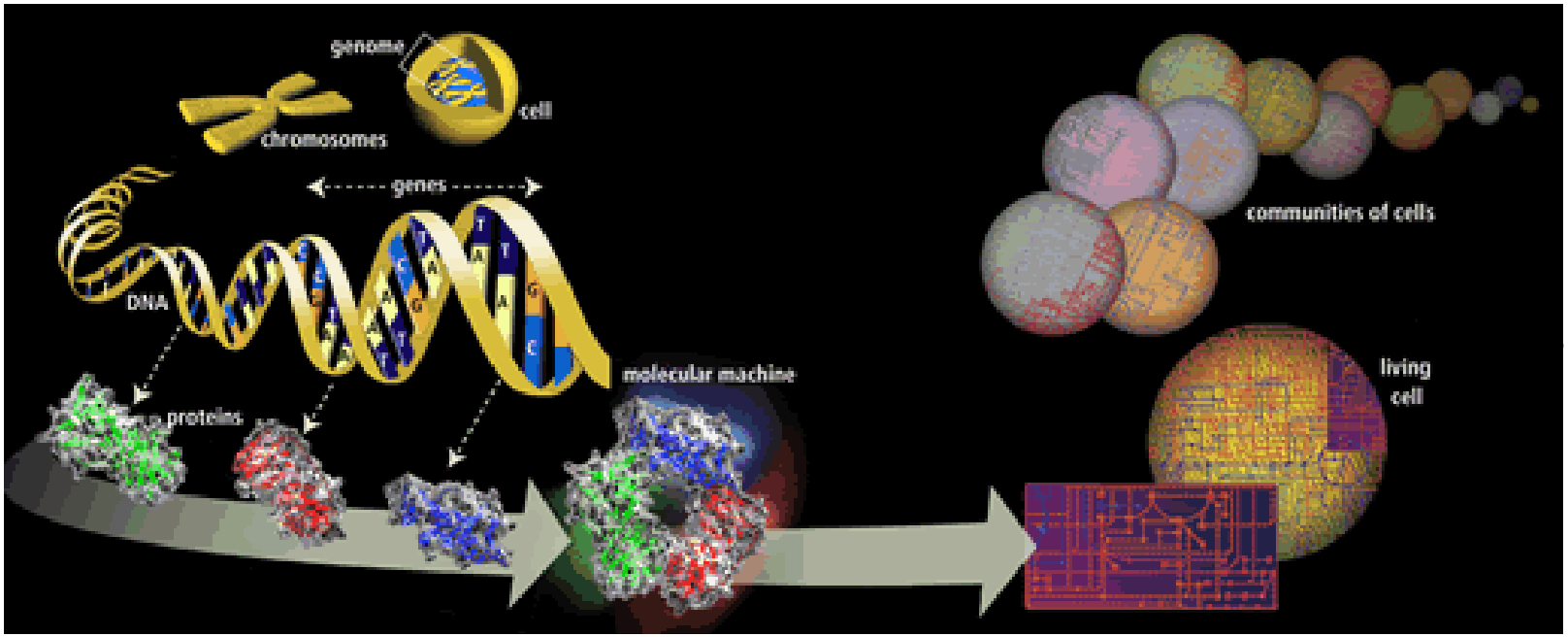
Introduction to Molecular Biology



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

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)

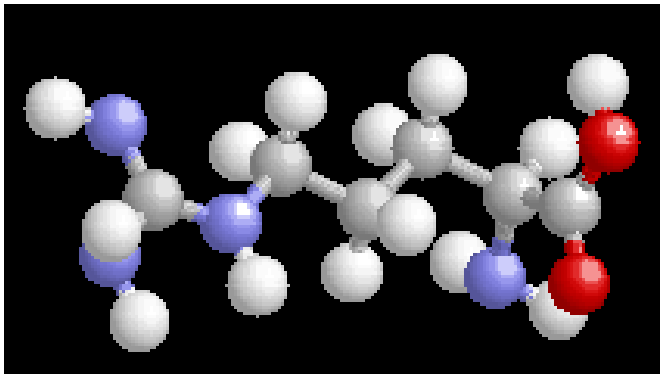
Mendel: The Father of Genetics

- **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

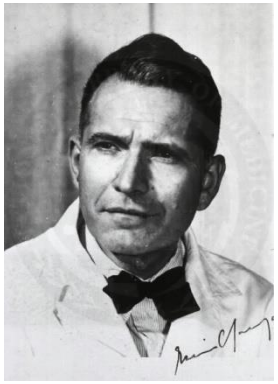
Major events in the history of Molecular Biology 1940 - 1950

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

George Beadle

Edward Tatum

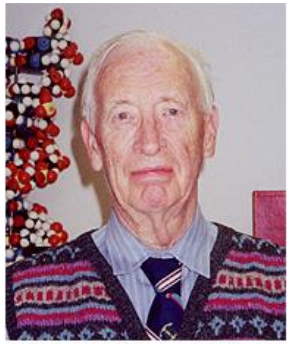
- **1950** – Edwin Chargaff find Cytosine complements Guanine and Adenine complements Thymine



Edwin Chargaff

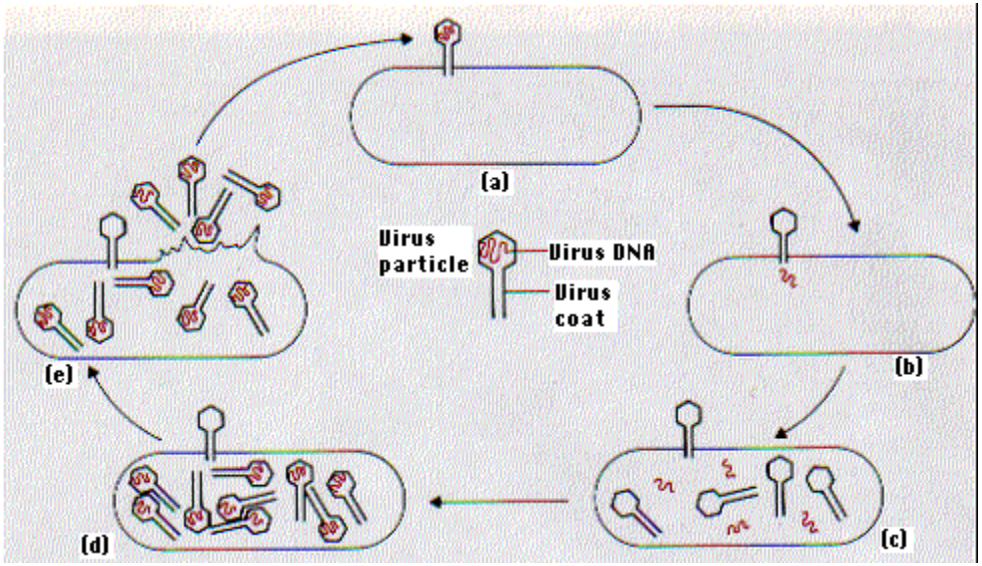
Major events in the history of Molecular Biology 1950 - 1952

- **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

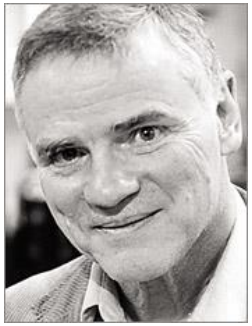
James Watson and Francis Crick

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

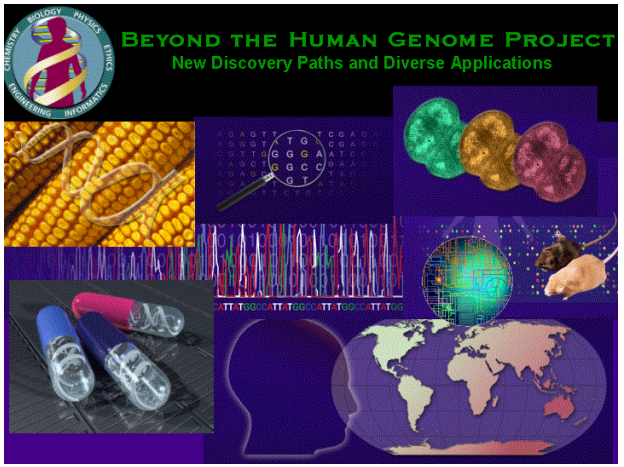
George Emil Palade

Major events in the history of Molecular Biology 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
- **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 bactierial genomes sequenced

John Craig Venter

- **1995** Automated fluorescent sequencing instruments and robotic operations

- **1996** First eukaryotic genome-yeast-sequenced

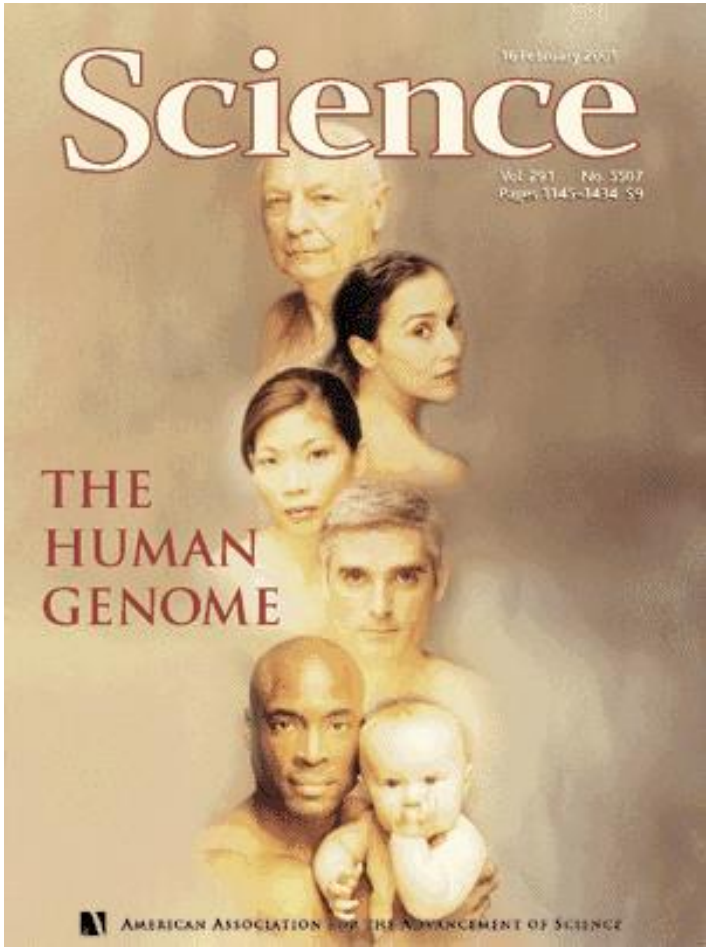
Major events in the history of Molecular Biology 1997 - 1999

- **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

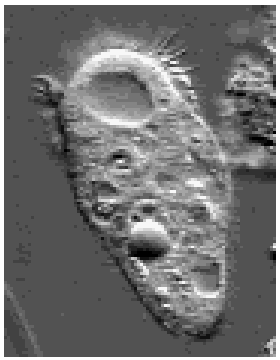
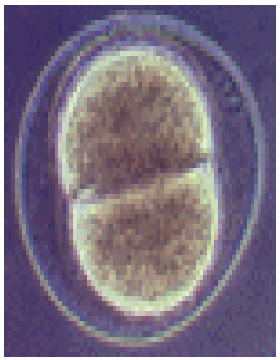
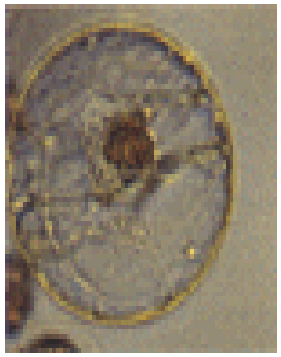
- **2001** International Human Genome Sequencing:first draft of the sequence of the human genome published

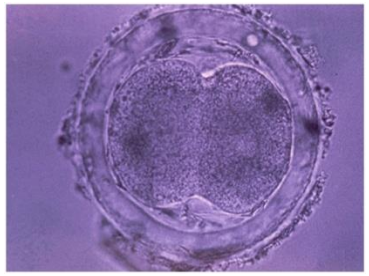
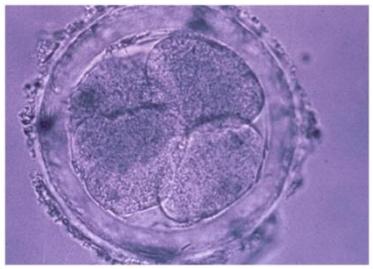
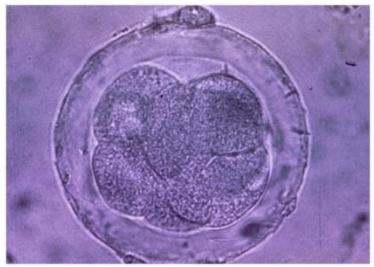


Major events in the history of Molecular Biology 2003- Present

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

Section1: 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:

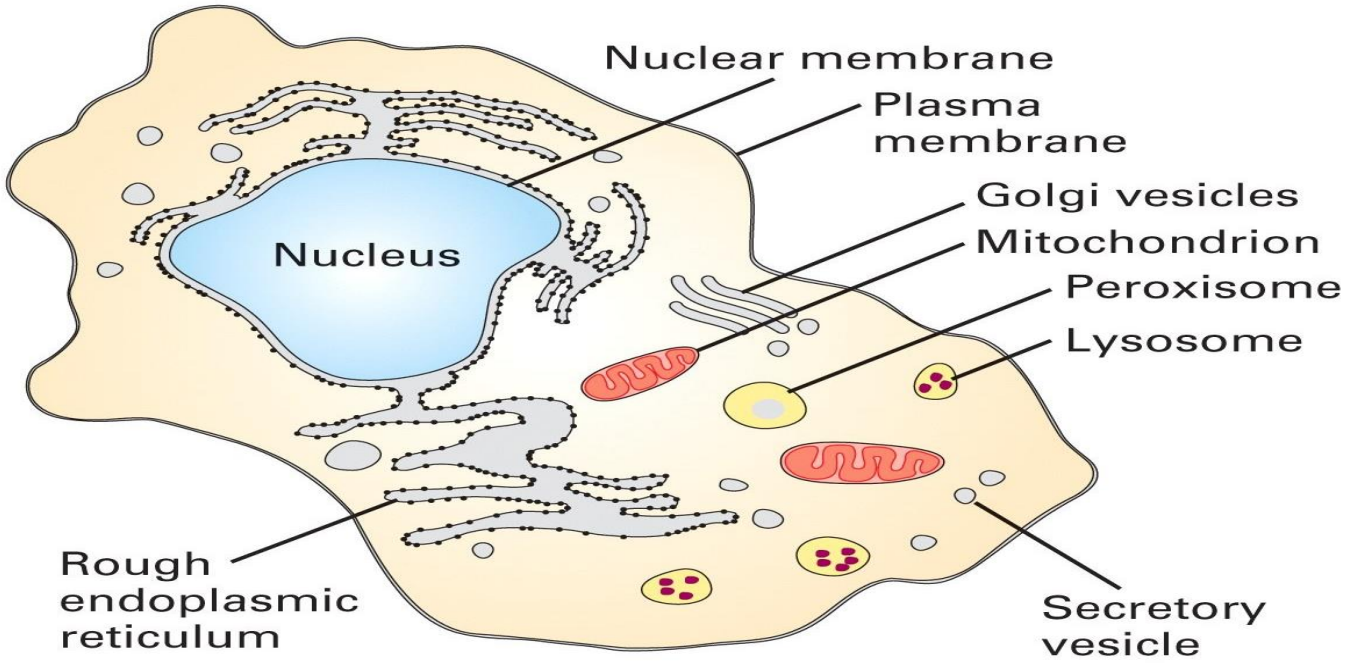
  prokaryotic cells or

  eukaryotic cells.
- **Prokaryotes** and **Eukaryotes** are descended from the same primitive cell.
  - All extant prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

Cells

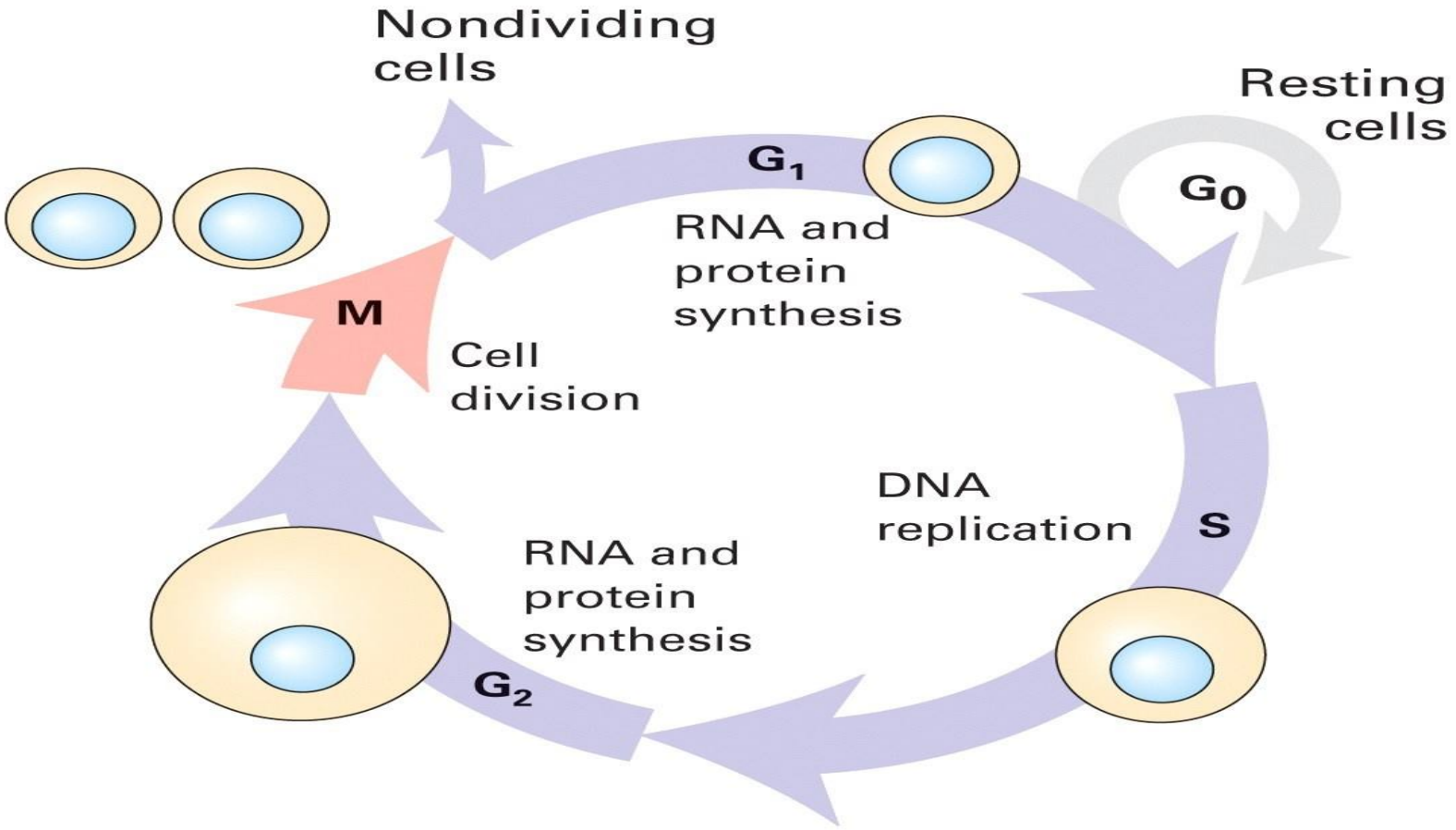
- *Chemical composition*-by weight
  - 70% water
  - 7% small molecules
    - salts
    - Lipids
    - amino acids
    - nucleotides
  - 23% macromolecules
    - Proteins
    - Polysaccharides
    - lipids
- *biochemical (metabolic) pathways*
- *translation of mRNA into proteins*

Life begins with Cell



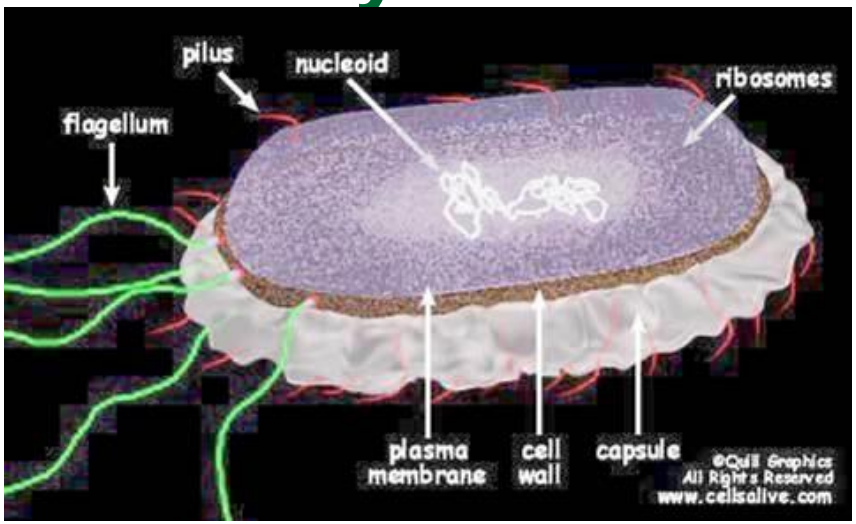
- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

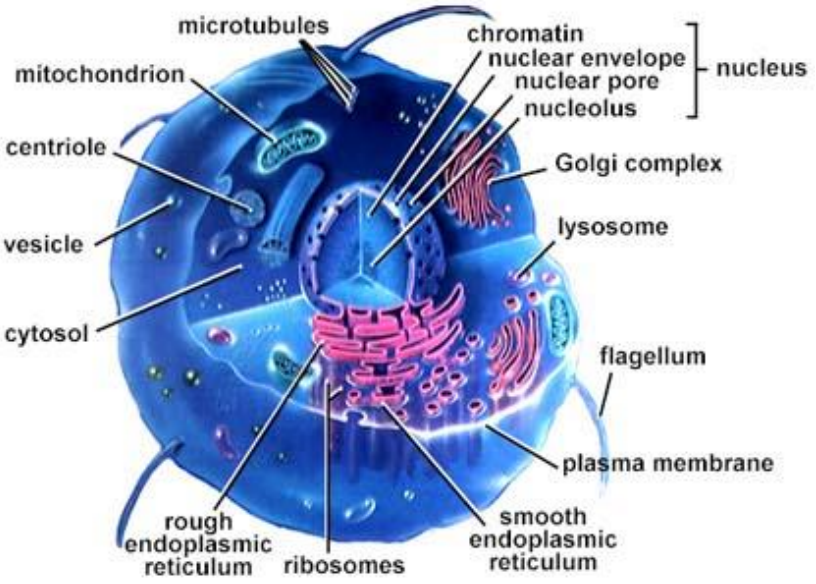
All Cells have common Cycles



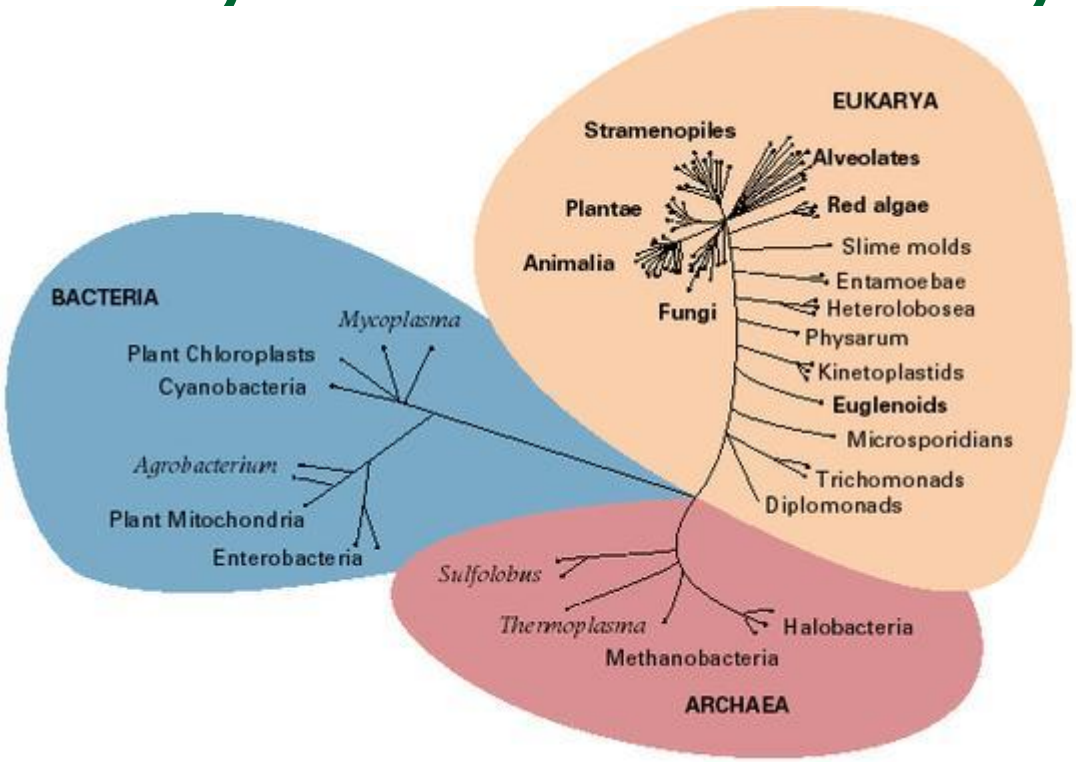
- Born, eat, replicate, and die

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.

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 vaious functions.

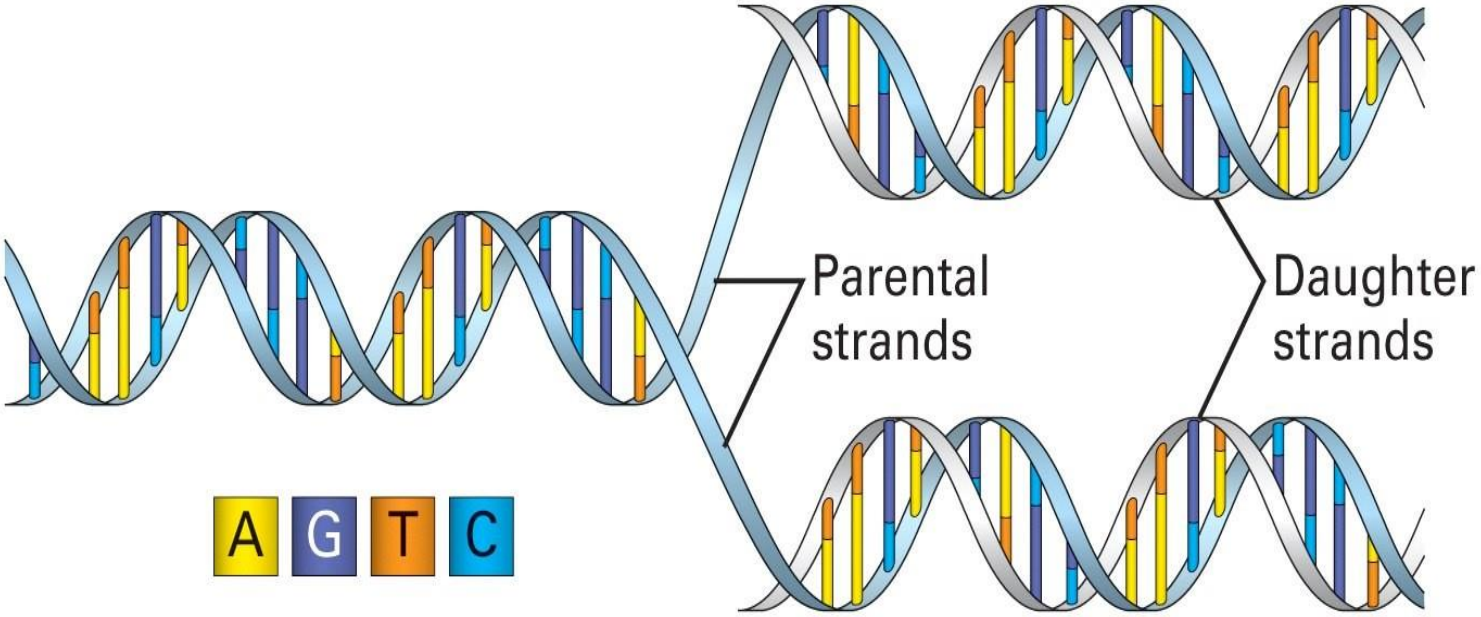
Terminology

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

All Life depends on 3 critical molecules

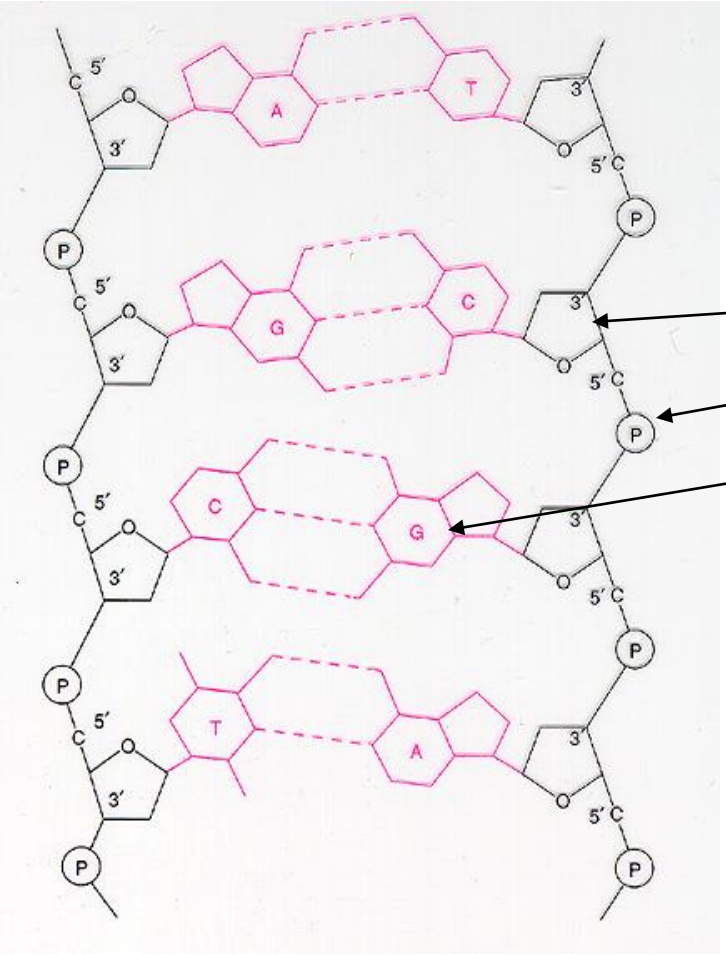
- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g.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



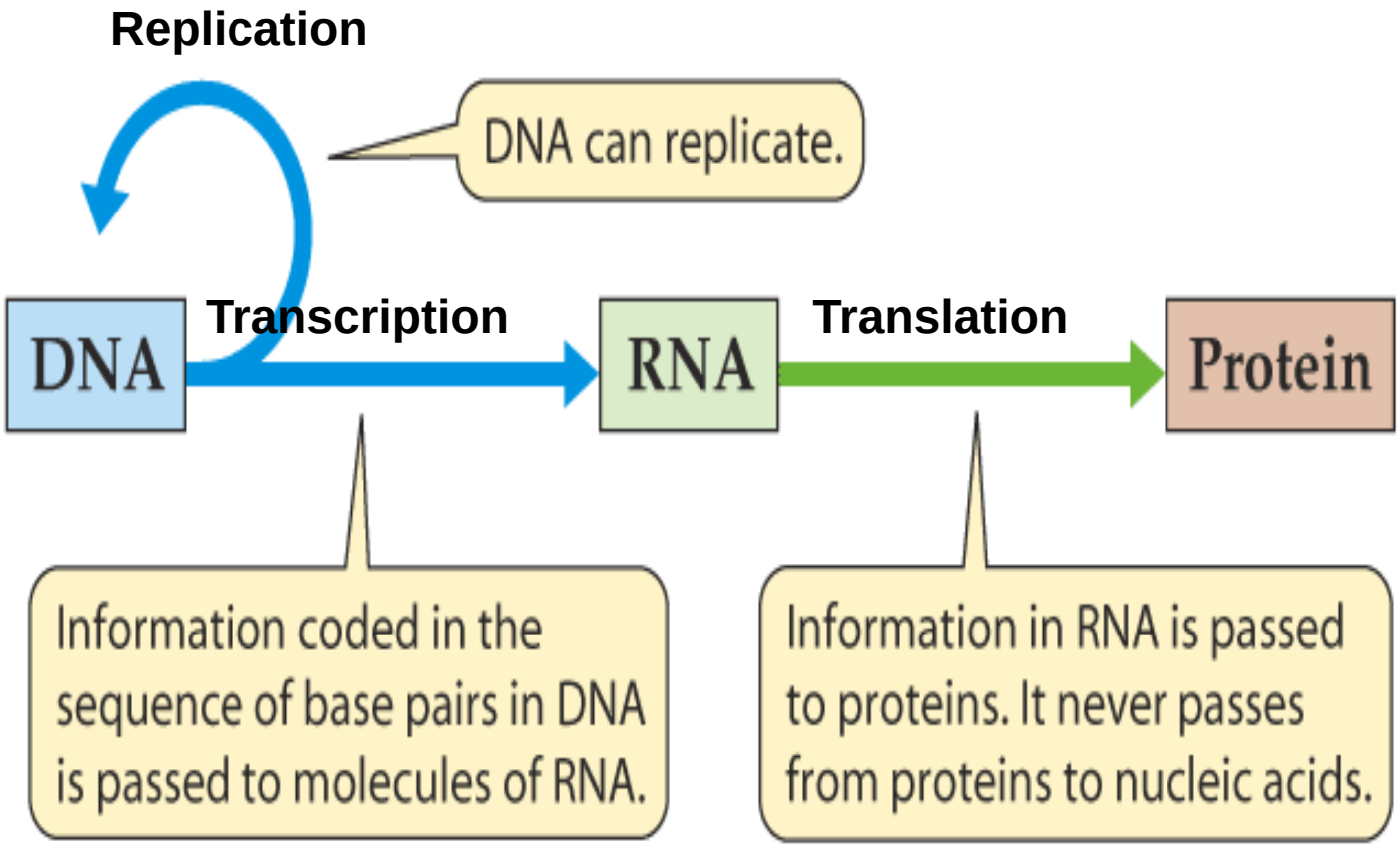
- DNA has a double helix structure which composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

- DNA always reads from 5’ end to 3’ end for transcription replication

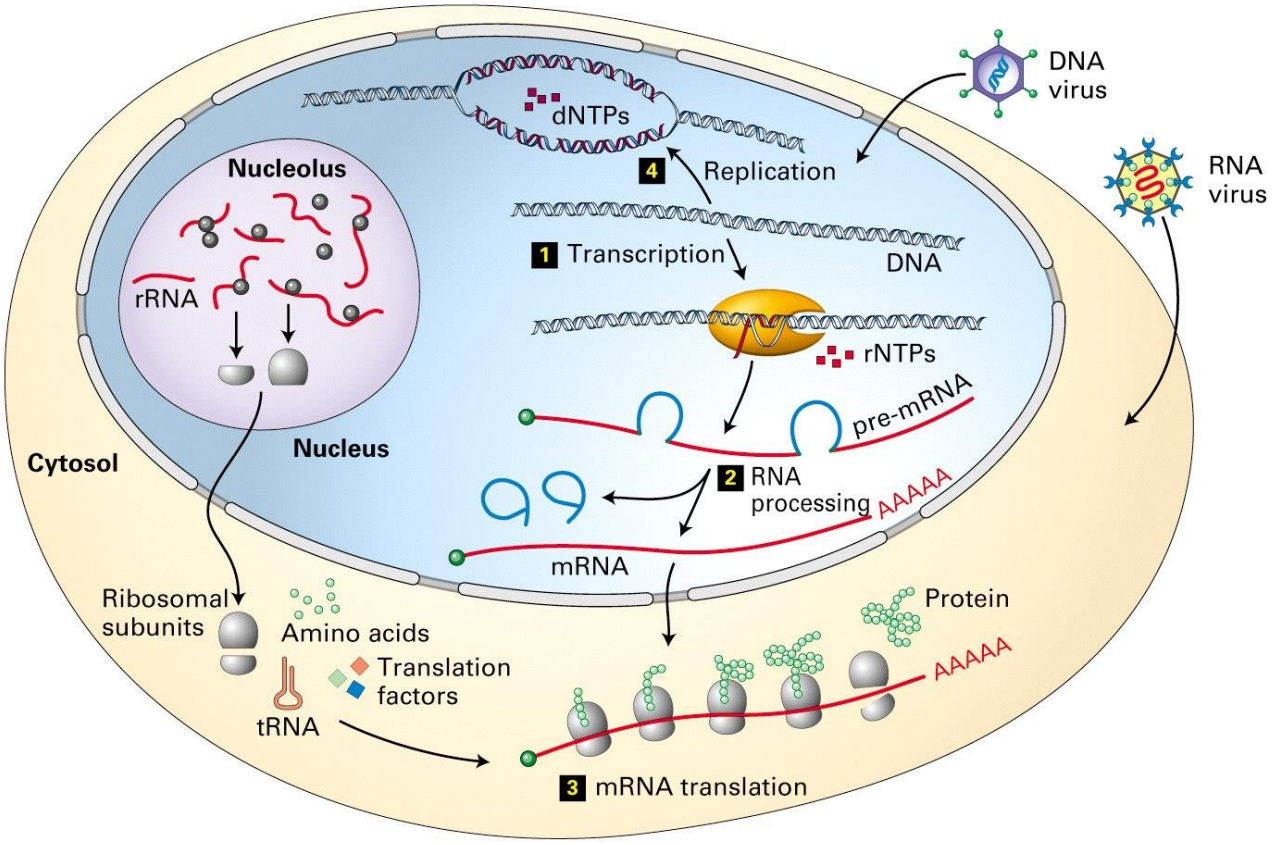
  5’ ATTTAGGCC 3’

  3’ TAAATCCGG 5’

DNA, RNA, and the Flow of Information



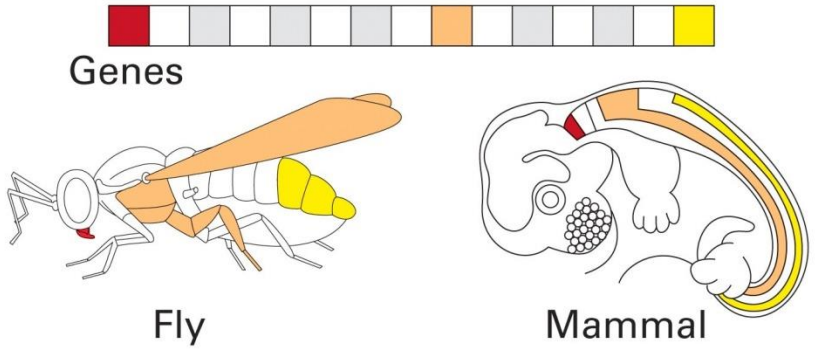
Overview of DNA to RNA to Protein



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

https://www.youtube.com/watch?v=NJxobgkPEAo

DNA the Genetics Makeup

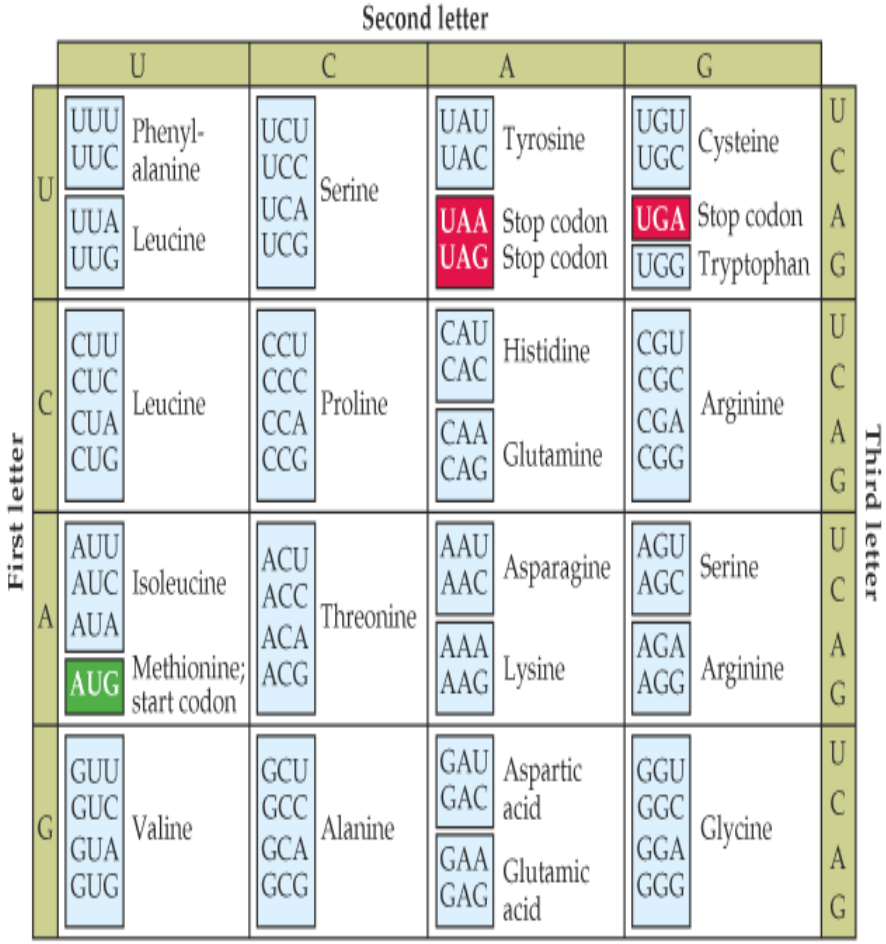


- Genes are inherited and are expressed
  - **genotype** (genetic makeup)
  - **phenotype** (physical expression)

- 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.)



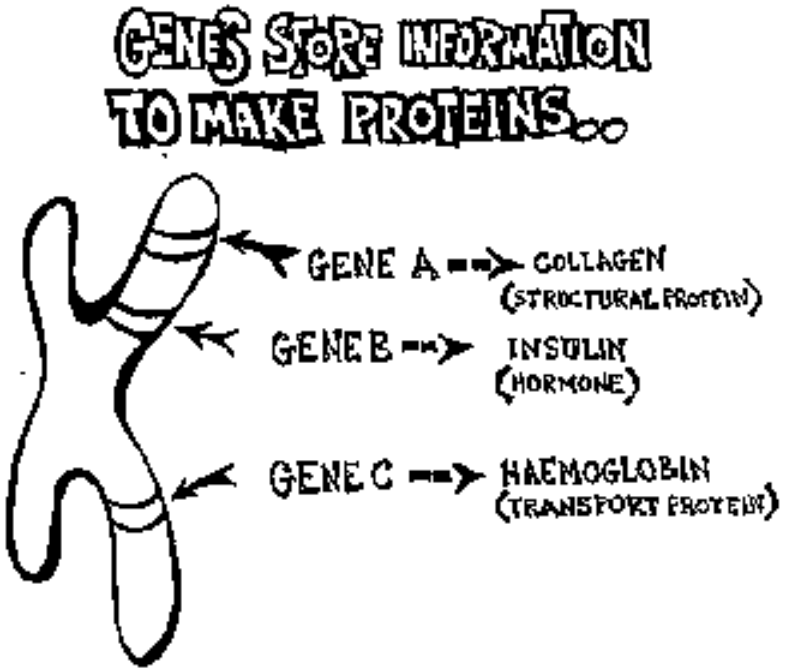
Chromosomes

Organism	Number of base pair	number of Chromosomes
Prokayotic		
Escherichia coli (bacterium)	$4x10^6$	1
Eukaryotic		
Saccharomyces cerevisiae (yeast)	$1.35x10^7$	17
Drosophila melanogaster(insect)	$1.65x10^8$	4
Homo sapiens(human)	$2.9x10^9$	23
Zea mays(corn)	$5.0x10^9$	10

Section 3: What Do Genes Do?

Genes Make Proteins

- genome-> genes ->protein(forms cellular structural & life functional)->pathways & physiology



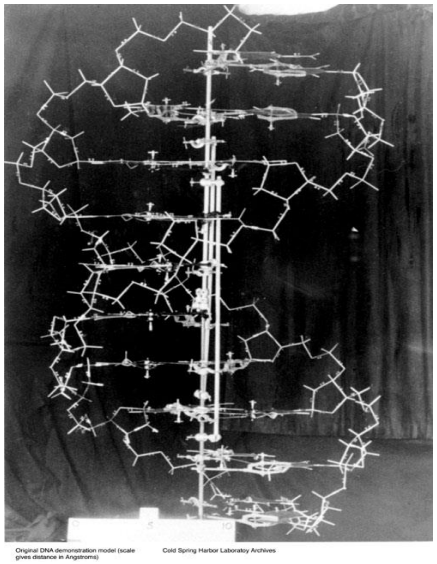
Proteins: Workhorses of the Cell

- 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.

Section 4: What Molecule Codes For Genes?

Discovery of DNA

- DNA Sequences
  - Chargaff and Vischer, 1949
    - DNA consisting of A, T, G, C
      - Adenine, Guanine, Cytosine, Thymine
  - Chargaff Rule
    - Noticing #A≈#T and #G≈#C
      - A "strange but possibly meaningless" phenomenon.
- Wow!! A Double Helix
  - Watson and Crick, *Nature,* April 25, 1953
  - **1 Biologist**
    1 Physics Ph.D. Student
    + 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 Backs Cold Spring Harbor Laboratory Archives

Crick Watson

Watson & Crick – *"…the secret of life"*

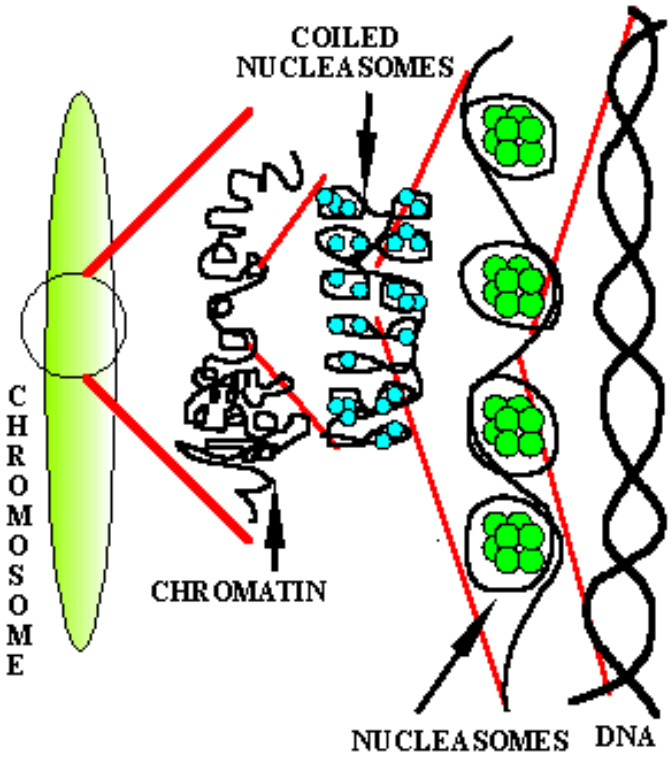
- Watson: a zoologist, Crick: a physicist
- "*In 1947 Crick knew no biology and practically no organic chemistry or crystallography..*" – www.nobel.se
- Applying Chagraff's rules and the X-ray image from Rosalind Franklin, they constructed a "tinkertoy" model showing the double helix
- Their 1953 *Nature* 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

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 where 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



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.

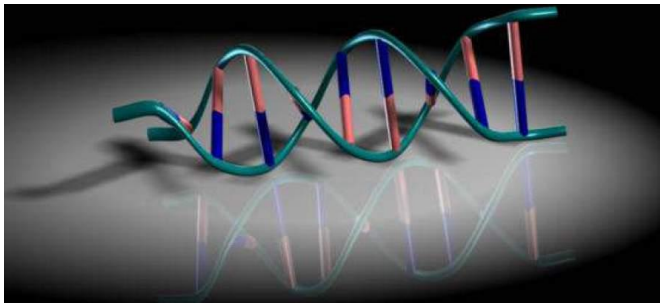
SOURCE: E. S. Lander et al., 2001, *Nature* 409:860.

Section 5: The Structure of DNA

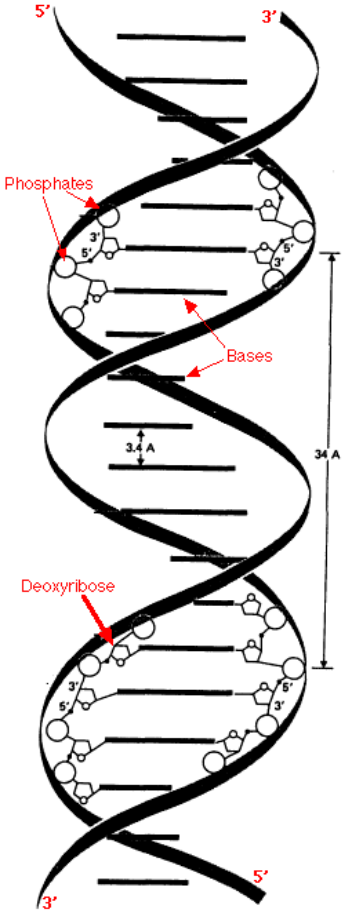
CSE 181

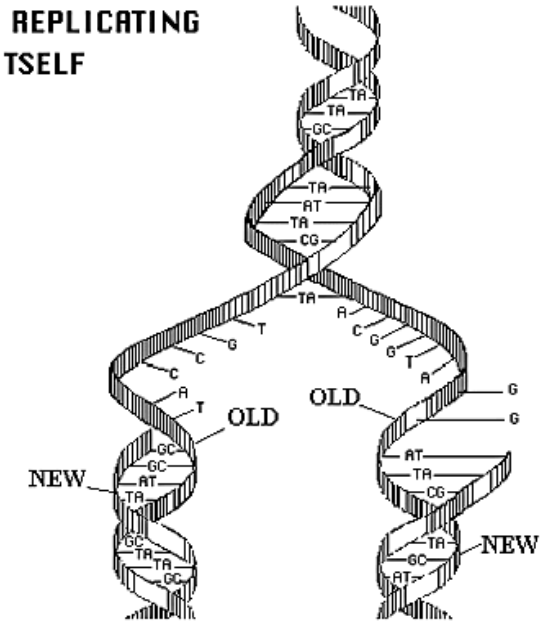
Raymond Brown

May 12, 2004

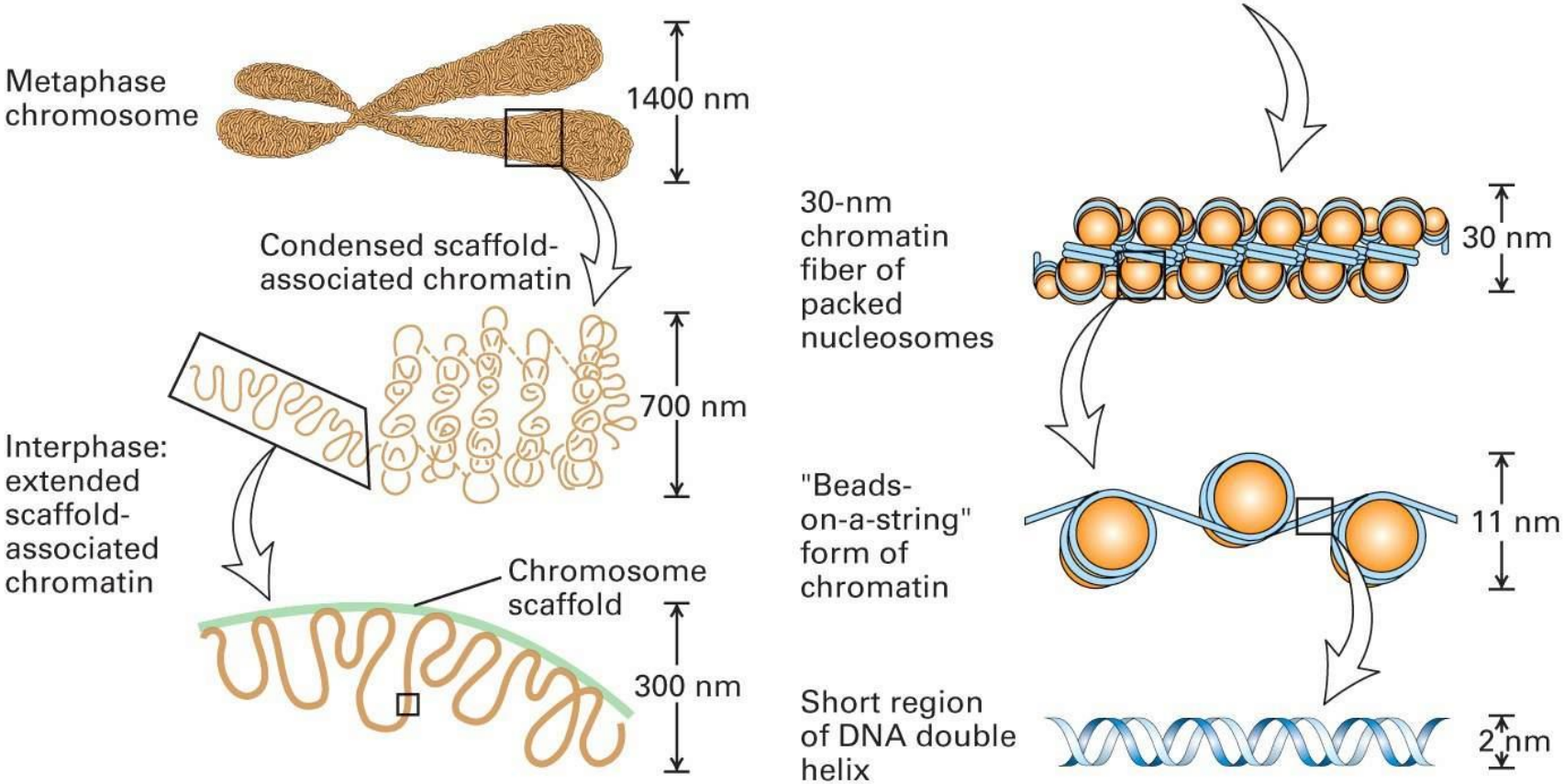


DNA Replication





Superstructure

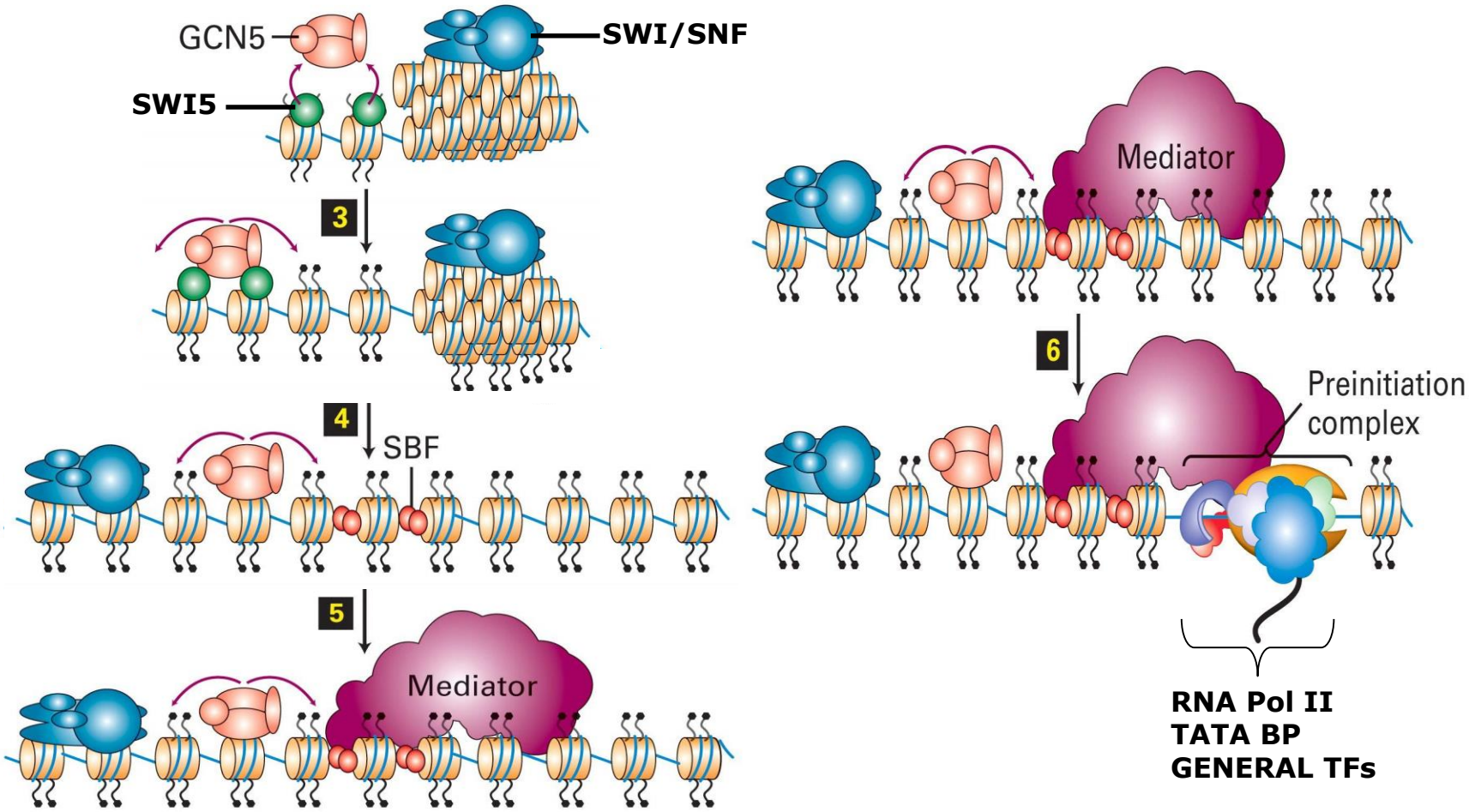


Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

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

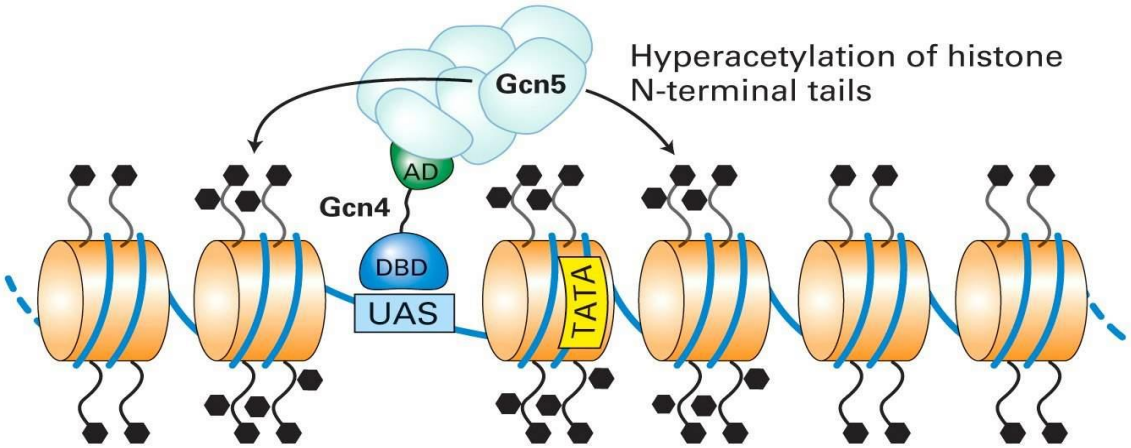
Transcriptional Regulation



Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

The Histone Code

- State of histone tails govern TF access to DNA
- State is governed by amino acid sequence and modification (acetylation, phosphorylation, methylation)



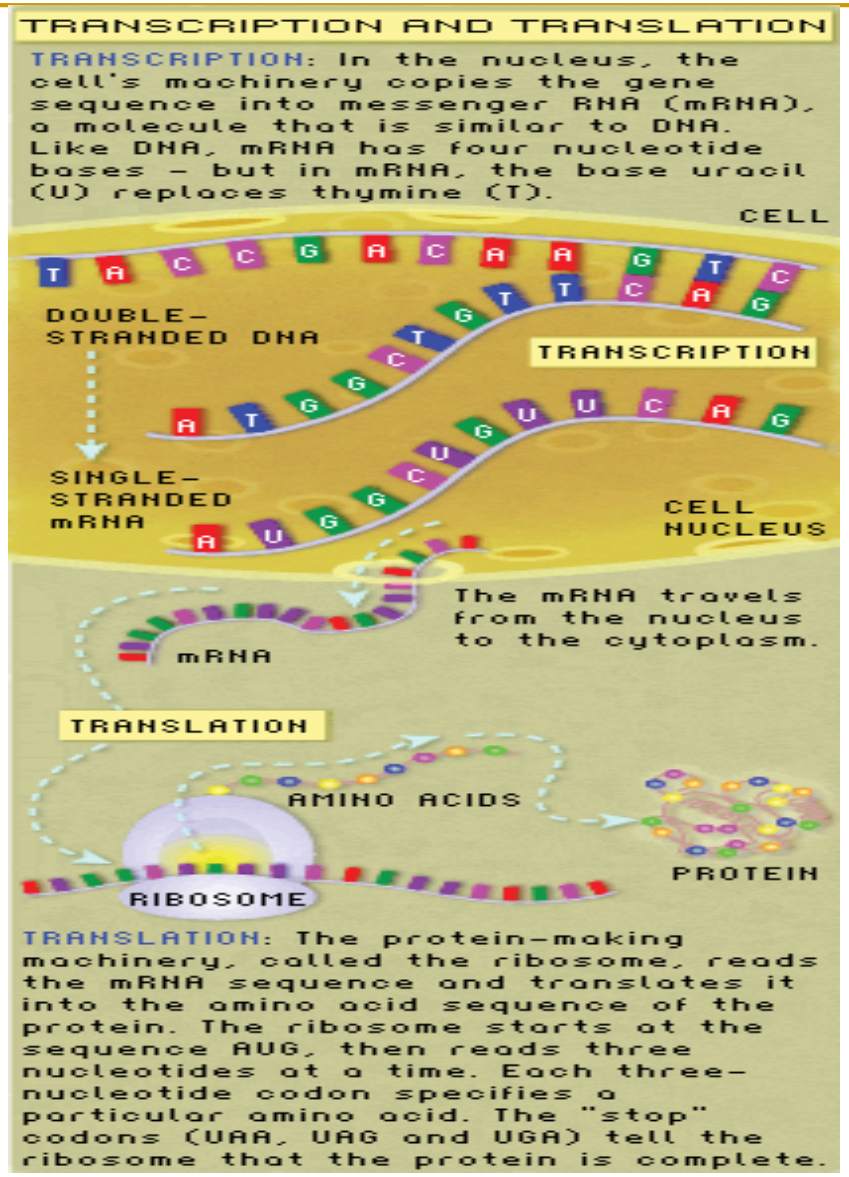
Lodish et al. *Molecular Biology of the Cell* (5th ed.). W.H. Freeman & Co., 2003.

Section 6: What carries information between DNA to Proteins

Outline For Section 6:

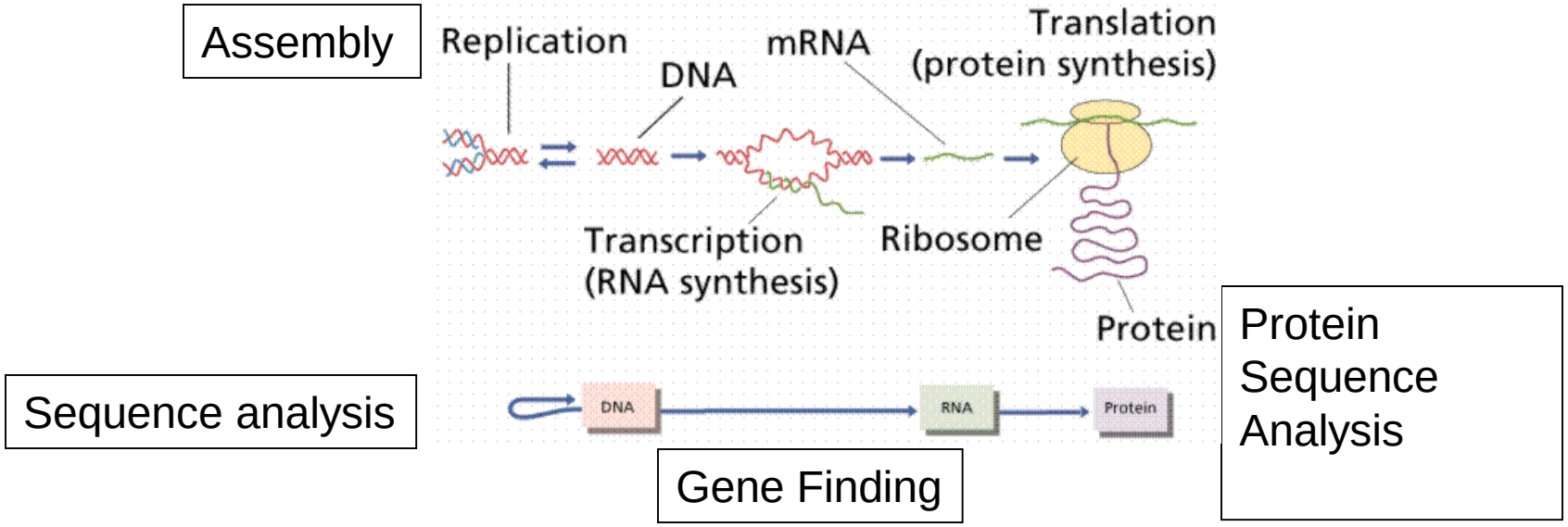
- *Central Dogma Of Biology*
- *RNA*
- *Transcription*
- *Splicing hnRNA-> mRNA*

- **Central Dogma**
  (DNA→RNA→protein) The paradigm that DNA directs its transcription to RNA, which is then translated into a protein.
- **Transcription**
  (DNA→RNA) The process which transfers genetic information from the DNA to the RNA.
- **Translation**
  (RNA→protein) The process of transforming RNA to protein as specified by the genetic code.



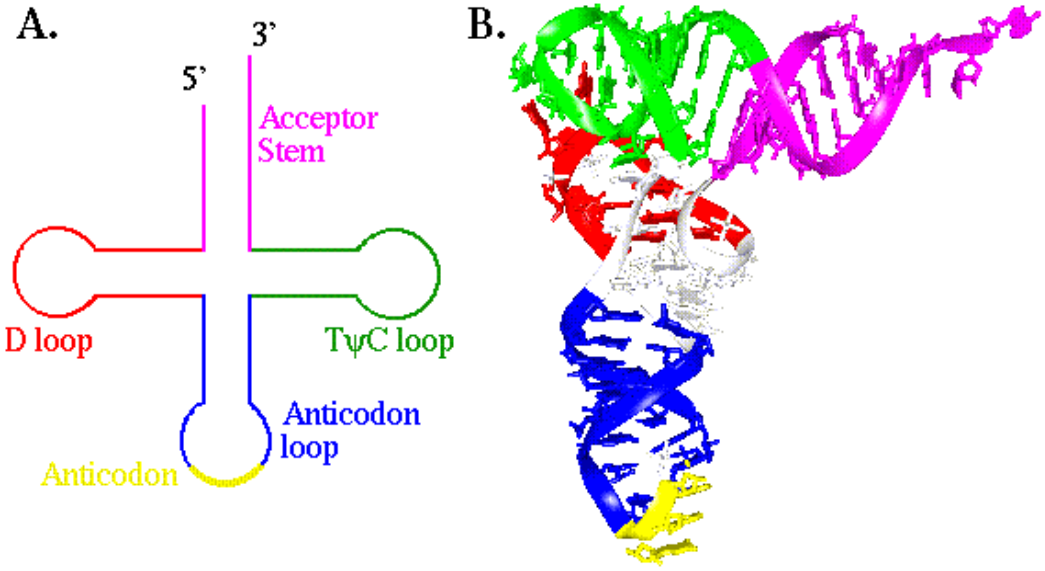
Central Dogma of Biology

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

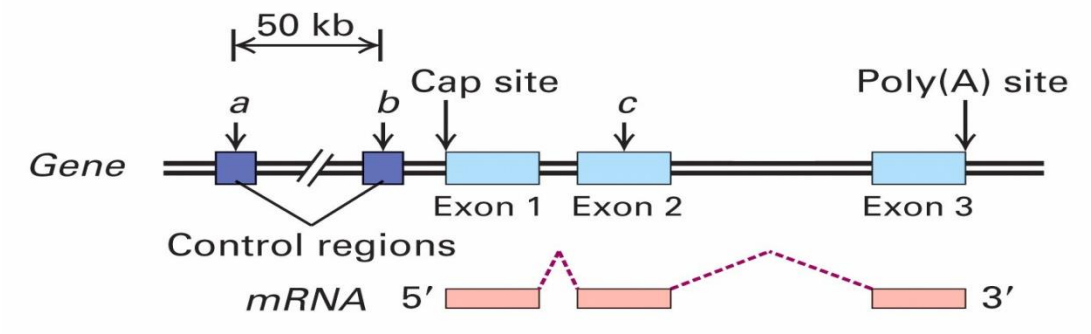
- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)
- Some forms of RNA can form secondary structures by "pairing up" with itself. This can have change its properties



DNA and RNA can pair with each other.

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

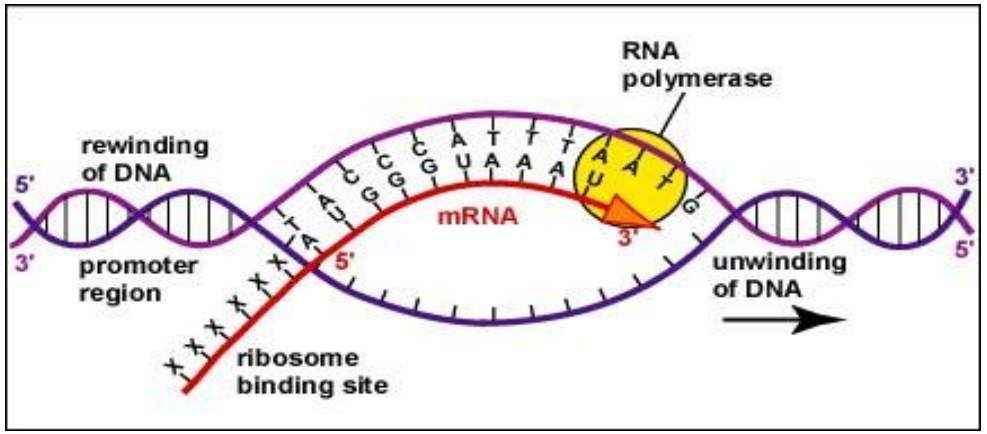
Definition of a Gene



- 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.

Transcription: DNA → hnRNA

- Transcription occurs in the nucleus.
- σ factor from RNA polymerase reads the promoter sequence and opens a small portion of the double helix exposing the DNA bases.

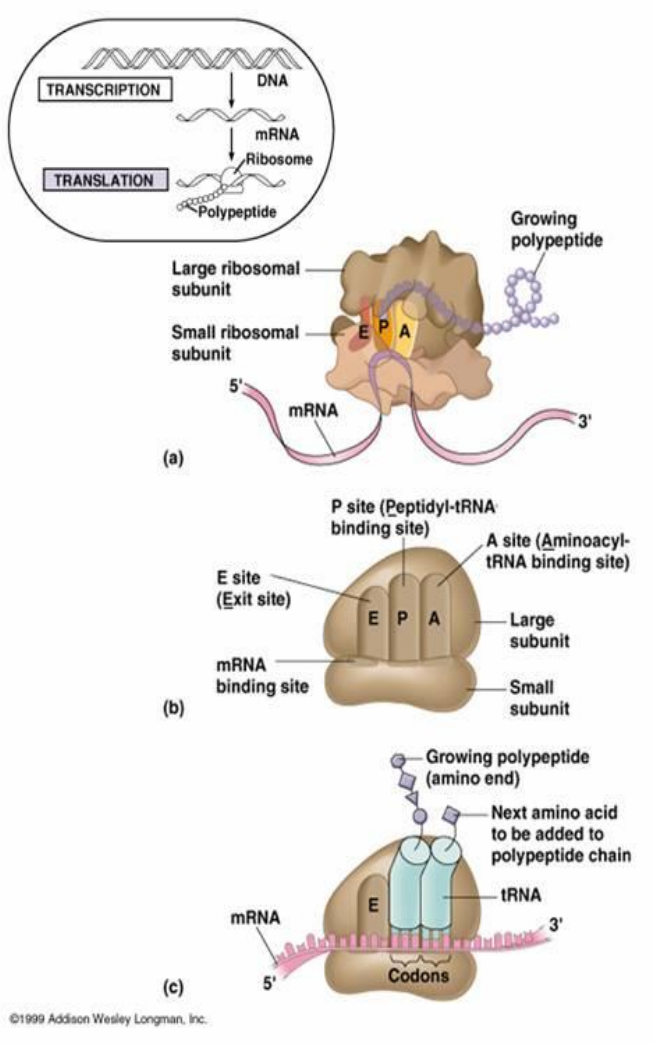


- RNA polymerase II catalyzes the formation of phosphodiester bond that link nucleotides together to form a linear chain from 5' to 3' by unwinding the helix just ahead of the active site for polymerization of complementary base pairs.
- The hydrolysis of high energy bonds of the substrates (nucleoside triphosphates ATP, CTP, GTP, and UTP) provides energy to drive the reaction.
- During transcription, the DNA helix reforms as RNA forms.
- When the terminator sequence is met, polymerase halts and releases both the DNA template and the RNA.

Section 7: How Are Proteins Made? (Translation)

mRNA → Ribosome

- mRNA leaves the nucleus via nuclear pores.
- Ribosome has 3 binding sites for tRNAs:
  - A-site: position that aminoacyl-tRNA molecule binds to vacant site
  - P-site: site where the new peptide bond is formed.
  - E-site: the exit site
- Two subunits join together on a mRNA molecule near the 5' end.
- The ribosome will read the codons until AUG is reached and then the initiator tRNA binds to the P-site of the ribosome.
- Stop codons have tRNA that recognize a signal to stop translation. Release factors bind to the ribosome which cause the peptidyl transferase to catalyze the addition of water to free the molecule and releases the polypeptide.

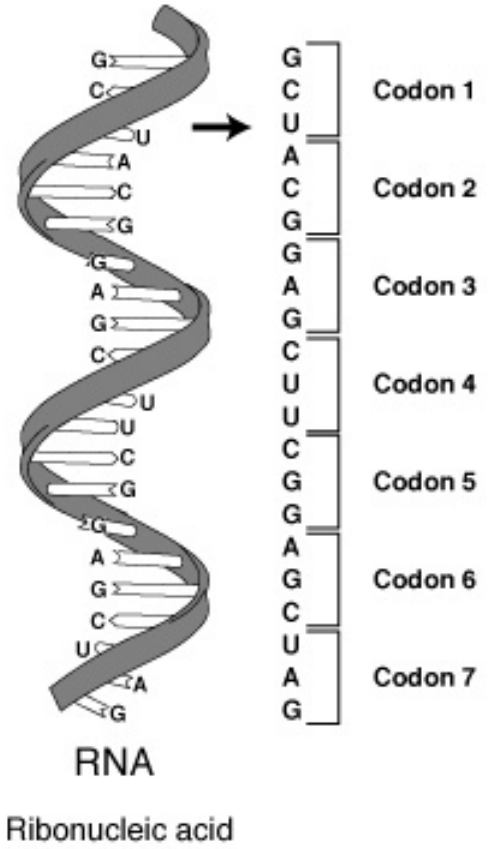


Uncovering the code

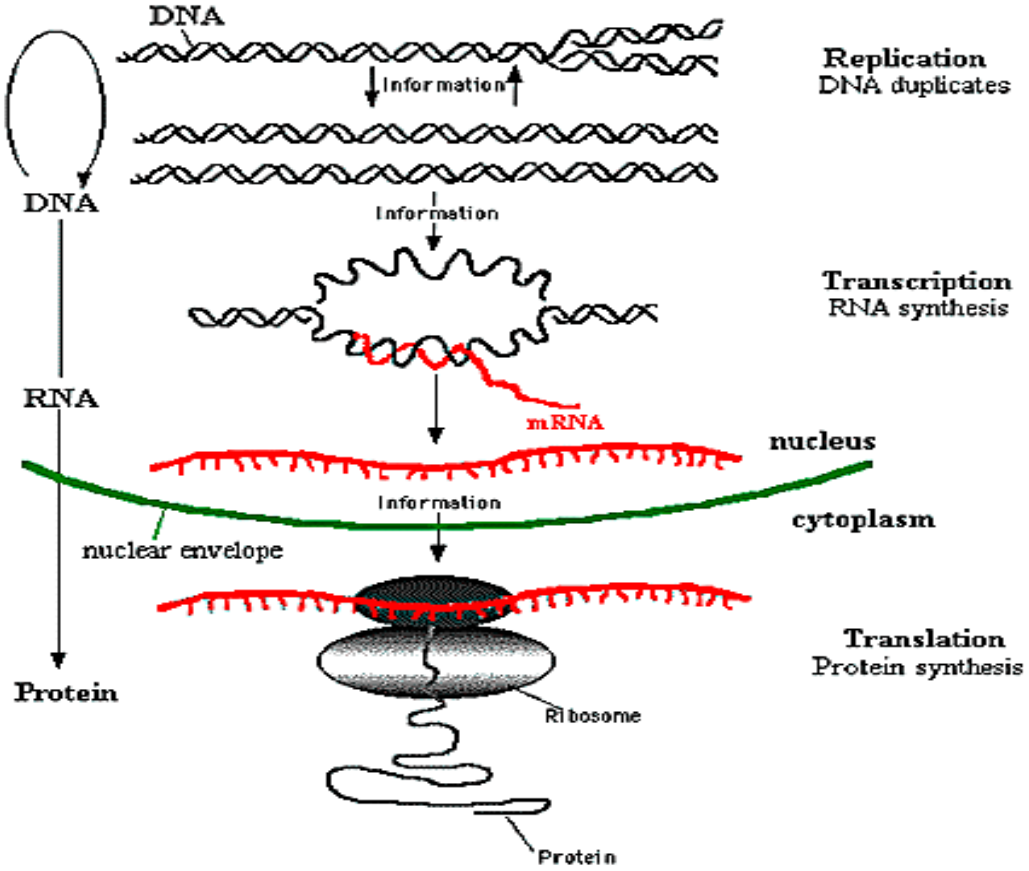
- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there'd be 4^1 amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$
  - This triplet of 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

Revisiting the Central Dogma

- In going from DNA to proteins, there is an intermediate step where mRNA is made from DNA, which then makes protein
  - This known as **The Central Dogma**
- Why the intermediate step?
  - DNA is kept in the nucleus, while protein sythesis happens in the cytoplasm, with the help of ribosomes



The Central Dogma (cont'd)



The Central Dogma of Molecular Biology

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

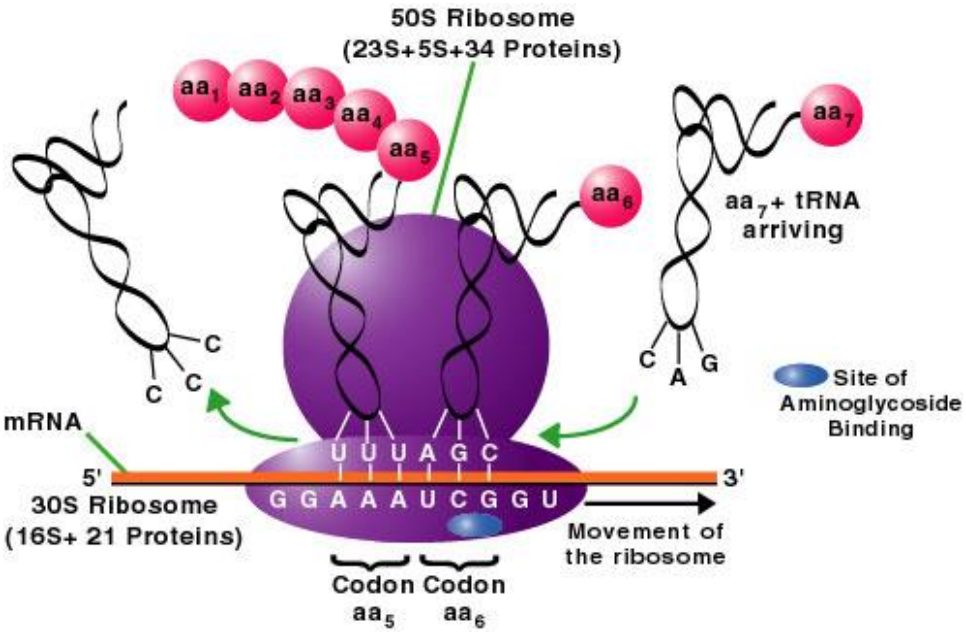
SECOND POSITION

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

* and start

Translation, continued

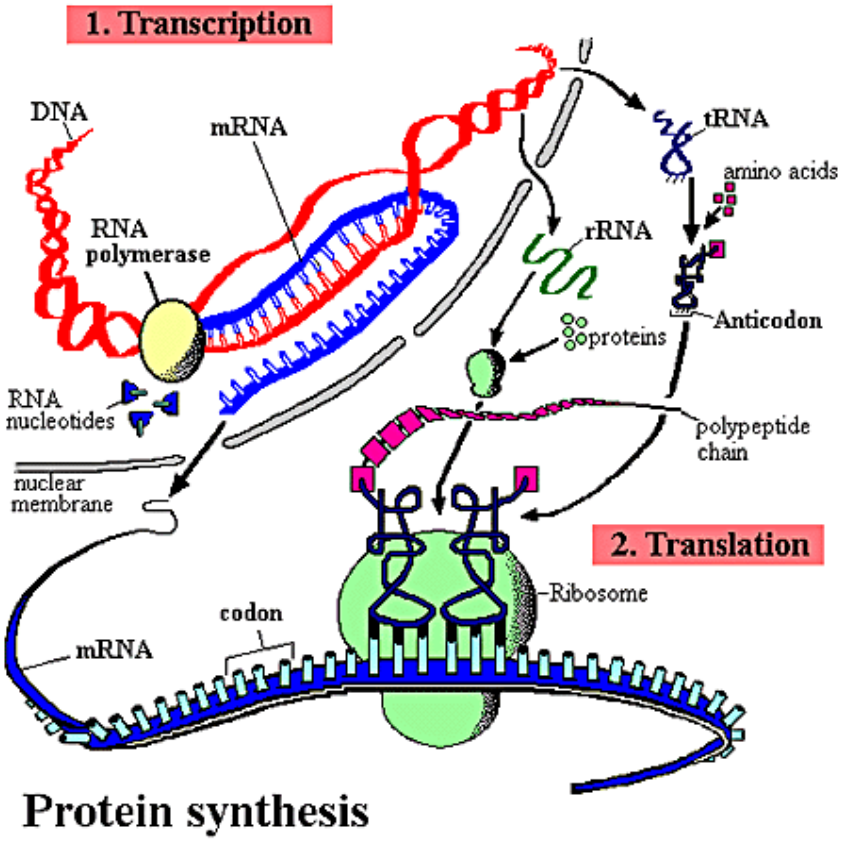
- Catalyzed by Ribosome
- Using two different sites, the Ribosome continually binds tRNA, joins the amino acids together and moves to the next location along the mRNA
- ~10 codons/second, but multiple translations can occur simultaneously



http://wong.scripps.edu/PIX/ribosome.jpg

Protein Synthesis: Summary

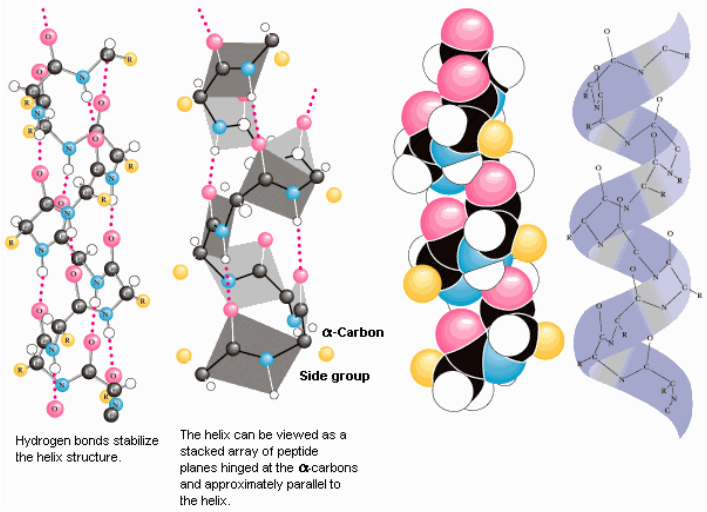
- There are twenty amino acids, each coded by three-base-sequences in DNA, called “codons”
  - This code is degenerate
- The **central dogma** describes how proteins derive from DNA
  - DNA → mRNA → (splicing?) → protein
- The protein adopts a 3D structure specific to it’s amino acid arrangement and function

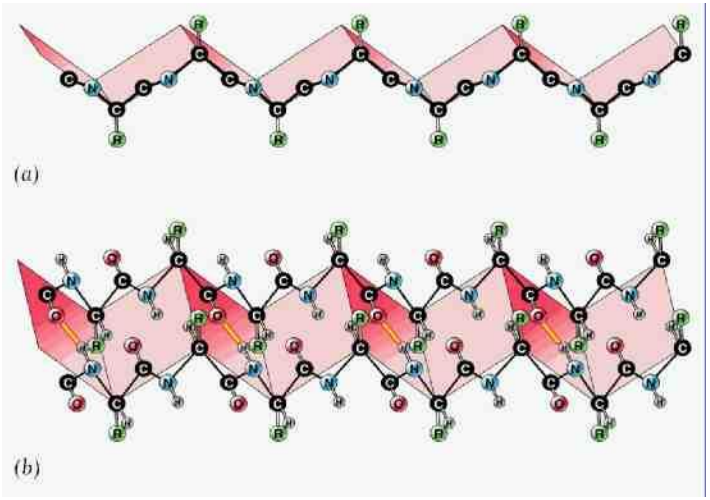


Protein synthesis

Protein Folding

- Proteins tend to fold into the lowest free energy conformation.
- Proteins begin to fold while the peptide is still being translated.
- Proteins bury most of its hydrophobic residues in an interior core to form an α helix.
- Most proteins take the form of secondary structures α helices and β sheets.
- Molecular chaperones, hsp60 and hsp 70, work with other proteins to help fold newly synthesized proteins.
- Much of the protein modifications and folding occurs in the endoplasmic reticulum and mitochondria.



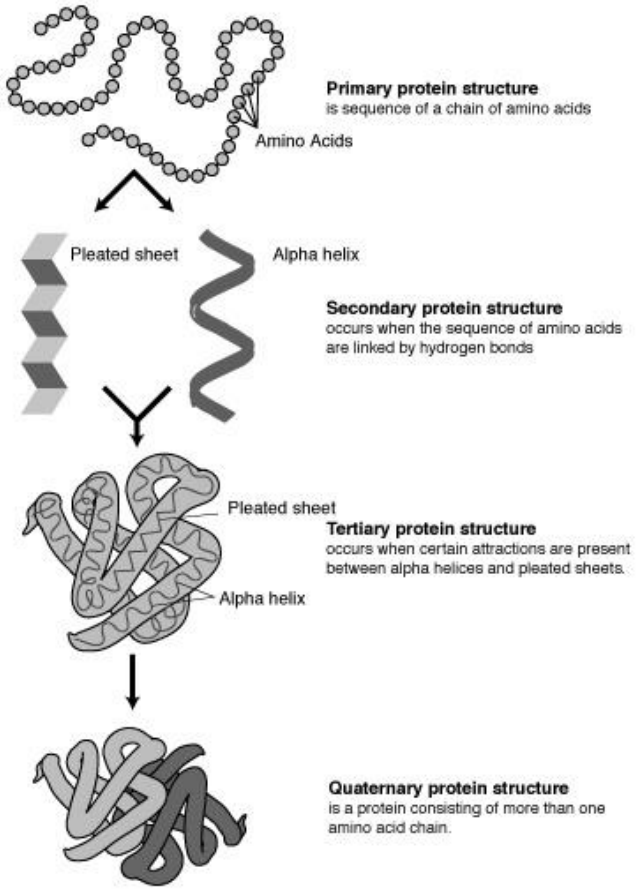


Protein Folding

- 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

Protein Folding (cont'd)

- The structure that a protein adopts is vital to it's chemistry
- Its structure determines which of its amino acids are exposed carry out the protein's function
- Its structure also determines what substrates it can react with.



https://www.youtube.com/watch?v=wJyUtbn0O5Y

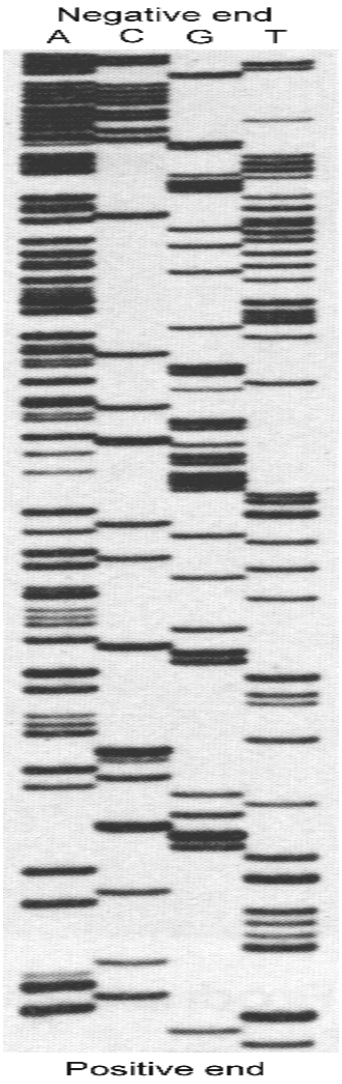
Section 8: How Can We Analyze DNA?

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.

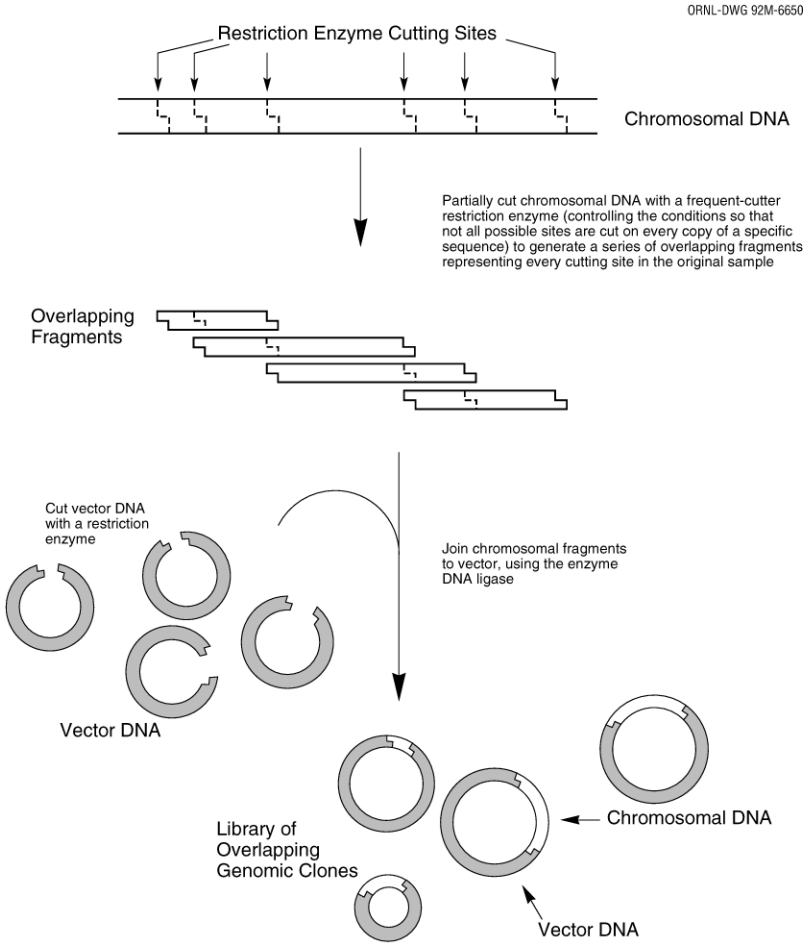
Reading DNA

- 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?

Section 9: How Do Individuals of a Species Differ?

How Do Individuals of Species Differ?

- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes
- While 0.1% of the 3 billion nucleotides in the human genome are the same, small variations can have a large range of phenotypic expressions
- These traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the truth behind several genetic diseases

The Diversity of Life

- 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 ($4^{3,000,000}$) 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.

The Genome of a Species

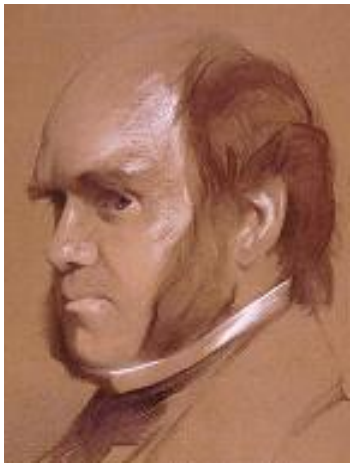
- It is important to distinguish between the genome of a species and the genome of an individual.
- The genome of a species is a representation of all possible genomes that an individual might have since the basic sequence in all individuals is more or less the same.
- The genome of an individual is 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.

Human Diversity Project

- The Human Diversity Project samples the genomes of different human populations and ethnicities to try and understand how the human genome varies.
- It is highly controversial both politically and scientifically because it involves genetic sampling of different human races.
- The goal is to figure out differences between individuals so that genetic diseases can be better understood and hopefully cured.

Section 10: How Do Different Species Differ?

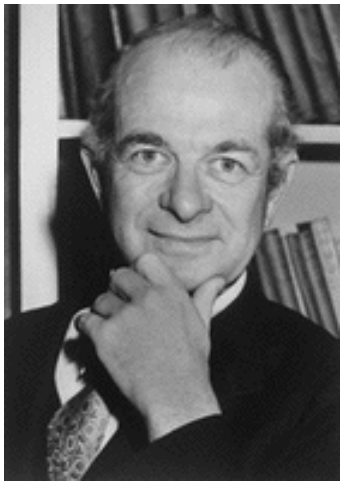
What is evolution?



- 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 Spieces* published in 1859.

Molecular Clock

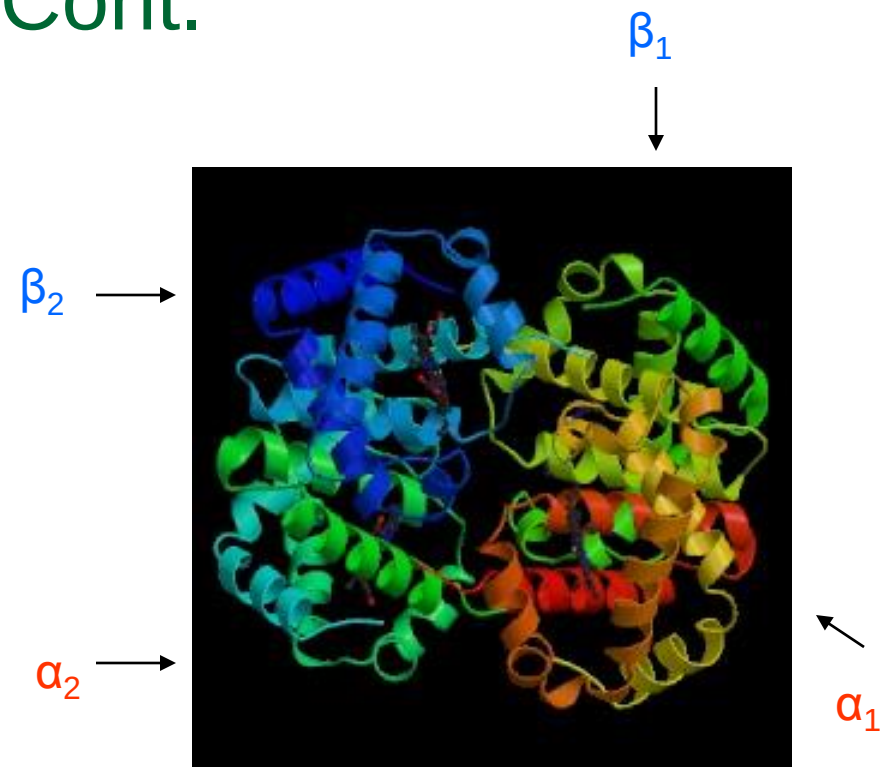
- Introduced by Linus Pauling and his collaborator Emile Zuckerkandl in 1965.
- They proposed that *the rate of evolution in a given protein (or later, DNA) molecule is approximately constant overtime and among evolutionary lineages.*



Linus Pauling

Molecular Clock Cont.

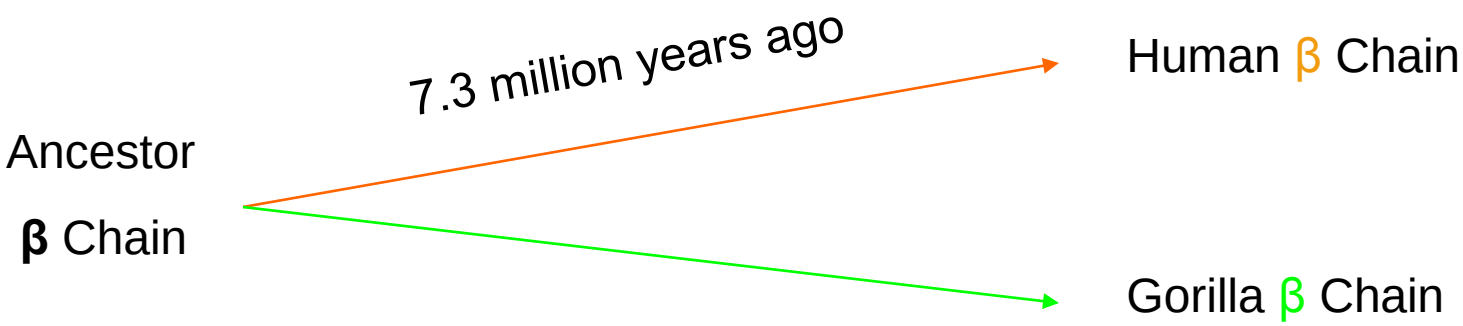
- Observing hemoglobin patterns of some primates, They found:
  - The gorilla, chimpanzee and human patterns are almost identical.
  - The further one gets away from the group of Primates, the primary structure that is shared with human hemoglobin *decreases*.
  - α and β chains of human hemoglobin are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



Molecular Evolution

- Pauling and Zuckerkandl research was one of the pioneering works in the emerging field of *Molecular Evolution.*
- *Molecular Evolution* is 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.

Molecular Evolution Cont.

- 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 evolved 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: MVHLT**PE**EK**S**AV**TA**LWGKV N**V**D**E**VGGEALGRLL

Mouse β chain: MVHLT**DA**EK**A**AV**NG**LWGKVN**P**D**D**VGGEALGRLL

Human β chain: VVYPWTQR**F**F**E**SFGDLS**TPD**A**V**MGNPKVKAHGKKV**LG**

Mouse β chain: VVYPWTQR**Y**F**D**SFGDLS**SAS**A**I**MGNPKVKAHGKK V**IN**

Human β chain: AF**S**DGL**A**HLDNLKGTFA**T**LSELHCDKLHVDPENFRLLGN

Mouse β chain: AF**N**DGL**K**HLDNLKGTFA**H**LSELHCDKLHVDPENFRLLGN

Human β chain: **VL**V**C**VL**A**HH**F**GKEFTP**PV**QAA**Y**QKVVAGVA**N**ALAHKYH

Mouse β chain: **MI** V**I** VL**G**HH**L**GKEFTP**CA**QAA**F**QKVVAGVA**S**ALAHKYH

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

Beta globins: Cont.

Human β chain: MVH L TPEEKSAVTALWGKVNVDEVGGEALGRLL

Chicken β chain: MVHWTAEEKQL I TGLWGKVNVAECGAEALARLL

Human β chain: VVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG

Chicken β chain: IVYPWTQRFF ASFGNLSSPTA I LGNPMVRAHGKKVLT

Human β chain: AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGN

Chicken β chain: SFGDAVKNLDNIK NTFSQLSELHCDKLHVDPENFRLLGD

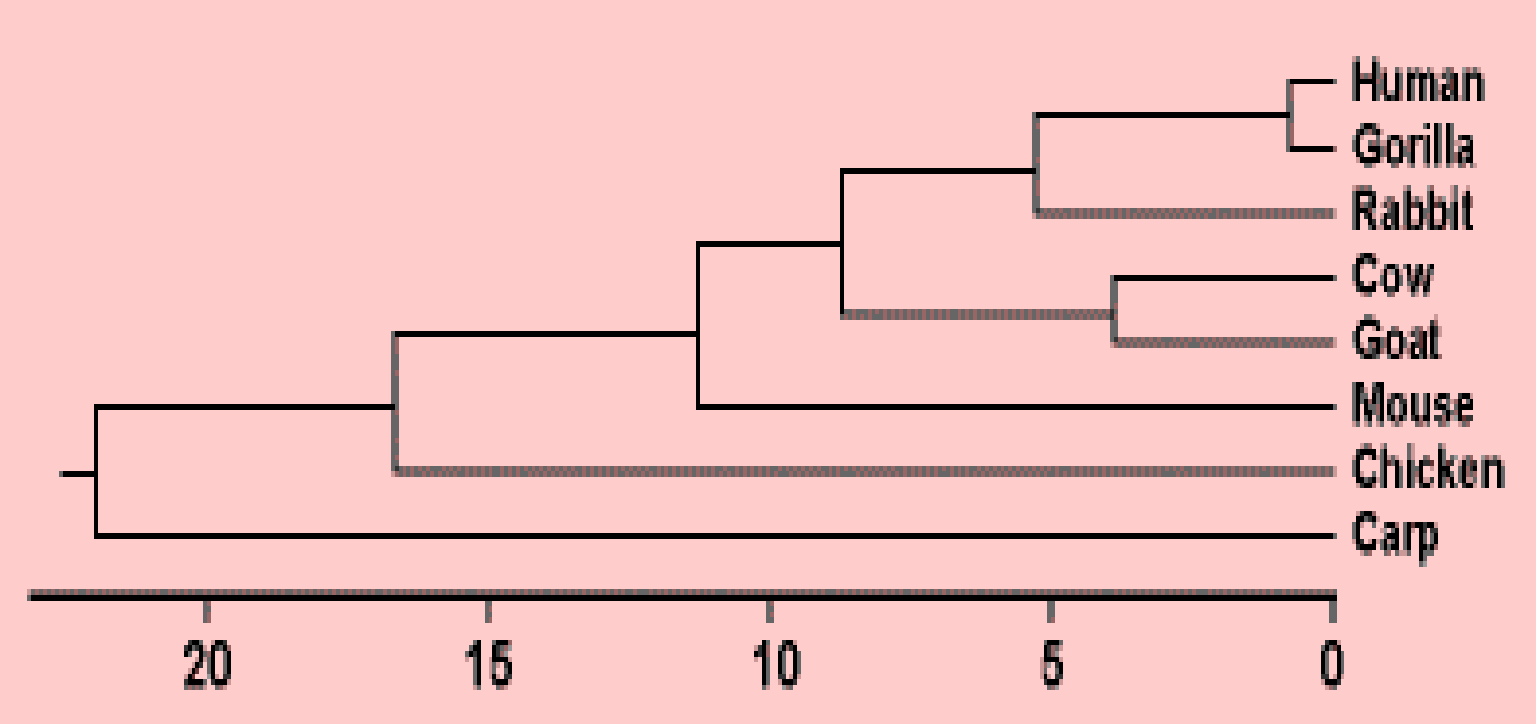
Human β chain: VLVCVLAHHFGKEFTPPVQAAY QKVVAGVANALAHKYH

Mouse β chain: I L I I VLAAHFSKDFTPECQAAWQKLVRVVAHALARKYH

-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)

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 x10^9 base pairs versus 2.9 x 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*

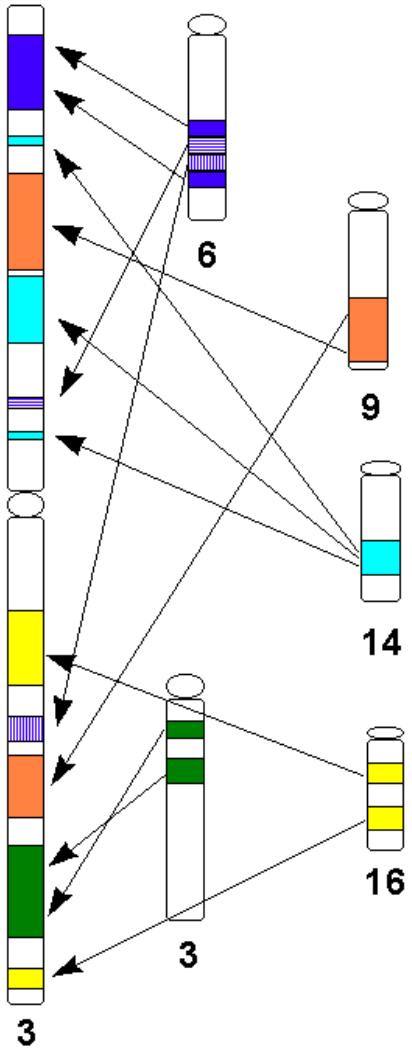
*Scientific American Dec. 5, 2002

Human and Mouse

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

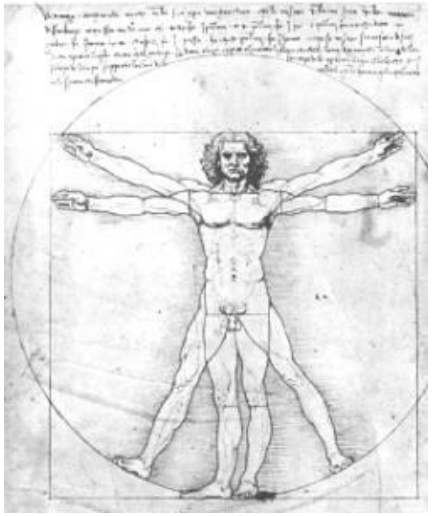
Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.



Comparative Genomics

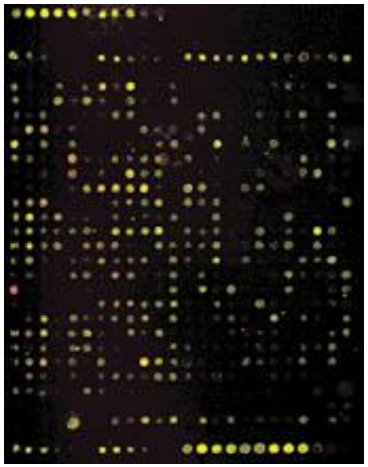
- What can be done with the full Human and Mouse Genome? One possibility is to create “knockout” mice – mice lacking one or more genes. Studying the phenotypes of these mice gives predictions about the function of that gene in both mice and humans.



Comparative Genomics

- By looking at the expression profiles of human and mouse (a recent technique using Gene Chips to detect mRNA as genes are being transcribed), the phenotypic differences can be attributed to genes and their expression.

A gene chip made by Affymetrix. The well can contain probes for thousands of genes.



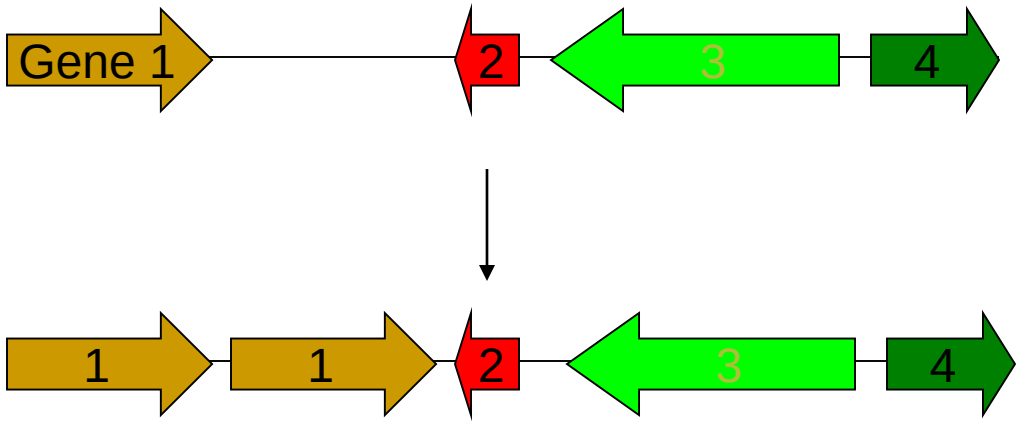
Imaging of a chip. The amount of fluorescence corresponds to the amount of a gene expressed.

Comparative Genome Sizes

- The genome of a protish 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

- 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

Cystic fibrosis

- Symptoms:
  - excessively salty sweat
  - The lungs, pancreas, and liver become clogged with thick mucus, which results in chronic infections and eventual malfunction
  - Mucus often builds up in the digestive tract, causing malnourishment
  - Patients often die from infections of the respiratory system.

Cystic Fibrosis

- In 1989, Francis Collins and Lap-Chee Tsui
  - identified the CF gene
  - characterized some of the mutation that cause this disease.
- A cDNA (complimentary DNA) library was prepared from mRNA isolated from sweat gland cells growing in culture and screened by colony hybridization
- CF gene product is similar to several ion channels protein,
  - which form pores between cells through which ions pass.
- Mutant CFTR protein does not function properly
  - salt accumulates in epithelial cells and mucus builds up on the surfaces of the cells.

Cystic Fibrosis

- Chromosome walking and jumping and complementary DNA hybridization were used to isolate DNA sequences, encompassing more than 500,000 base pairs, from the cystic fibrosis region on the long arm of human chromosome 7.
- neither gene therapy nor any other kind of treatment exists
- doctors can only ease the symptoms of CF
    1. antibiotic therapy combined with treatments to clear the thick mucus from the lungs.
    2. For patients whose disease is very advanced, lung transplantation may be an option.

Waardenburg's syndrome

- Genetic disorder
- Characterized by loss of hearing and pigmentary dysphasia
- Found on human chromosome 2

Waardenburg's syndrome

- A certain breed of mice (with splotch gene) that had similar symptoms caused by the same type of gene in humans
- Mice and Human genomes very similar → but easier to study mice
- Finding the gene in mice gives clues to where the same gene is located in humans
- Succeeded in identifying location of gene responsible for disorder in mice

Waardenburg's syndrome

- To locate where corresponding gene is in humans, we have to analyze the relative architecture of genes of humans and mouse
- About 245 genomic rearrangements
- Rearrangement operation in this case: reversals, translocation, fusion, and fission
- Reversal is where a block of genes is flipped within a genomic sequence

Section 10.3 Genome Rearrangements.

Turnip and Cabbage

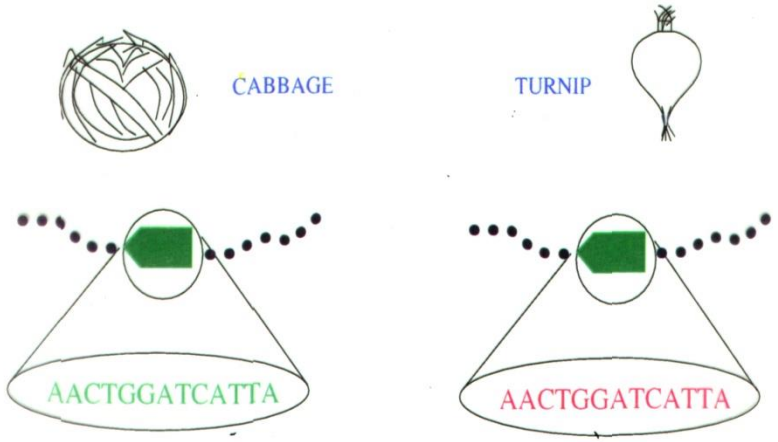
- Cabbages and turnips share a common ancestor

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

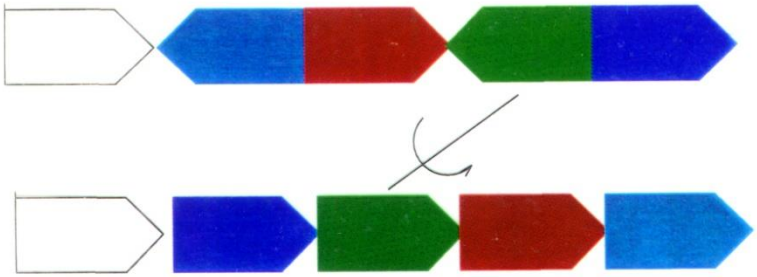
GENE SEQUENCE COMPARISON



AACTGGATCATTA
AACTGGATCATTA

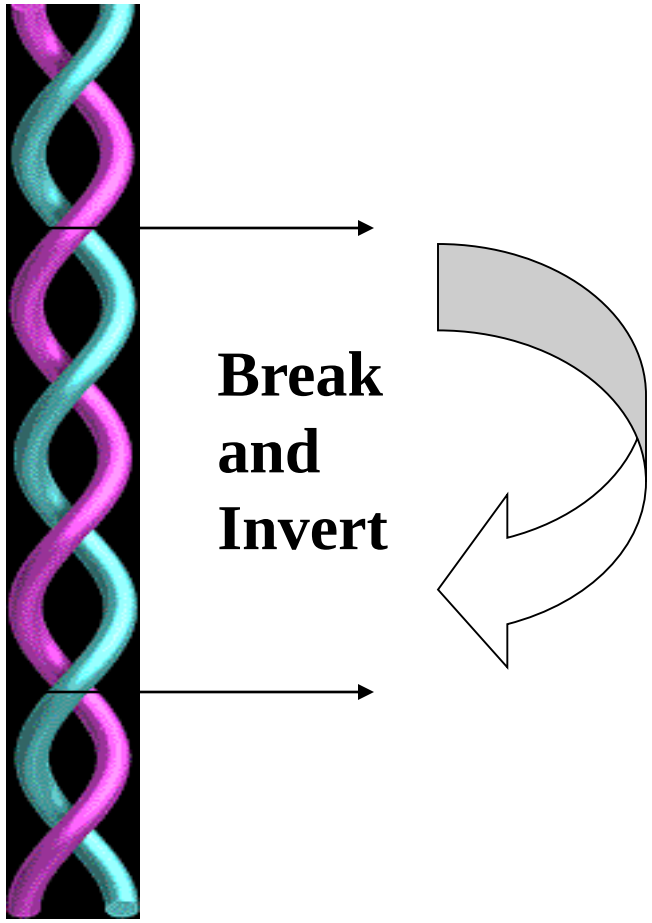
Comparing gene sequences yields no evolutionary information

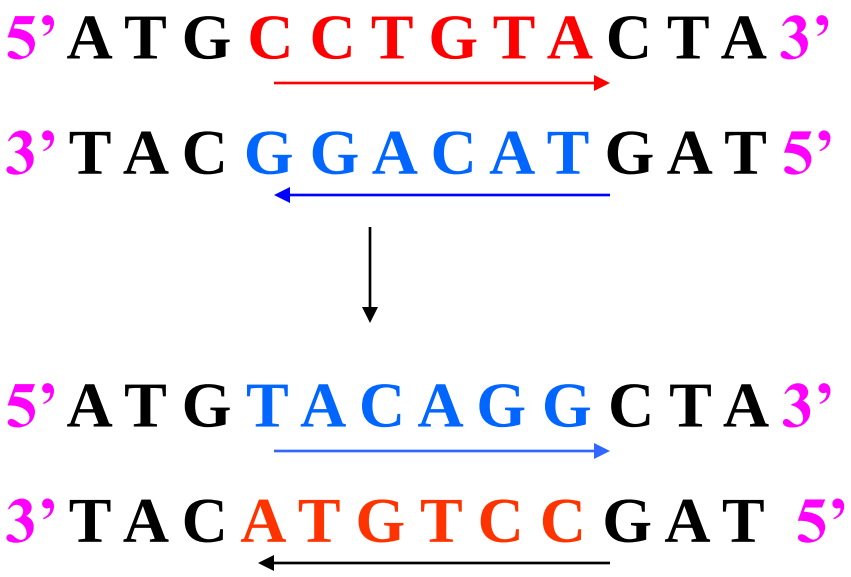
GENE ORDER COMPARISON



Evolution is manifested as the divergence in Gene Order

DNA Reversal





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.
    - N-P 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?

Outline For Section 11:

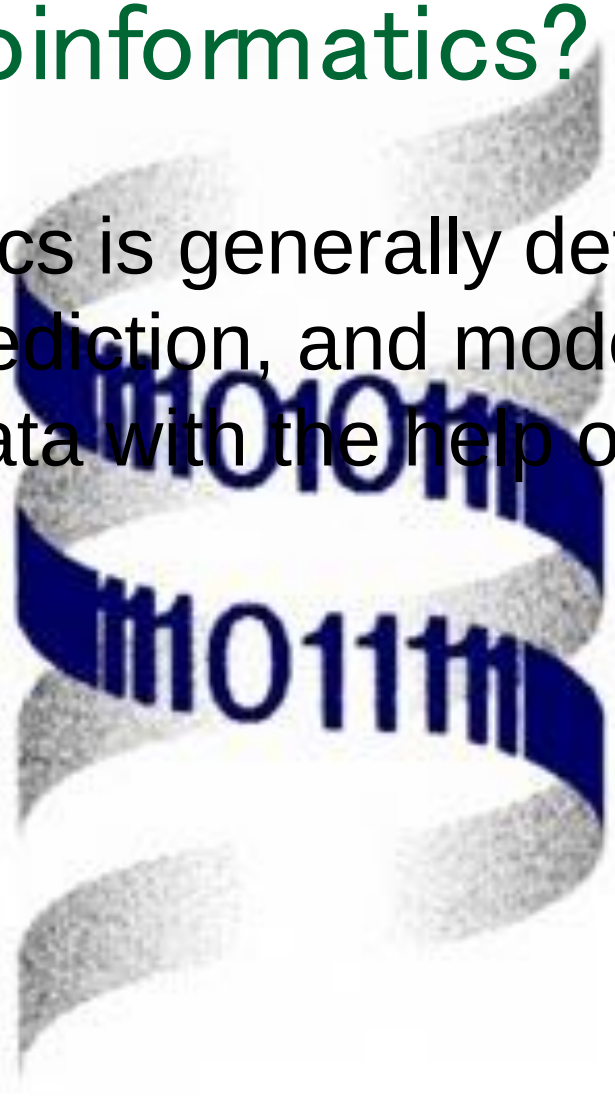
- *Sequence Driven Problems*
- *Human and Mouse*
- *Comparative Genomics*
- *Gene Mapping*
- *Cystic Fibrosis*

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
  - 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

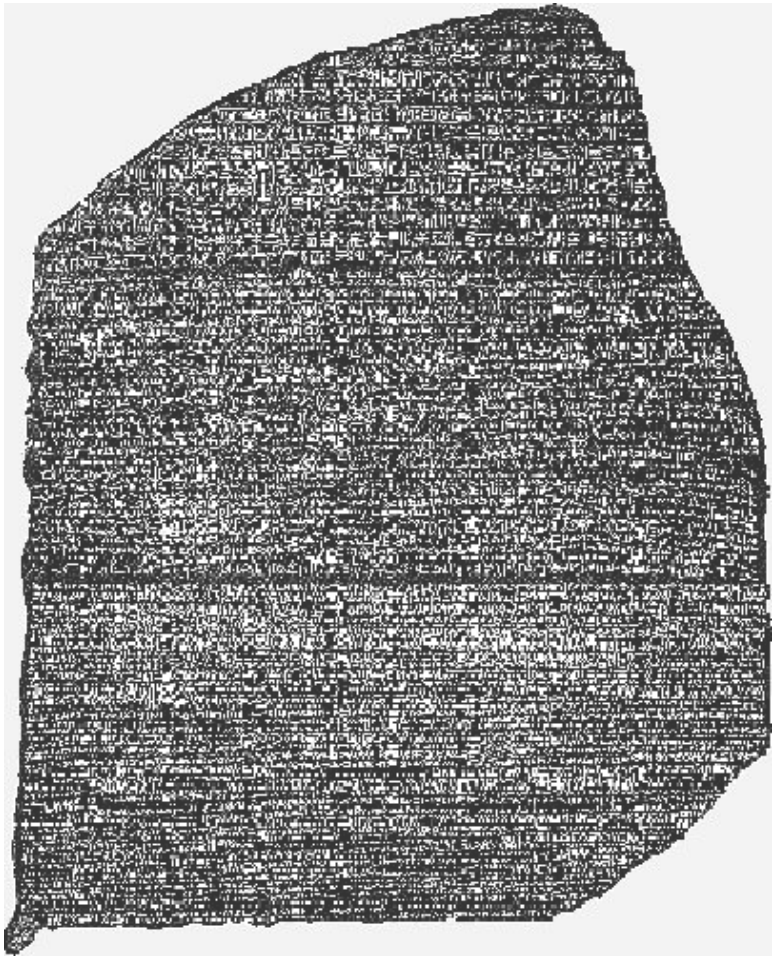
- Since discovering how DNA acts as the instructional blueprints behind life, biology has become an information science
- Now that many different organisms have been sequenced, we are able to find meaning in DNA through *comparative genomics,* not unlike comparative linguistics.
- Slowly, we are learning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

Sequence Information

- Many written languages consist of sequential symbols
- Just like human text, genomic sequences represent a language written in A, T, C, G
- Many DNA decoding techniques are not very different than those for decoding an ancient language

The *Rosetta Stone*

- The Rosetta Stone allowed linguists to solve the code of Egyptian Hieroglyphics
- The Greek language inscribed gave clues to what the Hieroglyphs meant.
- This is an example of *comparative linguistics*



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

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.

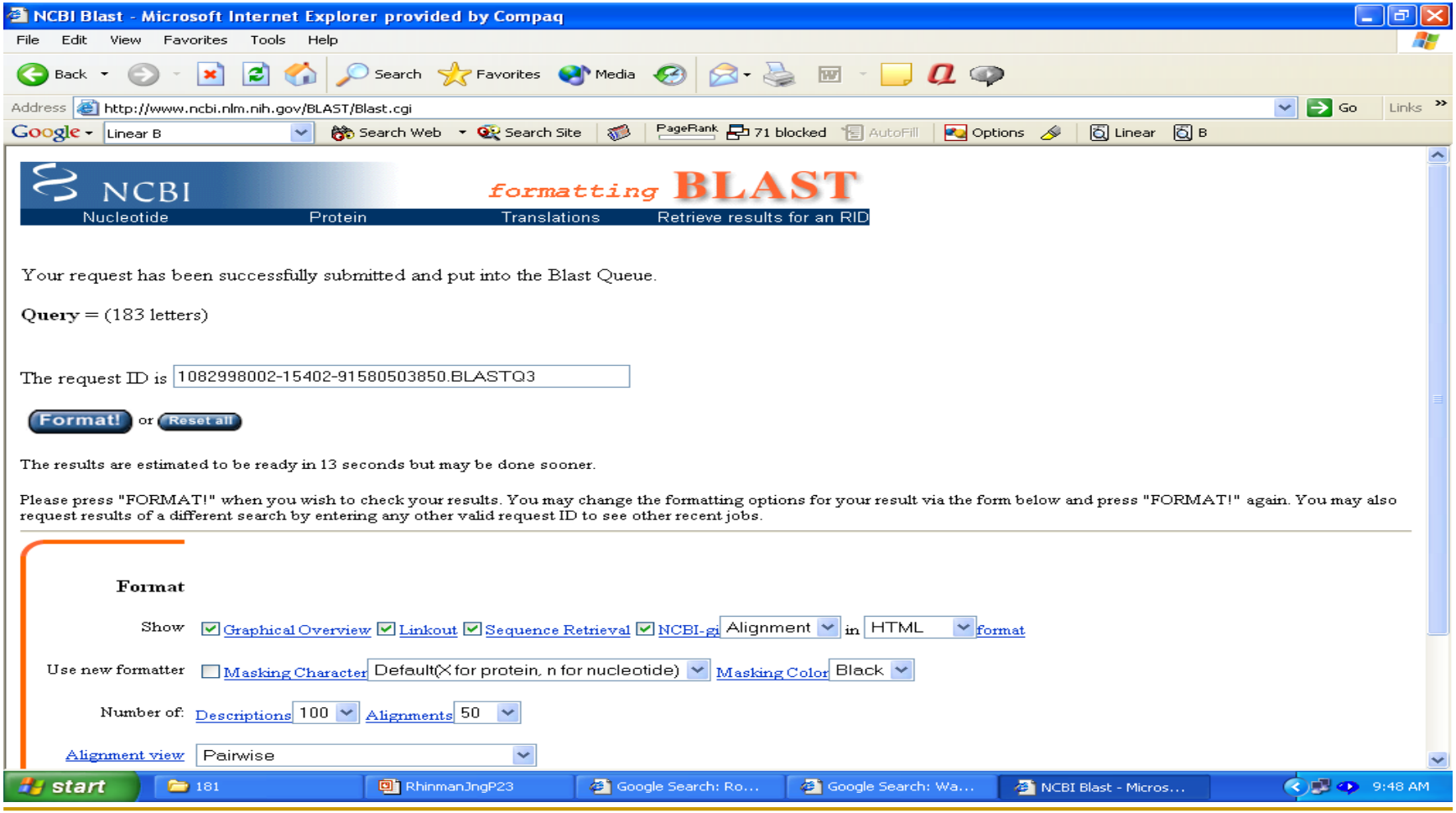
Two Quick Bioinformatics Applications

- 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.

BLAST


NCBI Blast - Microsoft Internet Explorer provided by Compaq
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Go Links
Google Linear B Search Web Search Site PageRank 71 blocked AutoFill Options Linear B
NCBI
formatting BLAST
Nucleotide Protein Translations Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = (183 letters)
The request ID is 1082998002-15402-91580503850.BLASTQ3
Format! or Reset all
The results are estimated to be ready in 13 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show Graphical Overview Linkout Sequence Retrieval NCBI-gi Alignment in HTML format
Use new formatter Masking Character Default(X for protein, n for nucleotide) Masking Color Black
Number of: Descriptions 100 Alignments 50
Alignment view Pairwise
start 181 RhinmanJngP23 Google Search: Ro... Google Search: Wa... NCBI Blast - Micros... 9:48 AM

Some Early Roles of **Bioinformatics**

- Sequence comparison
- Searches in sequence databases

Biological Sequence Comparison

- Needleman- Wunsch, 1970
  - Dynamic programming algorithm to align sequences

	A	D	C	N	S	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
S	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
S	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

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 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.

PROSITE


List of PROSITE documentation entries - Microsoft Internet Explorer provided by Compaq
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.expasy.org/cgi-bin/prosite-list.pl
Go Links
Google PROSITE Search Web Search Site PageRank 71 blocked AutoFill Options PROSITE
ExPASy Home page
Site Map
Search ExPASy
Contact us
PROSITE
Search PROSITE for Go Clear
PROSITE
Database of protein families and domains
Browse PROSITE documentation entries
Release 18.26, of 26-Apr-2004
[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] [Others]
- 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.
Example:
* PDOC00020 2 Kringle domain signature and profile
This documentation entry has been updated since the last release ('*'), its accession number is PDOC00020 and it describes two patterns.
start 181 RhinmanJngP23 Google Searc... Google Search... List of PROSIT... untitled - Paint 9:52 AM

Sequence Analysis

- Some algorithms analyze biological sequences for patterns
  - RNA splice sites
  - ORFs
  - 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.