Introduction to Bioinformatics

Learning objectives

- After studying these materials you should be able to do the following:
  - define the terms bioinformatics;
  - explain the scope of bioinformatics;
  - describe web-based versus command-line approaches to bioinformatics.

  - We seek to understand biological principles on a genome-wide scale using the tools of bioinformatics.

What is Bioinformatics?...

- A quick google search with the keyword *bioinformatics*
  - yields about **1.480.000** results (06.10.2008)
  - yields about **24.900.000** results (07.02.2016)
  - yields about **30.400.000** results (21. 02.2018)
- **Synonyms:**
  - Computational Biology
  - Computational Molecular Biology
  - Biocomputing

… What is Bioinformatics?...

- Bioinformatics
  - the study of how information is represented and transmitted in biological systems, starting at the molecular level
- According to a National Institutes of Health (NIH) definition, bioinformatics is
  - "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, analyze, or visualize such data."
    - The related discipline of computational biology is "the development and application of data-analytical and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological, behavioral, and social systems."

…What is Bioinformatics?...

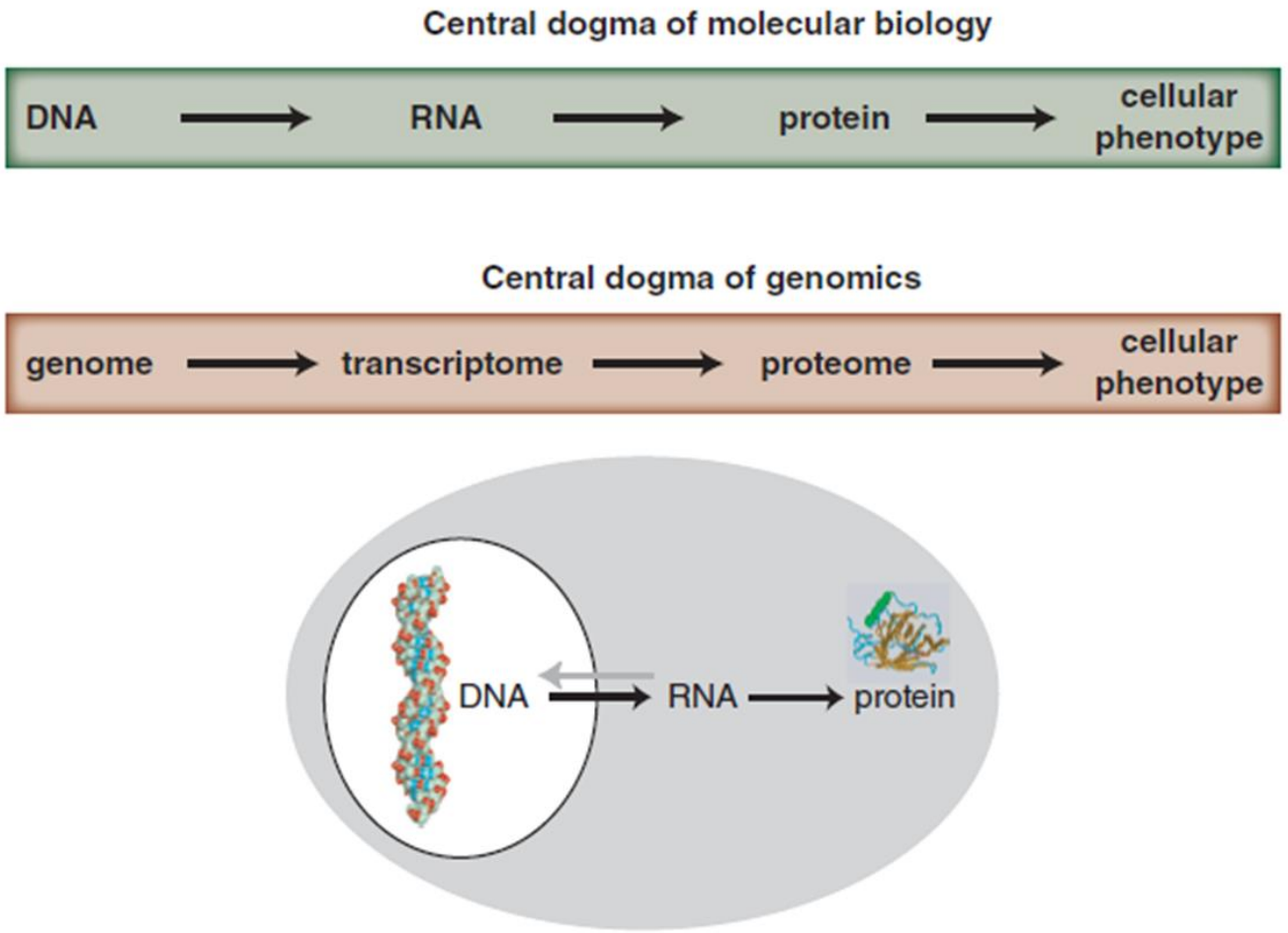
- **From Webopedia:**
  - The application of computer technology to the management of biological information.
  - Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research.
  - Bioinformatics is being used largely in the field of human genome research by the Human Genome Project that has been determining the sequence of the entire human genome (about 3 billion base pairs) and is essential in using genomic information to understand diseases.
  - It is also used largely for the identification of new molecular targets for drug discovery.

… What is Bioinformatics?...

- Another definition from the National Human Genome Research Institute (NHGRI) is that
  - "Bioinformatics is the branch of biology that is concerned with the acquisition, storage, display, and analysis of the information found in nucleic acid and protein sequence data."
- Russ Altman (1998) and Altman and Dugan (2003) offer two definitions of bioinformatics.
  - The first involves information flow following the central dogma of molecular biology (next slide)
  - The second definition involves information flow that is transferred based on scientific methods. This definition includes problems e.g.
    - designing, validating, and sharing software;
    - storing and sharing data;
    - performing reproducible research workflows;
    - interpreting experiments.

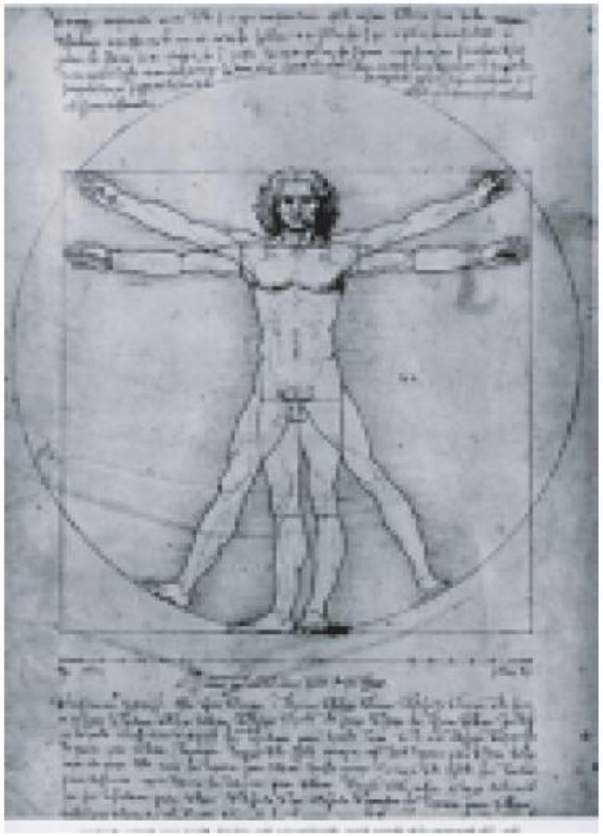
Altman, R.B. 1998. Bioinformatics in support of molecular medicine. *Proceedings of AMIA Symposium* **1998**, 53–61. PMID: 9929182.
Altman, R.B., Dugan, J.M. 2003. Defining bioinformatics and structural bioinformatics. *Methods of Biochemical Analysis* **44**, 3–14. PMID: 12647379.

… What is Bioinformatics?...



- A 1st perspective of the field of bioinformatics is the cell.
- Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data

… What is Bioinformatics?...

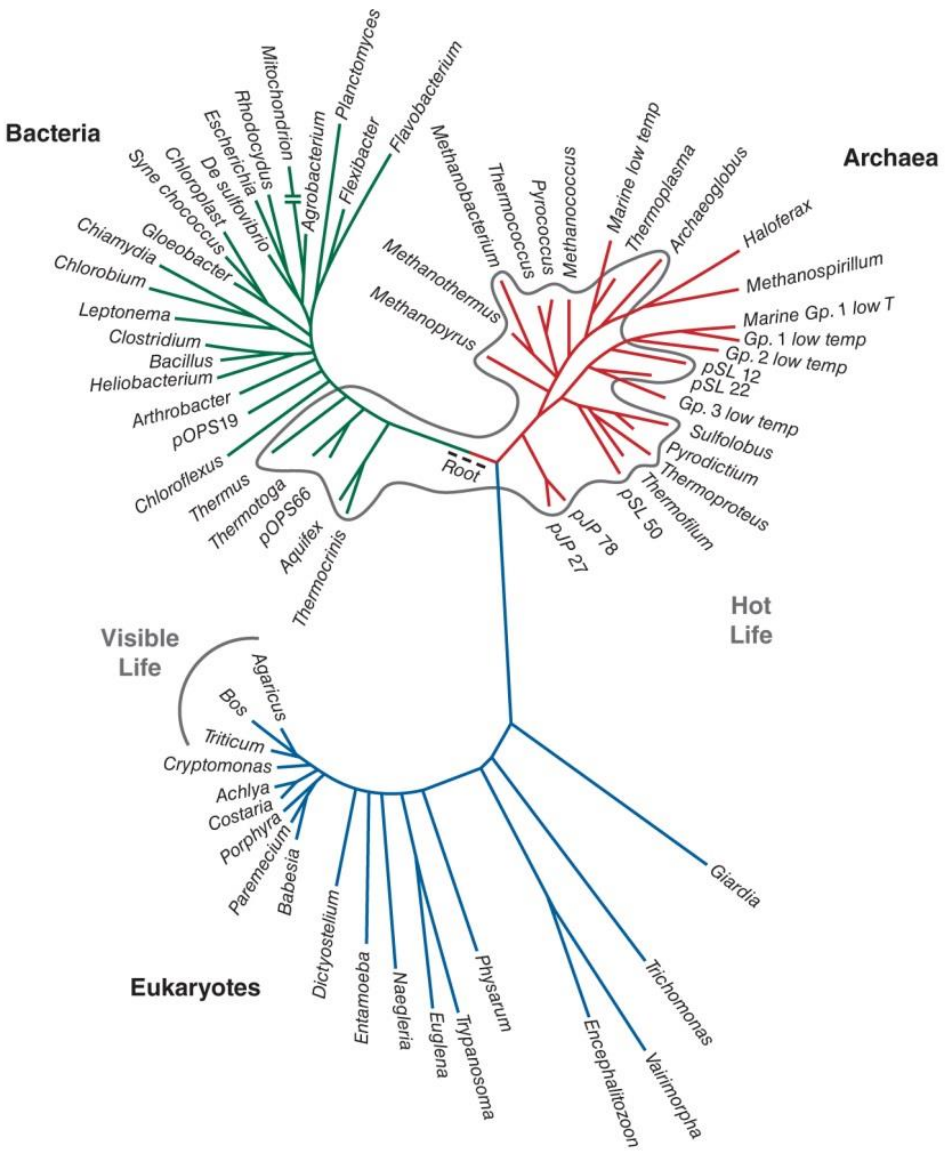


A 2nd perspective of bioinformatics is the organism.

Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products.

For an individual organism, bioinformatics tools can therefore be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.

… What is Bioinformatics?...



- A third perspective of the field of bioinformatics is represented by the tree of life.
- The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes.
- Viruses, which exist on the borderline of the definition of life, are not depicted here.
- For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome).
- We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth

…What is Bioinformatics?...

- From a practical sense, bioinformatics is a science that involves
  - collecting,
  - manipulating,
  - analyzing,
  - transmitting

  huge quantities of data,
- uses computers whenever appropriate.
- bioinformatics refers to computational bioinformatics.

Bioinformatics

- an interdisciplinary field that develops
  - methods and software tools for understanding biological data
- combines
  - computer science,
  - statistics,
  - mathematics,
  - engineering

  to analyze and interpret biological data

…What is Bioinformatics?...

- has been used for in silico analyses of biological queries using mathematical and statistical techniques.
  - [In silico (Latin for "in silicon") is an expression used to mean "performed on computer or via computer simulation.]
- primary goal is to increase the understanding of biological processes.
- focuses on developing and applying computationally intensive techniques to achieve this goal.

…What is Bioinformatics?...

- Techniques used include
  - pattern recognition,
  - data mining,
  - machine learning algorithms,
  - visualization

…What is Bioinformatics?...

- Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from
  - graph theory,
  - artificial intelligence,
  - soft computing,
  - data mining,
  - signal processing,
  - image processing,
  - computer simulation.

…What is Bioinformatics?...

- The algorithms in turn depend on theoretical foundations such as
  - discrete mathematics
  - control theory
  - system theory
  - information theory
  - statistics

...What is Bioinformatics?...

- Bioinformatics derives knowledge from computer analysis of biological data.
  - These can consist of the information
    - stored in the genetic code,
    - experimental results from various sources,
    - patient statistics,
    - and scientific literature.
- Research in bioinformatics includes method development for
  - storage,
  - retrieval,
  - analysis

  of the data.

Path to the Bioinformatics

- 1st,
  - Learn Biology.
- 2nd,
  - Decide and pick a problem that interests you for experiment.
- 3rd,
  - Find and learn about the Bioinformatics tools.
- 4th,
  - Learn the Computer Programming Languages.
    - Perl, Pyton, R, Java, etc.
- 5th,
  - Experiment on your computer and learn different programming techniques.

Why is Bioinformatics Important?

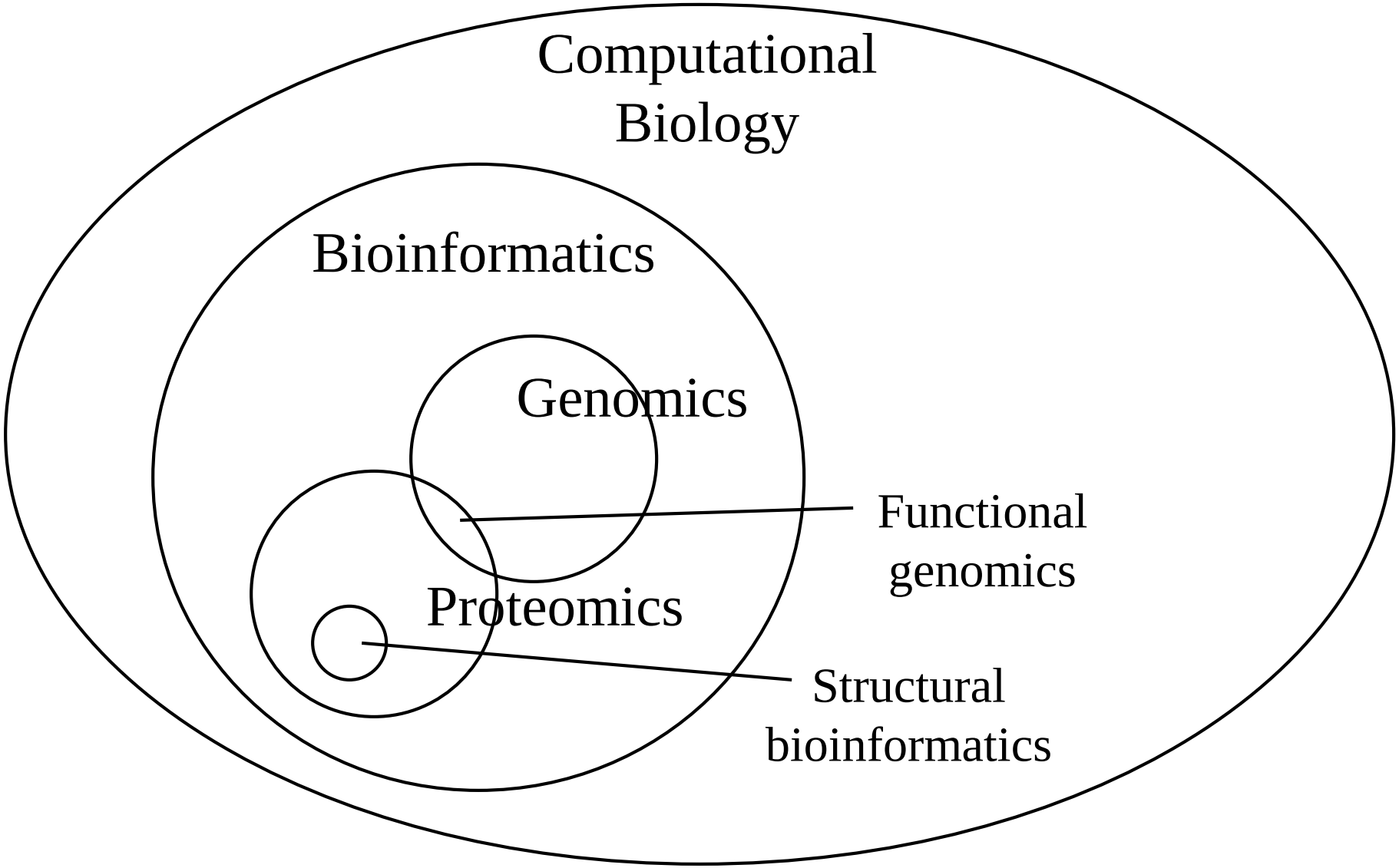
- Applications areas include
  - Medicine
  - Pharmaceutical drug design
  - Toxicology
  - Molecular evolution
  - Biosensors
  - Biomaterials
  - Biological computing models
  - DNA computing

What skills are needed?

- Well-grounded in one of the following areas:
  - Computer science
  - Molecular biology
  - Statistics

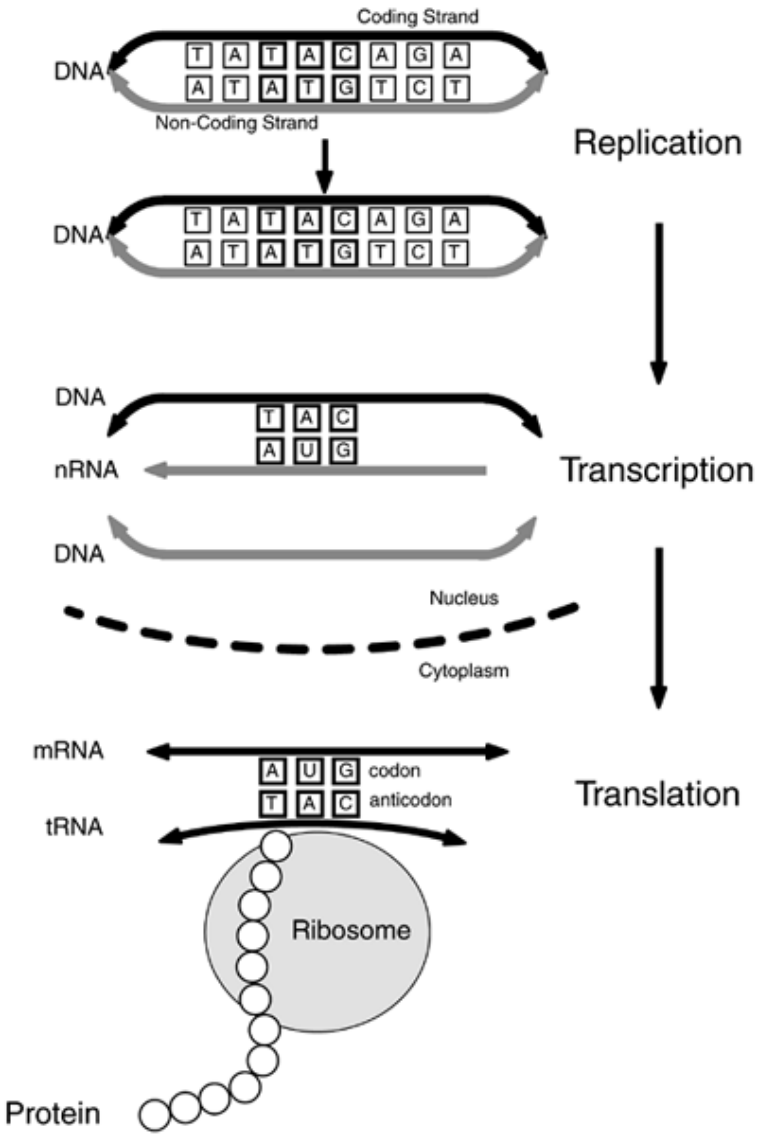
- Working knowledge and appreciation in the others!

Scope of Computational Biology



The Central Dogma of Molecular Biology

- DNA is transcribed to messenger RNA in the cell nucleus, which is in turn translated to protein in the cytoplasm.
- The Central Dogma, shown here from a structural perspective, can also be depicted from an information flow perspective



Genomics

- The study of the genome,
  - which is the complete set of the genetic material or DNA present in an organism.
- studies all genes and their inter relationships in an organism, so as to identify their combined influence on its growth and development.
- The field of genomics attracted worldwide attention in the late 1990s with the race to map the human genome.
  - The Human Genome Project (HGP), completed in April 2003, made available for the first time the complete genetic blueprint of a human being.

Proteomics

- large-scale study of proteomes,
  - which is a set of proteins produced in an organism, system, or biological context.
    - We may refer to, for instance, the proteome of a species (eg, Homo sapiens) or an organ (eg, the liver).
  - The proteome is not constant;
    - it differs from cell to cell and changes over time.
  - To some degree, the proteome reflects the underlying transcriptome.
    - However, protein activity (often assessed by the reaction rate of the processes in which the protein is involved) is also modulated by many factors in addition to the expression level of the relevant gene.

Proteomics

- is used to investigate:
  - when and where proteins are expressed;
  - rates of protein production, degradation, and steady-state abundance;
  - how proteins are modified (for example, post-translational modifications (PTMs) such as phosphorylation);
  - the movement of proteins between subcellular compartments;
  - the involvement of proteins in metabolic pathways;
  - how proteins interact with one another.

Proteomics

- can provide significant biological information for many biological problems, such as:
  - Which proteins interact with a particular protein of interest (for example, the tumor suppressor protein p53)?
  - Which proteins are localized to a subcellular compartment (for example, the mitochondrion)?
  - Which proteins are involved in a biological process (for example, circadian rhythm)?

Structural bioinformatics/genomics

- is the branch of bioinformatics
  - which is related to the analysis and prediction of the three-dimensional structure of biological macromolecules such as
    - proteins,
    - RNA,
    - DNA.

Structural bioinformatics/genomics

- deals with generalizations about macromolecular 3D structure such as
  - comparisons of overall folds and local motifs,
  - principles of molecular folding, evolution, and binding interactions,
  - structure/function relationships, working both from experimentally solved structures and from computational models.

Functional genomics

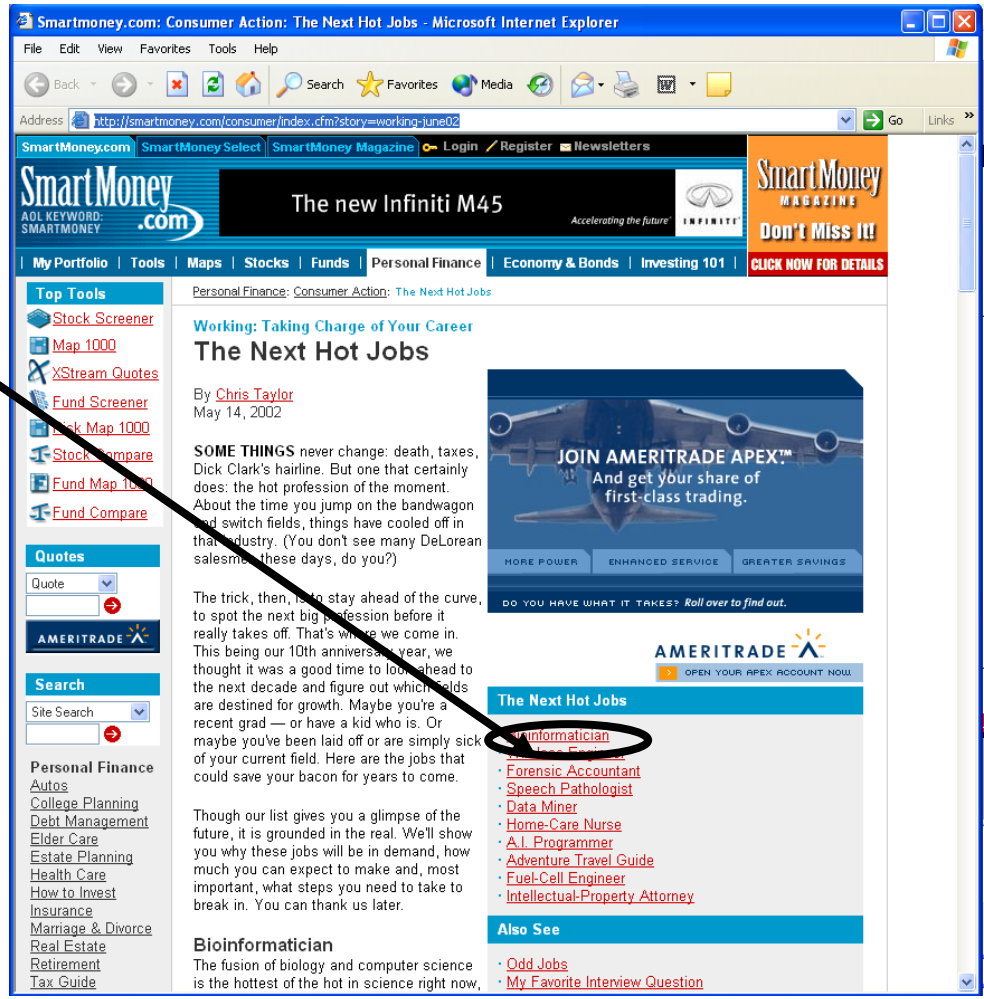
- is a field of molecular biology,
  - which attempts to make use of the vast wealth of data given by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing)
    - to describe gene (and protein) functions and interactions.

Functional genomics

- focuses on the dynamic aspects such as
  - gene transcription, translation,
  - regulation of gene expression
  - protein–protein interactions.
- attempts to answer questions about the function of DNA at the levels of genes, RNA transcripts, and protein products.

Why should I care?

- SmartMoney ranks Bioinformatics as #1 among next HotJobs
- Business Week 50 Masters of Innovation
- Jobs available, exciting research potential
- Important information waiting to be decoded!



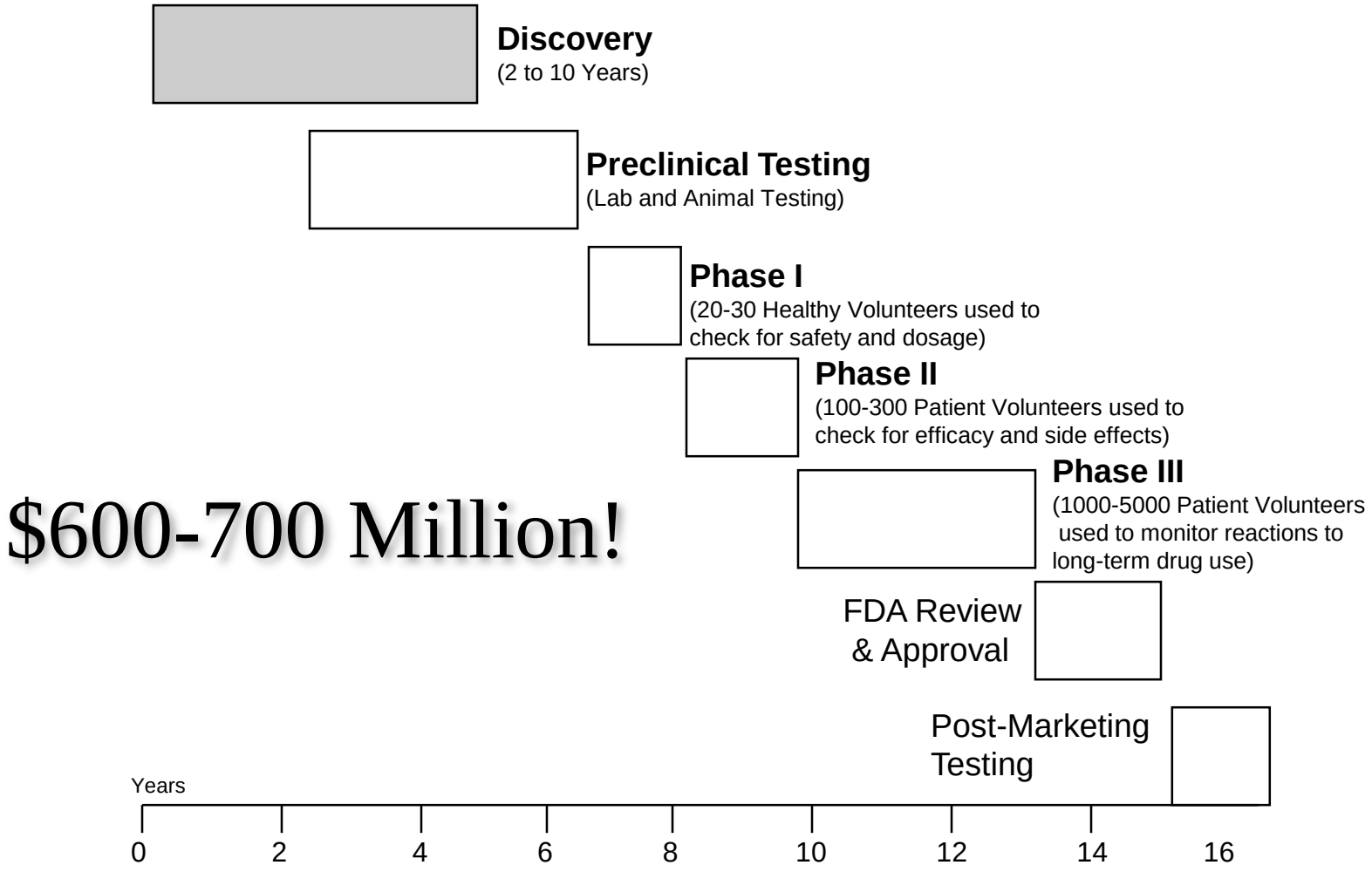
Why is bioinformatics hot?

- Supply/demand: few people adequately trained in both biology and computer science
- Genome sequencing, microarrays, RNA-sequencing, single cell-sequencing, etc lead to large amounts of data to be analyzed
- Leads to important discoveries
- Saves time and money

Fighting Human Disease

- Genetic / Inherited
  - Diabetes
- Viral
  - Flu, common cold
- Bacterial
  - Meningitis, Strep throat

Drug Development Life Cycle

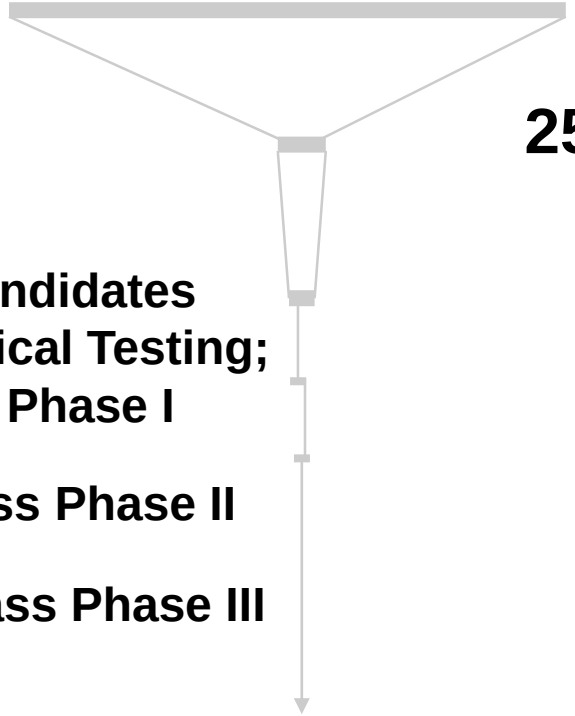


$600-700 Million!

7 – 15 Years!

Drug lead screening

5,000 to 10,000 compounds screened



250 Lead Candidates in Preclinical Testing

5 Drug Candidates enter Clinical Testing; 80% Pass Phase I

30%Pass Phase II

80% Pass Phase III

One drug approved by the FDA

Killer application

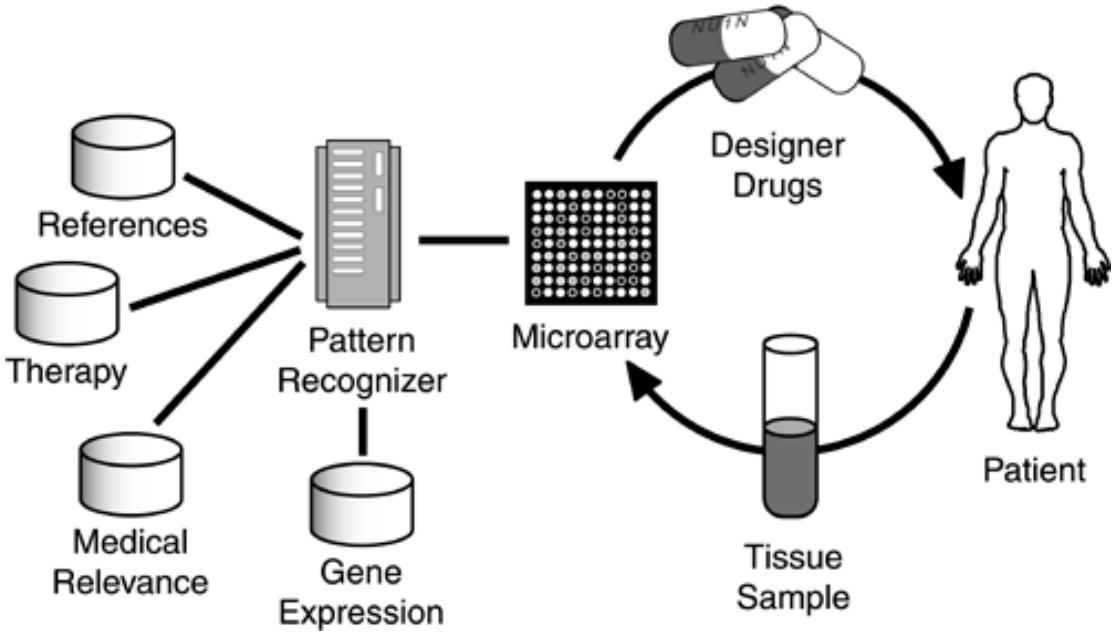
- In the biotechnology industry, every researcher and entrepreneur hopes to develop or discover the next "killer application”
  - the one application that will bring the world to his or her door and provide funding for R&D, marketing, and production.

Killer application

- For example, in general computing, the electronic spreadsheet and the desktop laser printer have been the notable killer apps.
  - The spreadsheet not only transformed the work of accountants, research scientists, and statisticians, but the underlying tools formed the basis for visualization and mathematical modeling.
  - The affordable desktop laser printer created an industry and elevated the standards of scientific communications, replacing rough graphs created on dot-matrix printers with high-resolution images.

Killer application

- "What might be the computer-enabled 'killer app' in bioinformatics?“
- Although there are numerous military and agricultural opportunities, one of the most commonly cited examples of the killer app is in personalized medicine, as illustrated in Figure



personalized medicine: the custom, just-in-time delivery of medications (popularly called "designer drugs") tailored to the patient‘s condition.

Killer application

- Instead of taking a generic or over-the-counter drug for a particular condition, a patient would submit a tissue sample, such as a mouth scraping, and submit it for analysis.
  - A microarray would then be used to analyze the patient's genome and the appropriate compounds would be prescribed.
- The drug could be a cocktail of existing compounds, much like the drug cocktails used to treat cancer patients today.

Killer application

- Alternatively, the drug could be synthesized for the patient's specific genetic markers—as in tumor specific chemotherapy, for example.
  - This synthesized drug might take a day or two to develop, unlike the virtually instantaneous drug cocktail.
  - The tradeoff is that the drug would be tailored to the patient's genetic profile and condition, resulting in maximum response to the drug, with few or no side effects.

Killer application

- How will this or any other killer app be realized?
  - The answer lies in addressing the molecular biology, computational, and practical business aspects of proposed developments such as custom medications.
- A practical system would include:
  - High throughput screening
    - The use of affordable, computer-enabled microarray technology to determine the patient's genetic profile.
      - The issue here is affordability, in that microarrays costs tens of thousands of dollars

Killer application

– Medically relevant information gathering

- Databases on gene expression, medical relevance of signs and symptoms, optimum therapy for given diseases, and references for the patient and clinician must be readily available.
- The goal is to be able to quickly and automatically match a patient's genetic profile, predisposition for specific diseases, and current condition with the efficacy and potential side effects of specific drug-therapy options.

Killer application

– Custom drug synthesis

- The just-in-time synthesis of patient-specific drugs, based on the patient's medical condition and genetic profile, presents major technical as well as political, social, and legal hurdles.
- For example, for just-in-time synthesis to be accepted by the FDA, the pharmaceutical industry must demonstrate that custom drugs can skip the clinical-trials gauntlet before approval.

Killer application

- Achieving this killer app in biotech is highly dependent on
  - computer technology,
    - especially in the use of computers to speed the process testing-analysis-drug synthesis cycle, where time really is money.
- For example, consider that for every 5,000 compounds evaluated annually by the U.S. pharmaceutical R&D laboratories, 5 make it to human testing, and only 1 of the compounds makes it to market.

Killer application

- In addition, the average time to market for a drug is over 12 years,
  - including several years of pre-clinical trials followed by a 3-phase clinical trial.
- These clinical trials progress from
  - safety and dosage studies in Phase I,
  - to effectiveness and side effects in Phase II,
  - to long-term surveillance in Phase III,

  with each phase typically lasting several years.

Killer application

- Most pharmaceutical companies view computerization as the solution to creating smaller runs of drugs focused on custom production.
- Obvious computing applications range from
  - predicting efficacy and side effects of drugs based on genome analysis,
  - to visualizing protein structures to better understand and predict the efficacy of specific drugs,
  - to illustrating the relative efficacy of competing drugs in terms of quality of life and cost, based on the Markov simulation of likely outcomes during Phase IV clinical trials.

Bioinformatics Software: Two Cultures

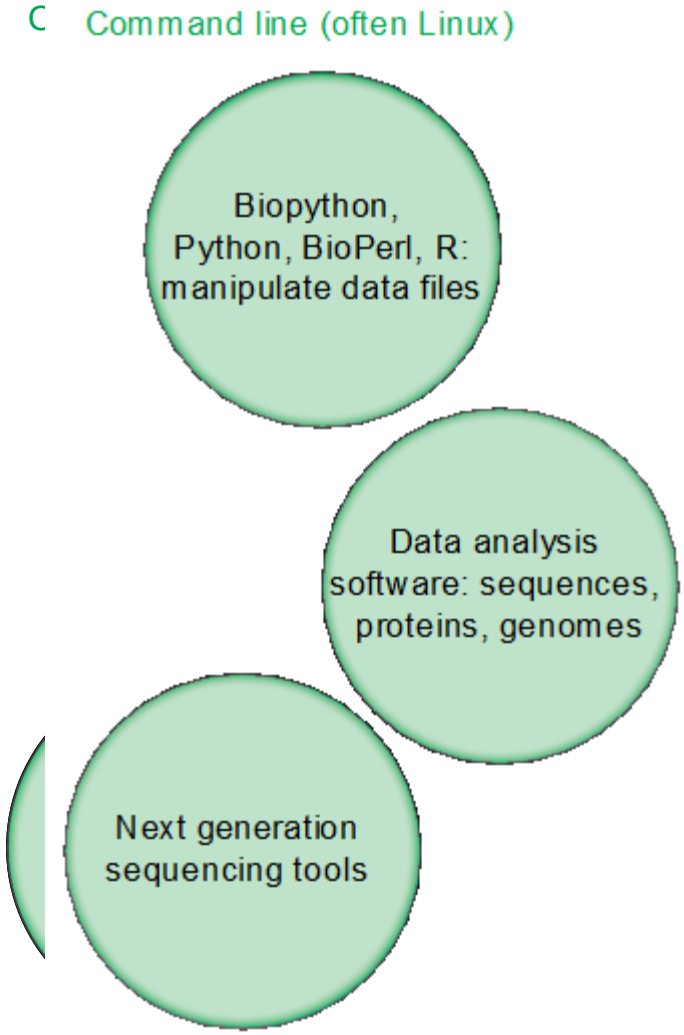
- Many bioinformatics tools and resources are available on the internet, such as major genome browsers and major portals (NCBI, Ensembl, UCSC).
- These are:
  - accessible (requiring no programming expertise)
  - easy to browse to explore their depth and breadth
  - very popular
  - familiar (available on any web browser on any platform)

Bioinformatics Software: Two Cultures

- Many bioinformatics tools and resources are available on the command-line interface (sometimes abbreviated CLI).
  - These are often on the Linux platform (or other Unix-like platforms such as the Mac command line).
  - They are essential for many bioinformatics and genomics applications.
  - Most bioinformatics software is written for the Linux platform.
    - Many bioinformatics datasets are so large (e.g. high throughput technologies generate millions to billions or even trillions of data points) requiring command-line tools to manipulate the data.

CLI

- Should you learn to use the Linux operating system?
  - Yes, if you want to use mainstream bioinformatics tools.
- Should you learn Python or Perl or R or another programming language?
  - It's a good idea if you want to go deeper into bioinformatics, but also, it depends what your goals are.
  - Many software tools can be run in Linux on the command-line without needing to program.
- Think of this figure like a map.
  - Where are you now?
  - Where do you want to go?



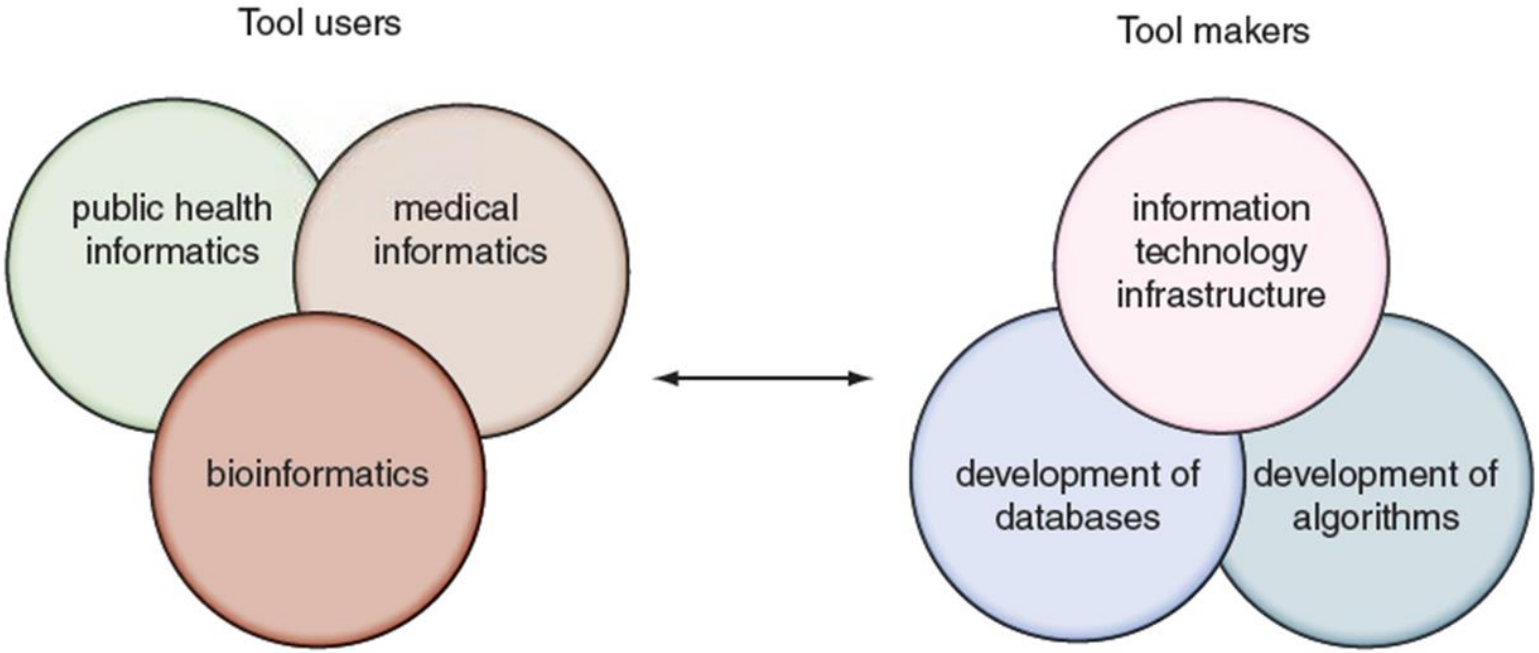
Some web-based (GUI) and command-line (CLI) software

Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	MrBayes
Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
RNA	RNAfam, tRNAscan	

Some web-based (GUI) and command-line (CLI) software

RNAseq	Galaxy	affy (R package), RSEM
Proteomics	ExPASy	pepstats (EMBOSS)
Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
Functional genomics	FLink, Cytoscape	
Tree of life		Velvet (assembly)
Viruses		MUMmer (alignment)
Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
Fungi	YGOB	Ensembl (variants)
Eukaryotic genomes		
Human genome		PLINK
Human disease	OMIM, BioMart	EDirect, MitoSeek

Tool makers and tool users across informatics disciplines



- Many informatics disciplines have emerged in recent years.
- Bioinformatics is distinguished by its particular focus on DNA and proteins (impacting its databases, its tools, and its entire culture).

Learning Programming for Bioinformatics

- In addition to available books and courses, many websites offer online training in the forms of tutorials or courses.
- Rules for online learning include:
  - make a plan;
  - be selective;
  - organize your learning environment;
  - do the readings;
  - do the exercises;
  - do the assessments;
  - exploit the advantages (e.g., convenience);
  - reach out to others;
  - document your achievements;
  - be realistic about your expectations for what you can learn.
- These rules also apply to reading a textbook.

https://doi.org/10.1371/journal.pcbi.1002631
https://doi.org/10.1371/journal.pcbi.1000589

Reproducible Research in Bioinformatics

- Science by its nature is cumulative and progressive.
- Whether you use web-based or command-line tools, research should be conducted in a way that is reproducible by the investigator and by others.
- This facilitates the cumulative, progressive nature of your work.
- In the realm of bioinformatics this means the following.

Reproducible Research in Bioinformatics

- A workflow should be well documented.
  - This may include keeping text documents on your computer in which you can copy and paste complex commands, URLs, or other forms of data.
- To facilitate your work, information stored on a computer should be well organized.
- Data should be made available to others.
  - Repositories are available to store high-throughput data in particular.
    - Examples are Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) at NCBI and ArrayExpress and European Nucleotide Archive (ENA) at EBI.
- Metadata can be equally as crucial as data.
  - Metadata refers to information about datasets.
    - For a bacterial genome that has been sequenced, the metadata may include the location from which the bacterium was isolated, the culture conditions, and whether it is pathogenic.

Reproducible Research in Bioinformatics

- Databases that are used should be documented.
  - Since the contents of databases change over time, it is important to document the version number and the date(s) of access.
- Software should be documented.
  - For established packages, the version number should be provided.
    - Further documenting the specific steps you use allows others to independently repeat your analyses.
      - In an effort to share software, many researchers use repositories such as GitHub.

https://doi.org/10.1371/journal.pcbi.1000424

Where Can I Learn More?

- ISCB: http://www.iscb.org/
- NBCI: http://ncbi.nlm.nih.gov/
- http://www.bioinformatics.org/
- Books
- Journals
- Conferences

Where Can I Learn More?

- https://www.codeschool.com/
- https://www.codecademy.com/
- https://www.datacamp.com/
- https://software-carpentry.org/
- https://github.com/

Introduction to Molecular Biology

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

Outline Cont.

- 8. How Can We Analyze DNA
  - 1. Copying DNA
  - 2. Cutting and Pasting DNA
  - 3. Measuring DNA Length
  - 4. Probing DNA
- 9. How Do Individuals of a Species Differ
- 10. How Do Different Species Differ
  - 1. Molecular Evolution
  - 2. Comparative Genomics
  - 3. Genome Rearrangement
- 11. 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

- 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)

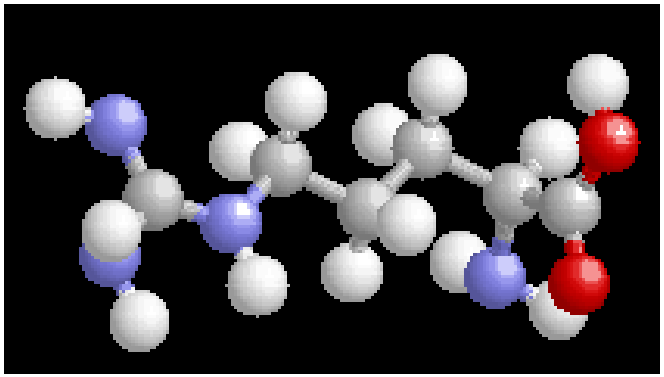
Mendel: The Father of Genetics

- **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

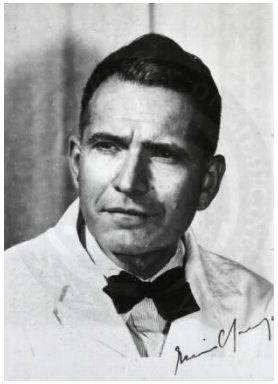
Major events in the history of Molecular Biology 1940 - 1950

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

George Beadle

Edward Tatum

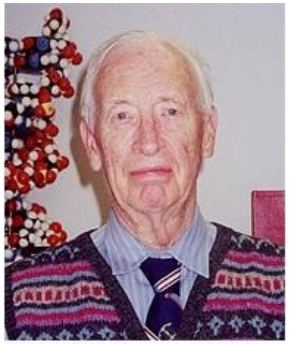
- **1950** – Edwin Chargaff find Cytosine complements Guanine and Adenine complements Thymine



Edwin Chargaff

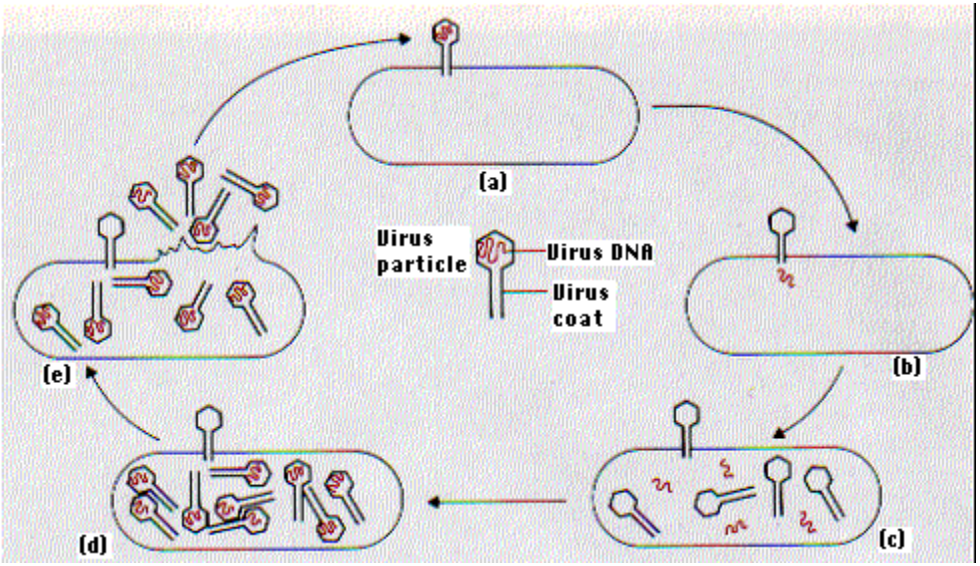
Major events in the history of Molecular Biology 1950 - 1952

- **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

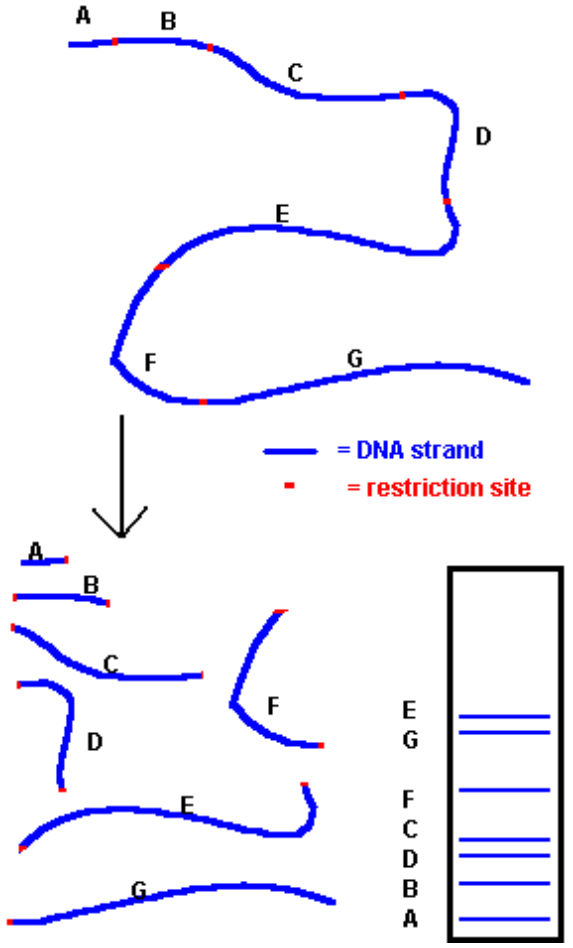
- **1956** George Emil Palade showed the site of enzymes manufacturing in the cytoplasm is made on RNA organelles called ribosomes.

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

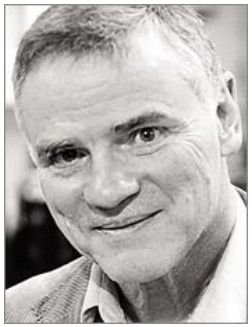
Richard Roberts

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

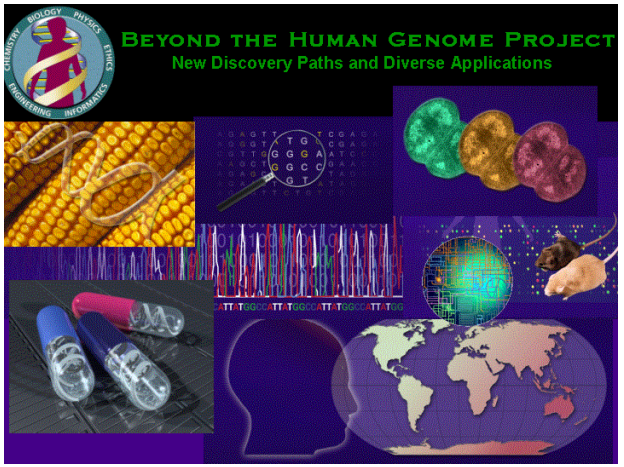
Joan Steitz

Major events in the history of Molecular Biology 1986 - 1995

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

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

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

John Craig Venter

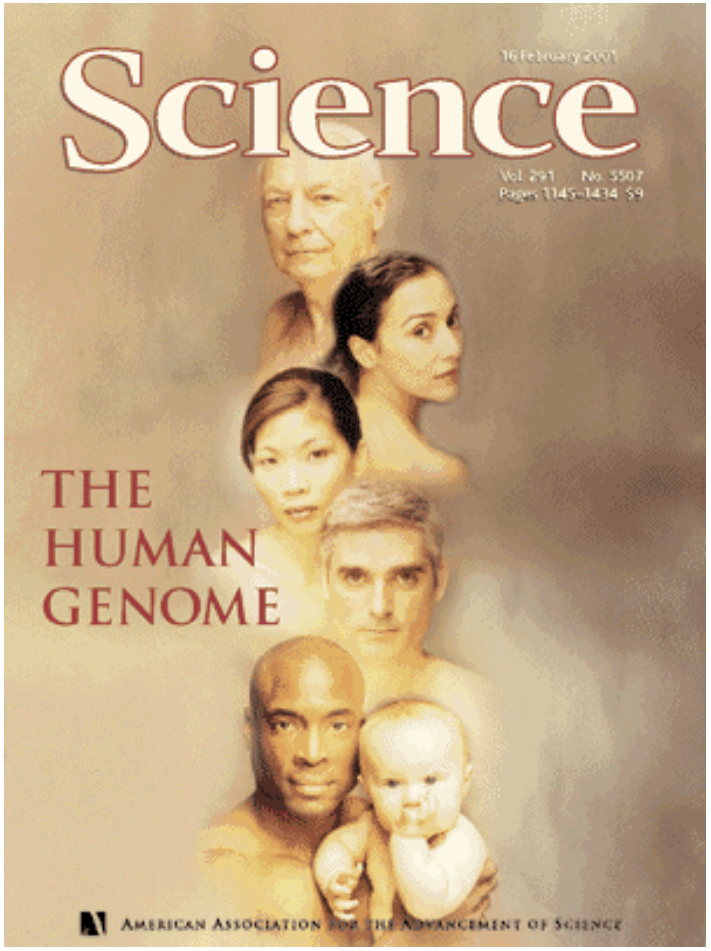
Major events in the history of Molecular Biology 1997 - 1999

- **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

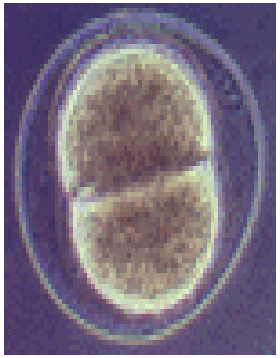
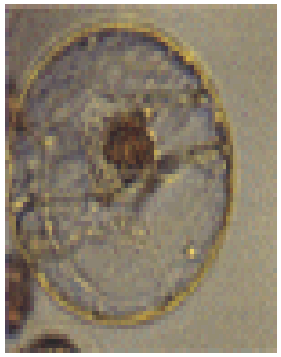
- **2001** International Human Genome Sequencing:first draft of the sequence of the human genome published

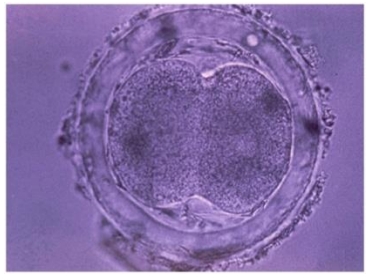
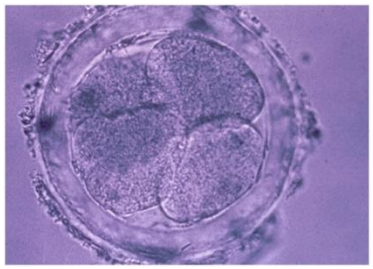
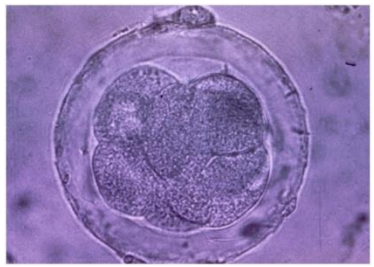


Major events in the history of Molecular Biology 2003-Present

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

What is Life made of?

Outline

- All living things are made of Cells
  - *Prokaryote, Eukaryote*
- *Cell Signaling*
- *What is Inside the cell: From DNA, to RNA, to Proteins*

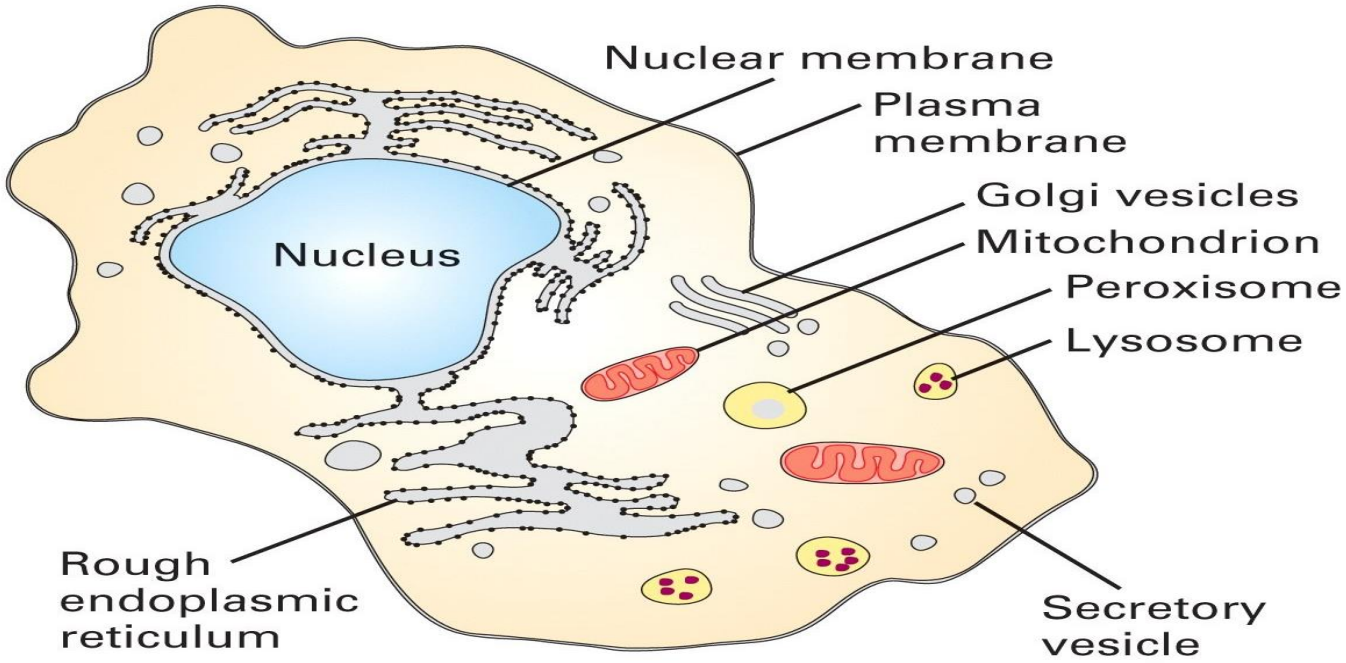
Cells

- **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 extant prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.

Cells

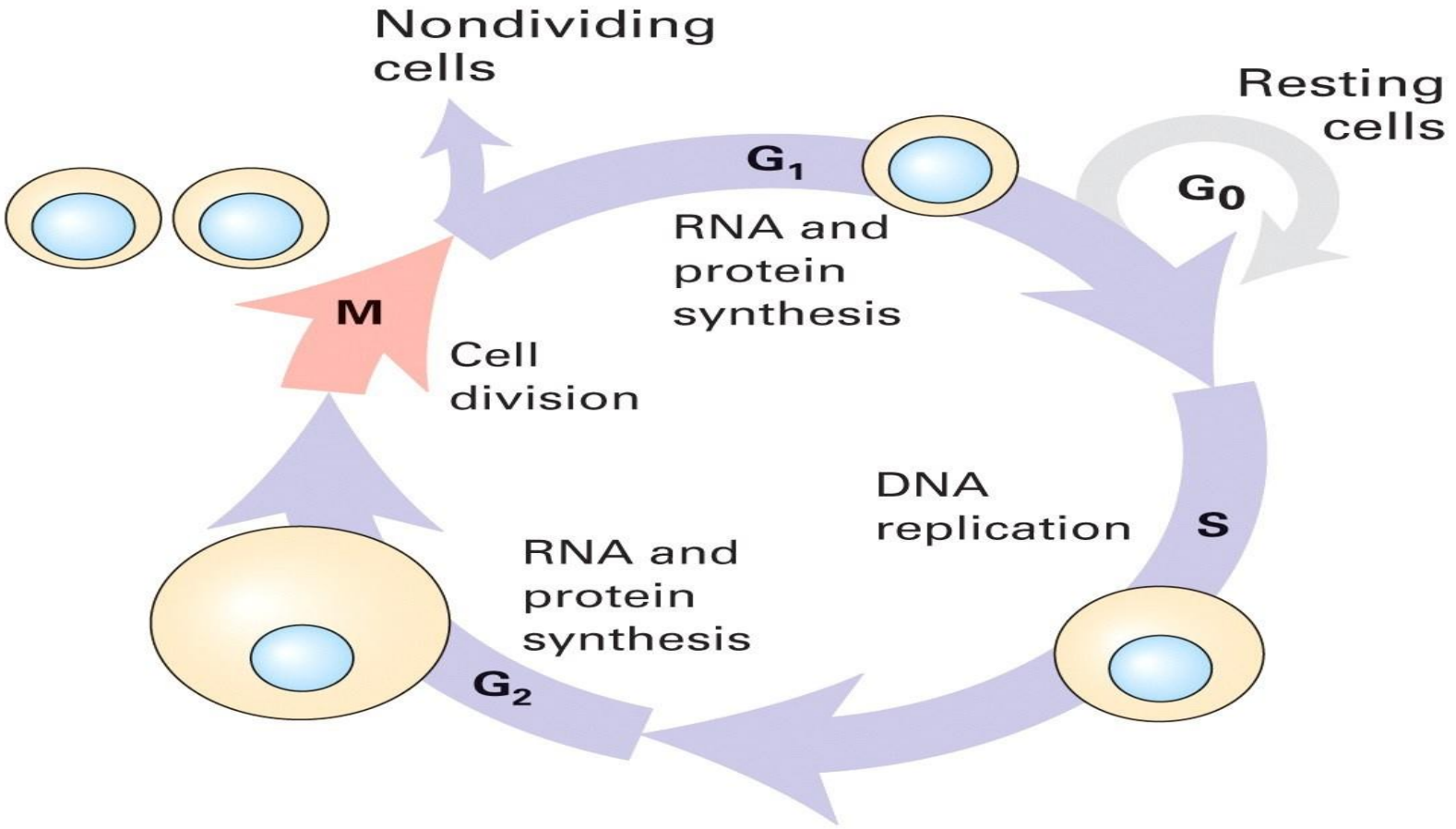
- *Chemical composition*-by weight
  - 70% water
  - 7% small molecules
    - salts
    - Lipids
    - amino acids
    - nucleotides
  - 23% macromolecules
    - Proteins
    - Polysaccharides
    - lipids
- *biochemical (metabolic) pathways*
- *translation of mRNA into proteins*

Life begins with Cell



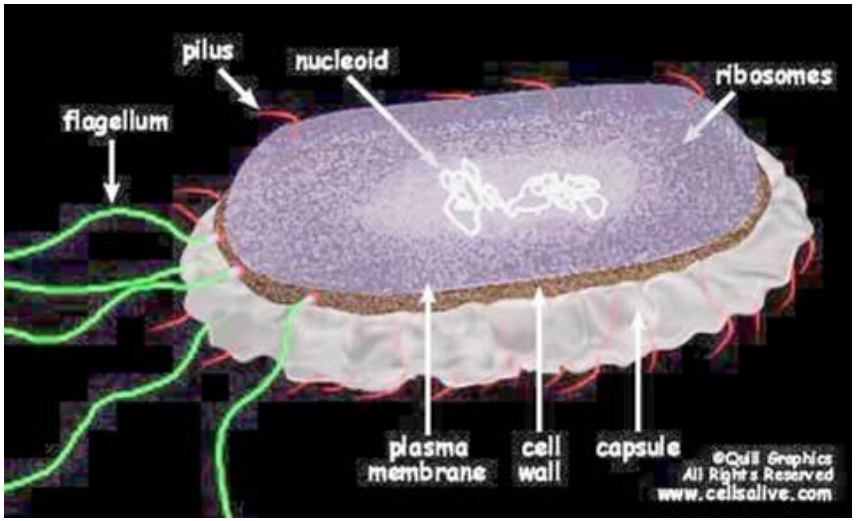
- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features

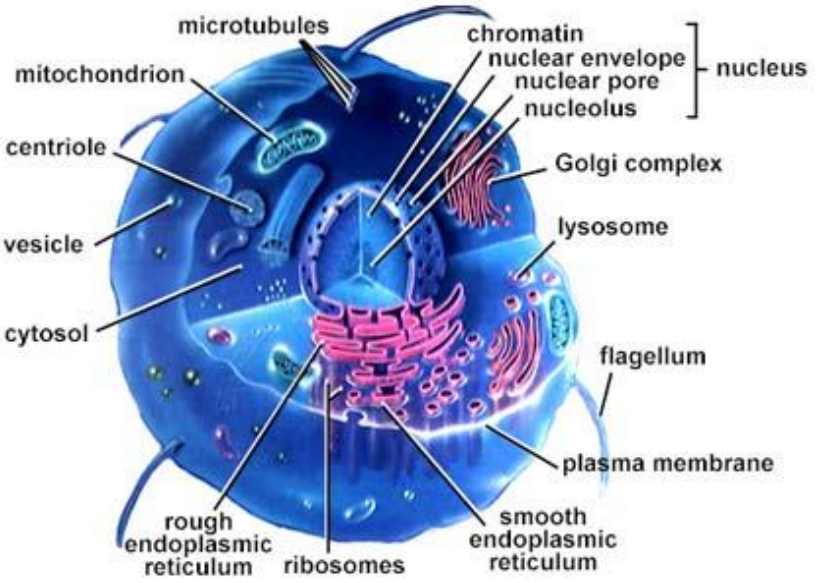
All Cells have common Cycles



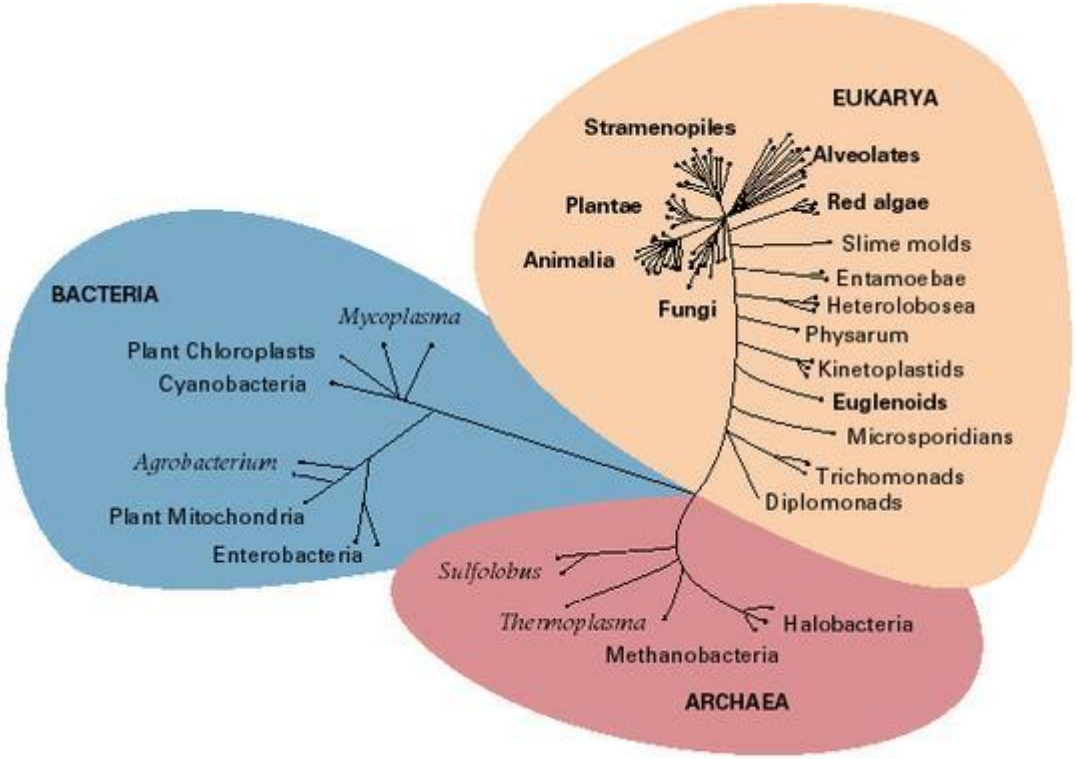
- Born, eat, replicate, and die

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.

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 archaebacteria
- only one type of membrane-- plasma membrane forms
  - the **boundary** of the cell proper
- The smallest cells known are bacteria
  - Ecoli cell
  - $3x10^6$ 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
  - $5x10^9$ protein molecules
  - 5000-10,000 polypeptide species

*Hela is an immortal cell line used in scientific research. It is the oldest and most commonly used human cell line. The line was derived from cervical cancer cells taken on February 8, 1951 from Henrietta Lacks, a patient who died of cancer on October 4, 1951.

Prokaryotic and Eukaryotic Cells
Chromosomal differences

Prokaryotes

- The genome of E.coli contains amount of t $4X10^6$ base pairs
- > 90% of DNA encode protein

- Lacks a membrane-bound nucleus.
  - Circular DNA and supercoiled domain
- Histones are unknown

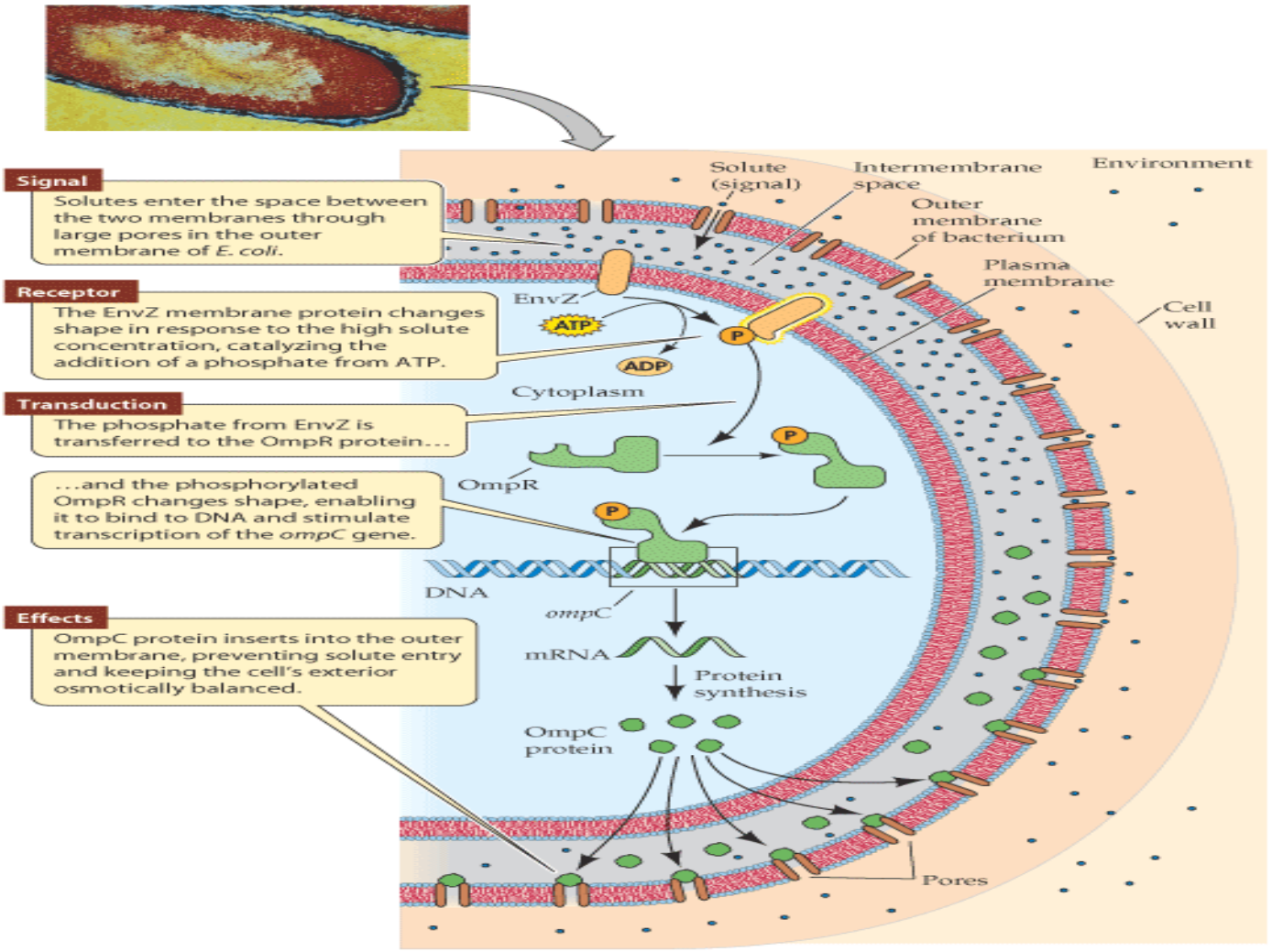
Eukaryotes

- The genome of yeast cells contains $1.35x10^7$ base pairs
- A small fraction of the total DNA encodes protein (~1.5%for human).
  - 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 acid**: Biological molecules(RNA and DNA) that allow organisms to reproduce;

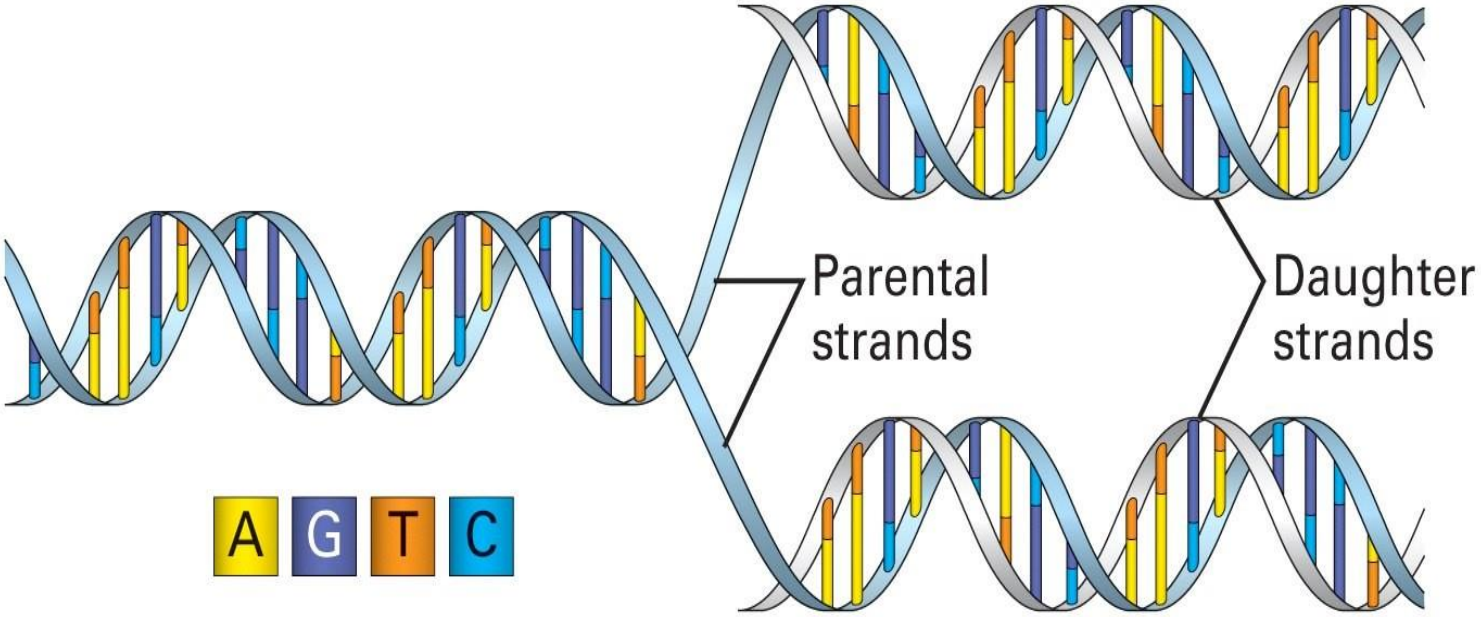
More Terminology

- The **genome** is an organism's complete set of DNA.
  - a bacterium contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion.
- human genome has 24 distinct chromosomes (22 pairs of autosomes + one pair of sex chromosomes [X and Y]).
  - 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

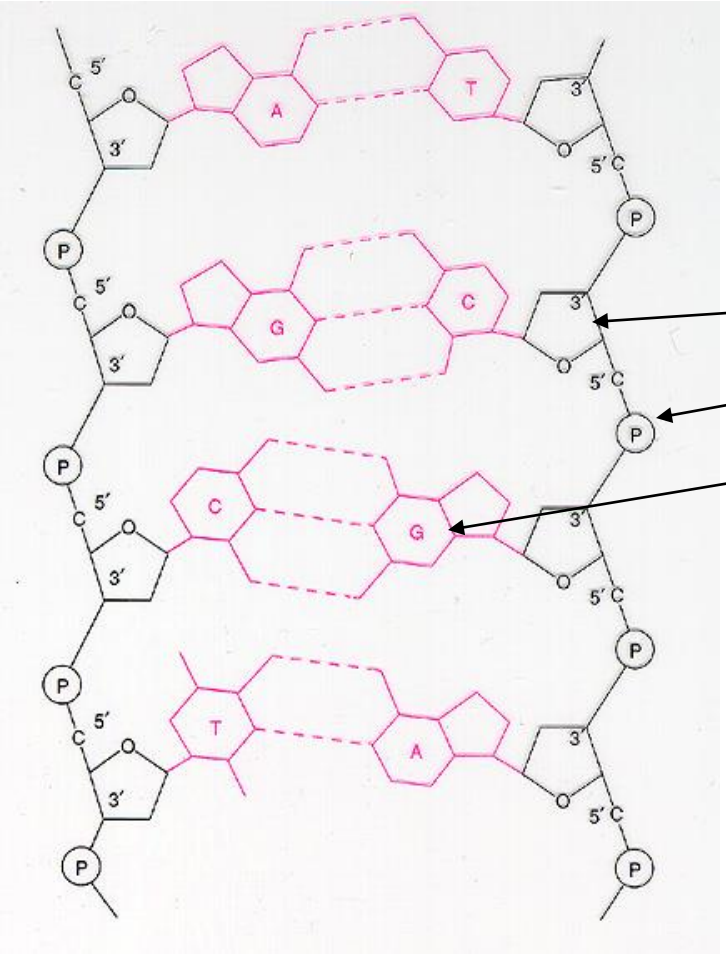
- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g.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



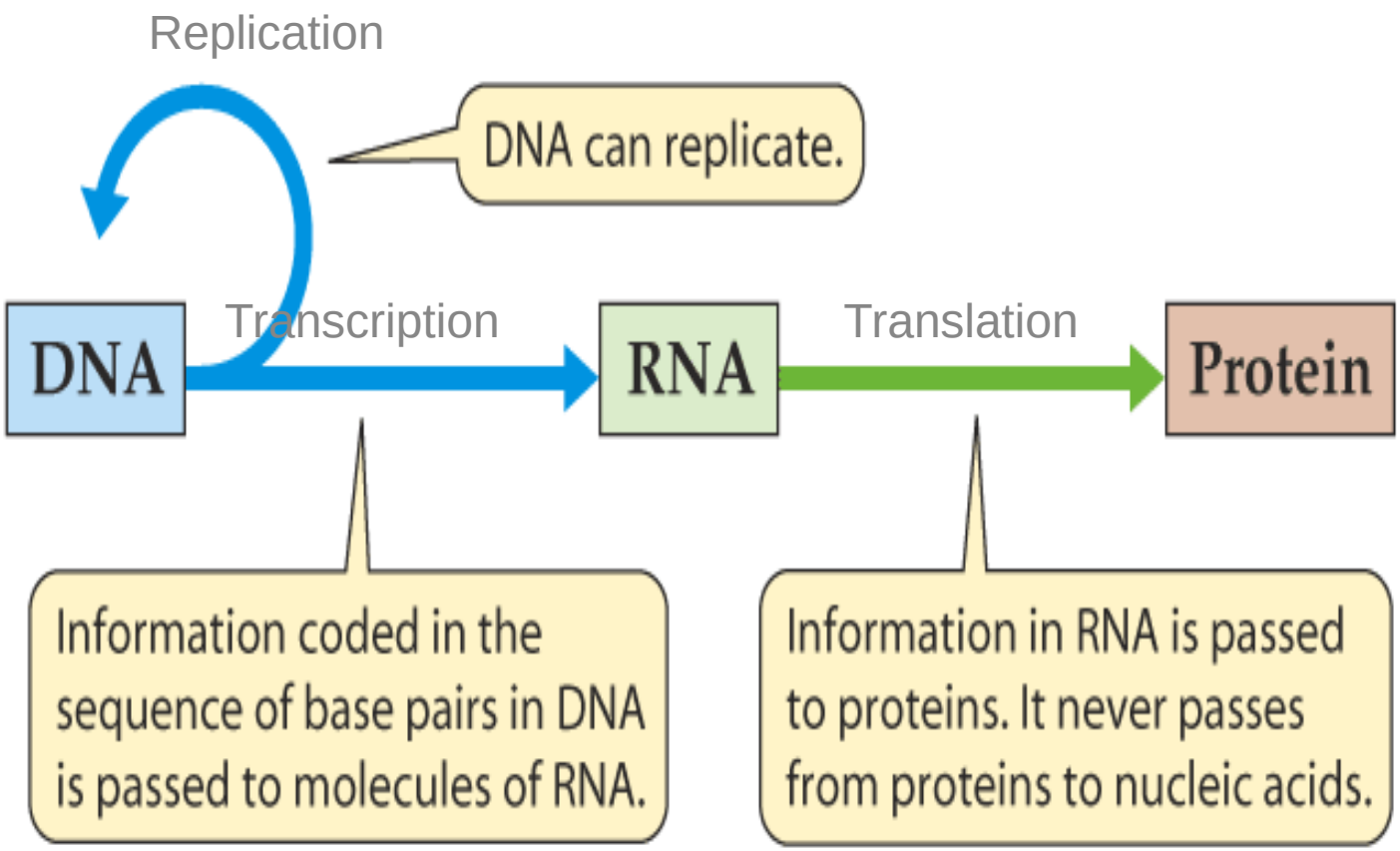
- DNA has a double helix structure which composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

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

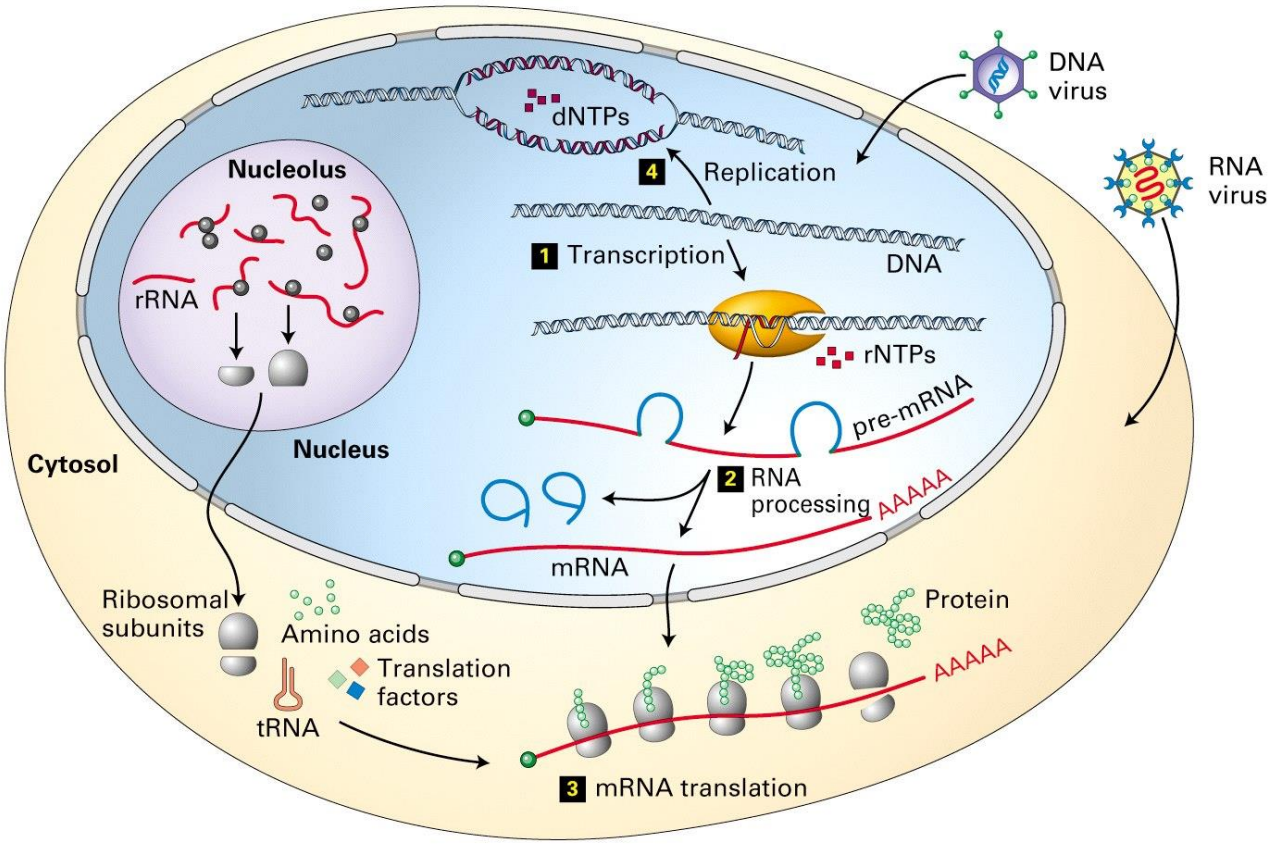
  5’ ATTTAGGCC 3’

  3’ TAAATCCGG 5’

DNA, RNA, and the Flow of Information

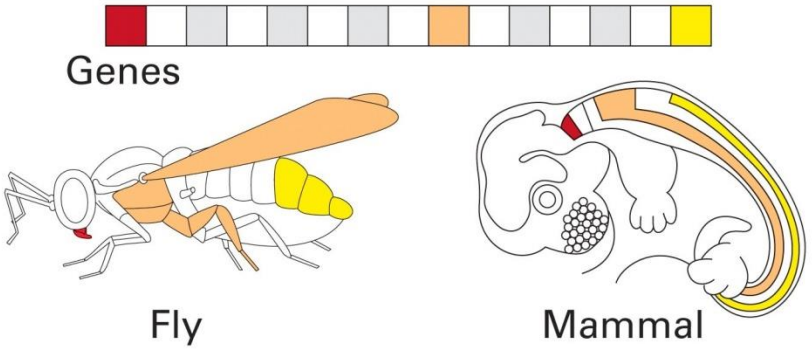


Overview of DNA to RNA to Protein



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

DNA the Genetics Makeup

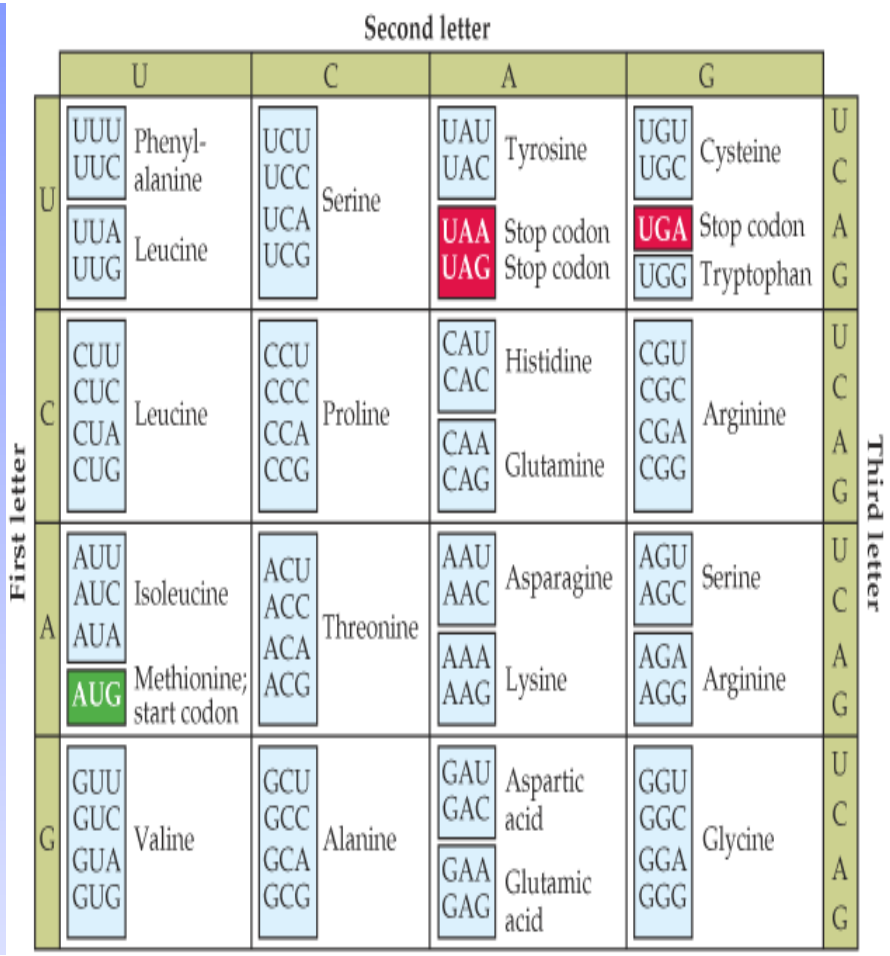


- Genes are inherited and are expressed
  - **genotype** (genetic makeup)
  - **phenotype** (physical expression)

- 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.)



Genetic Material of Life

- 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?

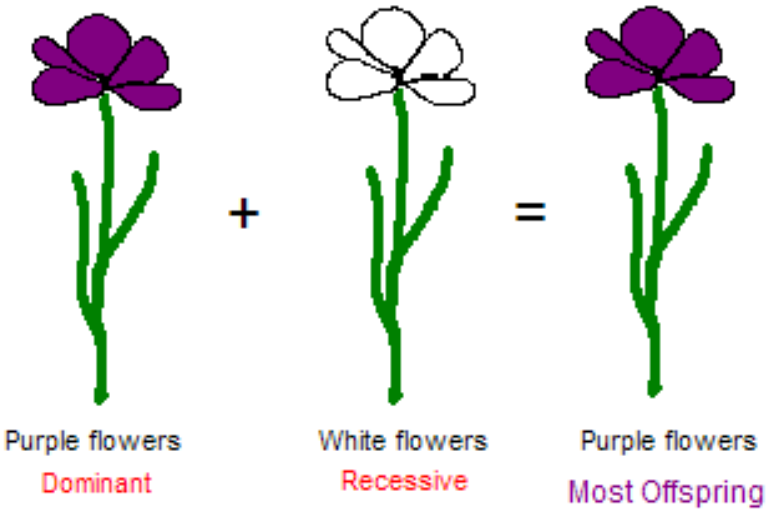
  -physical and functional traits that are passed on from one generation to the next.
- Genes were discovered by Gregor Mendel in the 1860s while he was experimenting with the pea plant. He asked the question:

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**

Mutations

- 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

The Good, the Bad, and the Silent

- Mutations can serve the organism in three ways:

- The Good : A mutation can cause a trait that enhances the organism's function:

  Mutation in the sickle cell gene provides resistance to malaria.

- The Bad : A mutation can cause a trait that is harmful, sometimes fatal to the organism:

  Huntington's disease, a symptom of a gene mutation, is a degenerative disease of the nervous system.

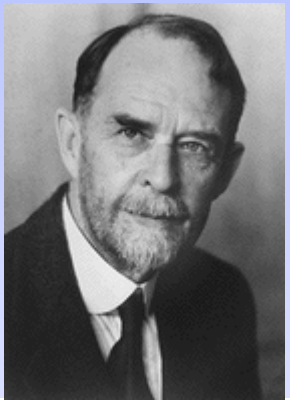
- The Silent: A mutation can simply cause no difference in the function of the organism.

Campbell, Biology, 5th edition, p. 255

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.



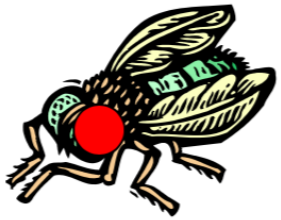
Portrait of Morgan

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

The White-Eyed Male

White-eyed male

X

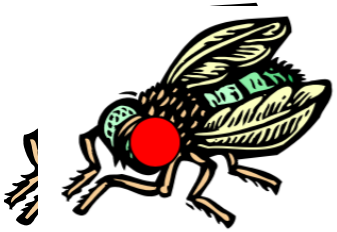


Red-eyed female
(normal)

white-eyed

Red-eyed

Mostly male progeny



Mostly female progeny

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:

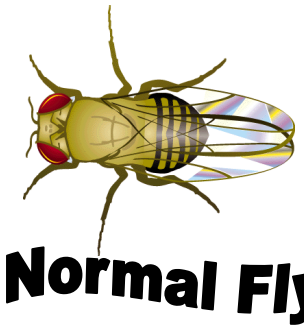
cn – eye color

Courtesy of the Archives, California Institue of Technology, Pasadena

Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html

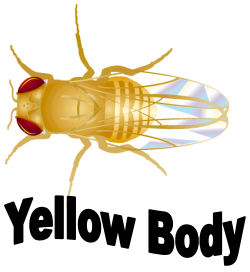
Linked Genes and Gene Order cont…

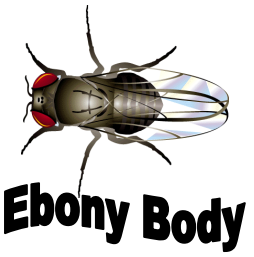
- By looking at the frequency that two genes are co-inherited, genetic maps can be constructed for the location of each gene on a chromosome.
- One of Morgan's students Alfred Sturtevant pursued this idea and studied 3 fly genes:



cn – eye color

b – body color

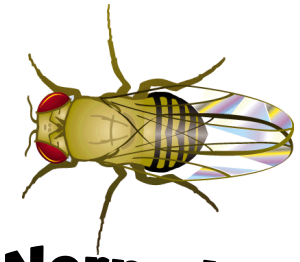




Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html

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:

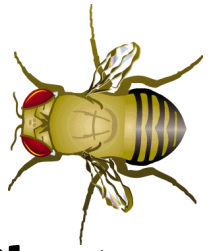


Normal Fly

cn –eye color

b –body color

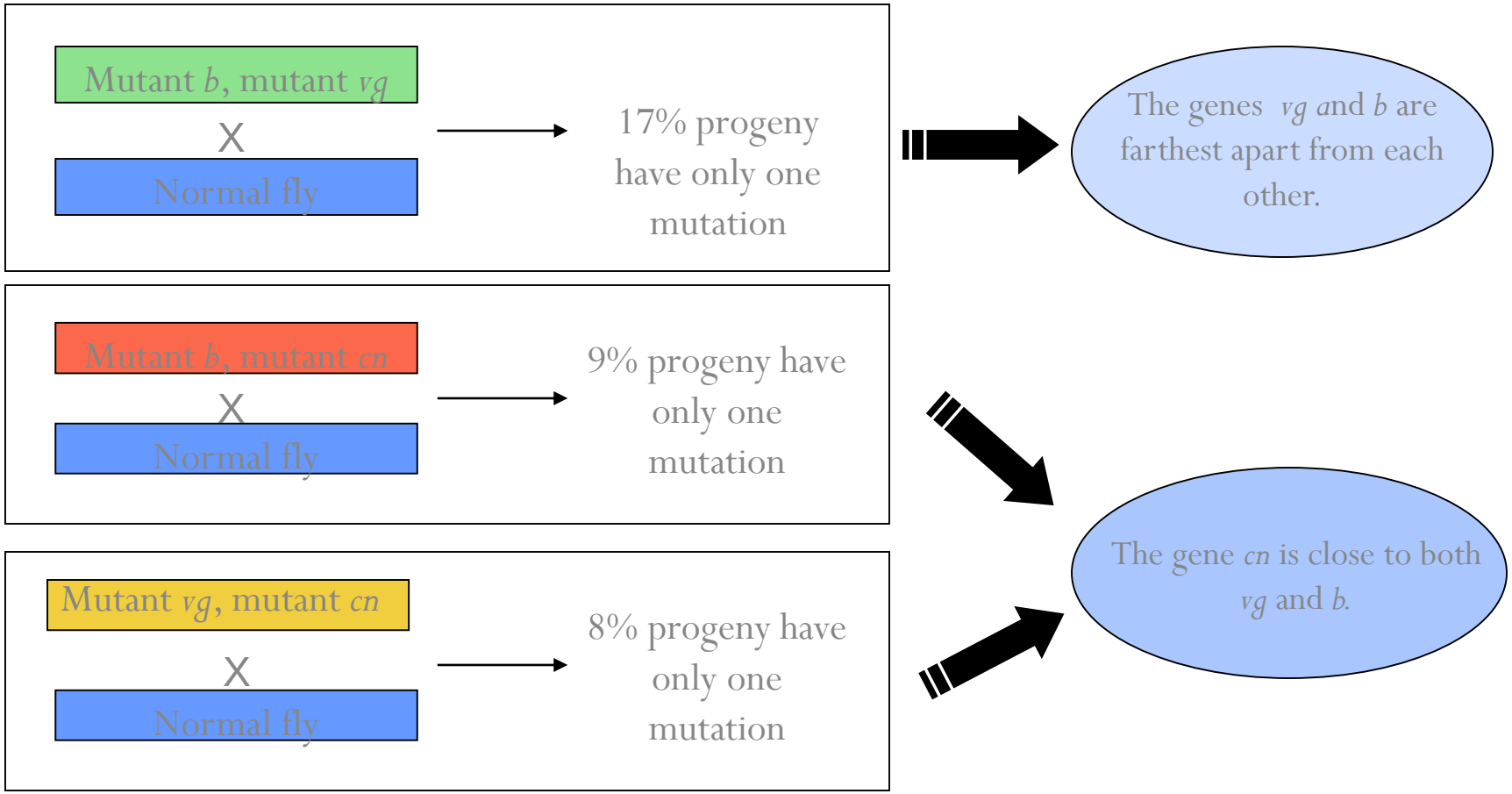
vg –wing size



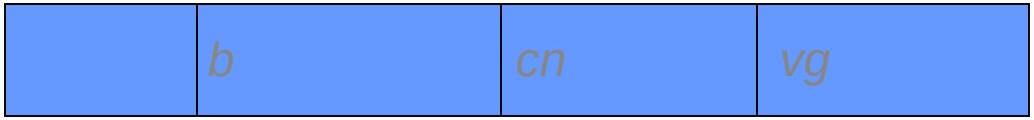
Short wings

Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html

What are the genes' order on the chromosome?

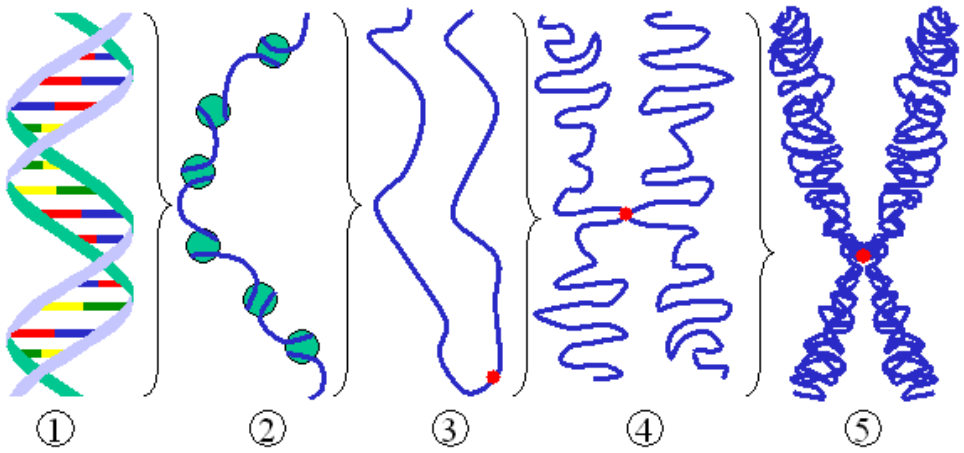


What are the genes' order on the chromosome?



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

Genetic Information: Chromosomes

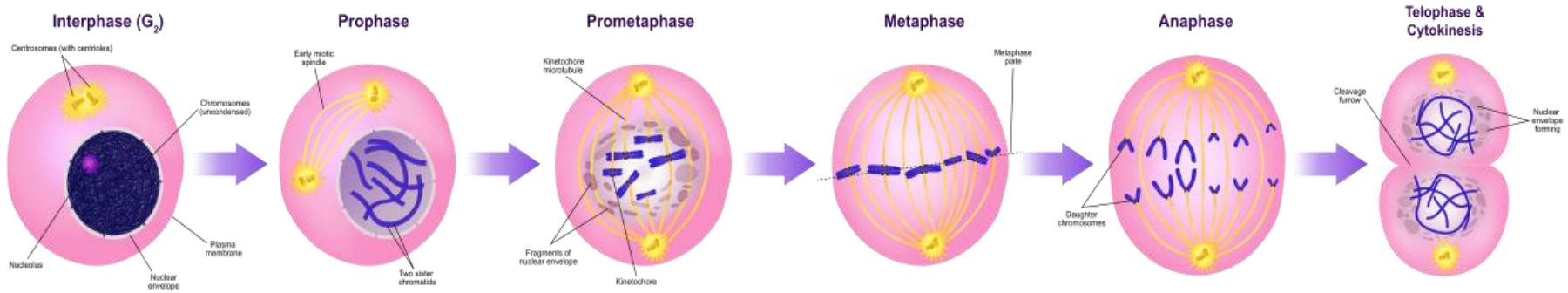


- (1) Double helix DNA strand.
- (2) Chromatin strand (**DNA** with **histones**)
- (3) Condensed chromatin during interphase with **centromere**.
- (4) Condensed chromatin during prophase
- (5) Chromosome during metaphase

Mitosis

- Interphase is the preparation step (G1, S, G2) before the mitosis starts.

- Steps of mitosis
  - **Prophase:** Each chromosome duplicates and remains closely associated. These are called sister chromatids.
  - **Metaphase**: Homologous chromosomes align at the equatorial plate.
  - **Anaphase**: Centromeres divide and sister chromatids migrate separately to each pole.
  - **Telophase**: Opposite of the prophase step. Two daughter cells are formed from one parent cell.



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

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.

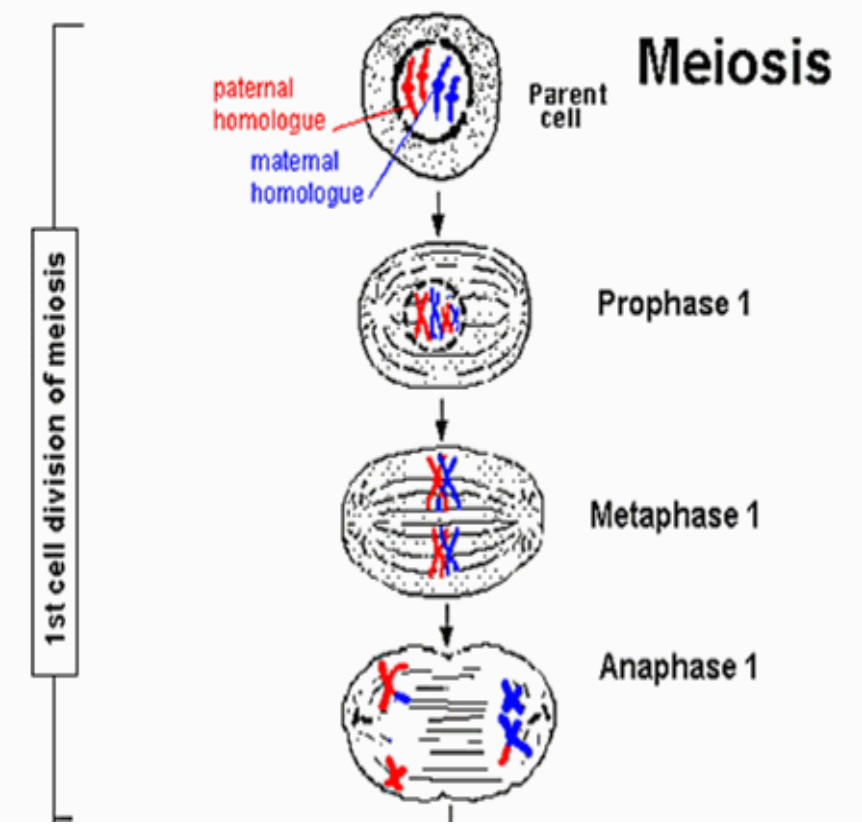
Meiosis

- 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.

Meiosis



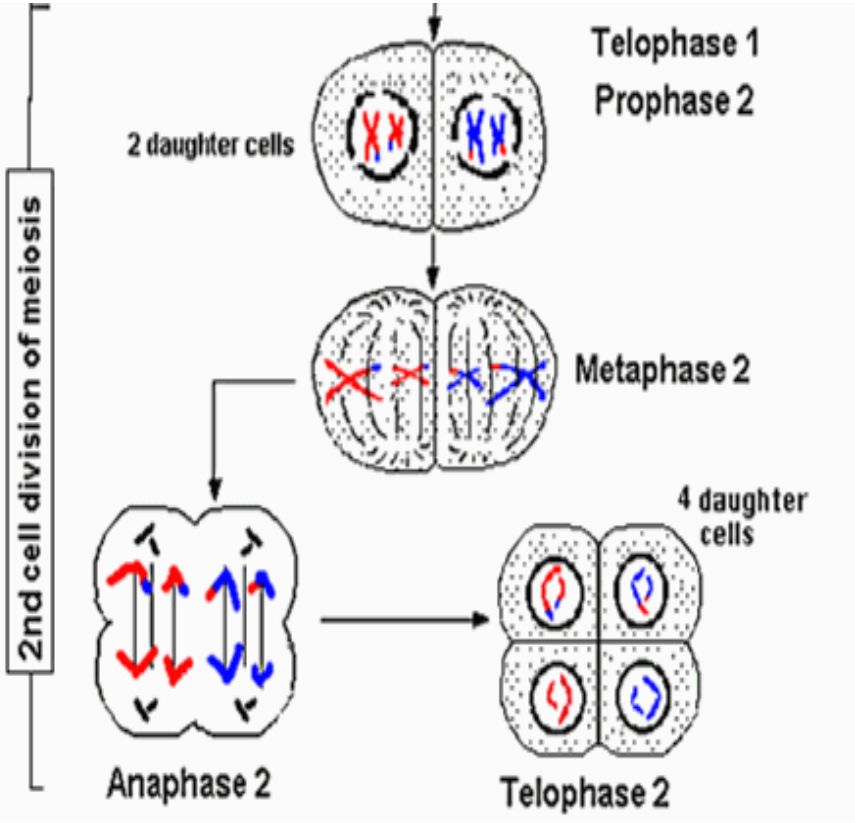


Diagram 1.