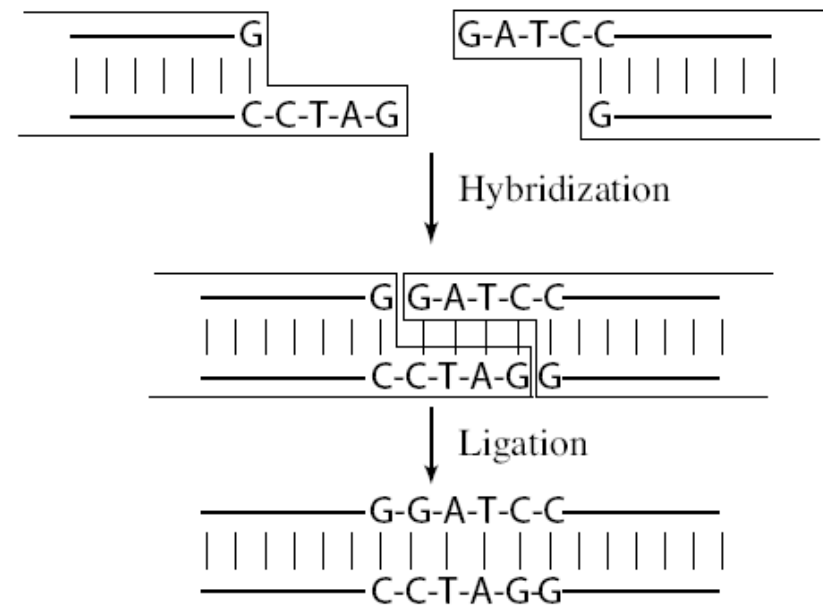
Restriction Enzyme “A” & Restriction Enzyme “B” Cutting Sites



Restriction enzyme	<i>Eco I</i>	<i>Bal I</i>	<i>Sma I</i>
Cuts at	GAATTC CTTAAG	TGGCCA ACCGGT	CCCGGG GGGCCC

Pasting DNA

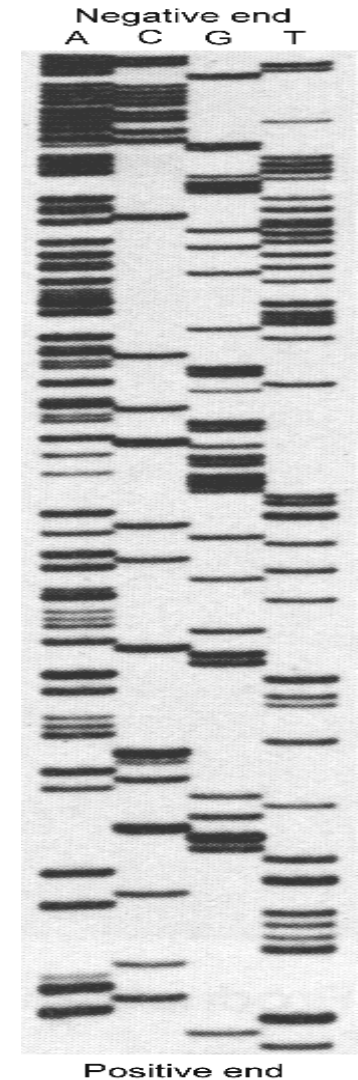
- Two pieces of DNA can be fused together by adding chemical bonds
 - **Hybridization** – complementary base-pairing
 - **Ligation** – fixing bonds with single strands



9.3 Measuring DNA Length

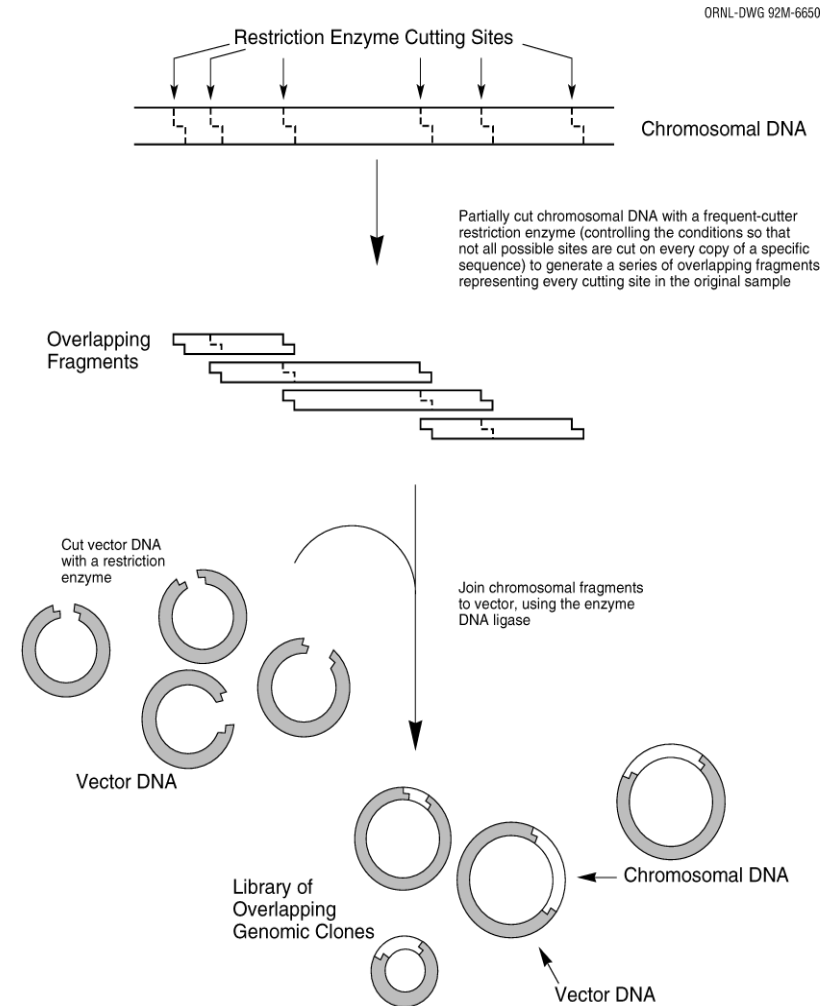
Reading DNA

- Electrophoresis
 - Reading is done mostly by using this technique. This is based on separation of molecules by their size (and in 2D gel by size and charge).
 - DNA or RNA molecules are charged in aqueous solution and move to a definite direction by the action of an electric field.
 - The DNA molecules are either labeled with radioisotopes or tagged with fluorescent dyes. In the latter, a laser beam can trace the dyes and send information to a computer.
 - Given a DNA molecule it is then possible to obtain all fragments from it that end in either A, or T, or G, or C and these can be sorted in a gel experiment.
- Another route to sequencing is direct sequencing using gene chips.



Assembling Genomes

- Must take the fragments and put them back together
 - **Not as easy as it sounds.**
- SCS Problem (Shortest Common Superstring)
 - Some of the fragments will overlap
 - Fit overlapping sequences together to get the shortest possible sequence that includes all fragment sequences



Assembling Genomes

- DNA fragments contain sequencing errors
 - Two complements of DNA
 - Need to take into account both directions of DNA
 - Repeat problem
 - 50% of human DNA is just repeats
 - If you have repeating DNA, how do you know where it goes?
-

9.4 Probing DNA

Che Fung Yung

May 12, 2004

DNA probes

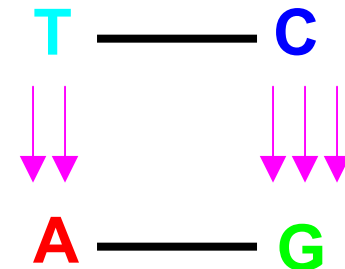
- Oligonucleotides: single-stranded DNA 20-30 nucleotides long
- Oligonucleotides used to find complementary DNA segments.
- Made by working backwards---AA sequence----mRNA---cDNA.
- Made with automated DNA synthesizers and tagged with a radioactive isotope.

DNA Hybridization

- Single-stranded DNA will naturally bind to complementary strands.
- Hybridization is used to locate genes, regulate gene expression, and determine the degree of similarity between DNA from different sources.
- Hybridization is also referred to as annealing or renaturation.

Create a Hybridization Reaction

1. Hybridization is binding two genetic sequences. The binding occurs because of the hydrogen bonds [pink] between base pairs.
2. When using hybridization, DNA must first be denatured, usually by using heat or chemical.

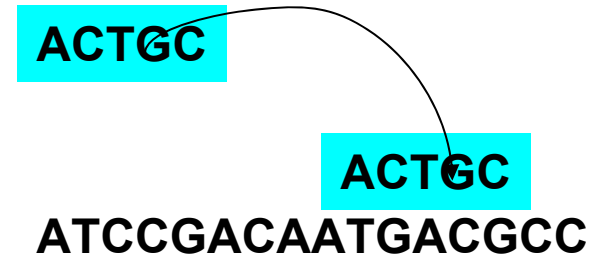


TAGGC T^GTTACT^GC
ATCCGACAATGACGCC

Create a Hybridization Reaction

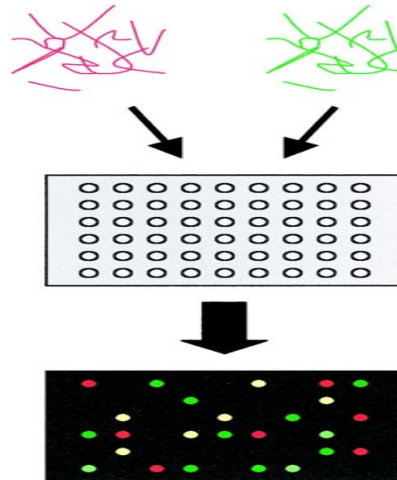
Cont.

- Once DNA has been denatured, a single-stranded radioactive probe [light blue] can be used to see if the denatured DNA contains a sequence complementary to probe.
- Sequences of varying homology stick to the DNA even if the fit is poor.

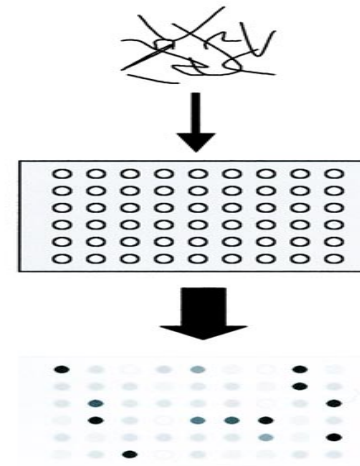


Labeling technique for DNA arrays

Two-color fluorescence



Radiolabeling



RNA samples are labeled using fluorescent nucleotides (*left*) or radioactive nucleotides (*right*), and hybridized to arrays. For fluorescent labeling, two or more samples labeled with differently colored fluorescent markers are hybridized to an array. Level of RNA for each gene in the sample is measured as intensity of fluorescence or radioactivity binding to the specific spot. With fluorescence labeling, relative levels of expressed genes in two samples can be directly compared with a single array.

DNA Arrays – Technical Foundations

- An array works by exploiting the ability of a given mRNA molecule to hybridize to the DNA template.
- Using an array containing many DNA samples in an experiment, the expression levels of hundreds or thousands genes within a cell by measuring the amount of mRNA bound to each site on the array.
- With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.

An experiment on a microarray

In this schematic:

GREEN represents **Control DNA**

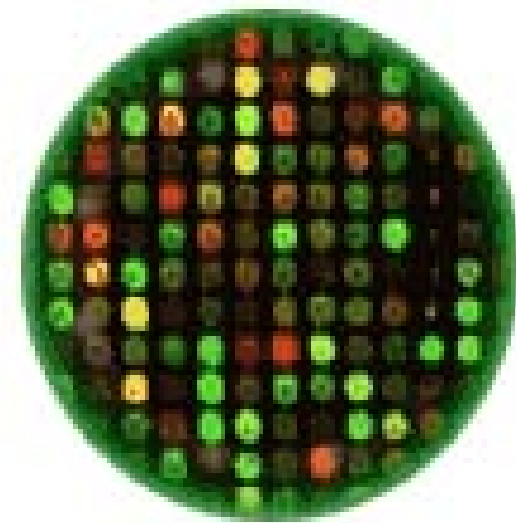
RED represents **Sample DNA**

YELLOW represents **a combination of Control and Sample DNA**

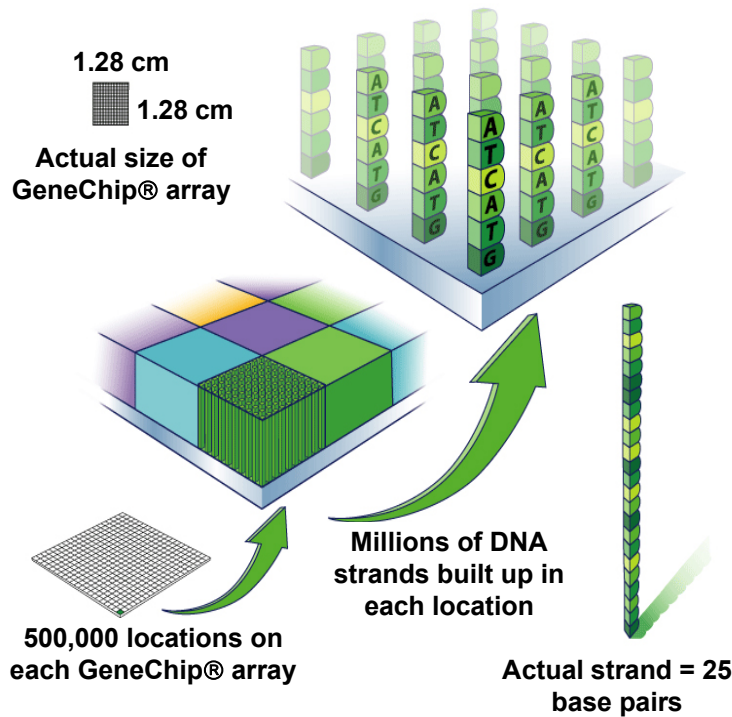
BLACK represents areas without **neither the Control nor Sample DNA**

Each color in an array represents either healthy (control) or diseased (sample) tissue.

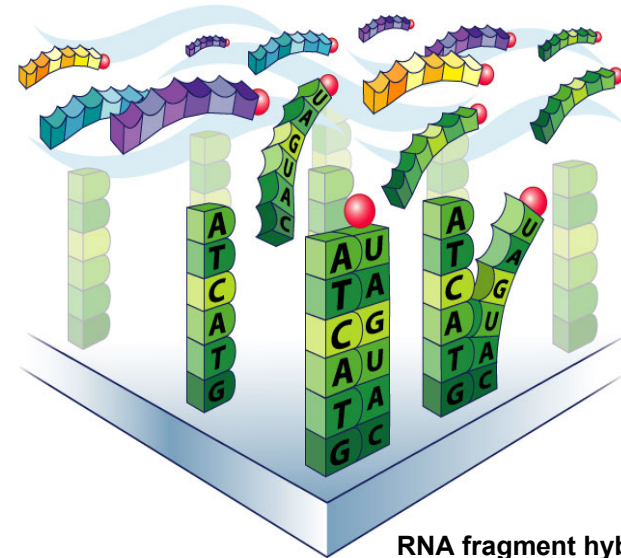
The location and intensity of a color tell us whether the gene, or mutation, is present in the control and/or sample DNA.



DNA Microarray



RNA fragments with fluorescent tags from sample to be tested



Millions of DNA strands build up on each location.

Tagged probes become hybridized to the DNA chip's microarray.

DNA Microarray



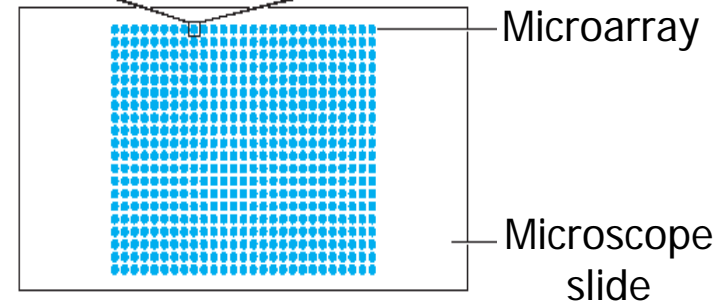
Affymetrix

Microarray is a tool for analyzing gene expression that consists of a glass slide.

Sequence of a gene

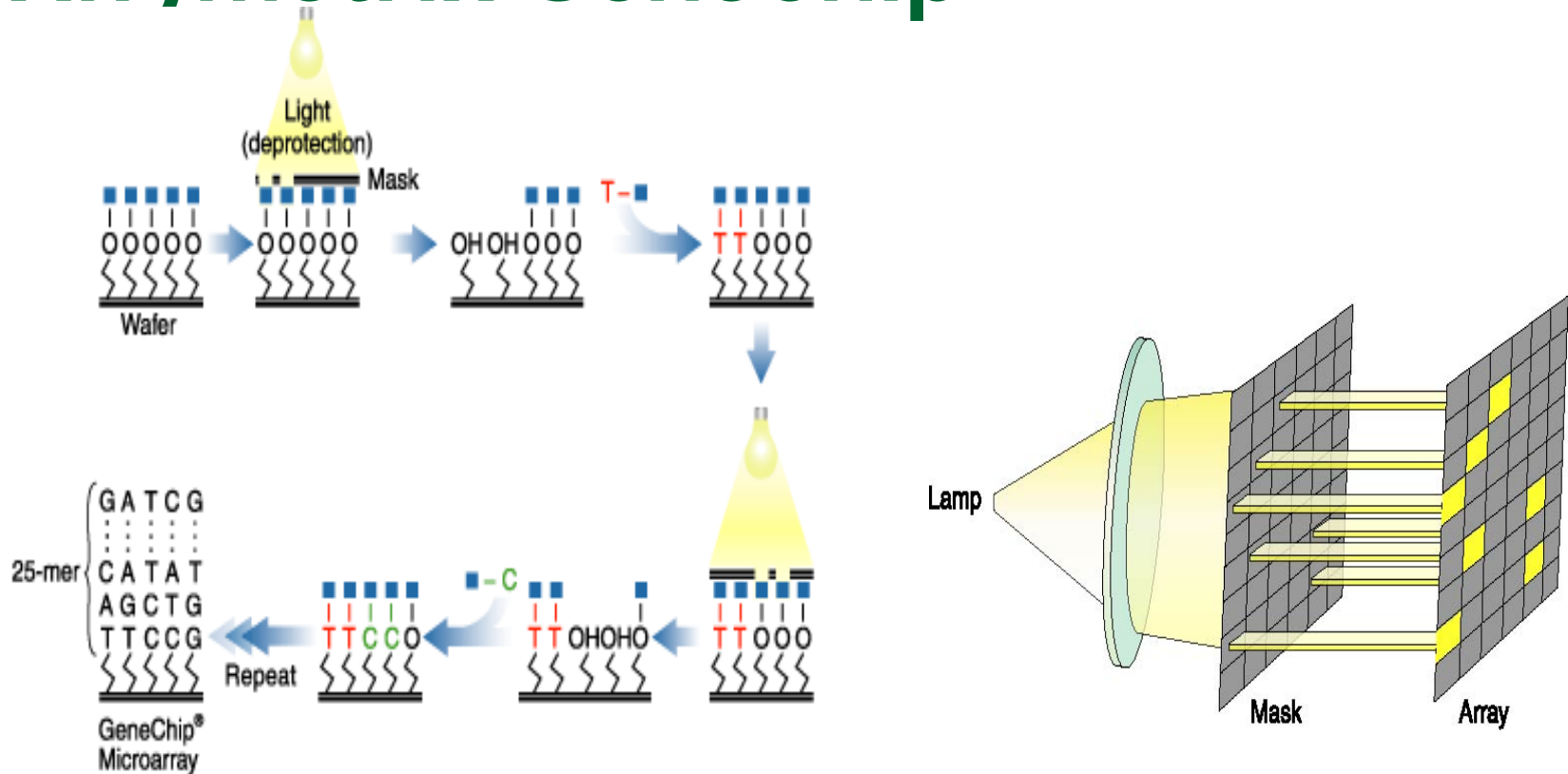
```

TCCTTTCCGG AACGGTTGGC GTCTGCGCAC GCGGGTGTGG GGCATGACAT
GCCGCCCCAG GAACAACCCC GACACGGCTT TAAGCCTCTC AAATCGCTGT
AGACATCATC TTTACGTGCT TGCCACCATT TGCCACCATT AGGGCTGTTC
CCGCGACGAC TCGCCATTCA ACCTCAGTCC TTCGGGTTGA GCGAGTGGGT
CGCGCGCAAG GTGCGAATGG GTCGCGCGCA AAGTGTTCGG CTGGCTGTAT
TATATGCTGC CTATAGCGAG ACTAACGACC CACACTTTCA CACAAGGATT
TCCCGCTAAT GGGTACCTCG CGTCAGGACC TTGACGCAAG CGCGCCTTCG
GTTGGCCCCA AGCTTGCTAG GACTACTTAT CTTGAGCTCA TTTAACATCC
CGGCGCCTCT CCGGGAGCGG TCGTCGCGAA GAAGTCAAAC CCGGAACGGC
GTTGACAAAG CGTGGAGACA TCGATACCTC TGTGTCAGCG GCCACAAATC
  
```



Each blue spot indicates the location of a PCR product. On a real microarray, each spot is about 100 μm in diameter.

Affymetrix GeneChip®



A combination of photolithography and combinatorial chemistry to manufacture GeneChip® Arrays. With a minimum number of steps, Affymetrix produces arrays with thousands of different probes packed at extremely high density. Enabled to obtain high quality, genome-wide data using small sample volumes.

END of SECTION 9

Section 10: How Do Individuals of a Species Differ?

- *Physical Variation and Diversity*
- *Genetic Variation*

How Do Individuals of Species Differ?

- Genetic makeup of an individual is manifested in traits, differences are caused by variations in genes
 - Not only do different species have different genomes, but also different individuals of the same species have different genomes.
 - No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
 - Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!
-

Physical Traits and Variances

- Individual variation among a species occurs in populations of all sexually reproducing organisms.
- Individual variations range from hair and eye color to less subtle traits such as susceptibility to malaria.
- Physical variation is the reason we can pick out our friends in a crowd, however most physical traits and variation can only be seen at a cellular and molecular level.



Sources of Physical Variation

- Physical Variation and the manifestation of traits are caused by
 - variations in the genes and
 - differences in environmental influences.
 - An example is height, which is dependent on genes as well as the nutrition of the individual.
 - Not all variation is inheritable – only genetic variation can be passed to offspring.
 - Biologists usually focus on genetic variation instead of physical variation because it is a better representation of the species.
-

Genetic Variation

- Despite the wide range of physical variation, genetic variation between individuals is quite small.
- Out of 3 billion nucleotides, only roughly 3 million base pairs (0.1%) are different between individual genomes of humans.
- Although there is a finite number of possible variations, the number is so high

$$\binom{3\,000\,000\,000}{3\,000\,000} 3^{3\,000\,000}$$

that we can assume no two individual people have the same genome.

- What is the cause of this genetic variation?

Sources of Genetic Variation

- **Mutations** are rare errors in the DNA replication process that occur at random.
 - When mutations occur, they affect the genetic sequence and create genetic variation between individuals.
 - Most mutations do not create beneficial changes and actually kill the individual.
 - Although mutations are the source of all new genes in a population, they are so rare that there must be another process at work to account for the large amount of diversity.
-

Sources of Genetic Variation

- **Recombination** is the shuffling of genes that occurs through sexual mating and is the main source of genetic variation.
 - Recombination occurs via a process called **crossing over** in which genes switch positions with other genes during meiosis.
 - Recombination means that new generations inherit random combinations of genes from both parents.
 - The recombination of genes creates a seemingly endless supply of genetic variation within a species.
-

How Genetic Variation is Preserved

- **Diploid** organisms (which are most complex organisms) have two genes that code for one physical trait – which means that sometimes genes can be passed down to the next generation even if a parent does not physically express the gene.
- **Balanced Polymorphism** is the ability of natural selection to preserve genetic variation. For example, natural selection in one species of finch keeps beak sizes either large or small because a finch with a hybrid medium sized beak cannot survive.

Variation as a Source of Evolution

- Evolution is based on the idea that variation between individuals causes certain traits to be reproduced in future generations more than others through the process of Natural Selection.
 - **Genetic Drift** is the idea that the prevalence of certain genes changes over time.
 - If enough genes are changed through mutations or otherwise so that the new population cannot successfully mate with the original population, then a new species has been created.
 - Do all variations affect the evolution of a species?
-

The Genome of a Species

- It is important to distinguish between the genome of a species and the genome of an individual.
 - The genome of a species is a representation of all possible genomes that an individual might have since the basic sequence in all individuals is more or less the same.
 - The genome of an individual is simply a specific instance of the genome of a species.
 - Both types of genomes are important – we need the genome of a species to study a species as a whole, but we also need individual genomes to study genetic variation.
-

Section 10: How Do Different Species Differ?

Outline For Section 10:

- *Section 10.1 – Molecular Evolution*
 - *What is Evolution*
 - *Molecular Clock*
 - *New Genes*
 - *Section 10.2 – Comparative Genomics*
 - *Human and Mouse*
 - *Comparative Genomics*
 - *Gene Mapping*
 - *Cystic Fibrosis*
 - *Section 10.3 – Genome Rearrangements*
 - *Gene Order*
 - *DNA Reversal*
-

Section 10.1 The Biological Aspects of Molecular Evolution

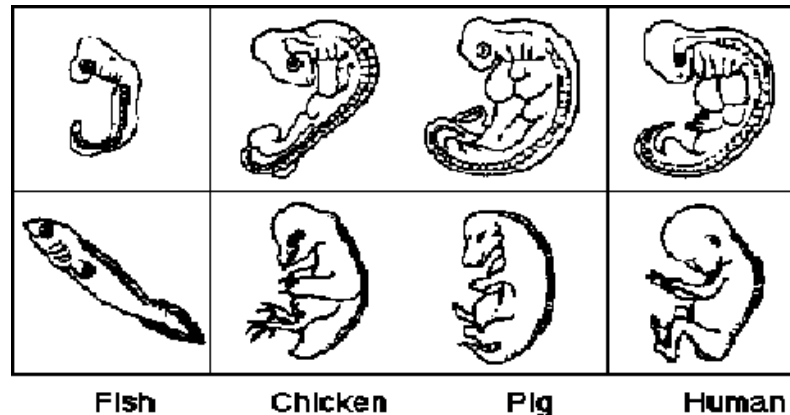
What is evolution?



- A process of change in a certain direction (*Merriam – Webster Online*).
- **In Biology**: The process of biological and organic change in organisms by which descendants come to differ from their ancestor (*Mc GRAW –HILL Dictionary of Biological Science*).
- **Charles Darwin** first developed the Evolution idea in detail in his well-known book *On the Origin of Species* published in 1859.

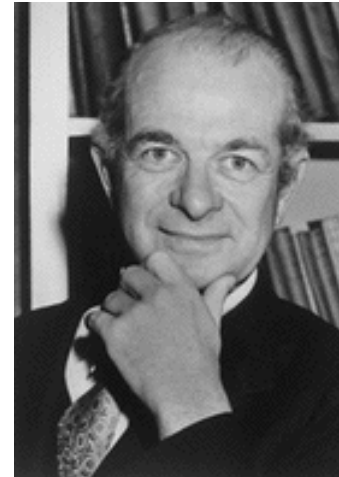
Some Conventional Tools For Evolutionary Studies

- **Fossil Record:** some of the biota found in a given stratum are the descendants of those in the previous stratum.
- **Morphological Similarity:** similar species are found to have some similar anatomical structure; For example: horses, donkeys and zebras.
- **Embryology:** embryos of related kinds of animals are astoundingly similar.



Molecular Clock

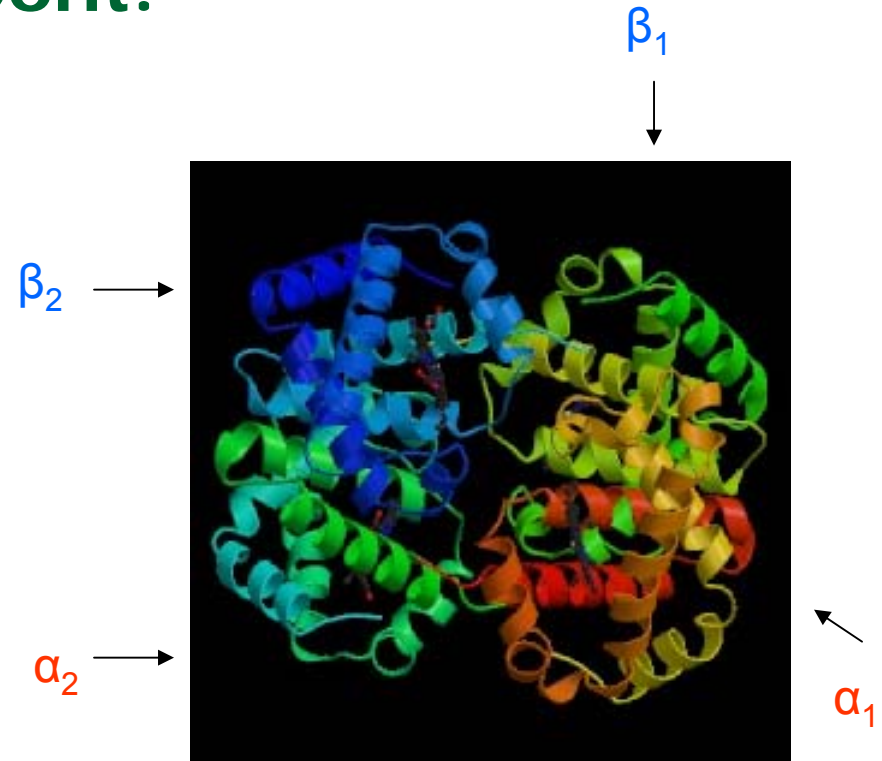
- Introduced by **Linus Pauling** and his collaborator **Emile Zuckerkandl** in 1965.
- They proposed that *the rate of evolution in a given protein (or later, DNA) molecule is approximately constant overtime and among evolutionary lineages.*



Linus Pauling

Molecular Clock Cont.

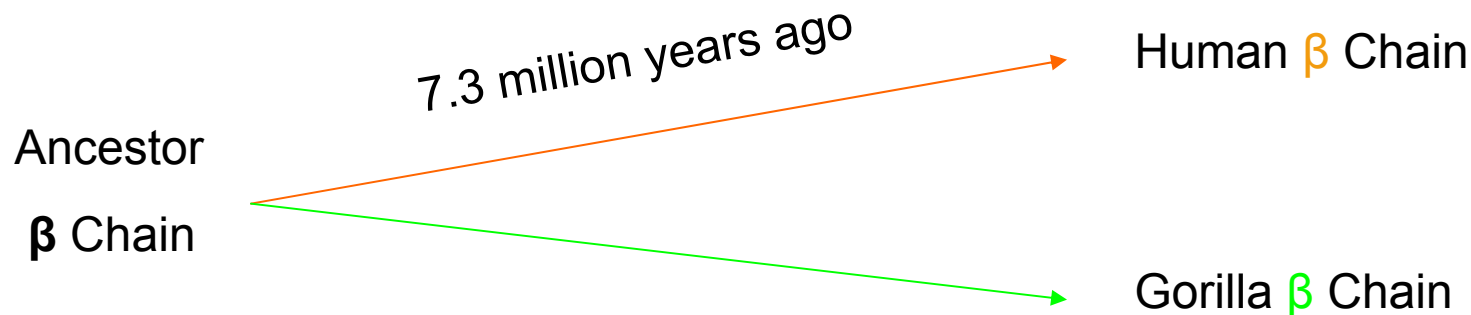
- Observing hemoglobin patterns of some primates, They found:
 - The gorilla, chimpanzee and human patterns are almost identical.
 - The further one gets away from the group of Primates, the primary structure that is shared with human hemoglobin *decreases*.
 - α and β chains of human hemoglobin are homologous, having a common ancestor.



Human Hemoglobin, A
2- α and 2- β tetramer.

Molecular Clock Cont.

- Linus and Pauling found that α -chains of human and gorilla differ by 2 residues, and β -chains by 1 residue.
- They then calculated the time of divergence between human and gorilla using evolutionary molecular clock.
- Gorilla and human β chain were found to diverge about 7.3 mil. years ago.



Molecular Evolution

- Pauling and Zuckerkandl research was one of the pioneering works in the emerging field of *Molecular Evolution* - the study of evolution at molecular level, genes, proteins or the whole genomes.
- Researchers have discovered that as somatic structures evolves (*Morphological Evolution*), so does the genes. But the *Molecular Evolution* has its special characteristics.
- Genes (and their proteins) products evolve at **different** rates.
 - For example, histones changes very slowly while fibrinopeptides very rapidly, revealing function conservation.
- Unlike physical traits which can evolve drastically, genes functions set **severe limits** on the amount of changes.

Thought Humans and Chimpanzees lineages separated at least 6 million years ago, many genes of the two species highly resemble one another.

Beta globins:

- Beta globin chains of closely related species are highly similar:
- Observe simple alignments below:

Human β chain: MVHLTPEEKSAVTALWGKV NVDEVGGEALGRLL

Mouse β chain: MVHLTDAEKAAVNGLWGKVNPDVVGGEALGRLL

Human β chain: VVYPWTQRFESFGDLSTPDVVMGNPKVKAHGKKV LGG

Mouse β chain: VVYPWTQRYFD SFGDLS SASAIMGNPKVKAHGKK VIN

Human β chain: AFSDGLAHL DNLKGTFA TLSELHCDKLHVDPENFRLLGN

Mouse β chain: AFNDGLKHL DNLKGTFA HLSELHCDKLHVDPENFRLLGN

Human β chain: VLVCVLAH HFGKEFTP PVQAAYQKVVAGVANALAHKYH

Mouse β chain: MI VI VLGHHLGKEFTP CAQA AFQKVVAGVA SALAHKYH

There are a total of **27** mismatches, or $(147 - 27) / 147 = 81.7\%$ identical

Beta globins: Cont.

Human β chain: MVH L TPEEKSAVTALWGKVNVDVGGGEALGRLL

Chicken β chain: MVHWTAEKQL I TGLWGKVNVAECGAEARLL

Human β chain: VVYPWTQRFFEESFGDLSTPDVVMGNPKVKAHGKKVLG

Chicken β chain: IVYPWTQRFF ASFGNLSPTA I LGNPMVRAHGKKVLT

Human β chain: AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLNGN

Chicken β chain: SFGDAVKNLDNIK NTFSQLSELHCDKLHVDPENFRLGDD

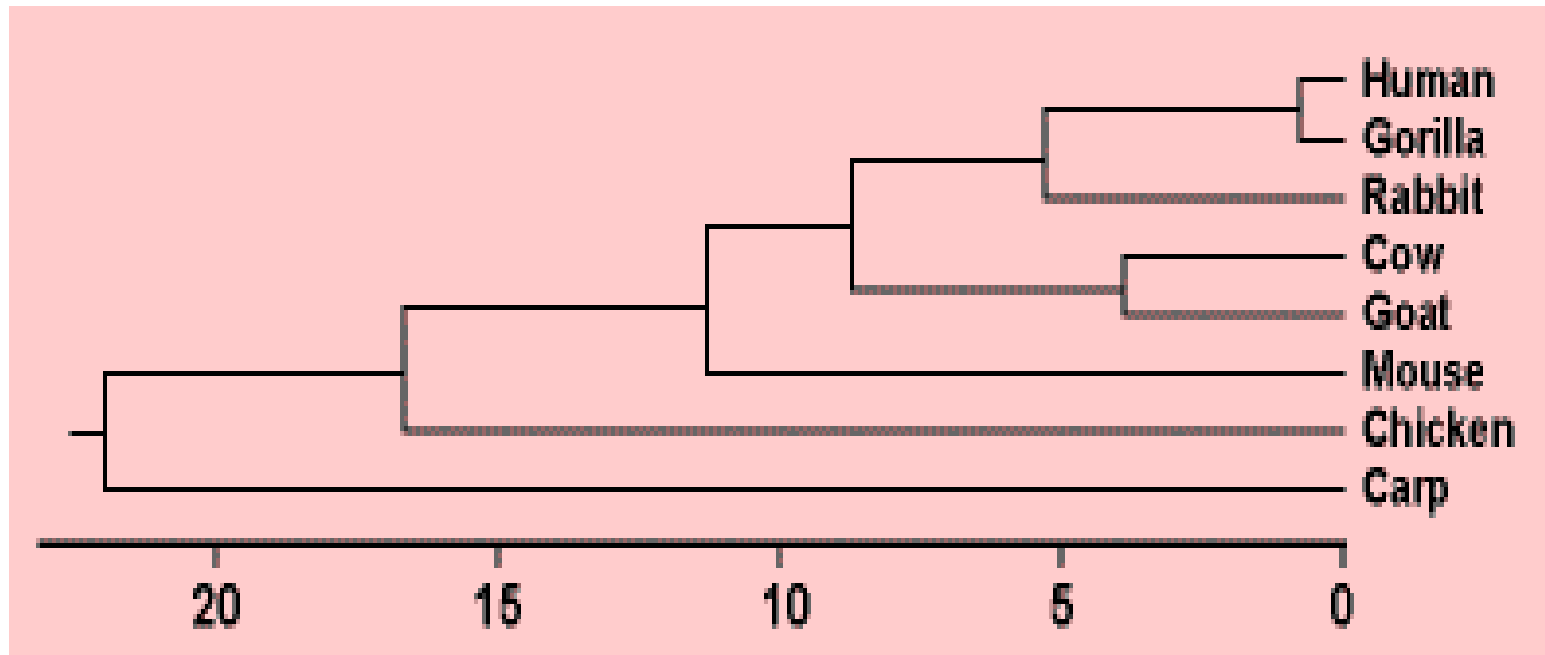
Human β chain: VLVCVLAHHFGKEFTPPVQAAY QKVAGVANALAHKYH

Mouse β chain: I L I I VLA AHFSKDFTPECAAWQKLVRVVAHALARKYH

- There are a total of **44** mismatches, or $(147 - 44) / 147 = 70.1$ % identical

- As expected, mouse β chain is '*closer*' to that of human than chicken's.

Molecular evolution can be visualized with phylogenetic tree.



Phylogenetic tree of Beta globin (Aligned using Clustal, PAM250)

Origins of New Genes

- All animals lineages traced back to a common ancestor, a protist about 700 million years ago.



Section 10.2: Comparative Genomics

How Do Different Species Differ?

- As many as 99% of human genes are conserved across all mammals
 - The functionality of many genes is virtually the same among many organisms
 - It is highly unlikely that the same gene with the same function would spontaneously develop among all currently living species
 - The theory of evolution suggests all living things evolved from incremental change over millions of years
-

Mouse and Human overview

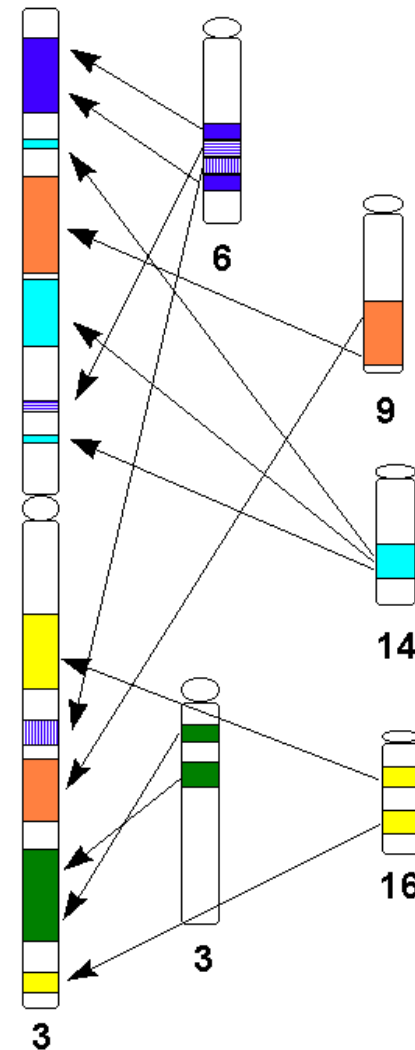
- Mouse has 2.1×10^9 base pairs versus 2.9×10^9 in human.
- About 95% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell and sex*

Human and Mouse

Significant chromosomal rearranging occurred between the diverging point of humans and mice.

Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.



Comparative Genomics

- What can be done with the full Human and Mouse Genome?
One possibility is to create “knockout” mice – mice lacking one or more genes. Studying the phenotypes of these mice gives predictions about the function of that gene in both mice and humans.

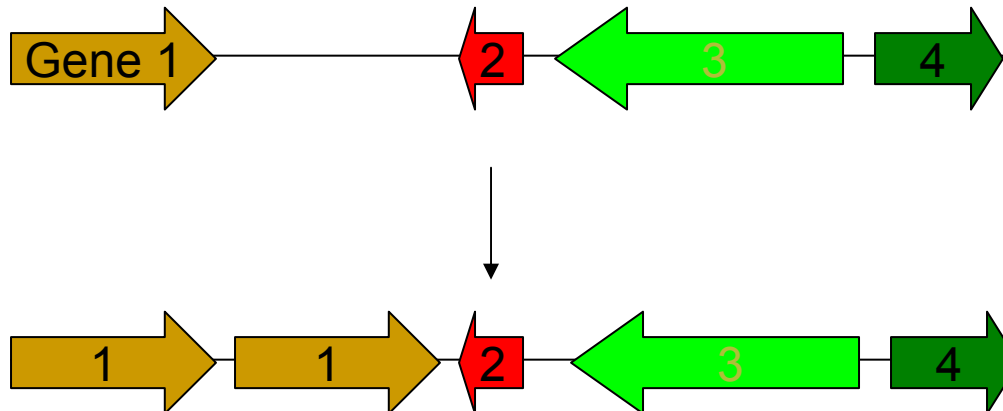


Comparative Genome Sizes

- The genome of a protist *Plasmodium falciparum*, which causes malaria, is 23 Mb long.
 - Human genome is approximately 150 times larger, mouse > 100 times, and fruit fly > 5 times larger.
 - Question: How genomes of old ancestors get bigger during evolution?
-

Mechanisms:

- Gene duplications or insertions



Comparative Genomics

- Knowing the full sequence of human and mouse genomes also gives information about gene regulation. Because the promoter regions tend to remain conserved through evolution, looking for similar DNA upstream of a known gene can help identify regulatory sites. This technique gets more powerful the more genomes can be compared.
-

Gene Mapping

- Mapping human genes is critically important
 - Insight into the evolutionary relationship of human to other vertebrate species
 - Mapping disease gene create an opportunity for researchers to isolate the gene and understand how it causes a disease.

Genomics: the sub discipline of genetics devoted to the mapping, sequencing, and functional analysis of genomes

Gene Mapping

- The procedure for mapping chromosomes was invented by Alfred H. Sturterant.
 - Analysis of experiment data from Drosophilia
 - Experimental data demonstrated that genes on the same chromosome could be separated as they went through meiosis and new **combination** of genes is formed.
 - Genes that are tightly linked seldom recombine, whereas genes that are loosely linked recombine
-

Gene Mapping

- Genetic maps of chromosomes are based on **recombination frequencies** between markers.
 - Cytogenetic maps are based on the location of markers within cytological features such as **chromosome banding** patterns observed by microscope.
 - Physical maps of chromosomes are determined by the molecular distances in base pairs, kilobase pairs, or mega base pairs separating markers.
 - High-density maps that integrate the genetic, cytological and physical maps of chromosomes have been constructed for all of human chromosomes and for many other organisms
-

Gene Mapping

- Recombinant DNA techniques have revolutionized the search for defective genes that cause human disease
 - Numerous major “disease genes” have already been identified by positional cloning
 - Huntington’s disease (HD gene)
 - Cystic fibrosis (CF gene)
 - Cancer
-

Section 10.3 Genome Rearrangements.

Turnip and Cabbage

- Cabbages and turnips share a common ancestor



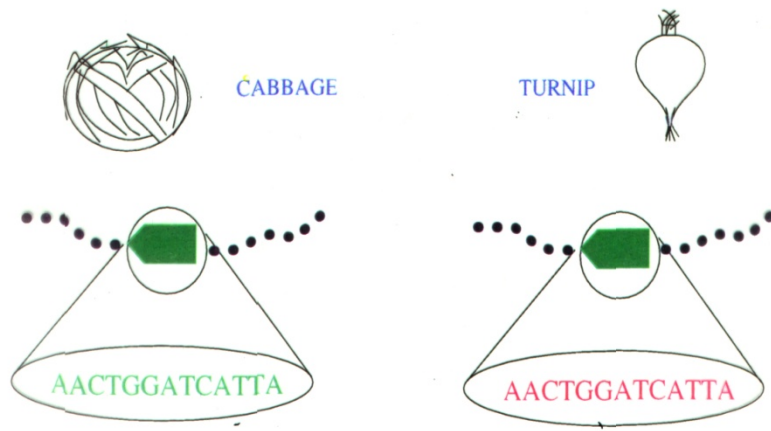
Turnip: po slovensky - kvaka; po česky - tuřín, vodnice

Jeffrey Palmer – 1980s

- discovered evolutionary change in plant organelles by comparing mitochondrial genomes of the cabbage and turnip
- 99% similarity between genes
- These more or less identical gene sequence surprisingly differed in gene order
- This finding helped pave the way to prove that genome rearrangements occur in molecular evolution in mitochondrial DNA

Important discovery

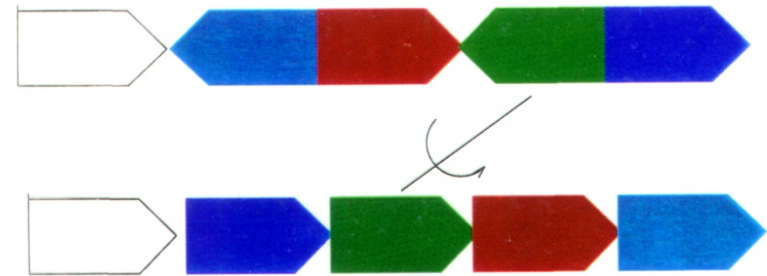
GENE SEQUENCE COMPARISON



AACTGGATCATT A
AACTGGATCATT A

Comparing gene sequences yields
no evolutionary information

GENE ORDER COMPARISON

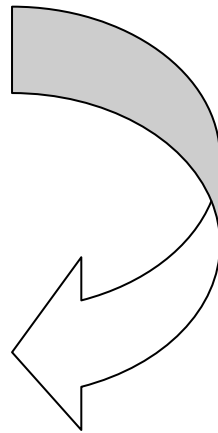


Evolution is manifested as the
divergence in Gene Order

DNA Reversal



Break
and
Invert



5' A T G C C T G T A C T A 3'

3' T A C G G A C A T G A T 5'



5' A T G T A C A G G C T A 3'

3' T A C A T G T C C G A T 5'

Bioinformatics

Sequence Driven Problems

- **Genomics**
 - Fragment assembly of the DNA sequence.
 - Not possible to read entire sequence.
 - Cut up into small fragments using restriction enzymes.
 - Then need to do fragment assembly. Overlapping similarities to matching fragments.
 - NP-complete problem.
 - Finding Genes
 - Identify open reading frames
 - Exons are spliced out.
 - Junk in between genes

Bioinformatics

Sequence Driven Problems

- **Proteomics**
 - Identification of functional domains in protein's sequence
 - Determining functional pieces in proteins.
 - Protein Folding
 - 1D Sequence → 3D Structure
 - What drives this process?
-

DNA... Then what?

- DNA → transcription → RNA → translation → Protein
- Ribonucleic Acid (RNA)
 - It is the messenger
 - a temporary copy
 - Why not DNA → Protein.
 - DNA is in nucleus and proteins are manufactured out of the nucleus
 - Adds a proofreading step. (Transcription = DNA→RNA)
- So actually... DNA → pre-mRNA → mRNA → Protein
 - Prokaryotes
 - The gene is continuous. Easy to translate.
 - Eukaryotes
 - Introns and Exons
 - Several Exons in different locations need to be spliced together to make a protein. (Splicing)
 - Pre-mRNA (unspliced RNA)
 - Splicisome cuts the introns out of it making processed mRNA.

Proteins

- Carry out the cell's chemistry
 - 20 amino acids
- A more complex polymer than DNA
 - Sequence of 100 has 20^{100} combinations
 - Sequence analysis is difficult because of complexity issue
 - Only a small number of the possible sequences are actually used in life. (Strong argument for Evolution)
- RNA Translated to Protein, then Folded
 - Sequence to 3D structure (Protein Folding Problem)
 - Translation occurs on Ribosomes
 - 3 letters of DNA → 1 amino acid
 - 64 possible combinations map to 20 amino acids
 - Degeneracy of the genetic code
 - Several codons to same protein

Section 11: Why Bioinformatics?

Julio Ng, Robert Hinman

CSE 181 Projects 2,3

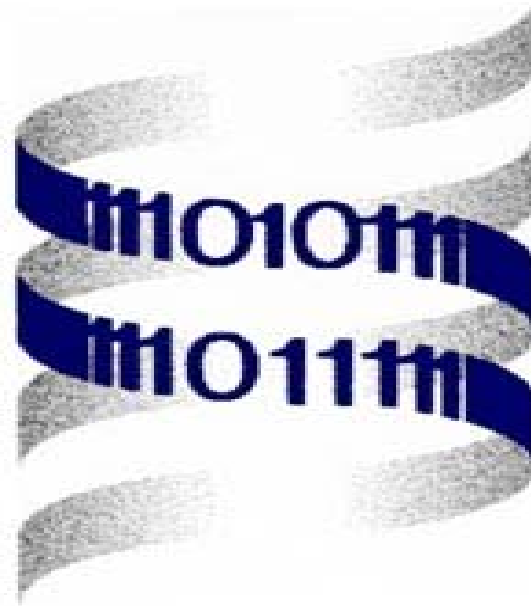
April 20, 2004

Why Bioinformatics?

- Bioinformatics is the combination of biology and computing.
 - DNA sequencing technologies have created massive amounts of information that can only be efficiently analyzed with computers.
 - So far >70 species sequenced (by 2004), >6000 species by 2010
 - Human, rat chimpanzee, chicken, and many others.
 - As the information becomes ever so larger and more complex, more computational tools are needed to sort through the data.
 - Bioinformatics to the rescue!!!
-

What is Bioinformatics?

- Bioinformatics is generally defined as the analysis, prediction, and modeling of biological data with the help of computers

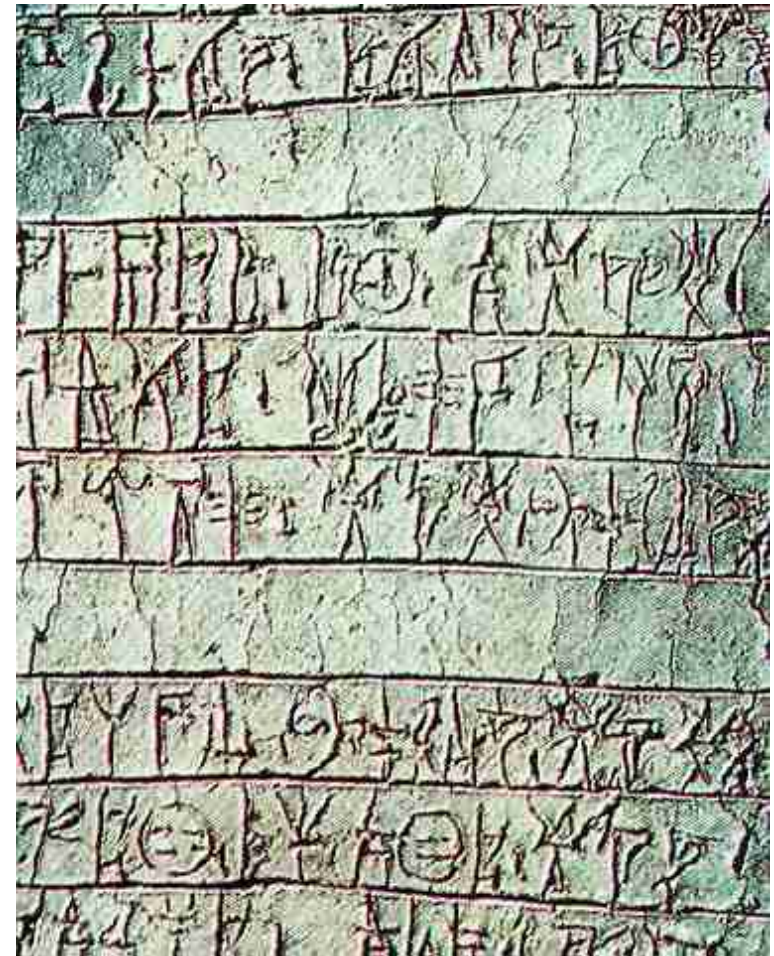


Bio-Information

- Since discovering how DNA acts as the instructional blueprints behind life, biology has become an information science
 - Now that many different organisms have been sequenced, we are able to find meaning in DNA through *comparative genomics*, not unlike comparative linguistics.
 - Slowly, we are learning the syntax of DNA
-

Linear B

- At the beginning of the twentieth century, archeologists discovered clay tablets on the island of Crete
- This unknown language was named "Linear B"
- It was thought to write in an ancient Minoan Language, and was a mystery for 50 years




Linear B

- The same time the structure of DNA is deciphered, Michael Ventris solves Linear B using mathematical code breaking skills
 - He notes that some words in Linear B are specific for the island, and theorizes those are names of cities
 - With this bit of knowledge, he is able to decode the script, which turns out to be Greek with a different alphabet
-

Amino Acid Crack

- Even earlier, an experiment in the early 1900s showed that all proteins are composed of sequences of 20 amino acids
 - This led some to speculate that polypeptides held the blueprints of life
-

Central Dogma

- DNA → mRNA → Proteins

- DNA in chromosome is transcribed to mRNA, which is exported out of the nucleus to the cytoplasm. There it is translated into protein
- Later discoveries show that we can also go from mRNA to DNA (retroviruses).
- Also mRNA can go through alternative splicing that lead to different protein products.

Structure to Function

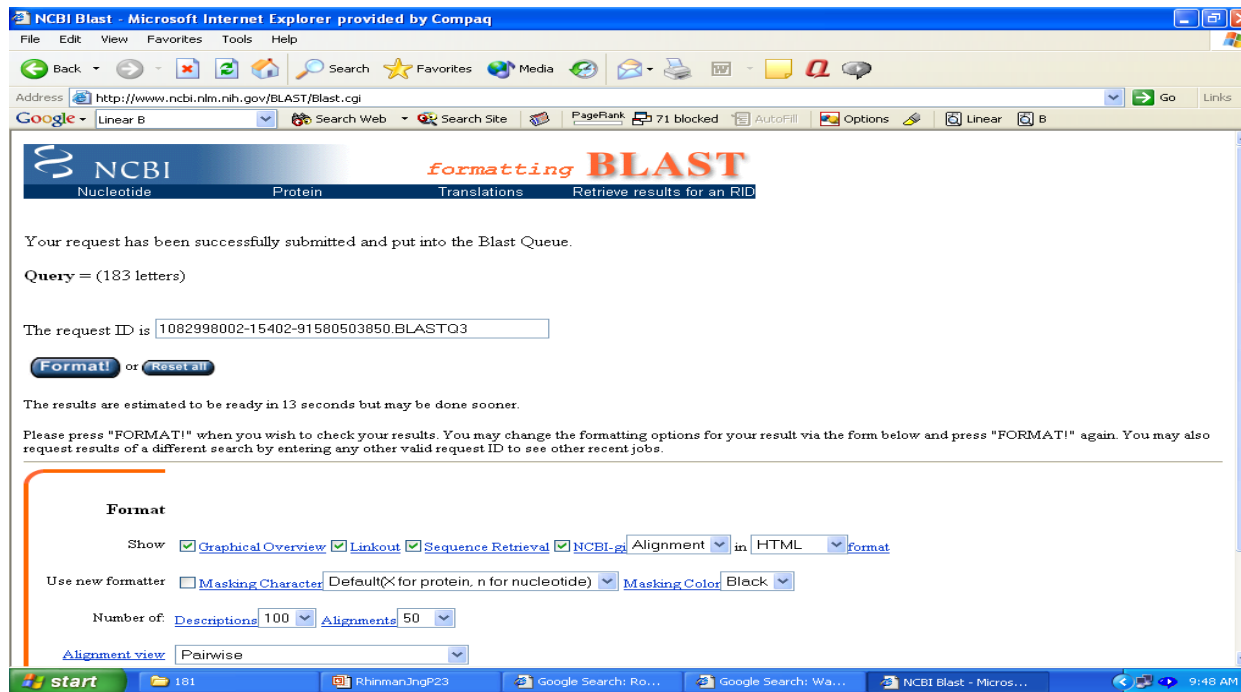
- Organic chemistry shows us that the structure of the molecules determines their possible reactions.
 - One approach to study proteins is to infer their function based on their structure, especially for active sites.
-

Two Quick Bioinformatics Applications

- BLAST (Basic Local Alignment Search Tool)
- PROSITE (Protein Sites and Patterns Database)

BLAST

- A computational tool that allows us to compare query sequences with entries in current biological databases.
- A great tool for predicting functions of a unknown sequence based on alignment similarities to known genes.



Some Early Roles of Bioinformatics

- Sequence comparison
- Searches in sequence databases



Early Sequence Matching

- Finding locations of restriction sites of known restriction enzymes within a DNA sequence (very trivial application)
 - Alignment of protein sequence with scoring motif
 - Generating contiguous sequences from short DNA fragments.
 - This technique was used together with PCR and automated high throughput sequencing (HT sequencing) to create the enormous amount of sequence data we have today
-

Biological Databases

- Vast biological and sequence data is freely available through online databases
- Use computational algorithms to efficiently store large amounts of biological data

Examples

- **NCBI GeneBank** <http://ncbi.nih.gov>
Huge collection of databases, the most prominent being the nucleotide sequence database
 - **Protein Data Bank** <http://www.pdb.org>
Database of protein tertiary structures
 - **SWISSPROT** <http://www.expasy.org/sprot/>
Database of annotated protein sequences
 - **PROSITE** <http://kr.expasy.org/prosite>
Database of protein active site motifs
-

PROSITE Database

- Database of protein active sites.
- A great tool for predicting the existence of active sites in an unknown protein based on primary sequence.

The screenshot shows a web browser window titled "List of PROSITE documentation entries - Microsoft Internet Explorer provided by Compaq". The address bar displays "http://www.expasy.org/cgi-bin/prosite-list.pl". The page content includes a navigation menu with "ExPASy Home page", "Site Map", "Search ExPASy", "Contact us", and "PROSITE". A search bar contains the text "PROSITE" and a "Go" button. Below the search bar is the PROSITE logo and the text "PROSITE Database of protein families and domains". A section titled "Browse PROSITE documentation entries" with the subtitle "Release 18.26, of 26-Apr-2004" contains a list of links for various protein categories: [\[Post-translational modifications\]](#), [\[Compositional biased regions\]](#), [\[Domains\]](#), [\[DNA or RNA associated proteins\]](#), [\[Enzymes\]](#), [\[Electron transport proteins\]](#), [\[Other transport proteins\]](#), [\[Structural proteins\]](#), [\[Receptors\]](#), [\[Cytokines and growth factors\]](#), [\[Hormones and active peptides\]](#), [\[Toxins\]](#), [\[Inhibitors\]](#), [\[Protein secretion and chaperones\]](#), and [\[Others\]](#). Below this is a list of documentation entries with explanatory text:

- The character in the first column is used to indicate if a documentation entry is new in this release '+', or has been modified '*' since the last major release (release 18.0 of July 2002).
- The numerical characters in positions 3 to 7 provide the documentation entry accession number.
- The numerical character in position 9 is used to indicate how many data entries (patterns, rules and profiles/matrices) are described by a documentation entry.

 An example entry is shown:

```
* PDOC00020 2 Kringle domain signature and profile
```

 A note states: "This documentation entry has been updated since the last release ('*'), its accession number is PDOC00020 and it describes two patterns." The Windows taskbar at the bottom shows the start button, a clock at 9:52 AM, and several open applications including "RhinmanJngP23", "Google Search...", "List of PROSITE...", and "untitled - Paint".

Sequence Analysis

- Some algorithms analyze biological sequences for patterns
 - RNA splice sites
 - ORFs (open reading frames)
 - Amino acid propensities in a protein
 - Conserved regions in
 - AA sequences [possible active site]
 - DNA/RNA [possible protein binding site]
- Others make predictions based on sequence
 - Protein/RNA secondary structure folding

It is Sequenced, What's Next?

- Tracing Phylogeny
 - Finding family relationships between species by tracking similarities between species.
 - Gene Annotation (cooperative genomics)
 - Comparison of similar species.
 - Determining Regulatory Networks
 - The variables that determine how the body reacts to certain stimuli.
 - Proteomics
 - From DNA sequence to a folded protein.
-

Modeling

- Modeling biological processes tells us if we understand a given process
 - Because of the large number of variables that exist in biological problems, powerful computers are needed to analyze certain biological questions
-

Protein Modeling

- Quantum chemistry imaging algorithms of active sites allow us to view possible bonding and reaction mechanisms
- Homologous protein modeling is a comparative proteomic approach to determining an unknown protein's tertiary structure
- Predictive tertiary folding algorithms are a long way off, but we can predict secondary structure with ~80% accuracy.

The most accurate online prediction tools:

PSIPred

PHD

Regulatory Network Modeling

- Micro array experiments allow us to compare differences in expression for two different states
 - Algorithms for clustering groups of gene expression help point out possible regulatory networks
 - Other algorithms perform statistical analysis to improve signal to noise contrast
-

Systems Biology Modeling

- Predictions of whole cell interactions.
 - Organelle processes, expression modeling
- Currently feasible for specific processes (eg. Metabolism in E. coli, simple cells)
 - Flux Balance Analysis



The future...

- Bioinformatics is still in it's infancy
 - Much is still to be learned about how proteins can manipulate a sequence of base pairs in such a peculiar way that results in a fully functional organism.
 - How can we then use this information to benefit humanity without abusing it?
-

Sources Cited

- Daniel Sam, "Greedy Algorithm" presentation.
 - Glenn Tesler, "Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes" presentation.
 - Ernst Mayr, "What evolution is".
 - Neil C. Jones, Pavel A. Pevzner, "An Introduction to Bioinformatics Algorithms".
 - Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell. New York: Garland Science. 2002.
 - Mount, Ellis, Barbara A. List. Milestones in Science & Technology. Phoenix: The Oryx Press. 1994.
 - Voet, Donald, Judith Voet, Charlotte Pratt. Fundamentals of Biochemistry. New Jersey: John Wiley & Sons, Inc. 2002.
 - Campbell, Neil. Biology, Third Edition. The Benjamin/Cummings Publishing Company, Inc., 1993.
 - Snustad, Peter and Simmons, Michael. Principles of Genetics. John Wiley & Sons, Inc, 2003.
-