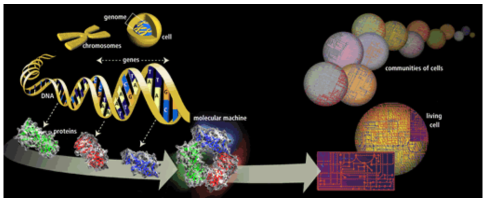
Molecular Biology Primer



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
Edited for Introduction to Bioinformatics (Autumn 2006) by Esa Pitkänen
http://www.cs.helsinki.fi/mbi/courses/06-07/itb/

Outline:

1. How molecular biology came about?
2. Similarities: What is life made of?
3. Differences: Variation in genomes

1. 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

- Robert Hooke

- Matthias Schleiden

- Theodor Schwann

Major events in the history of Molecular Biology 1800 - 1870

- **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)

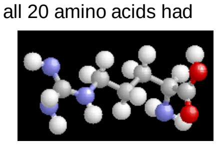
Mendel: The Father of Genetics

- **1869** Johann Friedrich Miescher discovered DNA and named it nuclein.

Johann Miescher

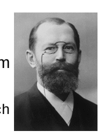
Major events in the history of Molecular Biology 1880 - 1900

- **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



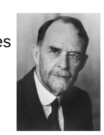
Major events in the history of Molecular Biology 1900-1911

- **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



Emil Fischer

- **1911 –** Thomas Hunt Morgan discovers genes on chromosomes are the discrete units of heredity

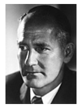


Thomas Morgan

- **1911** Pheobus Aaron Theodore Lerene discovers RNA

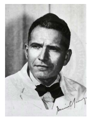
Major events in the history of Molecular Biology 1940 - 1950

- **1941** – George Beadle and Edward Tatum identify that genes make proteins


George Beadle

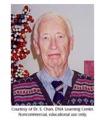

Edward Tatum

- **1950** – Edwin Chargaff find Cytosine complements Guanine and Adenine complements Thymine

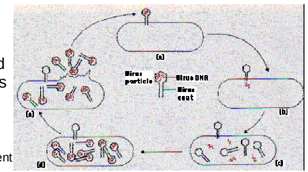

Edwin Chargaff

Major events in the history of Molecular Biology 1950 - 1952

- **1950s** – Mahlon Bush Hoagland first to isolate tRNA


Mahlon Hoagland

- **1952** – Alfred Hershey and Martha Chase make genes from DNA



Hershey Chase Experiment

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

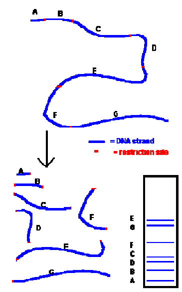

James Watson and Francis Crick

- **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.


Phillip Sharp


Richard Roberts

- Joan Steitz determined that the 5' end of snRNA is partially complementary to the consensus sequence of 5' splice junctions.

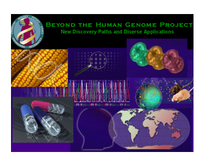

Joan Steitz

Major events in the history of Molecular Biology 1986 - 1995

- **1986** Leroy Hood: Developed automated sequencing mechanism
- **1986** Human Genome Initiative announced


Leroy Hood

- **1990** The 15 year Human Genome project is launched by congress
- **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)



Major events in the history of Molecular Biology 1995-1996

- **1995** John Craig Venter: First bactierial 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

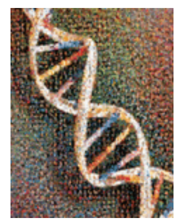
- **1997** E. Coli sequenced
- **1998** PerkinsElmer, Inc.. Developed 96-capillary sequencer
- **1998** Complete sequence of the Caenorhabditis elegans genome
- **1999** First human chromosome (number 22) sequenced

Major events in the history of Molecular Biology 2000-2001

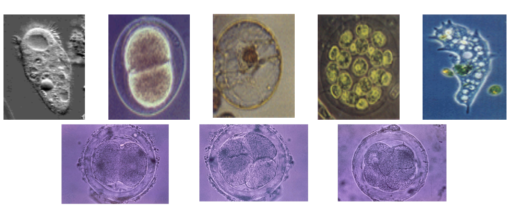
- **2000** Complete sequence of the euchromatic portion of the Drosophila melanogaster genome
- **2001** International Human Genome Sequencing:first draft of the sequence of the human genome published

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.



2. What is Life made of?



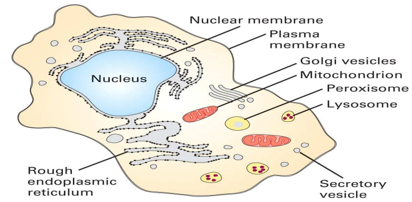
Cells

- **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 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

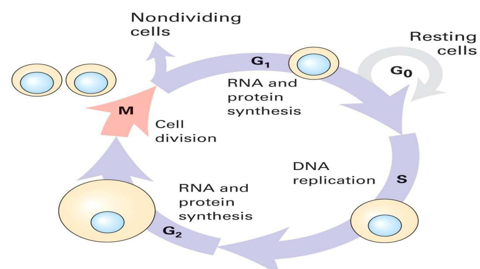


- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

Common features of organisms

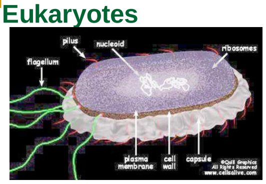
- Chemical energy is stored in ATP
- Genetic information is encoded by DNA
- Information is transcribed into RNA
- There is a common triplet genetic code
- Translation into proteins involves ribosomes
- Shared metabolic pathways
- Similar proteins among diverse groups of organisms

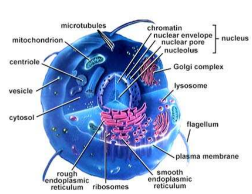
All Cells have common Cycles



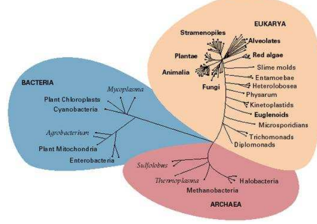
- Born, eat, replicate, and die

Two types of cells: Prokaryotes and 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

Signaling Pathways: Control Gene Activity

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

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.

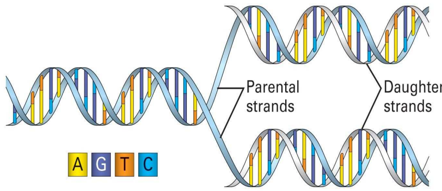
Terminology

- 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 and function
  - large, complex molecules made up of smaller subunits called **amino acids.**

All Life depends on 3 critical molecules

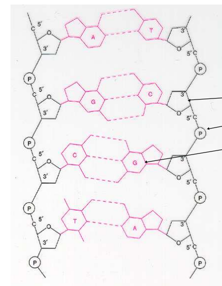
- DNAs (Deoxyribonucleic acid)
  - Hold information on how cell works
- RNAs (Ribonucleic acid)
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - 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

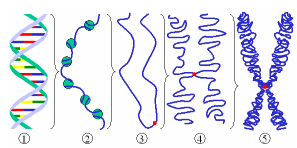


- DNA has a double helix structure which is composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)
- By convention, we read DNA strings in direction of transcription: from 5' end to 3' end

5' ATTTAGGCC 3'

3' TAAATCCGG 5'

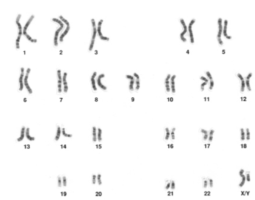
DNA is packed into chromosomes



- (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

Human chromosomes

- Somatic cells in humans have 2 pairs of 22 chromosomes + XX (female) or XY (male) = total of 46 chromosomes
- Germline cells have 22 chromosomes + either X or Y = total of 23 chromosomes



Karyogram of human male using Giemsa staining (http://en.wikipedia.org/wiki/Karyotype)

Length of DNA and number of chromosomes

Organism	#base pairs	#chromosomes (germline)
Prokayotic		
Escherichia coli (bacterium)	4×10^6	1
Eukaryotic		
Saccharomyces cerevisia (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

Human Genome Composition

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

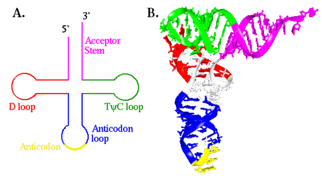
Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	≈15* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–≈1000	≈15* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–≈100	≈0.4
Unclassified spacer DNA	Variable	n.a.‡	≈25

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.

SOURCE: E. S. Lander et al., 2001, *Nature* 409:860.

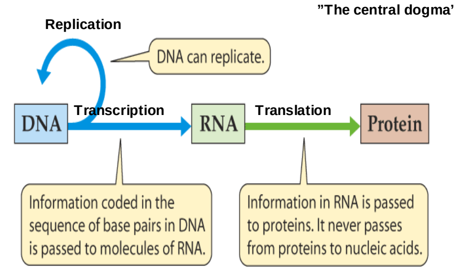
RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Several types of RNA exist for different functions in the cell.

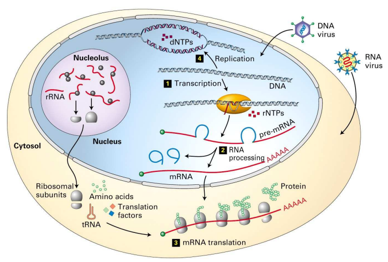


tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif

DNA, RNA, and the Flow of Information



Overview of DNA to RNA to Protein



- A gene is expressed in two steps
  1) Transcription: RNA synthesis
  2) Translation: Protein synthesis

Proteins

- Proteins are polypeptides (strings of amino acid residues)
- Represented using strings of letters from an alphabet of 20: AEGLV…WKKLAG
- Typical length 50…1000 residues

Urease enzyme from Helicobacter pylori

How DNA/RNA codes for protein?

- DNA alphabet contains four letters but must specify protein, or polypeptide sequence of 20 letters.
- Dinucleotides are not enough: 4^2 = 16 possible dinucleotides
- Trinucleotides (triplets) allow 4^3 = 64 possible trinucleotides
- Triplets are also called *codons*

Second letter

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

How DNA/RNA codes for protein?

- Three of the possible triplets specify "stop translation"
- Translation usually starts at triplet AUG (this also codes for methionine)
- Most amino acids may be specified by more than triplet
- How to find a gene? Look for start and stop codons (not that easy though)

Second letter

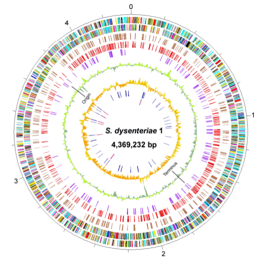
First letter	U	C	A	G	Third letter
U	UUU UUC Phenyl-alanine; UUA UUG Leucine	UCU UCC UCA UCG Serine	UAU UAC Tyrosine; UAA Stop codon; UAG Stop codon	UGU UGC Cysteine; UGA Stop codon; UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine	CCU CCC CCA CCG Proline	CAU CAC Histidine; CAA CAG Glutamine	CGU CGC CGA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

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.

3. Where does the variation in genomes come from?

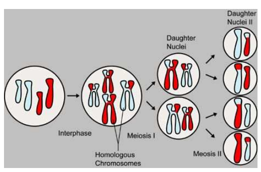
- Prokaryotes are typically haploid: they have a single (circular) chromosome
- DNA is usually inherited vertically (parent to daughter)
- Inheritance is clonal
  - Descendants are faithful copies of an ancestral DNA
  - Variation is introduced via mutations, transposable elements, and horizontal transfer of DNA



Chromosome map of *S. dysenteriae*, the nine rings describe different properties of the genome http://www.mgc.ac.cn/ShiBASE/circular_Sd197.htm

Mitosis and meiosis

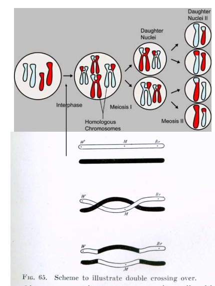
- Sexual organisms are usually diploid
  - Germline cells (gametes) contain N chromosomes
  - Somatic (body) cells have 2N chromosomes
- Meiosis: reduction of chromosome number from 2N to N during reproductive cycle
  - One chromosome doubling is followed by two cell divisions
- Mitosis: growth and development of the organism
  - One chromosome doubling is followed by one cell division



Major events in meiosis
http://en.wikipedia.org/wiki/Meiosis
http://www.ncbi.nlm.nih.gov/About/Primer

Recombination and variation

- Allele is a viable DNA coding occupying a given locus (position in the genome)
- In recombination, alleles from parents become suffled in offspring individuals via chromosomal crossover over
- Allele combinations in offspring are usually different from combinations found in parents
- Recombination errors lead into additional variations

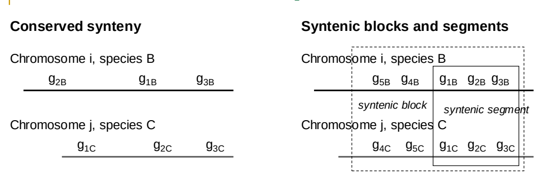


Chromosomal crossover as described by T. H. Morgan in 1916

Recombination frequency and linked genes

- Genetic marker: some DNA sequence of interest (e.g., gene or a part of a gene)
- Recombination is more likely to separate two distant markers than two close ones
- Linked markers: "tend" to be inherited together
- Marker distances measured in centimorgans: 1 centimorgan corresponds to 1% chance that two markers are separated in recombination

Recombination and longer time scales



- Assume that species B and C are descendants of A
- Conserved synteny: group of genes linked in both B and C
- Conserved segment: conserved synteny with same gene order
- Syntenic segment: group of markers (!) linked in both B and C
- Syntenic block: set of syntenic segments which may contain set inversions and duplications

Biological string manipulation

- Errors get introduced to DNA during replication
  - Deletion: removal of one or more contiguous bases (substring)
  - Insertion: insertion of a substring
  - Segmental duplication: insertion of a copy of a DNA region into a different location
  - Inversion: reversal of substring
  - Translocation: removal and insertion of a substring
  - Point mutation: substitution of a base

References

- Richard C. Deonier, Simon Tavaré and Michael S. Waterman. Computational Genome Analysis, An Introduction. Springer, 2005.
- 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.