CSE 5290: Algorithms for Bioinformatics

Suprakash Datta

datta@cse.yorku.ca

September 8, 2011

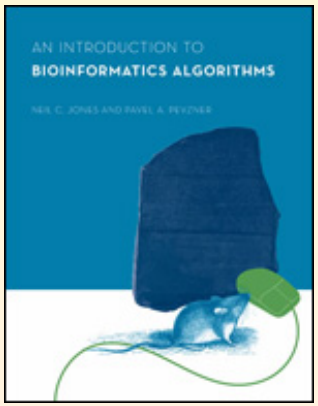
Administrivia

Essential information:

- Office: CSEB 3043
- Phone: 416-736-2100 ext 77875
- Course page: http://www.cse.yorku.ca/course/5290 – All announcements/handouts will be published on the webpage – check regularly for updates.
- Lectures: Thu 7:00 - 10:00 pm (CB 122)
- Office hours: By appointment.
- TA: none.

Administrivia - 2

- Text: An Introduction to Bioinformatics Algorithms: Neil C. Jones and Pavel A. Pevzner, MIT Press, August 2004. [should be available at the York bookstore]



- Grading: grades will be on ePost (linked from webpage).
  - Midterm : 15%
  - Homework : 30%
  - Final: 30%
  - Project: 25% (Project details will be posted on the class webpage).

Course Objectives

- Familiarity with computational problems in Biology
- Applying algorithmic ideas
- Understand real-life computational challenges
- Improve understanding of algorithms

My expectations

- Familiarity with undergraduate algorithms
- Interest in computational problems
- Willingness to pick up a little Biology
- Active interest in your project and assignments

My research

- Computer Networks.
- Analysis of Biological data, e.g.
  - Flow cytometry data
  - Microarray data
- Genomic Signal Processing: Convert biological sequences to numerical sequences and apply signal processing tools:
  - exon prediction
  - transposable elements
  - repeat detection

Outline

What is Bioinformatics?

No consensus!

- Genomics
- Proteomics
- Evolutionary biology
- Clinical trial informatics
- Epidemiology?
- Medical image processing?
- Artificial life?

From `http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html`:
"Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline."

Why Bioinformatics?

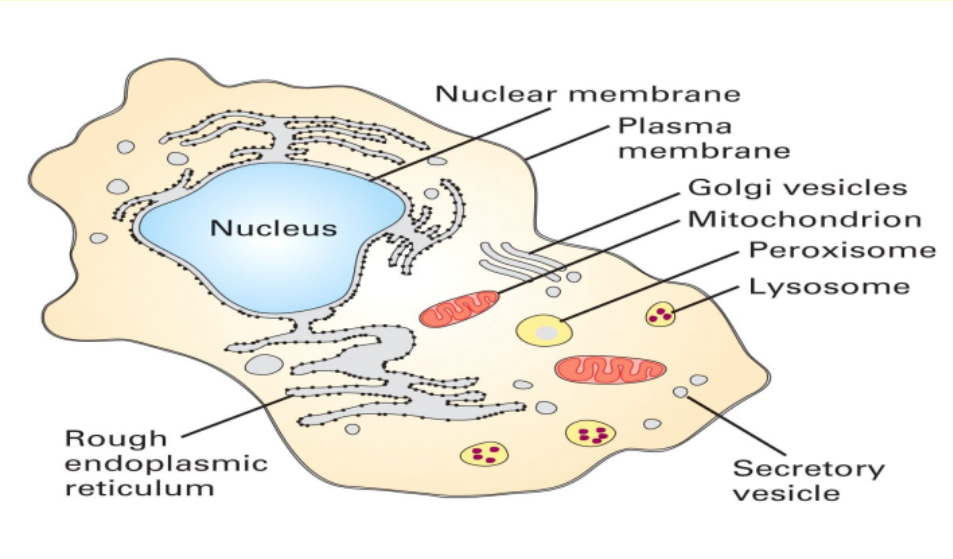
- Make an impact!
- Interdisciplinary work
- Work with real data sets
- Use algorithmic skills

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.

Understanding cell biology leads to understanding of life, including replication, self-regulation, self-repair, death, viruses, cancer. Many parts are not fully understood.

Cells - 2



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

Chemical composition (by weight)

- 70% water
- 7% small molecules (salts, lipids, amino acids, nucleotides)
- 23% macromolecules (Proteins, Polysaccharides, lipids)

Prokaryotes vs Eukaryotes

Prokaryotes

- Single cell
- No nucleus
- No organelles
- 1 piece of circular DNA
- No mRNA post transcriptional modification
- Genome sizes often smaller (e.g. E. Coli: 4 million nucleotides)
- Almost all of the DNA encodes protein

Eukaryotes

- Single or multi cell
- Nucleus
- Organelles
- Chromosomes
- Exons/Introns splicing
- Genomes often larger (e.g. yeast: 13.5 million nucleotides)
- A small fraction of DNA encodes protein

Fundamental questions

Life begins from a single cell. Cells divide, grow, differentiate, inherit characteristics.

- How is information stored?
- How is information transmitted?
- How is information translated to function?
- How does self-regulation and replication happen?

Information Storage: Cells store all information to replicate itself

- Human genome is around 3 billions nucleotides long
- Almost every cell in human body contains same set of genes
- But not all genes are used or expressed by those cells

Information organization in a cell

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.
Machinery:

- Collect and manufacture components
- Carry out replication
- Kick-start its new offspring

Control: 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

Some terminology

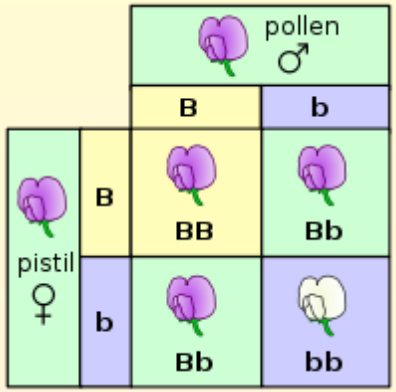
- Genome: an organisms genetic material (human and mouse genomes have some 3 billion).
- The human genome has 24 distinct chromosomes. Each chromosome contains many genes.
- Gene: a discrete unit of hereditary information located on the chromosomes and consisting of DNA.
  - basic physical and functional units of heredity
  - specific sequences of DNA bases that encode instructions on how to make proteins.
- Genotype: The genetic makeup of an organism
- Phenotype: the physical expressed traits of an organism

Terminology - continued

All Life depends on 3 critical molecules - DNA, RNA and proteins.

- Nucleic acid: Biological molecules(RNA and DNA) that allow organisms to reproduce;
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form bodys major components (e.g. hair, skin, etc.)
  - large, complex molecules made up of smaller subunits called amino acids.
- RNA: similar to DNA in composition, help transfer information.

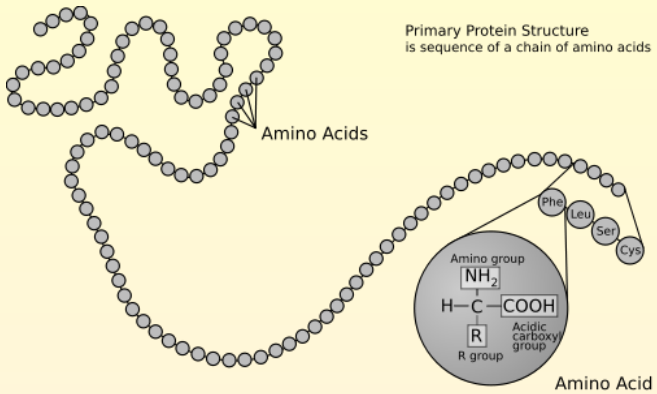
What is a gene? – Part 1



A **gene** is the basic unit of inheritance. It contains information about how to build and maintain a living organism's cells, and it conveys traits from parent to child. An **allele** is one possible variant of a gene.

image by Madprime on Wikimedia Commons

What is a protein?



A **protein** is a molecule made up of a chain of amino acids. Most proteins have between 200-300 amino acids, though some are smaller and some contain tens of thousands of amino acids. A protein can be

- an **enzyme** which drives a biochemical reaction
- a structural component of cells or tissues
- important to cell processes like cell signaling or immune response
- a dietary source of amino acids

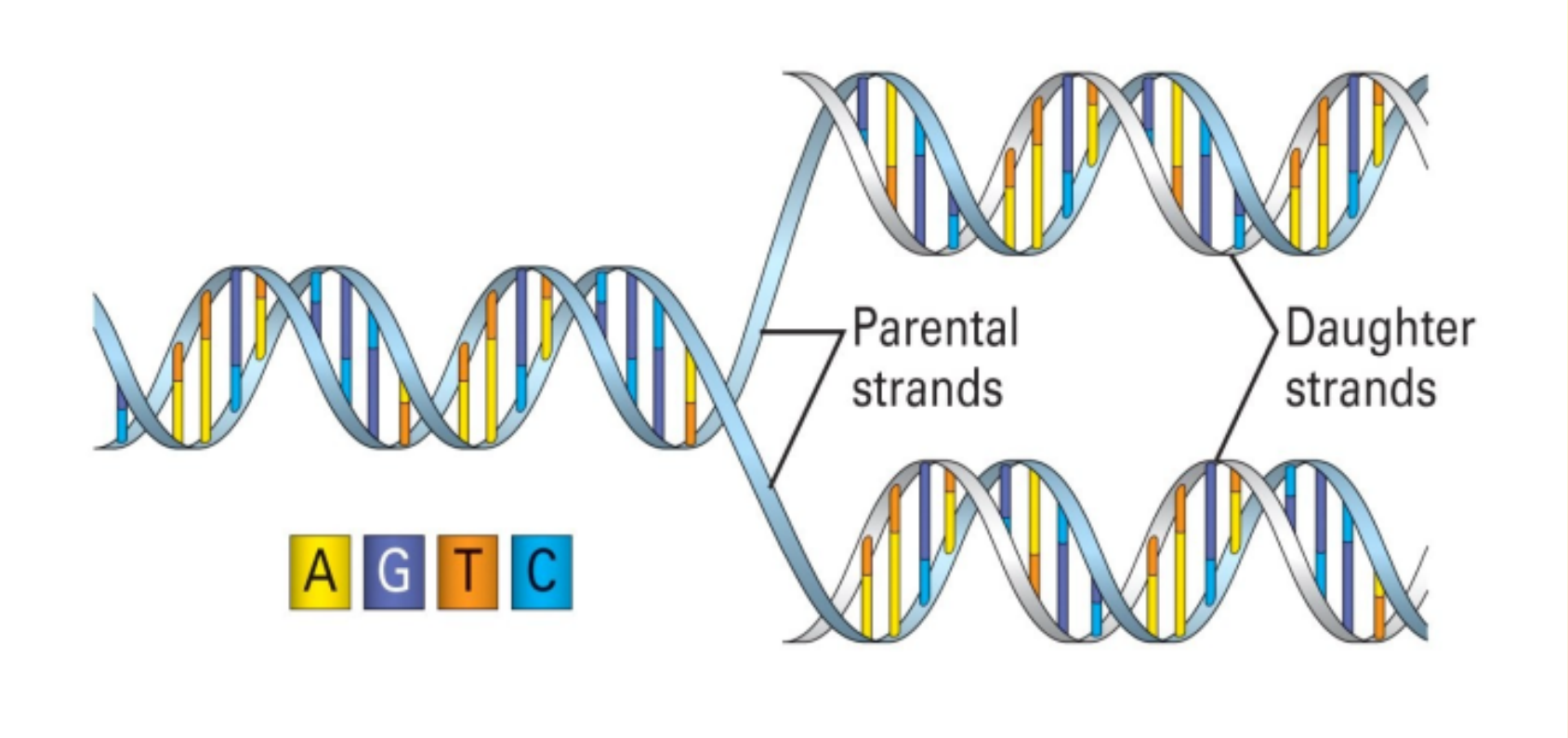
image courtesy: National Human Genome Research Institute

One Gene/One Protein Hypothesis

The first major advance in molecular biology occurred in the early 1940's when George Beadle and Edward Tatum discovered a connection between genes and proteins. They mutated bread mold by exposing it to X-rays. The mutants couldn't grow without the addition of vitamins. They had lost the ability to produce the enzyme required for synthesis of the vitamin. They verified that only one gene had been mutated, and, as a result of the experiment, formulated the

One Gene/One Protein Hypothesis.

What are genes made of?

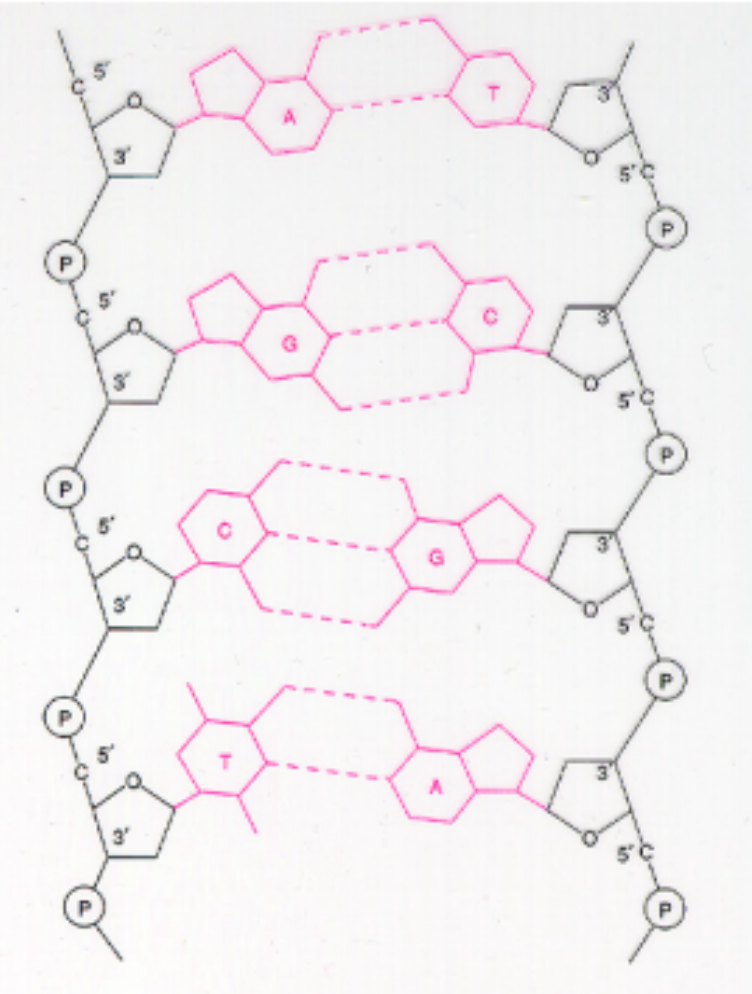


- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

Facts about DNA

- DNA is made up of nucleotides.
- Each nucleotide has three parts: a phosphate group, a sugar group, and a nitrogen base.
- There are four types of bases: adenine (A), cytosine (C), guanine (G), and thymine (T).
- A strand of DNA is a chain of nucleotides linked by alternating phosphate and sugar groups.
- DNA has two strands.
- Opposite strands are complementary: A-T and G-C.
- DNA is found in the cell nucleus, mitochondria, chloroplasts, and plasmids.

Facts about DNA



DNA has a double helix structure which is composed of

- sugar molecule
- phosphate group
- a base (A,C,G,T)

DNA always reads from 5 end to 3 end for transcription, replication

5 ATTTAGGCC 3

3 TAAATCCGG 5

Genetic Code

In 1961, the Crick, Brenner et al Experiment demonstrated the true nature of the genetic code. They induced mutations that inserted or deleted single nucleotides in T4 bacteriophage gene rIIB causing frameshift mutations.

- 1, 2, or 4 mutations rendered the protein produced non-functional
- 3 mutations – still functional

Conclusion:

- Code is non-overlapping and is in groups of three.
- 3 bases $\Rightarrow$ 64 possibilities.
- Since there are 20 amino acids, this means the code is degenerate.

Genetic Code - 2

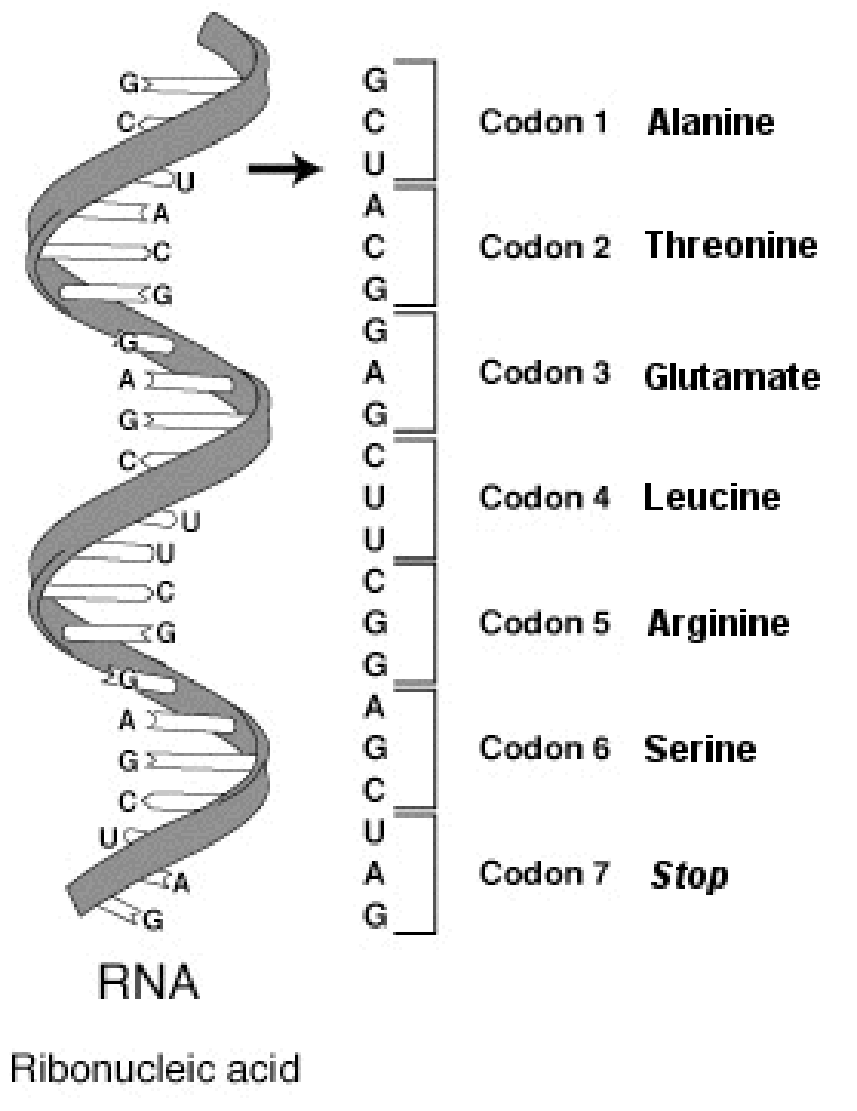


image courtesy: National Human Genome Research Institute

Genetic Code - 3

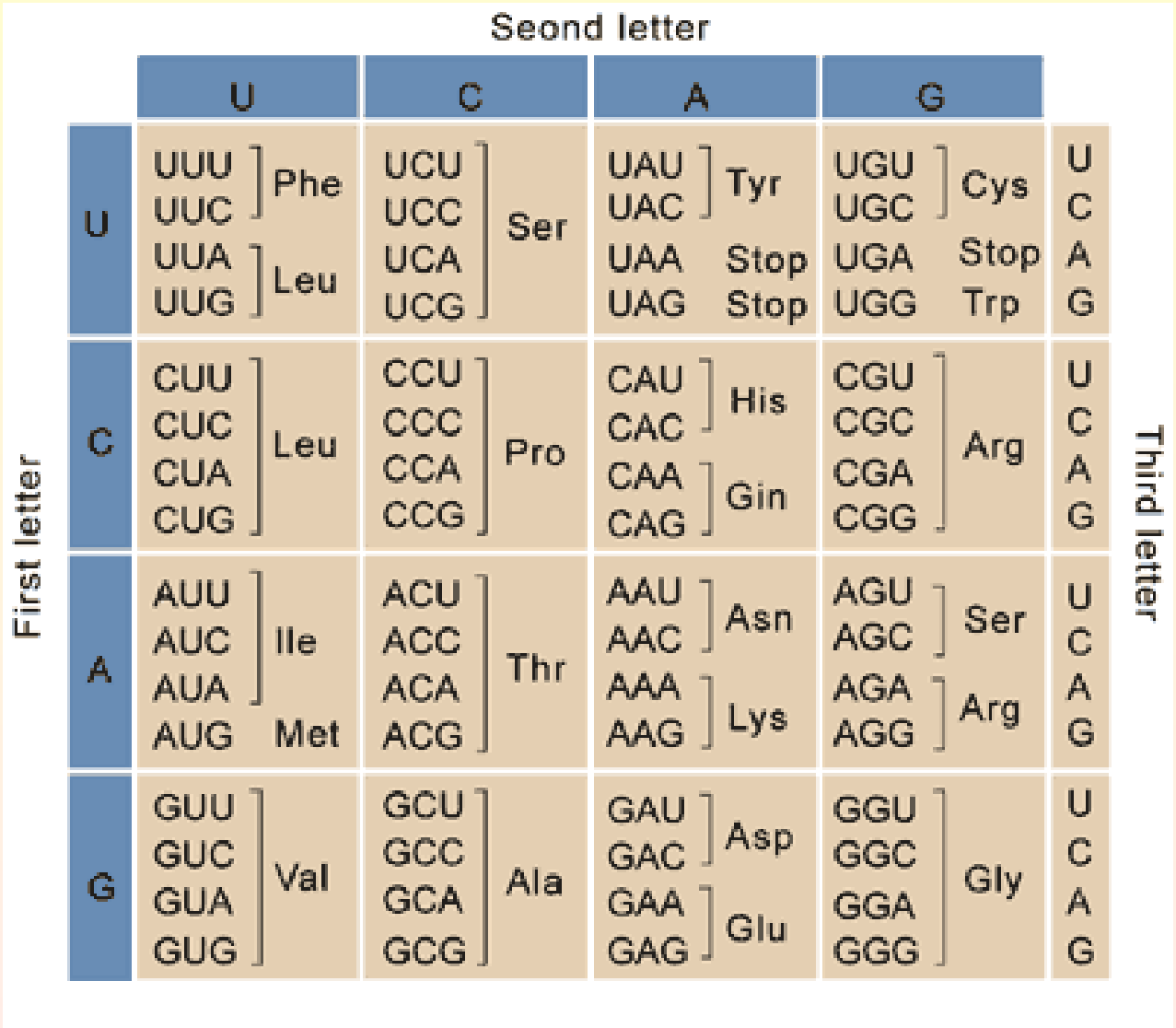
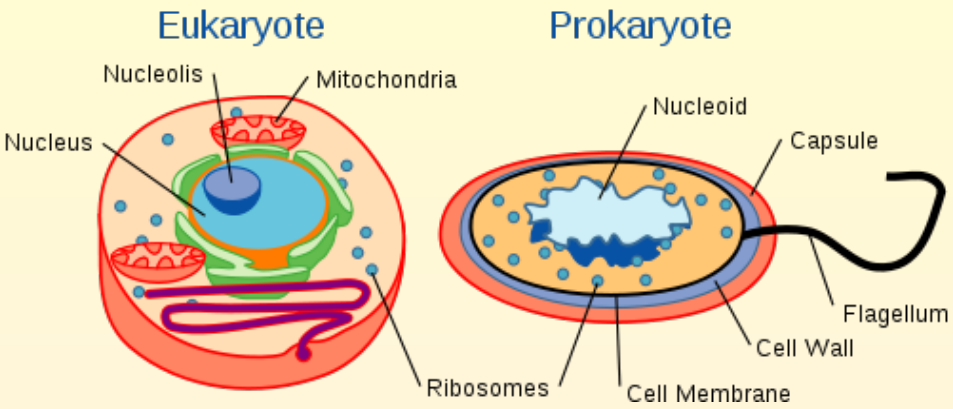


image courtesy: NIH

How does the genetic code work?



RNA is a molecule similar to DNA, except less stable and used for short term work. Three types of RNA.

- Messenger RNA (mRNA) copies information from DNA and brings it to the ribosomes.
- Transfer RNA (tRNA) attaches to amino acids and brings them to the ribosomes.
- Ribosomal RNA (rRNA) molecules synthesize protein.

image courtesy: NCBI

Central Dogma of Molecular Biology

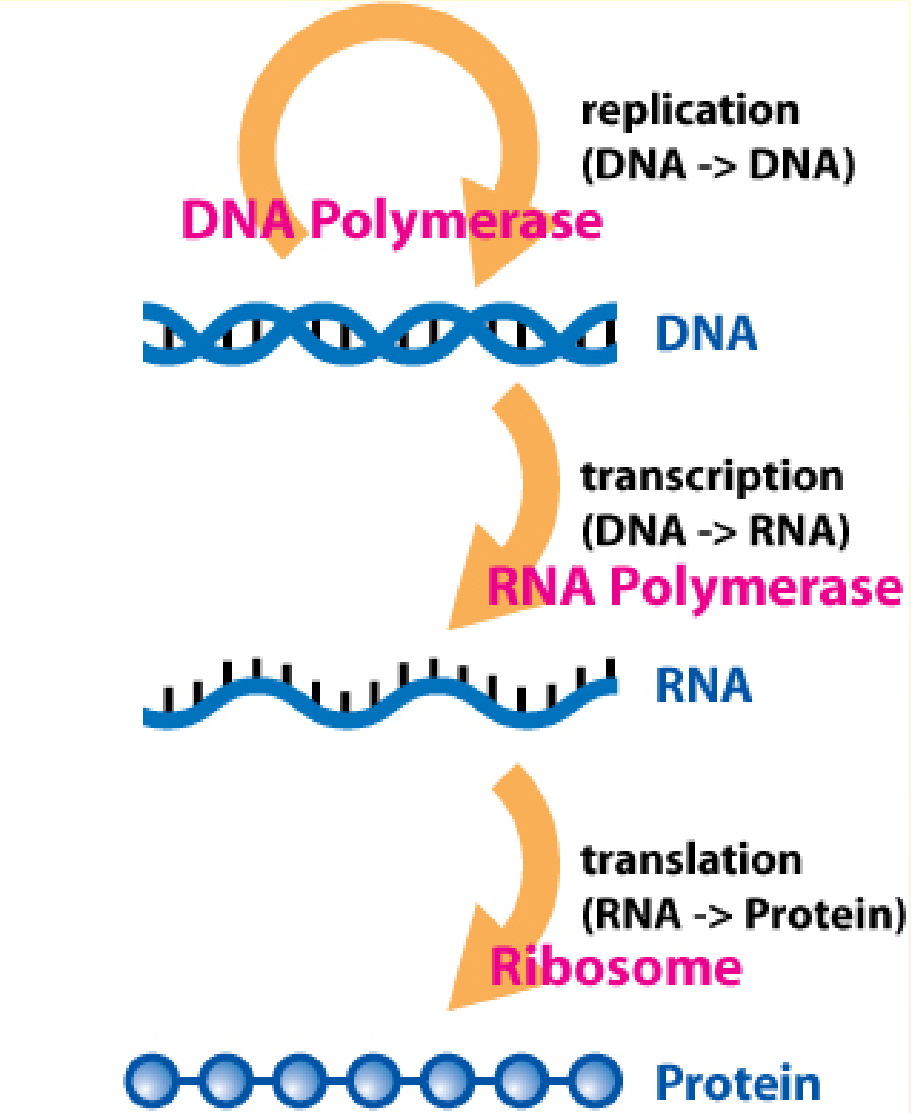


image by Daniel Horspool, Wikimedia Commons

Outline

Human Genome Project

The Human Genome Project (1990-2003) was a multinational effort coordinated by the US Dept. of Energy and the National Institutes of Health to

- identify the 20,000-25,000 genes in human DNA
- determine the sequence of the 3 million base pairs of the human genome
- store this information in public databases

Most of the genome does not code for proteins – only about 2.5% of the genome consists of genes

NCBI

The National Center for Biotechnology Information is the public database promised by the Human Genome Project. It can be found at http://www.ncbi.nlm.nih.gov/. It has a wealth of information and resources beyond the human genome sequence.

- Genbank
- BLAST
- Map Viewer
- Conserved Domain Database
- Special purpose tools and data for studying influenza, retroviruses, etc.
- Educational Resources

Some animals whose genomes have been sequenced

image by Thomas Lersch

image by John Gould

- chimpanzee
- lab mouse
- rat
- duck-billed platypus
- opossum
- cattle
- dog
- zebrafish
- chicken
- zebra finch

image by Robert Merkel

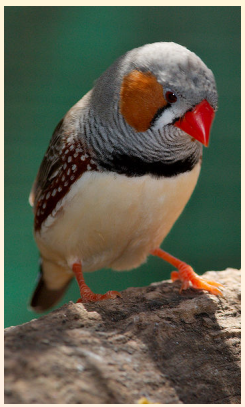


image by Peripitus

All images from Wikimedia Commons.

Some invertebrates whose genomes have been sequenced

image by James K. Lindsey

image by Bob Goldstein,

UNC Chapel Hill

- honey bee
- African malaria mosquito
- fruit fly
- red flour beetle
- pea aphid
- jewel wasp
- nematode (roundworm)
- purple sea urchin (embryology)
- hydra (model organism)
- sea squirt (shares 80% of genes with humans)

image by Eric Day

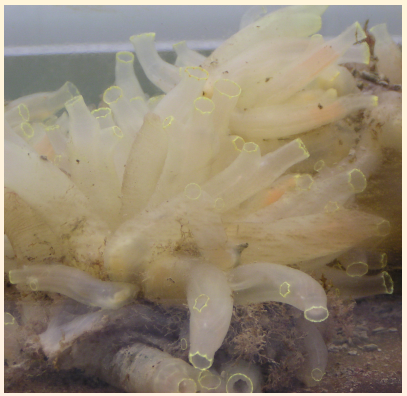


image by Perezoso

All images from Wikimedia Commons.

Some plants whose genomes have been sequenced

image by Roepers at nl.wikipedia

image by Fir0002

- thale cress (first plant sequenced)
- asparagus
- tomato
- rice
- wheat
- California poplar
- oak
- cocoa
- green algae
- moss

image by Walter Siegmund

public domain image

All images from Wikimedia Commons.

What is a gene? – Part 2

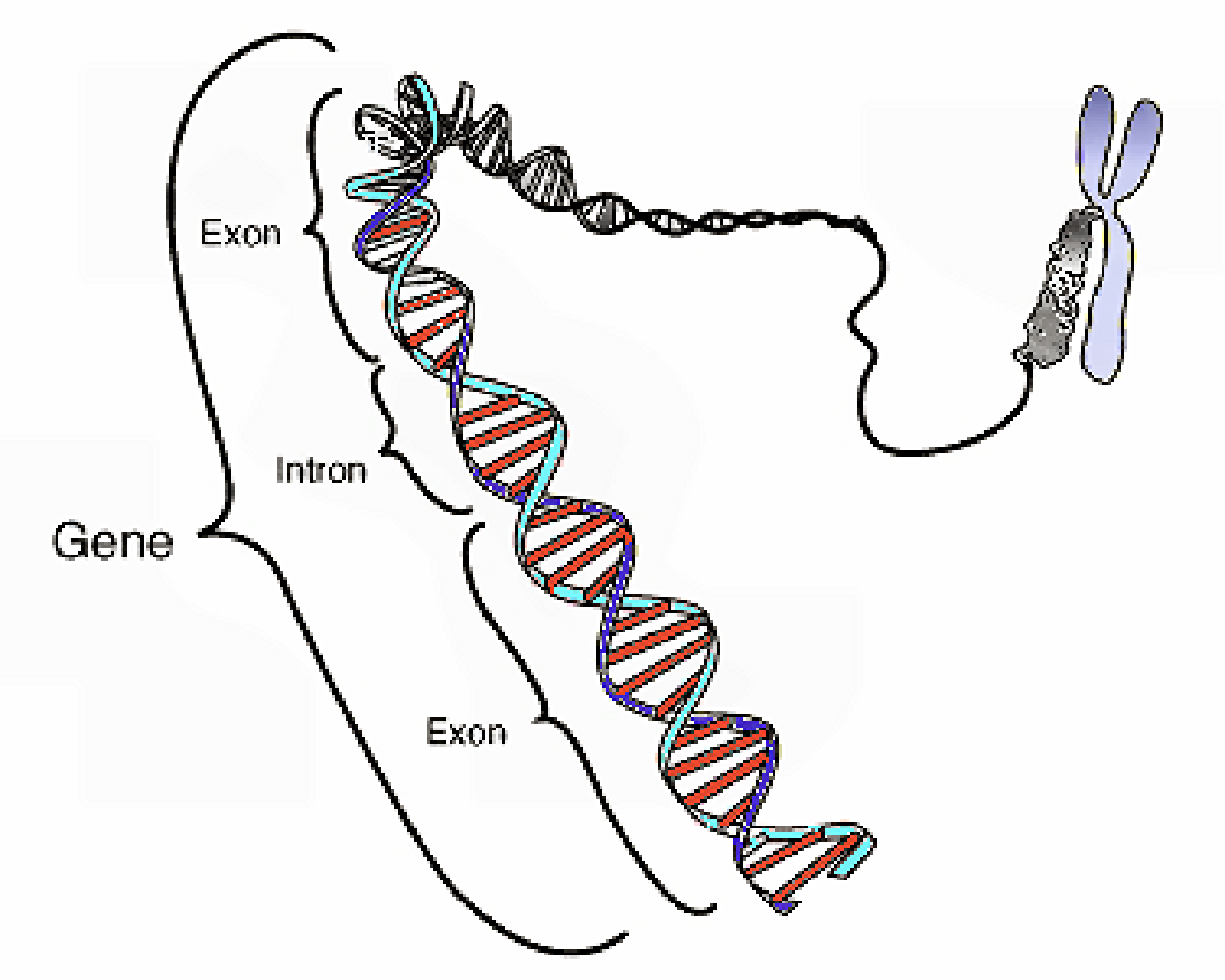


image courtesy: NIH

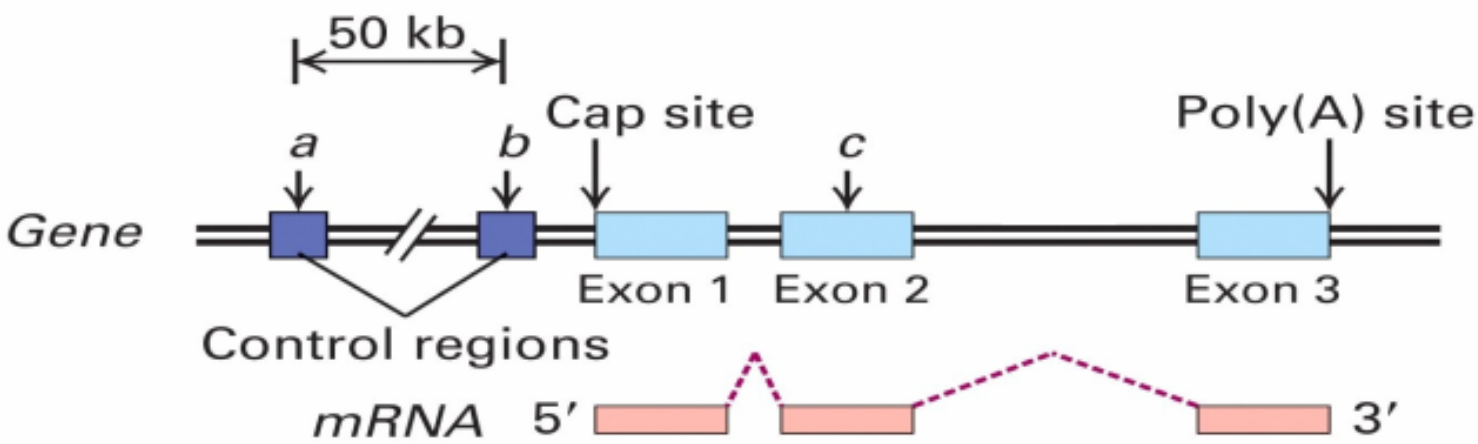
What is a gene?

Model of genes as beads on a string is too simple.

- alternative splicing (one gene codes for several proteins)
- fused transcripts (two genes "merge")
- exons combine with other exons hundreds of thousands of base pairs away
- RNA genes (produce RNA that is never translated into protein) 63% of the mouse genome is transcribed; only 2% of the mouse genome is protein-coding exons

Many geneticists are substituting more precise terms like **transcript** and **exon**.

A closer look at genes



- 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, 1 kb - 50 kb per intron
- Gene size: Largest 2.4 Mb (Dystrophin). Mean 27 kb.

The rest of the genome

- RNA genes
- *Cis*-regulatory elements (control transcription)
- introns
- pseudogenes
- microsatellites/tandem repeats
- retrotransposons (25% of human genome)
- endogenous retroviruses (8% of human genome)
- telomeres (repetitive ends of chromosomes)

Differences Between Individuals

NCBI only has a complete genome for a single individual, James Watson. A private corporation, Celera, has a complete sequence of another individual, Craig Venter. Both are white and of European descent. HapMap is studying the genomes of 270 individuals:

- 30 trios (two parents and an adult child) from the Yoruba people in Nigeria
- 45 unrelated people from Tokoyo, Japan
- 45 unrelated people from Beijing, China
- 30 trios from the US of northern and western European ancestry

Differences occur in many forms:

- single nucleotide polymorphisms (SNPs) – approx. 10 million
- insertions and deletions and inversions
- variable numbers of repetitive sequences

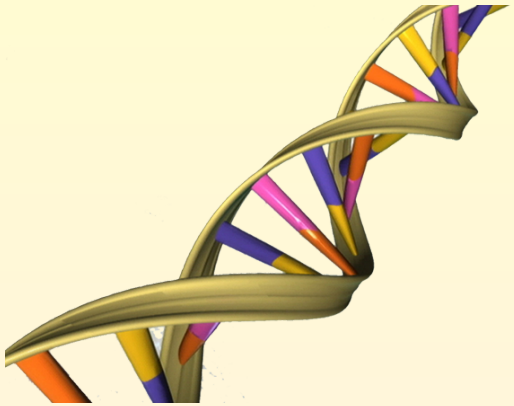
Genome Length

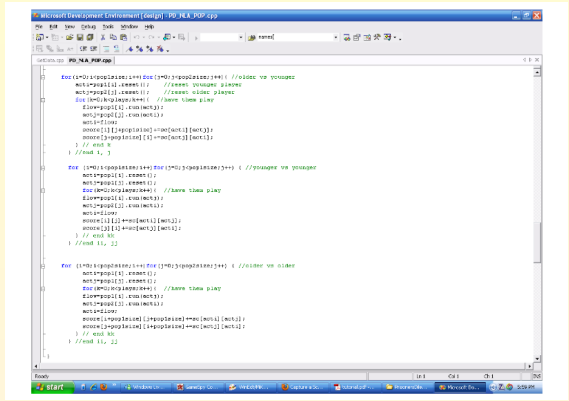
Genome length is unrelated to the complexity of the organism.

- Genomes range in size from half a million base pairs (bacteria) to a trillion base pairs (amoeba).
- The marbled lungfish has the largest animal genome (129.9 billion base pairs).
- Birds have shorter genomes than mammals, and salamanders and lungfish have longer ones.
- Some algae have longer genomes than all known mammal genomes.

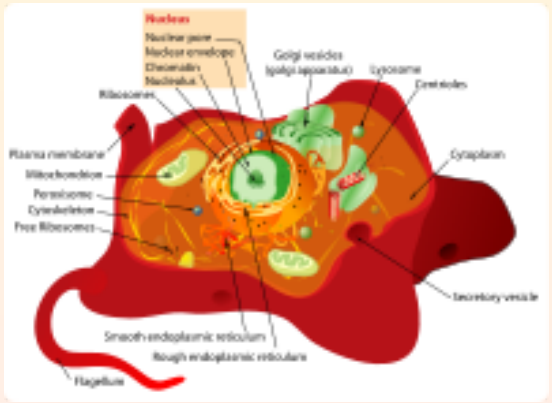
Resource: Gregory, T.R. (2005). Animal Genome Size Database. http://www.genomesize.com.

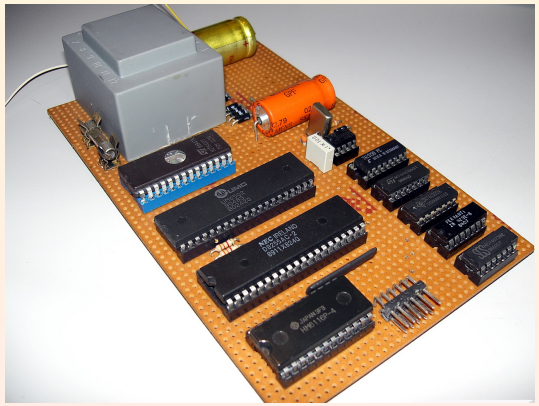
Bad Analogy





DNA = software





cells = hardware

Better Analogy

$$\text{genotype} \longmapsto \text{phenotype}$$

$$\text{DNA} = \text{data}$$

$$\text{metabolism in cells and cell interactions} = \text{software}$$

$$\text{distributed system of finite state machines}$$

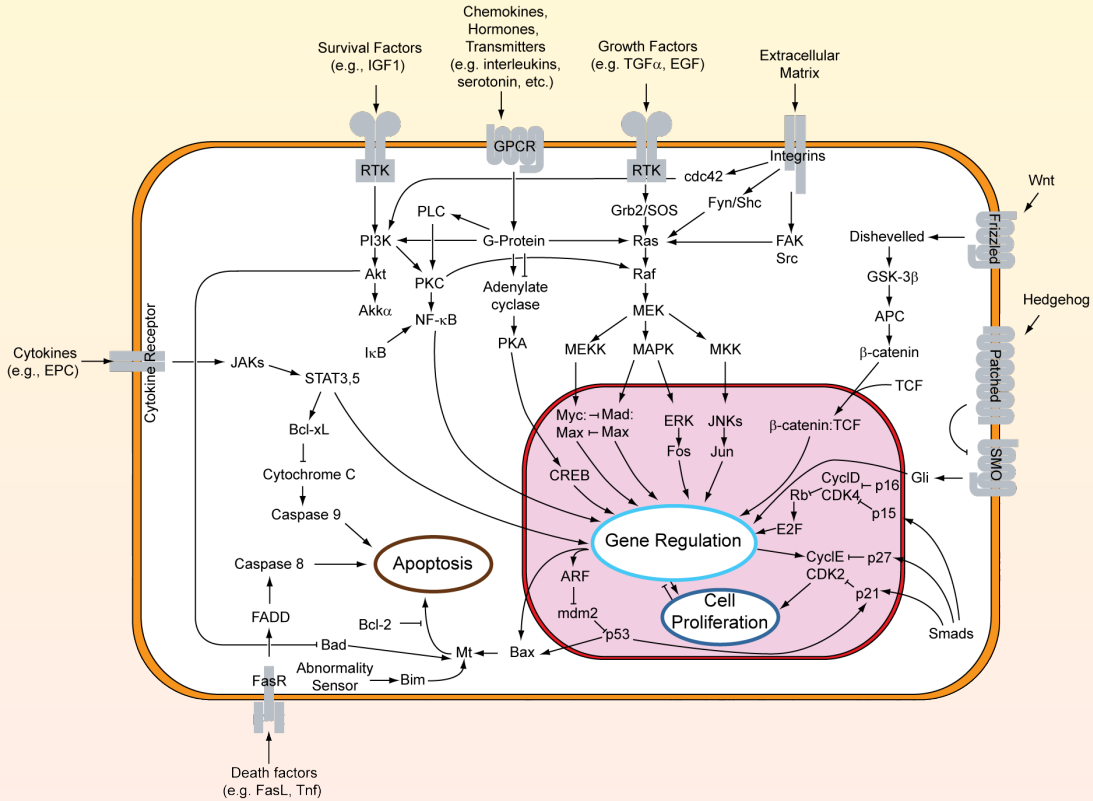


image by Roadnottaken on Wikimedia Commons

Differences between humans

- 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
- 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.

Genetic variation

- Sequencing and studying only one genome is not enough because every individual is genetically different!
- 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?
  - recombination
  - mutation
  - retroviruses

Outline

Function Determined by Structure

Protein function relies on its ability to bind tightly to other specific molecules.

- Enzymes bind to substrates to catalyze chemical reactions.
- Antibodies bind to foreign substances in the body.
- Ligand transport proteins (e.g. hemoglobin) bind to small biomolecules where they are common and release them where they are rare.

Sometimes protein structure IS its function.

- Fibrous proteins make up the cytoskeleton, which allows a cell to maintain its shape and size.
- Connective tissue contains proteins, collagen and elastin.
- Filamentous structures (hair) get their structure from the protein keratin.
- Proteins capable of generating mechanical force are important to cell motility.

Proteins

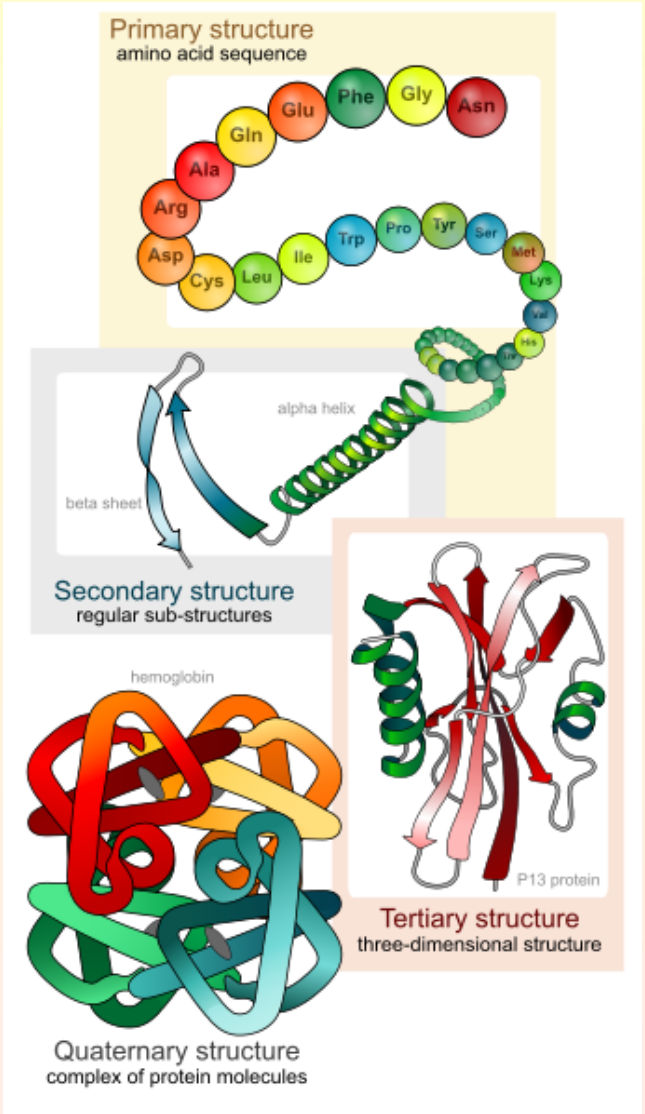
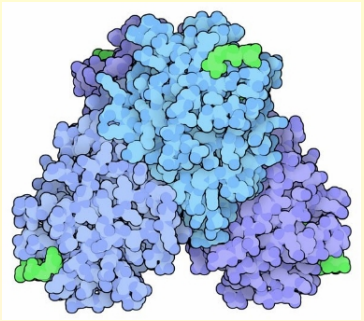
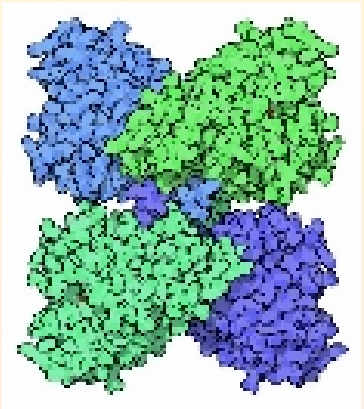


image courtesy: LadyofHats on Wikimedia Commmons

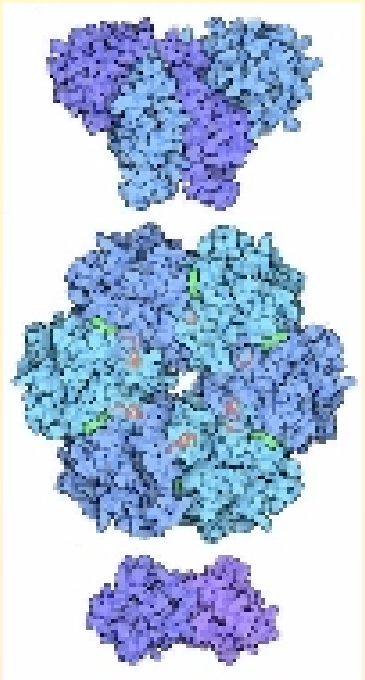
Protein Structures



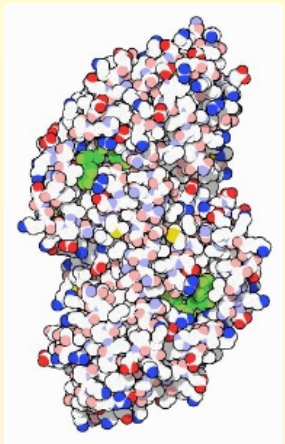
Concanavalin A (binds carbohydrates)



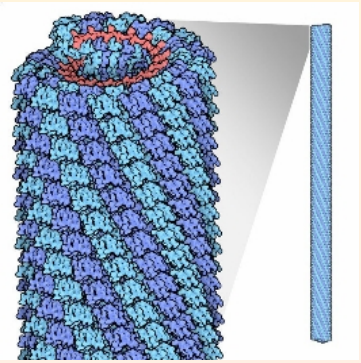
Enzyme that converts $F \mapsto Y$



Circadian Clock Proteins



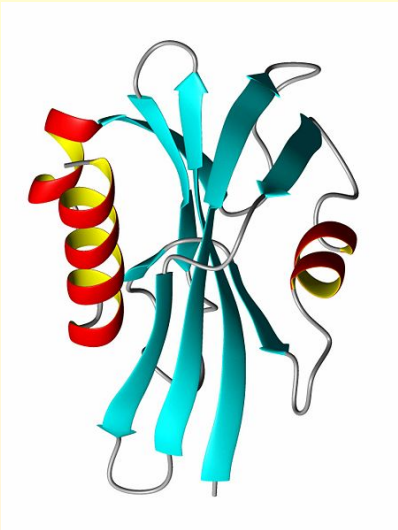
Alcohol Dehydrogenase



Tobacco Mosaic Virus

From the Molecule of the Month feature of the Protein Data Bank.
images ©David S. Goodsell and RCSB PDB

Secondary Structure Prediction



- alpha helix
- beta sheet

Best methods achieve about 80% accuracy.

- Chou-Fasman Method – based on relative frequency of occurrence of amino acids
- GOR Method – uses conditional probabilities based on neighbors
- Machine Learning – uses training sets of solved structures

Tertiary Structure Prediction

Very hard problem!

- Search space is huge.
- Energy function is unknown/circumstantial.
- Direct simulation is not computationally feasible.
- Some proteins fold with the aid of other proteins.
- Some proteins have multiple possible configurations.
- Biological configurations are not always optimal in a thermodynamic sense.
- Optimal thermodynamic configuration may be unachievable (steric hindrance).

Tertiary Structure

Biological Methods:

- X-ray crystallography (only some proteins)
- high-field NMR spectroscopy

Computational Methods:

- Protein structure more conserved than amino acid structure.
- Start with known structure with similar sequence ($> 40\%$).
- *ab initio* techniques
- Critical assessment of techniques for protein structure prediction (CASP).

This problem is well-suited to computational intelligence approaches.

Outline

Cell division and information transmission

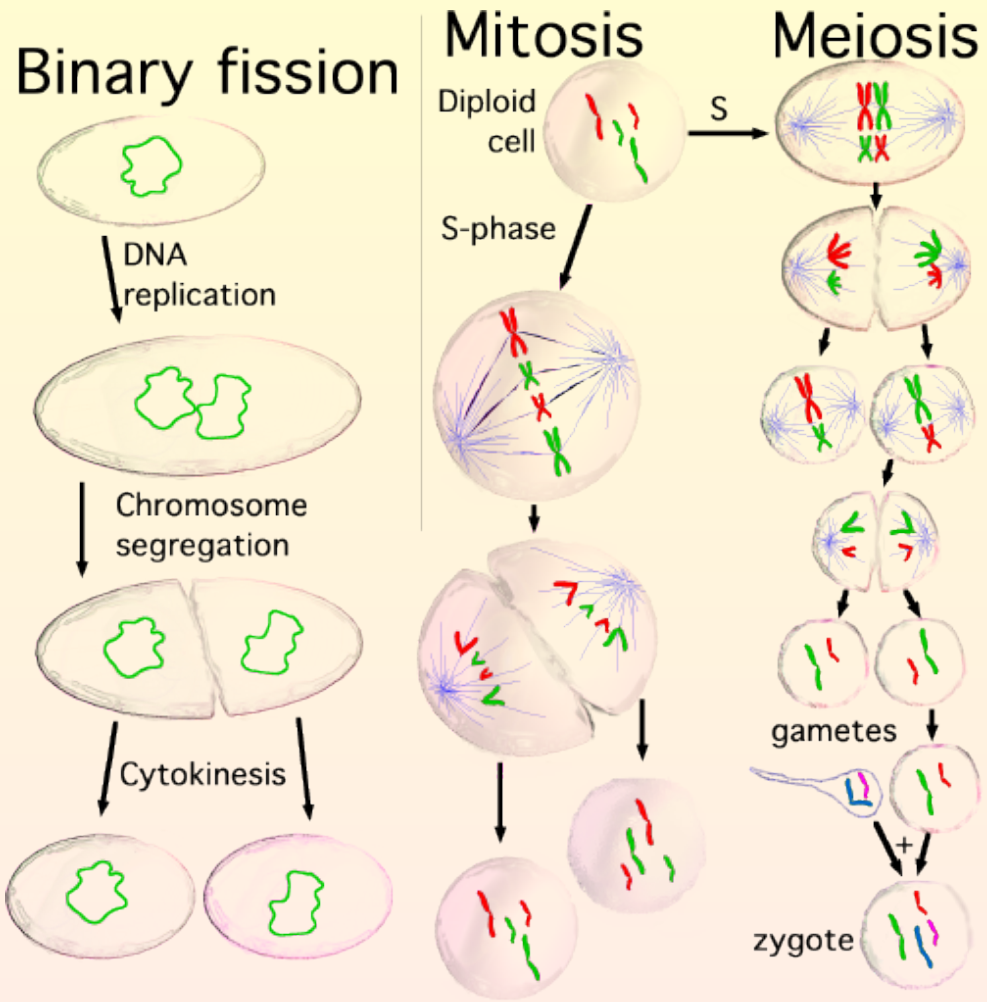


image from J. Schmidt/Wikimedia Commons

- Crucial process
- Evolution
- Regulation

Evolution and Phylogenetics

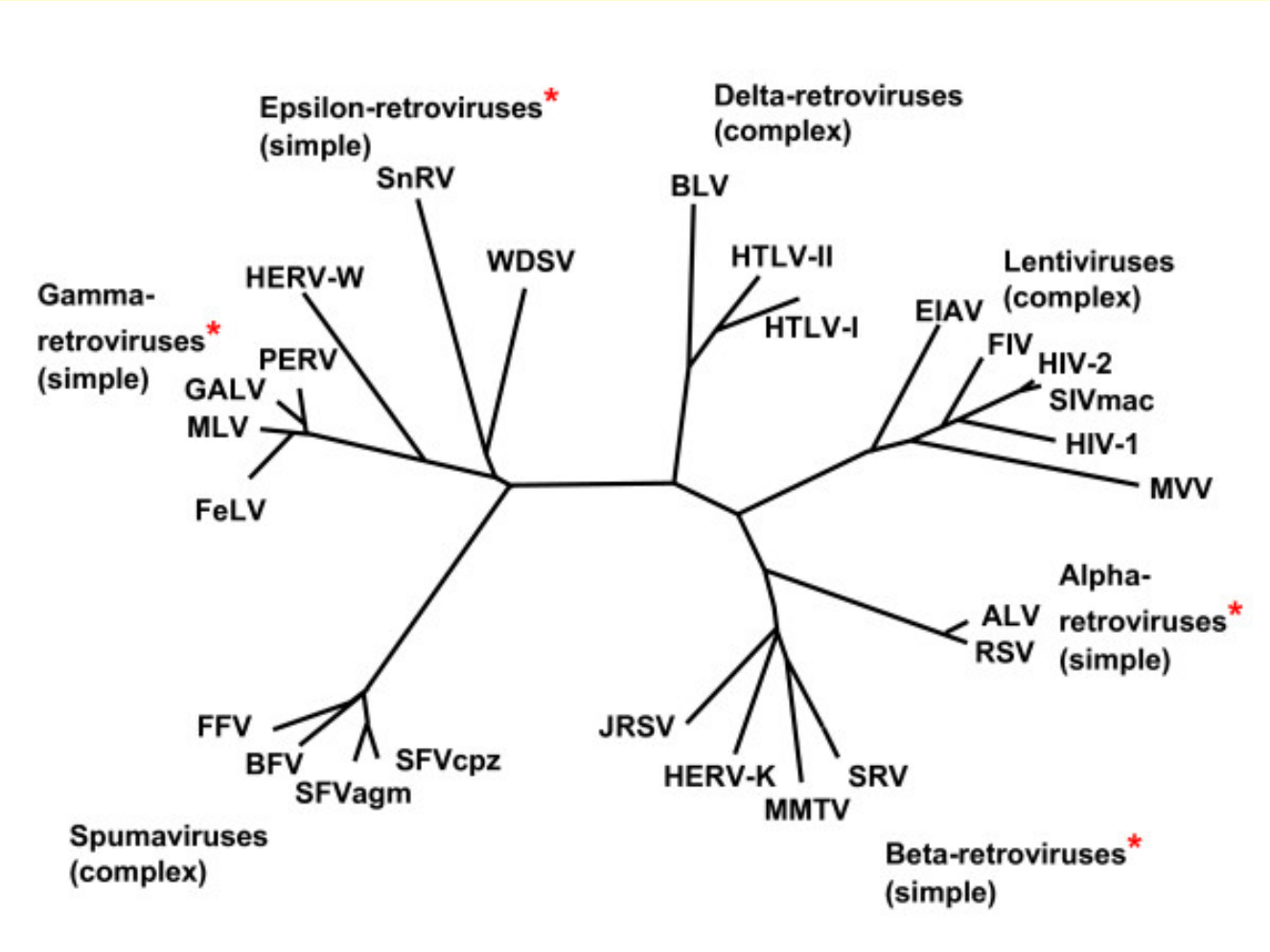
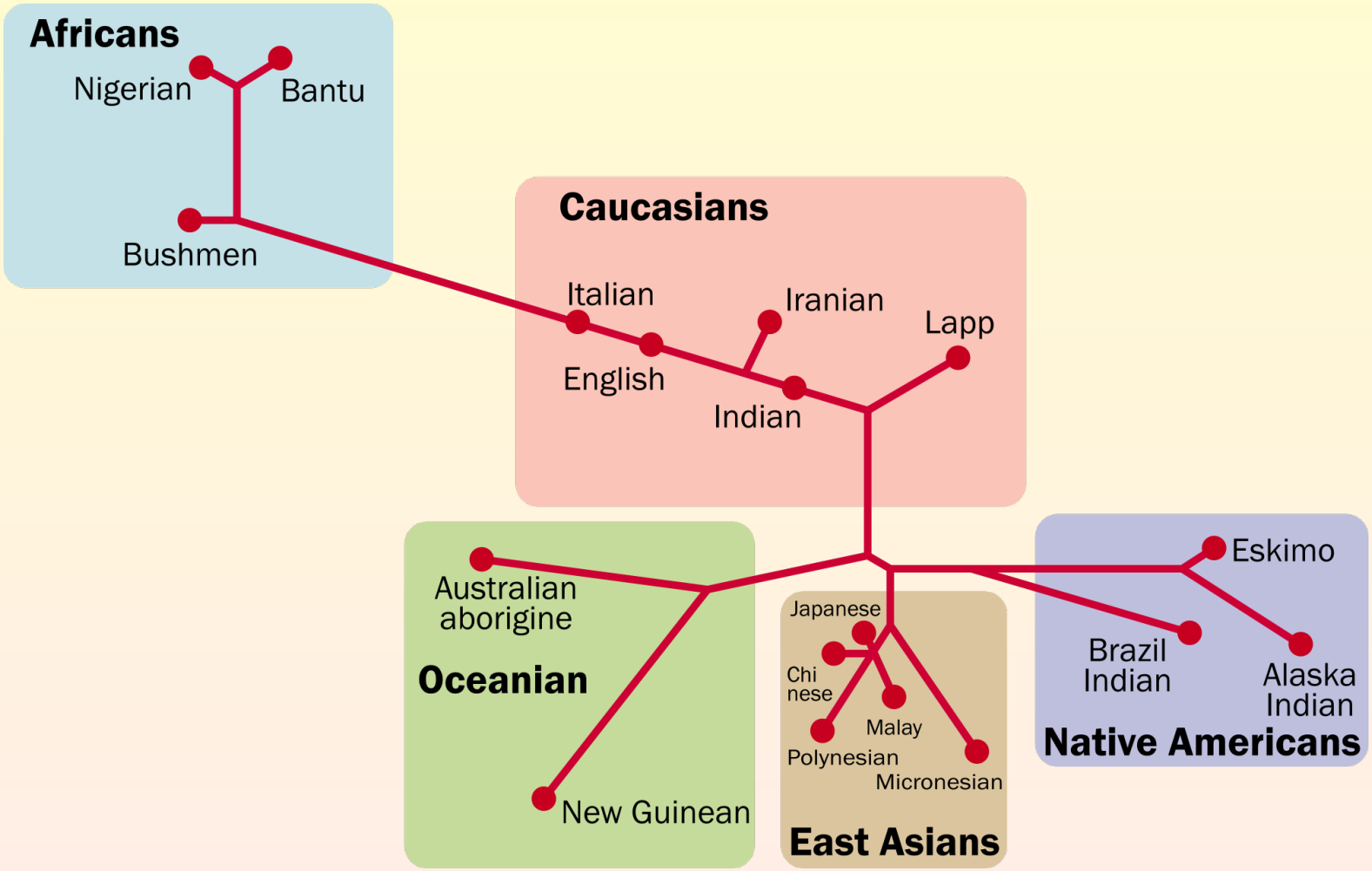


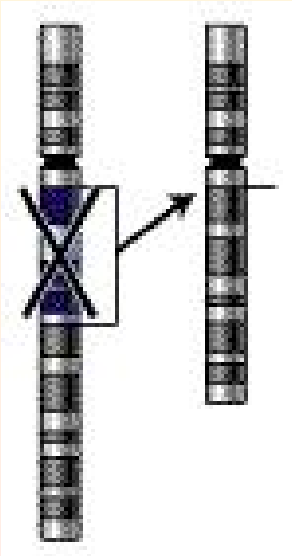
image from R. Weiss/Wikimedia Commons

Evolution and Phylogenetics - 2



Note: 90% of the nodes should be in **Africans**. image from Wikimedia Commons

DNA replication mistakes

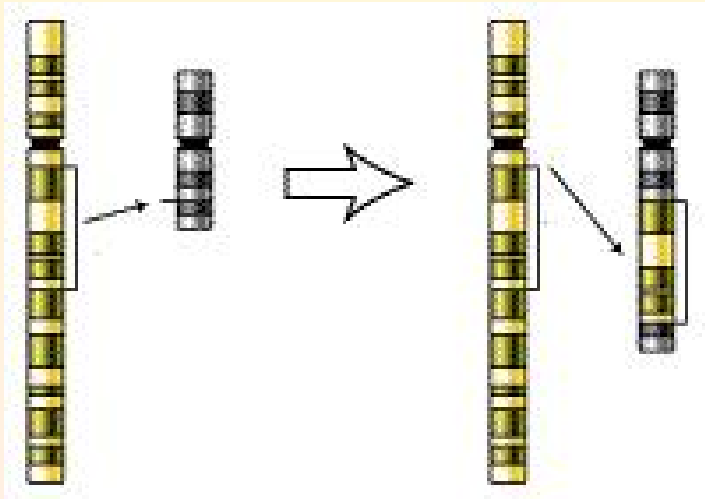


deletion

Two main causes:

- Environmental effects (e.g. chemicals, radiation)
- Spontaneous

Insertion:



Most mutations are self-repaired!

images from Wikimedia Commons

Frameshift mutations

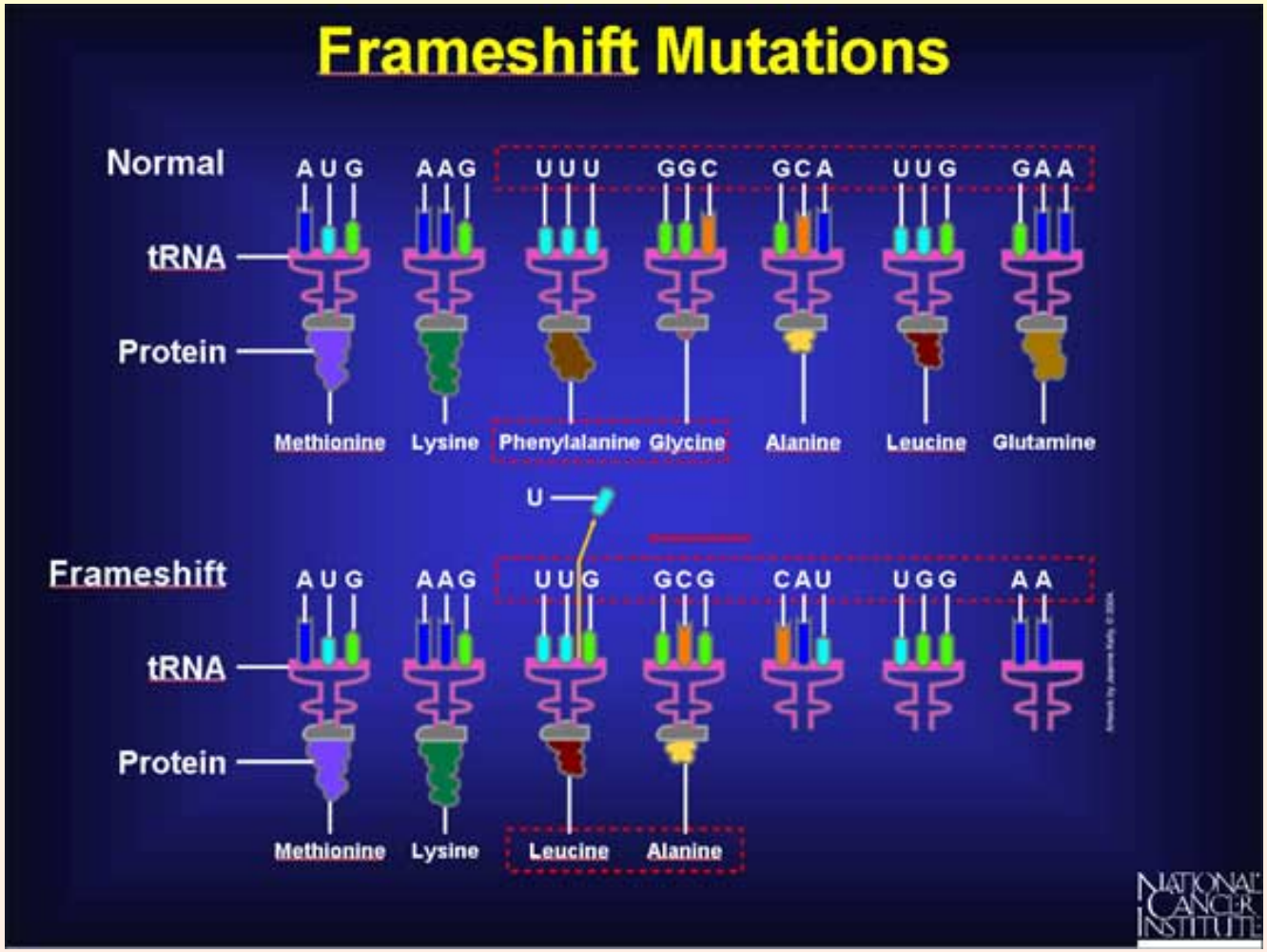


image from `http://www.cancer.gov/cancertopics/understandingcancer/cancergenomics/Slide18`

Outline

Viruses

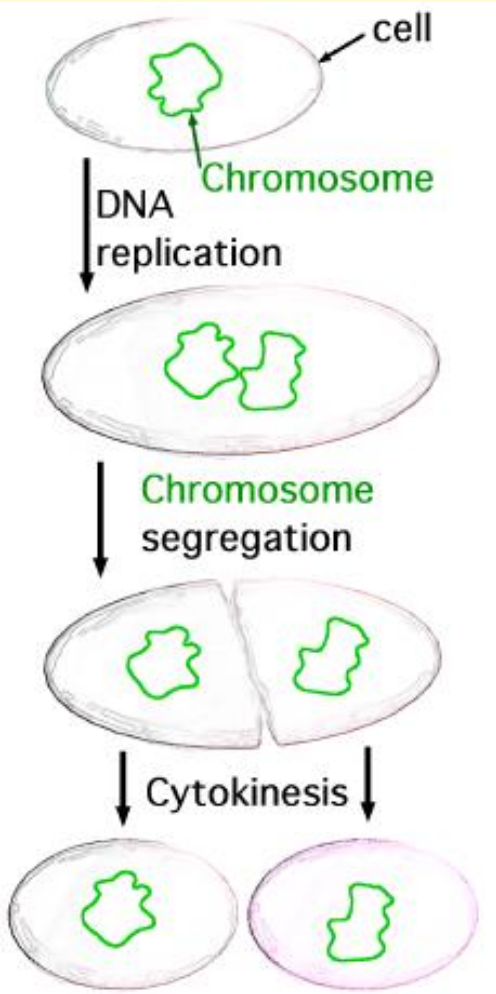


image from http://www.wikihow.com/Know-the-Difference-Between-Bacteria-and-Viruses

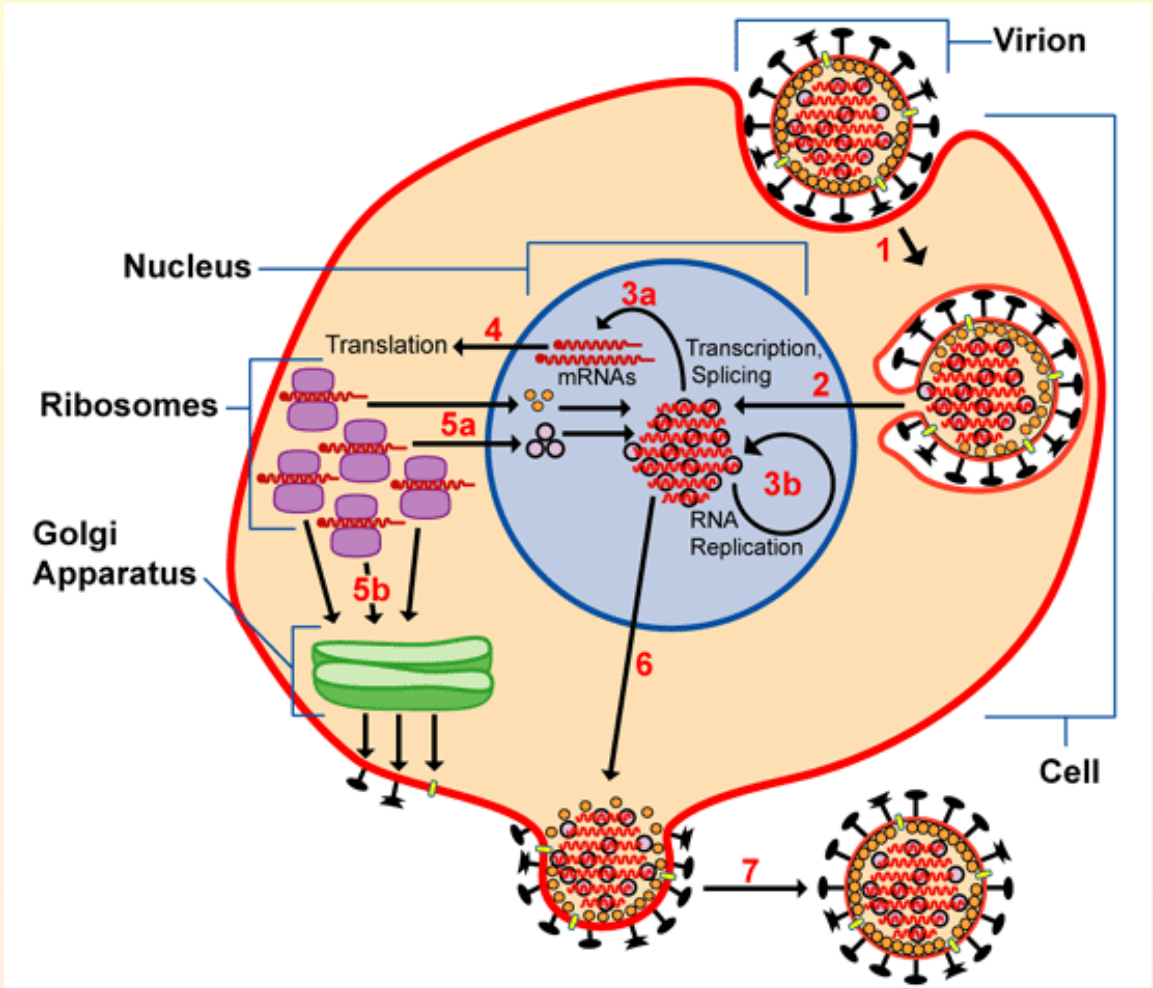


image from Wikimedia Commons

Retroviruses

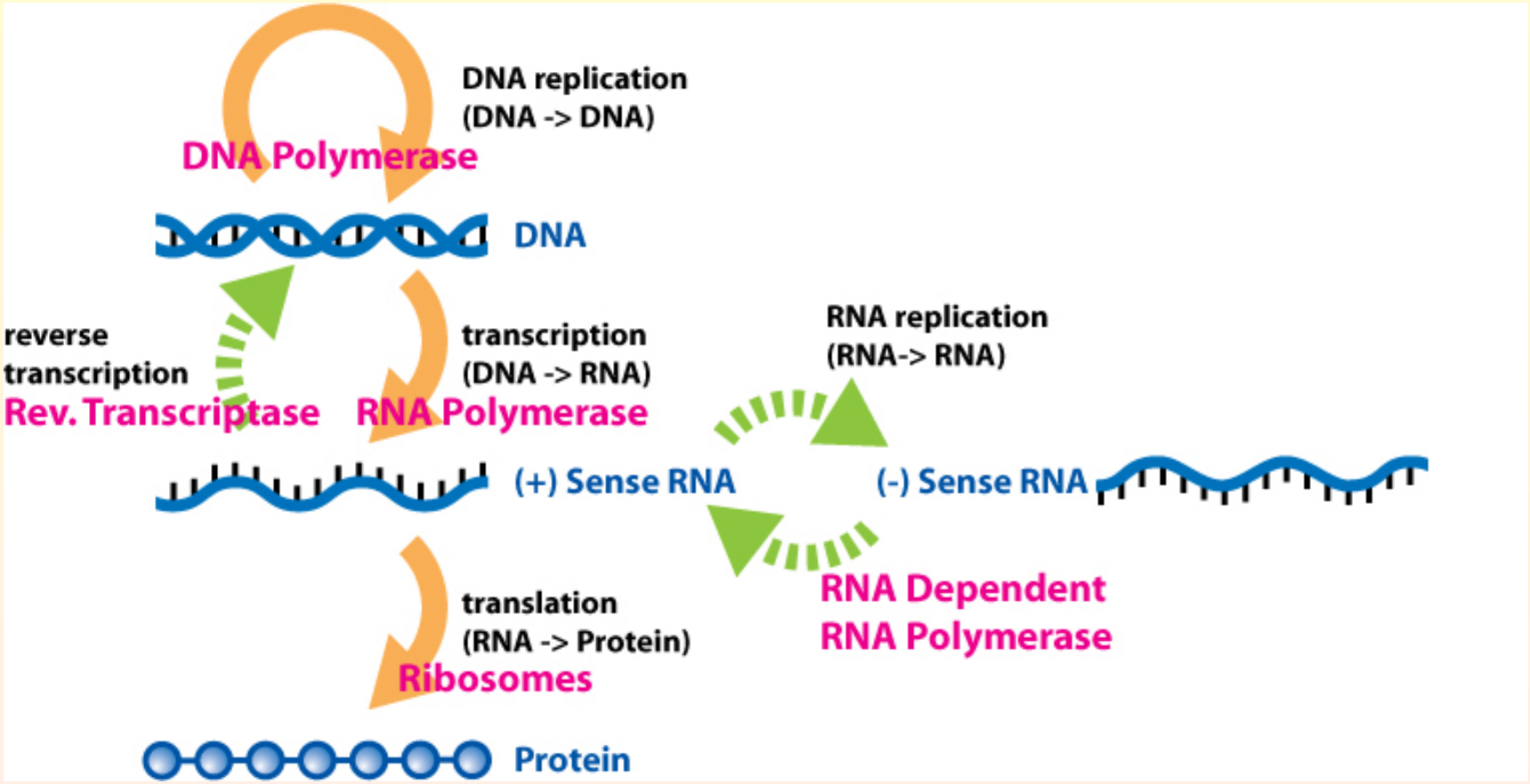
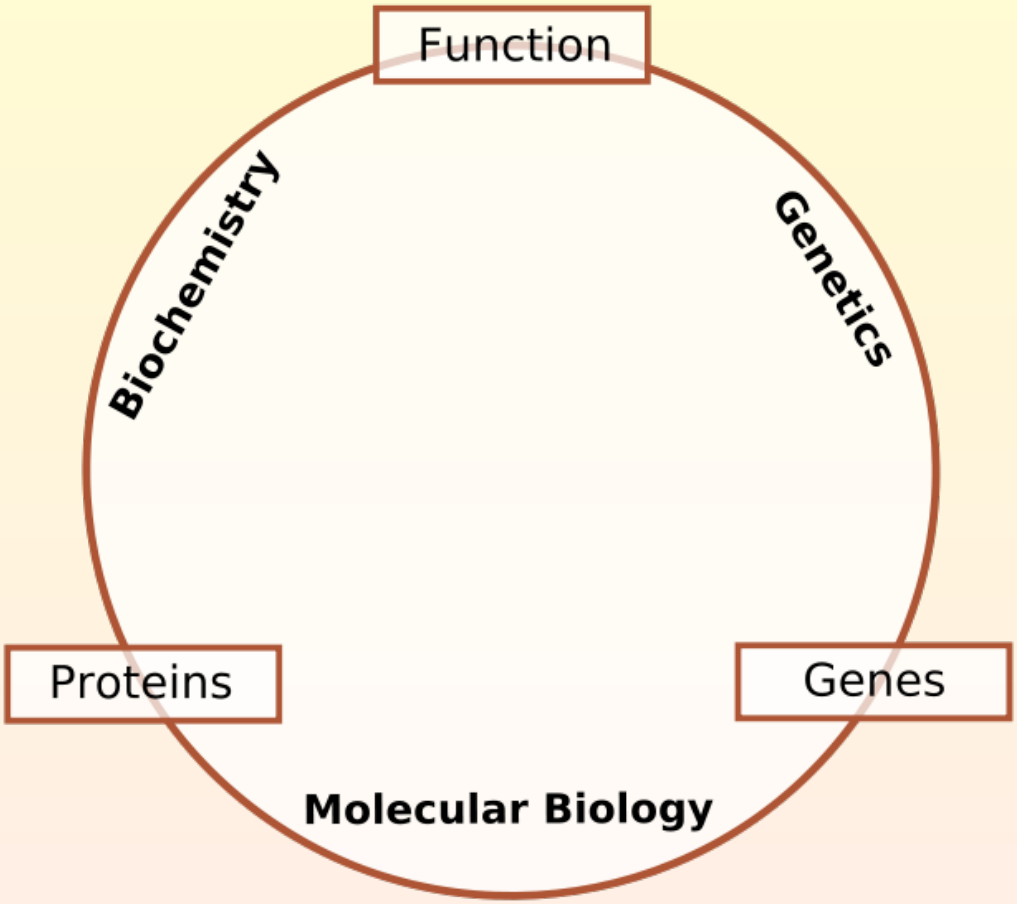


image from D. Horspool/Wikimedia Commons

Outline

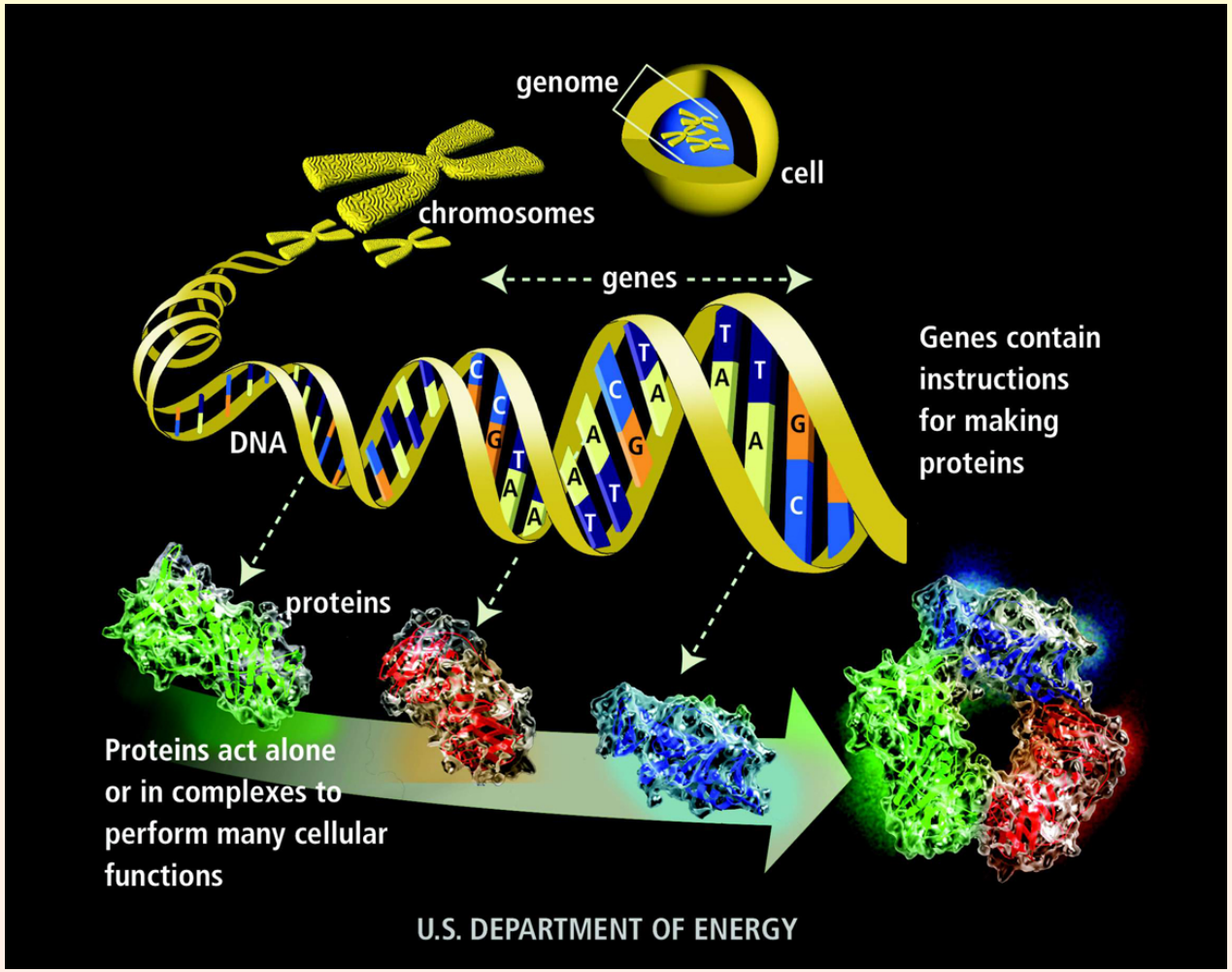
From Structure to Function



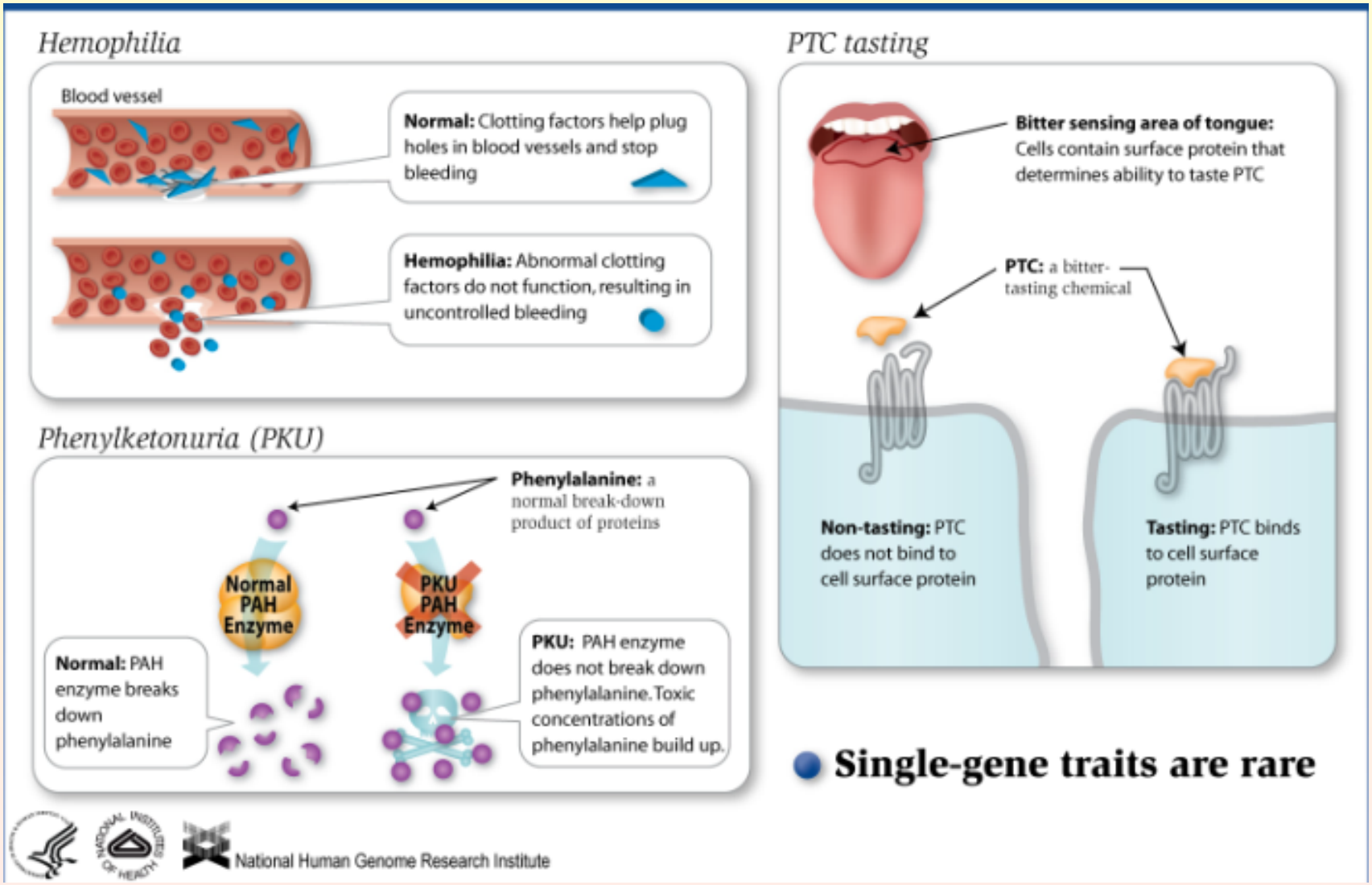
Objectives

- Understand the function of genes
- Use in medicine, genetic engineering,...

Recall that ....


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

The One-Gene-One-Function Theory



The more common scenario

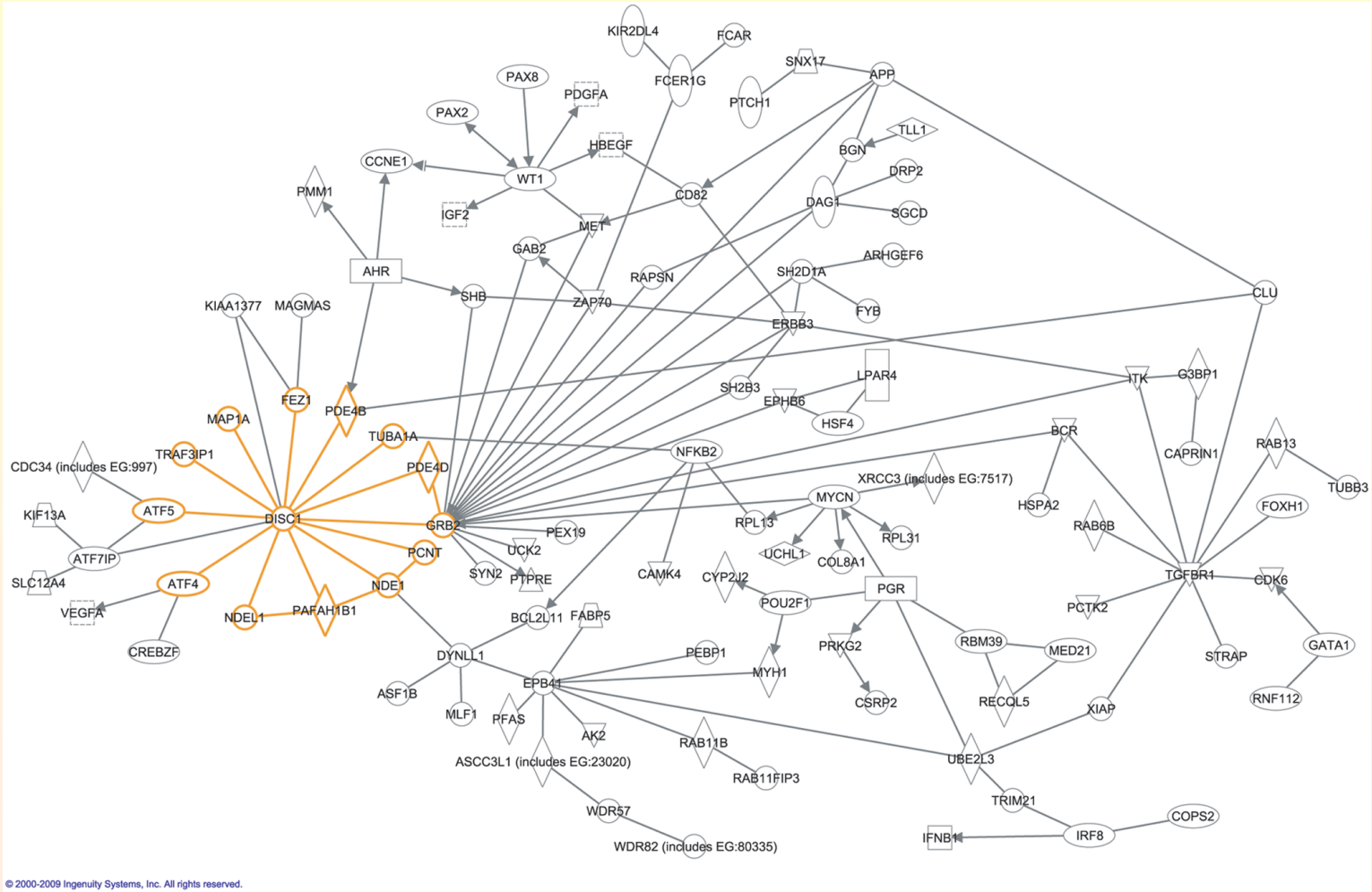


image credits: Hennah, Porteous/Wikimedia Commons

What do the edges stand for?

- Many different interactions
- Broadly two types

Gene network inference is a non-trivial problem.

Engineering done by evolution is very different from engineering done by human beings.

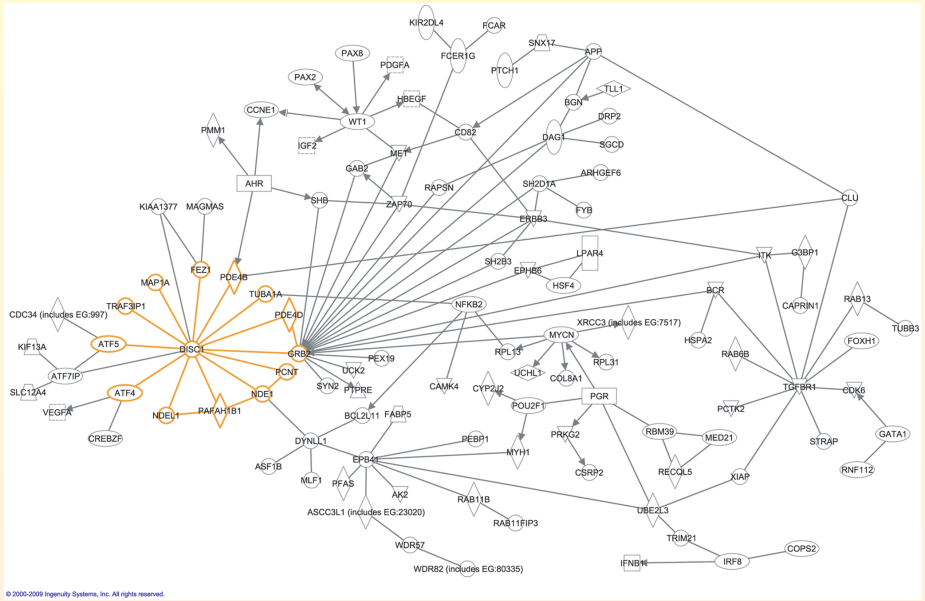
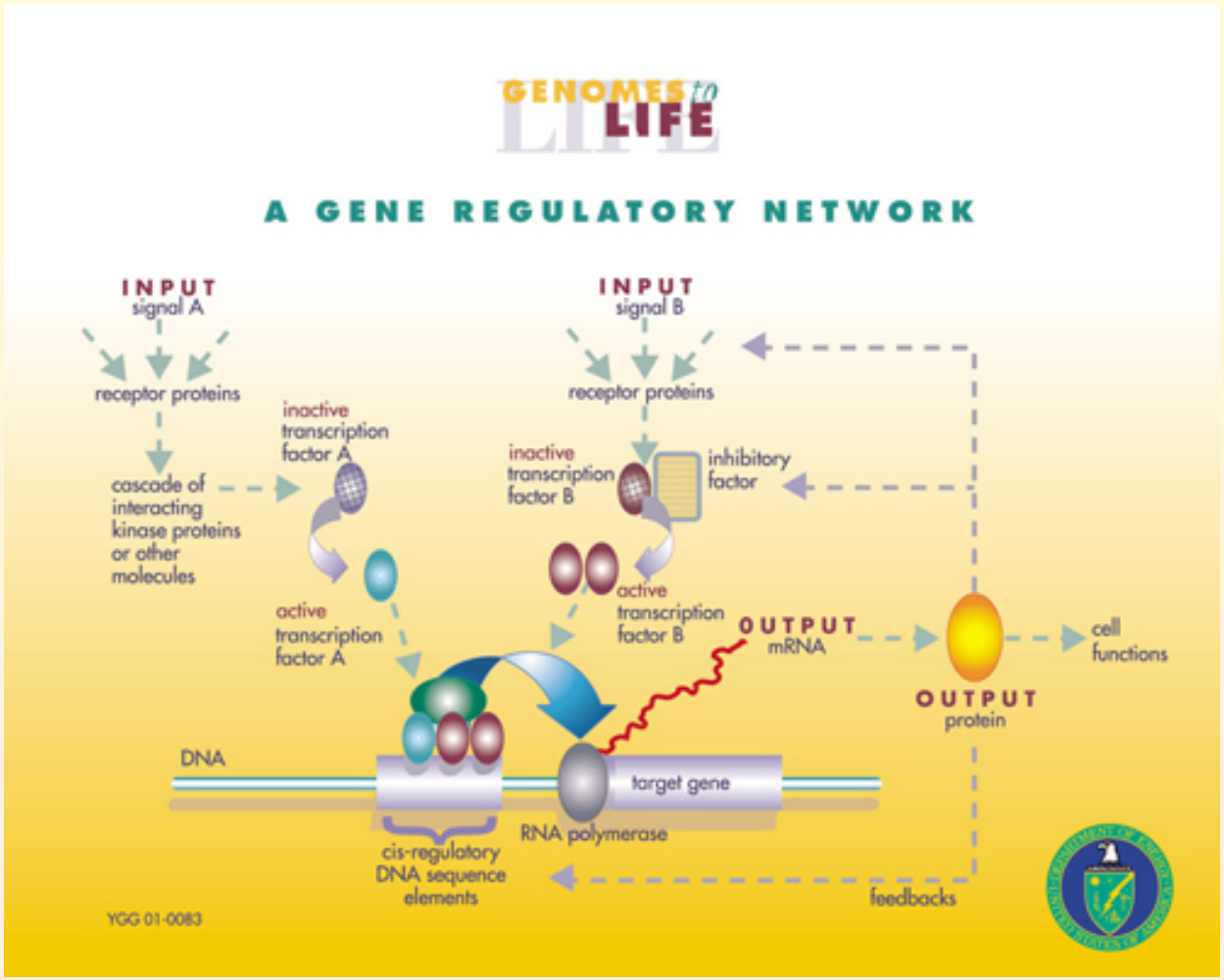


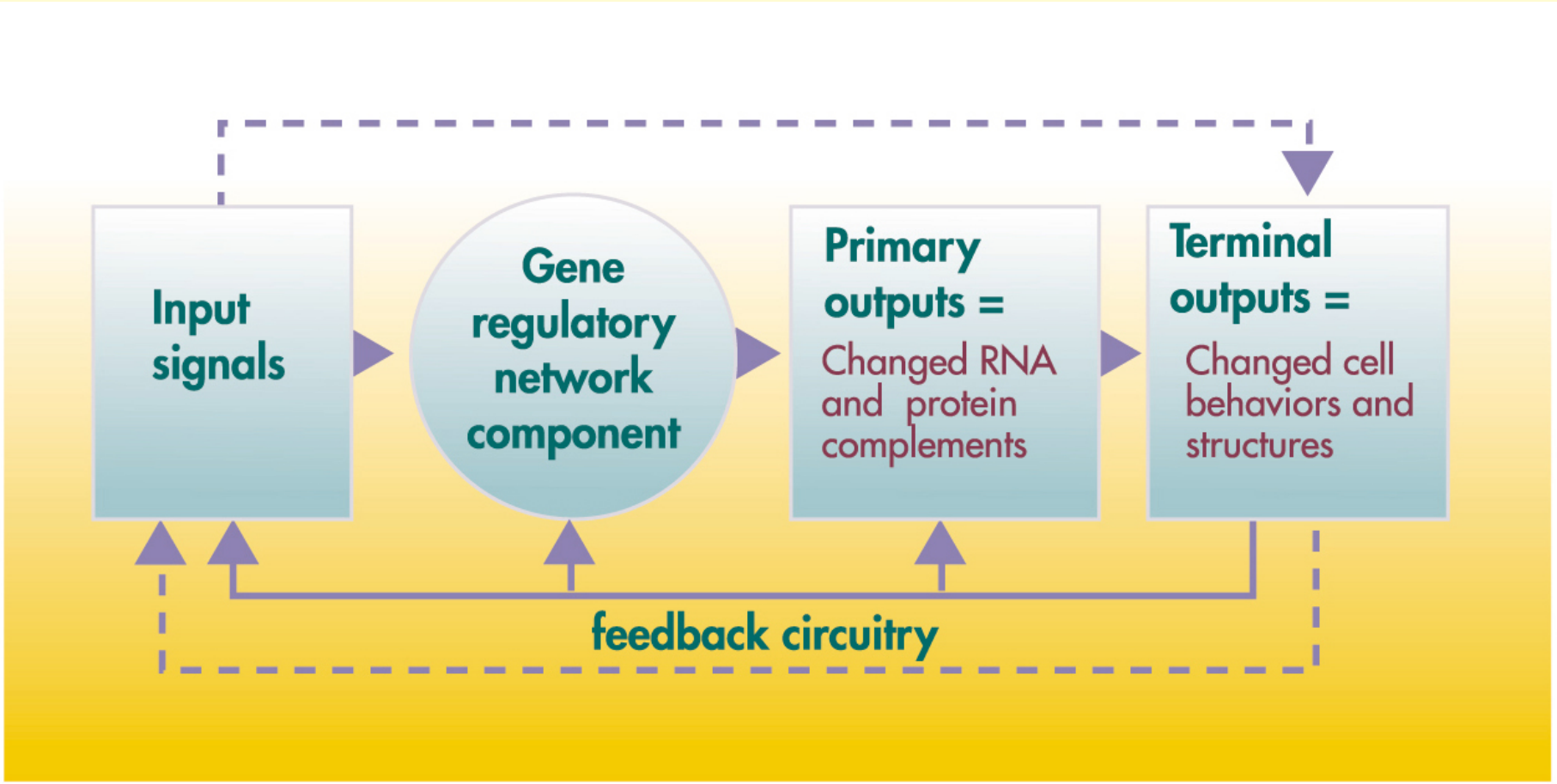
image credits: Hennah, Porteous/Wikimedia Commons

Gene regulatory networks



Understanding the functions of genes cannot happen without understanding the regulation process.

Gene regulatory networks - uses



Inferring Gene functions from sequence structure

- Homology
  - Orthologous
  - Paralogous
- Experimental analysis

Note: The most common way to determine the function of a gene is by knocking it out and observing the effects.

Pathways

Cellular events happen through complicated sequence of events (Pathways).

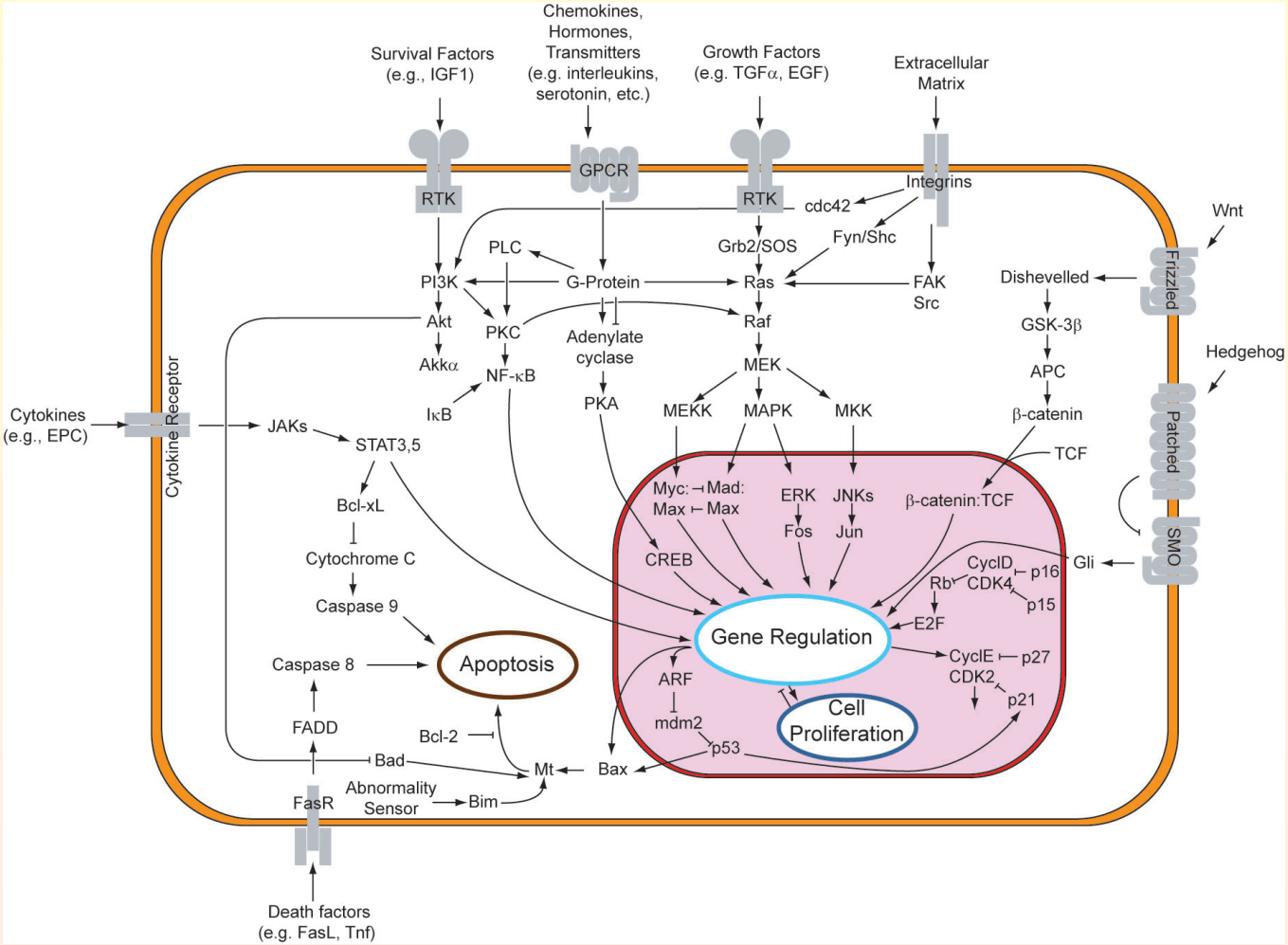


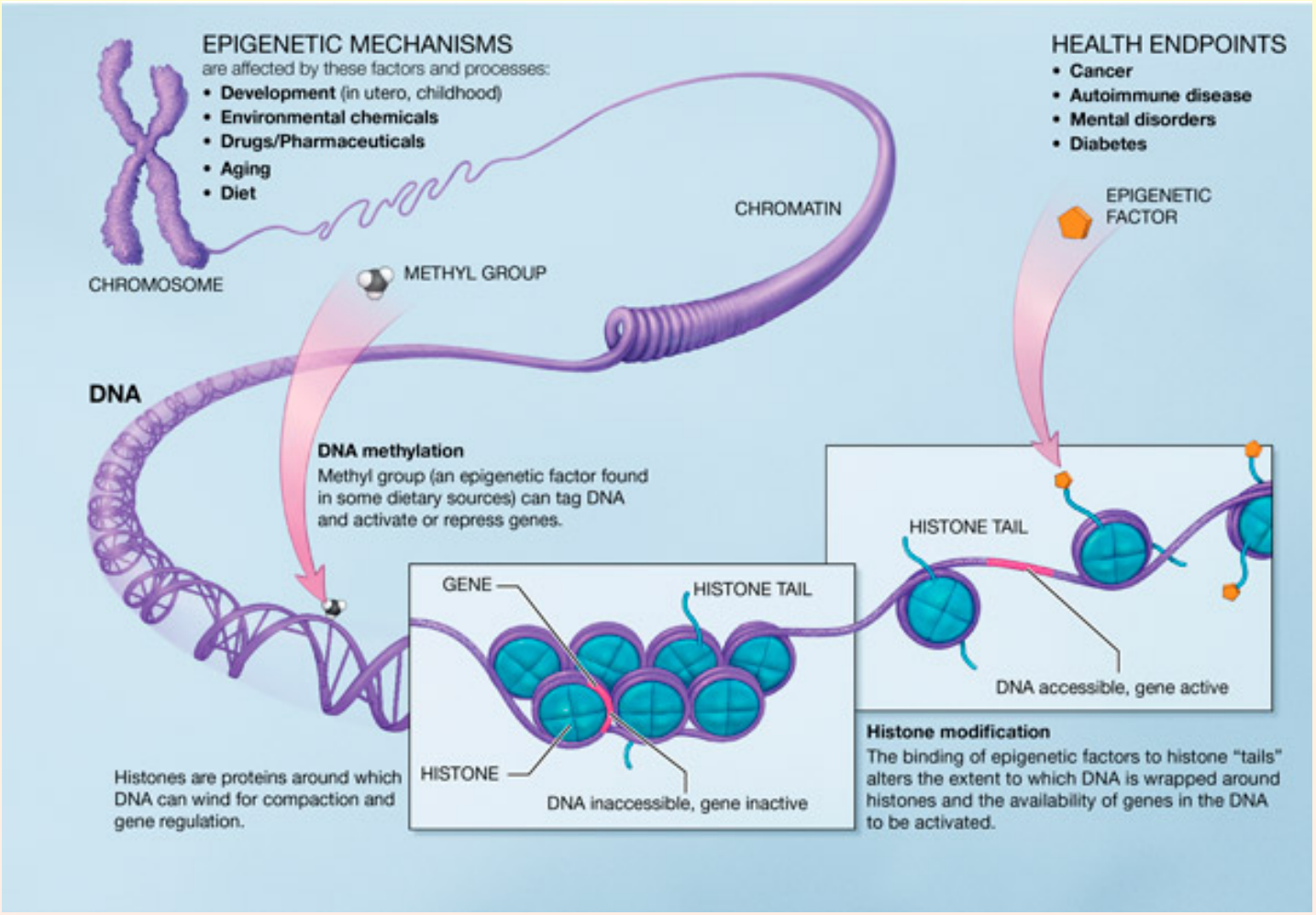
image by Roadnottaken on Wikimedia Commons

Is that the whole story?

Q: Does the genome tell us all that we wish to know?
Q: Since identical twins have the same DNA, why do they not have identical vulnerability to diseases?
Q: What role does the environment play and how?

Epigenetics provides some answers.

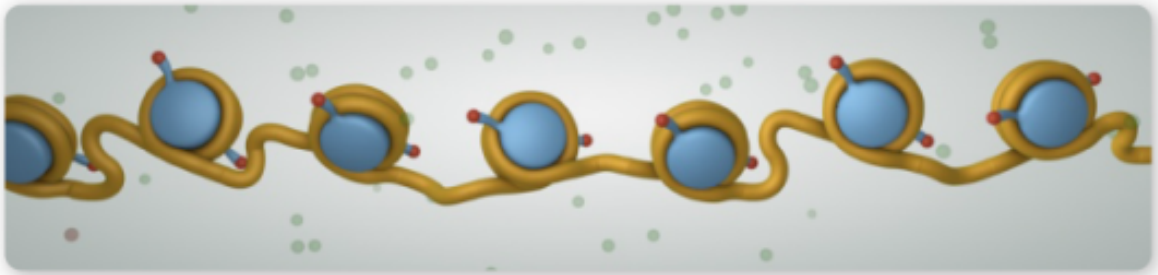
Epigenetics



picture from `http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg`

Epigenetics - continued

The epigenome tightly wraps **inactive** genes, making them unreadable.

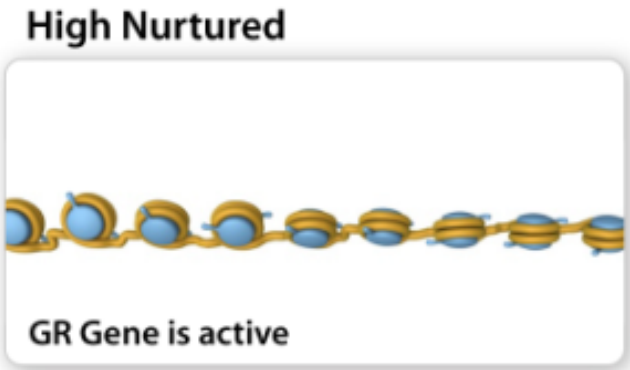


The epigenome relaxes **active** genes, making them easily accessible.

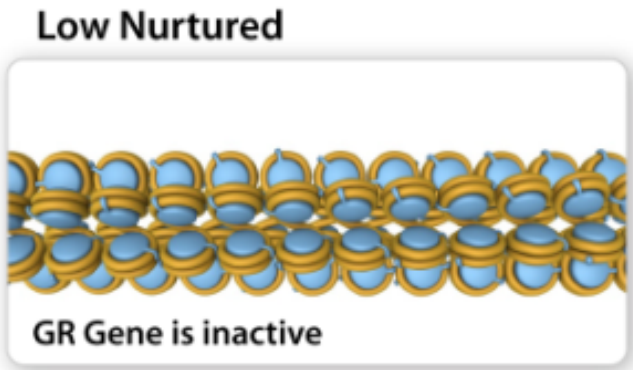
National Human Genome Research Institute

Epigenetics

Gene expression is affected by diet, toxins, physical activity, stress.



In high nurtured rat pups, the GR gene is active. These rats have an easy time relaxing after stress.



In low nurtured rat pups, the GR gene is epigenetically silenced. These rats have a hard time relaxing after stress.

When it's active, the GR gene produces a protein that helps the body relax after stress. Mom's nurturing during the first week of life shapes her pups' epigenomes.

National Human Genome Research Institute

(Wever et al, 2004. Nature Neuroscience, 7, 847)

Other -omes and -omics

- Metagenomes
- Interactomes
- Metabolomes

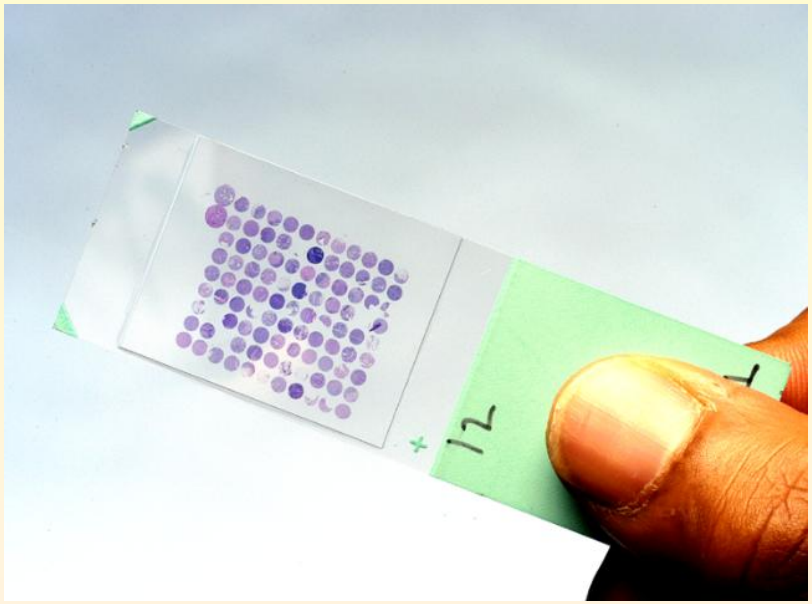
Outline

High throughput experiments

- Many modern methods produce enormous amounts of data.
- Computational scientists, in collaboration with biologists, are needed to analyze the data.

Understanding the Biological basis of the experiments is necessary.

Microarrays

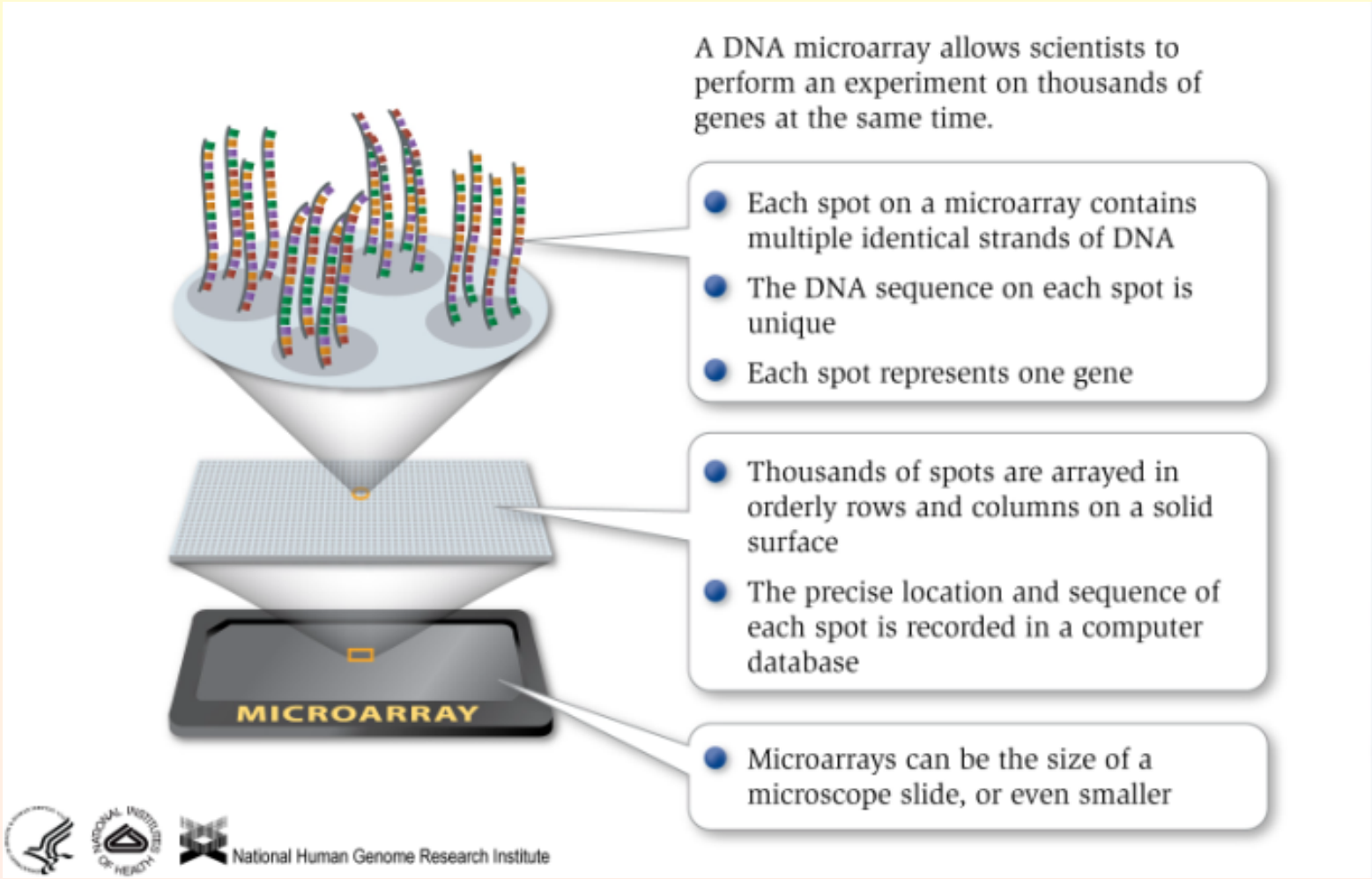


images from Ricardipus on Wikimedia Commons and Johns Hopkins Tissue MicroArray Core Facility

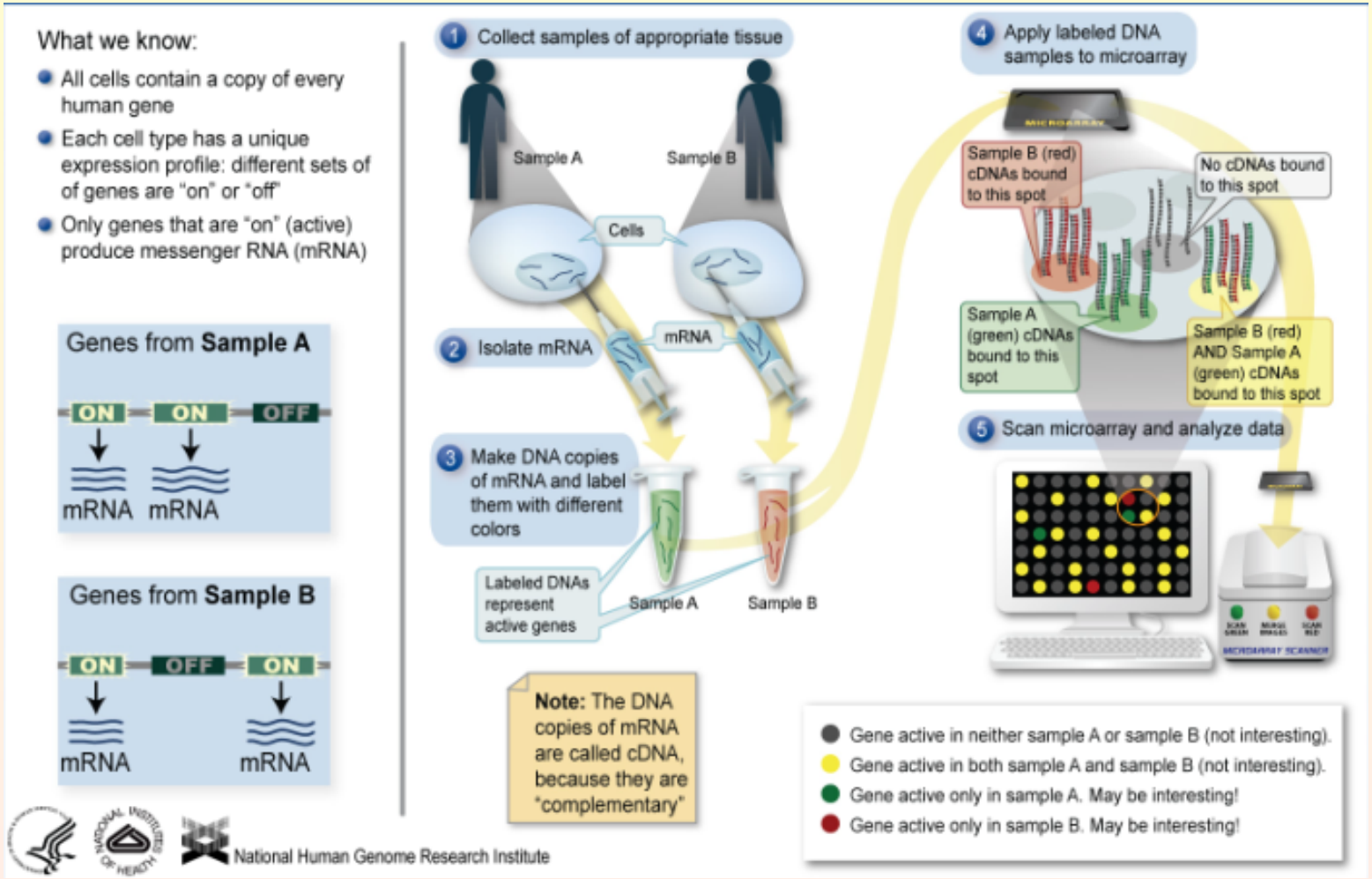
Basic idea: Each dot contains genetic material; many dots tested simultaneously (up to 22K per chip)

Approach: Measure gene expression values.

Microarrays - continued



Microarray usage - example



Microarrays

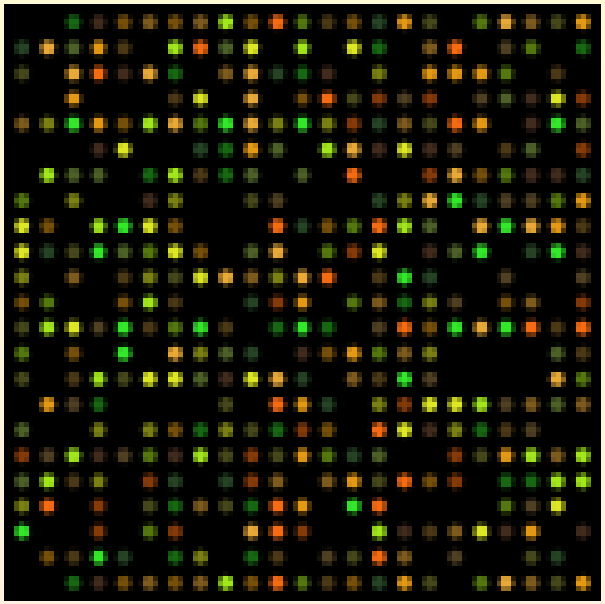


image from Guillaume Paumier/Wikimedia Commons

What does this tell us?

- Gene distance via profiles
- Gene distance via dendrograms
- Correlation with diseases

Data processing challenges

- Similarity measure
- Normalization
- Interpreting the dots
- Dimension reduction

Protein-protein interactions

- Graph representing interactions
- Many different notions of interaction
- Signal transduction
- Protein complexes

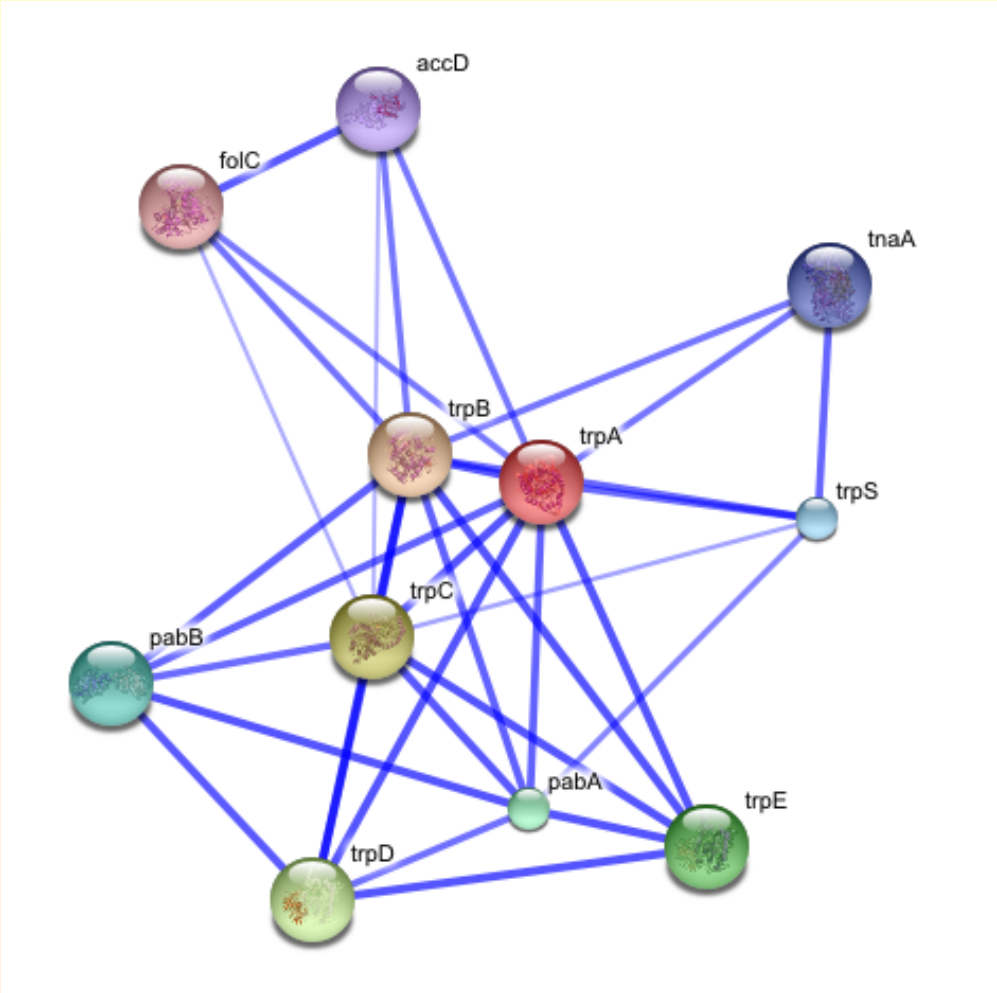


image from Wikimedia Commons

Outline

Computational problems

- Gene prediction
- Pseudogenes, non-coding RNA prediction
- RNA/Protein Structure prediction
- Retrovirus prediction
- Finding CpG Islands
- Detecting repeats

CpG Islands

CpG site: –C–phosphate–G–

- high frequency of CpG sites
- no consensus definitions
- 70% of human promoters have a high CpG content

Detecting Repeats

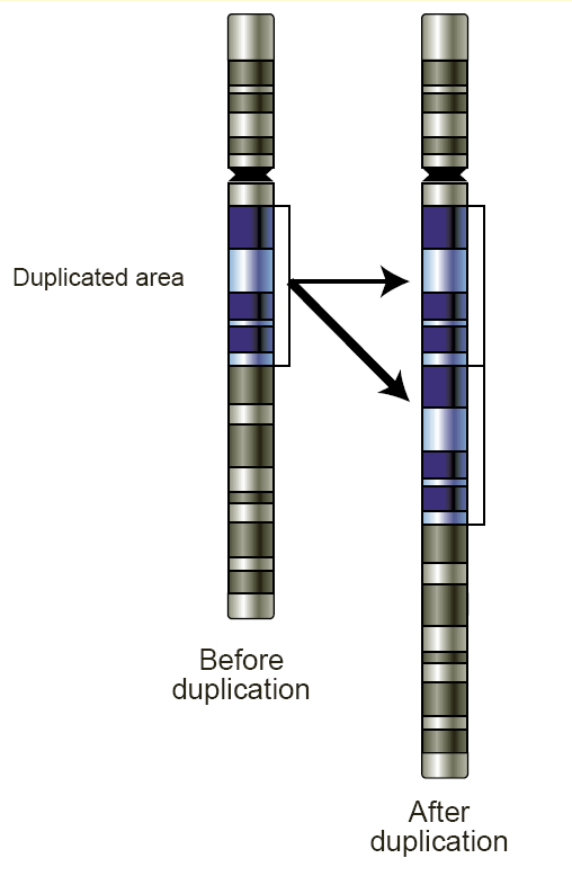


image from Wikimedia Commons

- Many different types of repeats
- Different copies are not exact replicas
- Copy number variation

Forensic DNA analyses focus on short tandem repeats.

Computational Intelligence in use

- Robust string matching : alignment, repeat finding, Motif finding
- Statistical techniques (Bayesian inference, MCMC, MLE): Phylogeny (tutorial this afternoon)
- Genomic Signal Processing : Gene prediction, retrovirus prediction
- Finding structures in data (SVD, PCA, mixture models) : microarray data analysis
- Unsupervised learning: microarray data analysis, CpG island detection
- Supervised learning (HMM, Neural Nets) : Motif finding, Gene prediction
- Many other techniques will be presented in the talks at this conference.

Credits

- Some slides are modified from a tutorial given by Wendy Ashlock and I at CIBCB 2010
- Some slides are modified from those at the book website.