Lecture 1: Course Preliminaries & Information in Biological Systems

Bioalgorithms Fall 2013

Read Chapter 1
and Chapter 3.1-3.7

Comp 555 Intended Audience

- Comp 555: Bioalgorithms
  - Suitable for undergraduate and graduate students
  - CS majors who want to learn bioinformatics
  - Non CS majors from the statistical of biological sciences who are interested in the algorithms used in bioinformatics.
  - BCB/BBSP students

Why?

- Benefits for Computer Scientists
  - See CS fundamentals applied to real problems
  - What computer scientists can learn from biology
    - Robust, parallel, self-repairing, and energy efficient
- Benefits for Biologist
  - Help to close the CS-Bio "language" gap
  - Appreciate CS as more than "coding"
  - What is a correct algorithm? An efficient one?
- Growth Potential
  - Bioinformatics is a very marketable skill
  - Future of CS and Biology

What Will We Learn?

Algorithm and data structures

Data Abstraction
Classic Data Structures
Lists, Queues, Heaps, Graphs, Trees, Hash tables
Program Correctness and Efficiency
Time and Space Complexity
Intractable Problems

Molecular Biology Basics

Biological systems as machines
Information in biological systems
DNA, nucleotides, codons, & genes
mRNA transcription and translation
Protein folding and function
Genetic variation
Gene expression and regulation

Algorithm Design Approaches

Exhaustive Search,
Branch & Bound,
Greedy Algorithms,
Dynamic Programming,
Divide-and-Conquer,
Data-driven Probabilistic Modeling,
Randomized Algorithms

Bioinformatics Problems

Restriction Mapping
Motif Finding
Sequence Alignment
Gene Prediction
Sequencing by Hybridization
Spectrum Graphs
Gene Expression Analysis

Course Logistics

- Website:

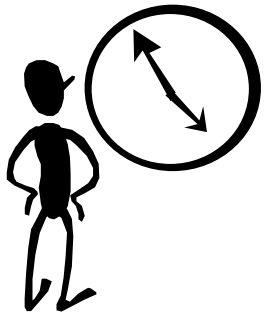
  http://csbio.unc.edu/mcmillan/?run=Comp555F13

  look here first for
  - News, hints, and helpful resources
  - Revisions, solutions, and corrections to problem sets
- Office Hours: TBA
- Grading

  5 – Problem sets (worth 10% each)

  Midterm Exam (worth 25%)

  Final Exam (worth 25%)
- Problem Sets
  - Roughly one every three weeks
  - Will include a short program to write

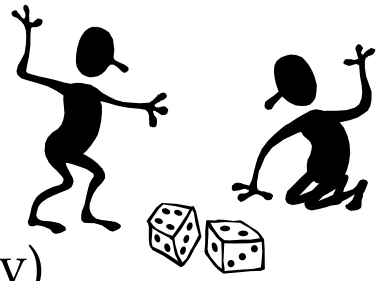


Bioinformatics = Biology + Information

- What is information?
  - Information: that which resolves uncertainty
  - We measure information in bits
  - Information = $-\log_2(\text{probability})$
    - The coin I tossed landed heads. How many bits?
    - You rolled a 7 on a pair of dice. How many bits? You roll a 3?

6 ways out of 36 to roll a 7. Thus, 7 conveys -log(6/36) = 2.58 bits. A roll of 3 conveys -log(2/36)= 4.17 bits

  - Concrete systems need mechanisms for
    - Reliably storing information (memory)
    - Reliably processing information (logic)
    - Reliably transporting information (connectivity)



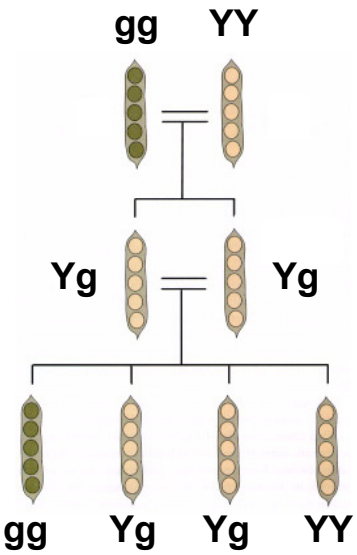
  - The focus of computer science is information
- How about biological systems?

There Must Be Information!

In biological systems…

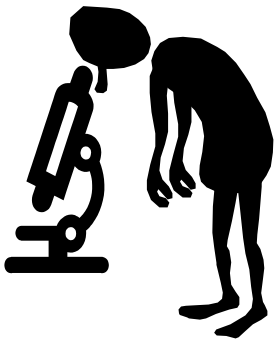
- Information is somehow passed between successive generations of plants and animals
- Distinguishable traits are inherited (phenotypes)
- Latent (recessive) traits can be masked by dominant traits, yet reappear in later generations
- Heredity

Gregor Mendel
1822-1884



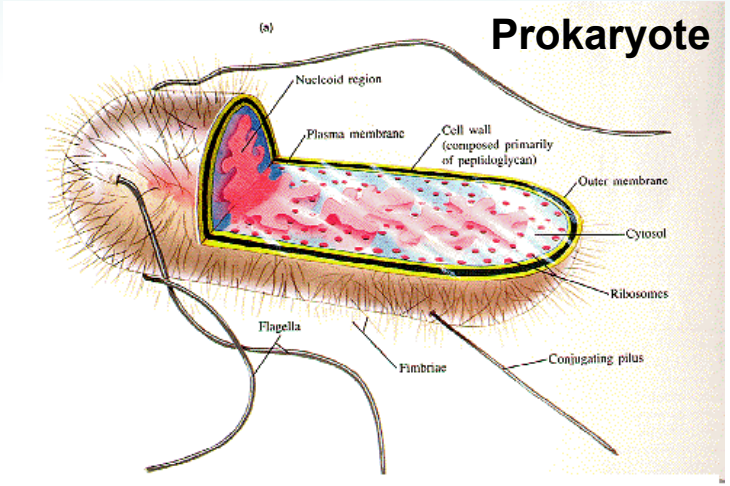
Where is the Information?

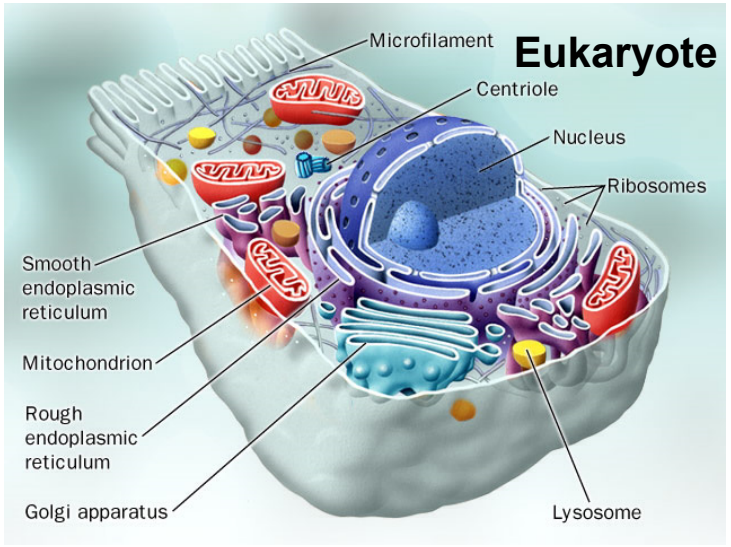
- Enter the Age of Microbiology
- Cells: the smallest structural unit of a living organism capable of functioning independently
- Cell composition by weight
  - 70% water
  - 7% small molecules
    salts, amino acids, nucleotides, lipids (fats, oils, waxes)
  - 23% larger polymers
    proteins, polysaccharides
- Two types
  - Prokaryotes: bacteria (no nucleus)
  - Eukaryotes: yeast, plants, and animals (with nucleus)



Looking for the Source of Heredity

- In 1879 Walther Fleming, a German anatomist, discovered threadlike structures clearly visible during cell division. He called this material '*chromatin*,' which was later called '*chromosomes*'
- Zoologists Oskar Hertwig and Herman Fol first observed the process of fertilization in detail in the early 1880s. In 1881, Edward Zacharias showed that chromosomes contained *nucleic acids*. In 1884, Hertwig wrote "*nucleic acid is the substance that is responsible not only for fertilization but also for the transmission of hereditary characteristics.*"





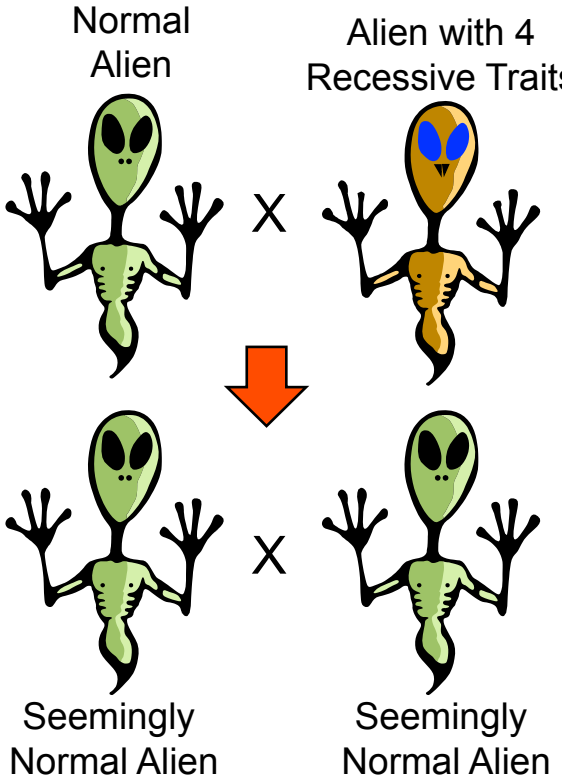
Early 20th Century Genetics

- In 1908, Thomas Hunt Morgan and Alfred H. Sturtevant showed that *genes* were located on chromosomes. Experimenting with Drosophila (fruit flies) they found sex chromosomes, sex-linked traits, and crossing-over. They were able to associate mutations to specific chromosomal regions, thus mapping gene locations.

- By the 1930's biochemists knew that the nucleic acid present in chromosomes was DeoxyriboNucleic Acid, DNA. They also knew that chromosomes contained proteins in addition to DNA. DNA appeared to be long repetitive chains, and therefore, it seemed unlikely to carry information. Proteins, however, looked more interesting and were generally assumed to contain genetic materials. DNA was considered as just some sort of glue.

Inferring Genetic Maps

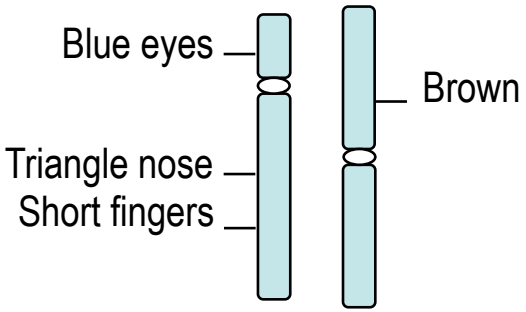
- Even without knowing the mechanisms of how heredity information is represented, clever scientists (Morgan) were able to "map" genes…



Normal	201	Short-fingered	9
Brown	64	Brown, blue-eyed, & triangle nosed	6
Blue-eyed	58	Triangle-nosed	5
Short fingered & triangle-nosed	54	Short-fingered & Brown	5
Short fingered, triangle-nosed, & brown	21	Brown, short fingered, triangle-nosed, & blue-eyed	4
Short fingered, triangle-nosed, & blue-eyed	20	Short-fingered & blue-eyed	4
Brown & blue-eyed	19	Triangle-nosed & brown	1
Blue-eyed & triangle-nosed	12	Brown, short fingered, & blue-eyed	1

Steps to Infer a Genetic Map

- Verify Mendelian ratios
  - Brown (64 + 21 + 19 + 6 + 5 + 4 + 1 + 1) /484 = 0.250
  - Blue-eyes (58 + 20 + 19 + 12 + 6 + 4 + 4 + 1) / 484 = 0.256
  - Triangle-nose (54 + 21 + 20 + 12 + 6 + 5 + 4 + 1) / 484 = 0.254
  - Short-finger (54 + 21 + 20 + 9 + 5 + 4 + 4 + 1) / 484 = 0.244
- Test for pairwise linkages
  (we'd expect ¼ × ¼ × 484 ≈ 30 if independent)
  - Short-finger & triangle-nose 54 + 21 + 20 + 4 = 99
  - Triangle nose & brown 21 + 6 + 4 + 1 = 32
  - Short-finger & brown 21 + 5 + 4 + 1 = 31
  - Blue-eyes & triangle-nose 20 + 12 + 6 + 4 = 42
  - Short-finger & blue-eyes 20 + 4 + 4 + 1 = 29
  - Brown & blue-eyes 19 + 6 + 4 + 1 = 30
- Indicates
  - Short-fingers & triangle-nose are closely linked
  - Blue-eyes & triangle-nose are probably linked
  - Short-finger & blue-eyes appear independent, thus the triangle nose gene should lie between them
  - Brown gene is likely to be on another chromosome.

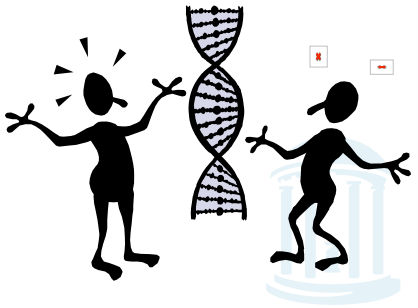


Morgan came up with even more clever techniques that were able to precisely locate the relative positions of genes on chromosomes. Even today chromosomal gene positions are measured in units of centiMorgans

DNA's Central Role

- In 1944, Oswald Avery showed that DNA, not proteins, carries hereditary information.
- In the late 1940's and early 50's Linus Pauling and associates develop modeling methods for simultaneously determining structure and chemical make-up of proteins and other large molecules.
- In 1952, James Watson and Francis Crick, are able to determine the structure and chemical makeup of DNA, using X-ray crystallography data collected by Rosalind Franklin and Maurice Wilkins.

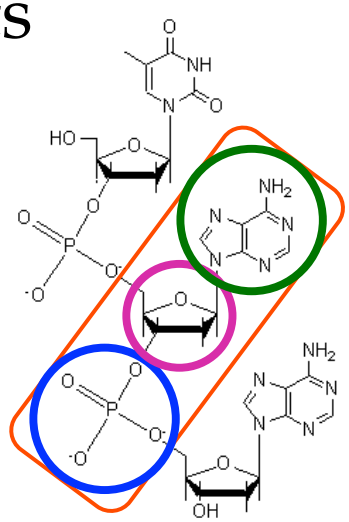
Beginning of Molecular Biology!



Details of DNA

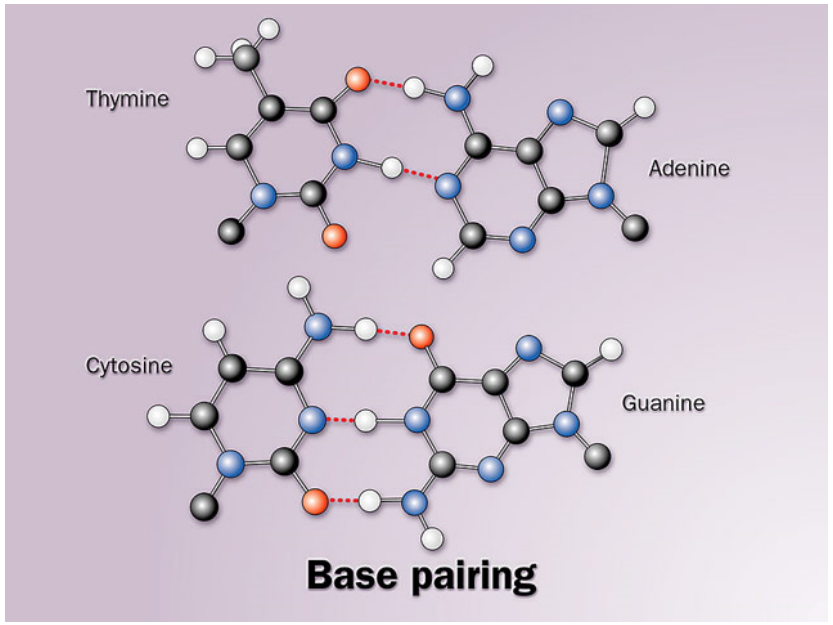
- The information stored in DNA organizes inanimate molecules into living organisms and orchestrates their lifelong function
- A long complementary chain of nucleotides

- Each nucleotide has 3 components
  - A phosphate group
  - A ribose sugar
  - One of four nitrogenous bases
- The information resides in the variation of bases



DNA Components

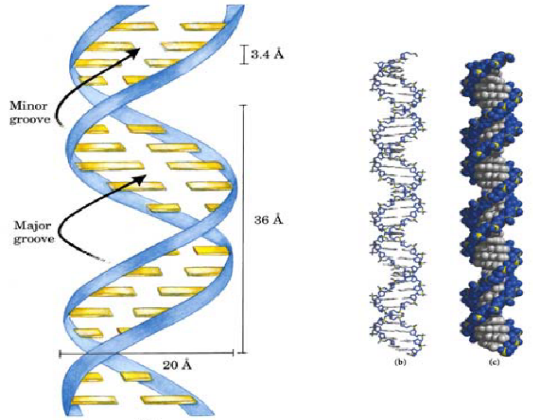
- The same DNA encodes all living organisms
- Different code sequences distinguish
  - Plants from animals
  - Species
  - Individuals
- A complete DNA sequence for an organism is called its *genome*
- Code sequences are composed of 4 bases (Adenine, Cytosine, Guanine, Thymine)
- Each base binds with another specific base (Thymine with Adenine and Cytosine with Guanine)



- A DNA molecule is comprised of primary sequence and a redundant "complementary" copy that allows it to self replicate (each acts like a template for the other sequence)

Schematic DNA

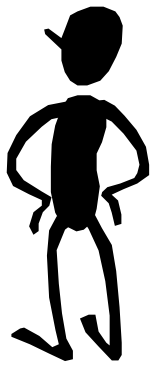
- Many more details are required to give a complete picture of DNA
  - Complementary strands are antiparallel and, thus, oriented (5′-bbbb-3′)
  - Not a simple twist, but has a major and minor grooves which are important for interacting with proteins



- Rather than keep track of all the details we will often consider DNA as a string of nucleotides

Biological Computing Machines

- DNA is an "Operating System" with "programs"
  - Collect raw materials and covert chemicals to energy
  - Perform specialized functions (neurons, muscle, retinal cones)
  - Protect and repair itself
  - Replicate itself, or duplicate entire organism



- How are these "programs" encoded?
- What biological machinery "executes" this program?
- How is the program's execution sequenced?

Genes are the Programs

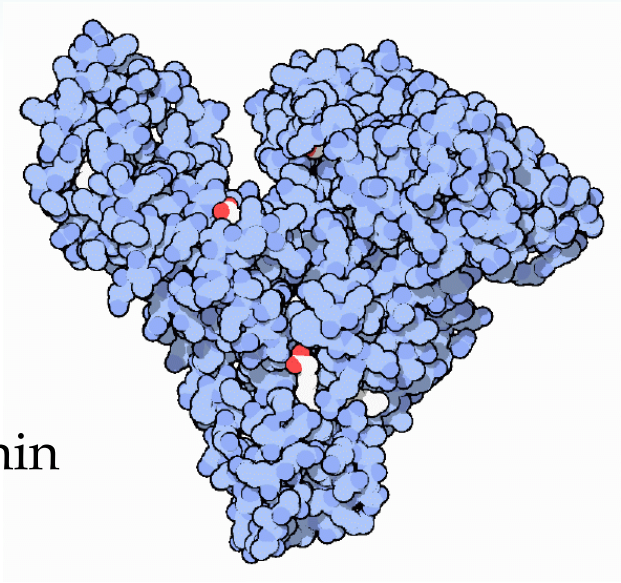
- Specific subsequences of DNA bases determine specific functions (programs) of a cell, these subsequences have commandeered the name "gene"
- Genes are distributed throughout a genome
- Not all DNA sequence sections contain genes
- Genes might not be entirely contiguous within the DNA sequence
- Genes can be either active or inactive
- Genes provide *instructions* for assembling proteins, which are the machinery of life

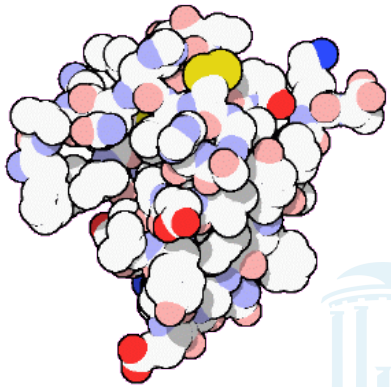
How are these instructions encoded?

What is the output of these programs?

What are Proteins

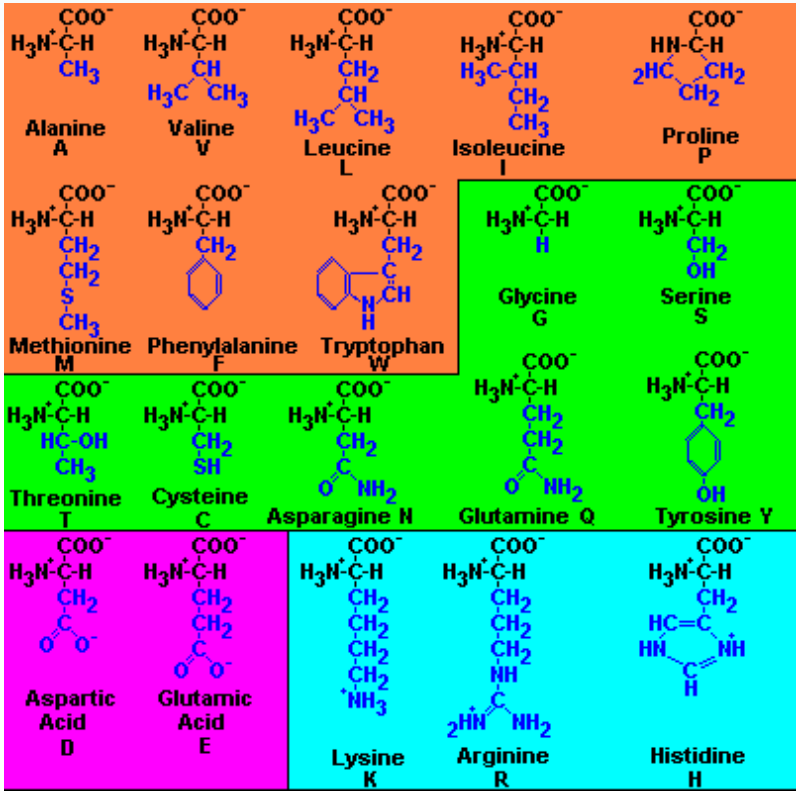
- Proteins are incredibly diverse
  - Structural proteins (collagen) provide structural support and rigidity
  - Enzymes act as biological catalysts (pepsin) that hasten critical reactions without taking part in them
  - Proteins transport small molecules and minerals to where they are needed within an organism (hemoglobin)
  - Used for signaling and intercellular communication (insulin)
  - Absorb photons to enable vision (rhodopsin)
- Proteins are assembled from simple molecules, called amino acids.





Amino Acids

- 20 ingredients of proteins
- Varying side chain is shown in blue
- Orange indicates non-polar and hydrophobic, the remainder are polar or hydrophilic
- Magenta indicates acidic
- Cyan indicates a base



A hydrophobic amino acid avoids water whereas a hydrophillic amino acid is attracted to water. This influences the shape that proteins fold into.

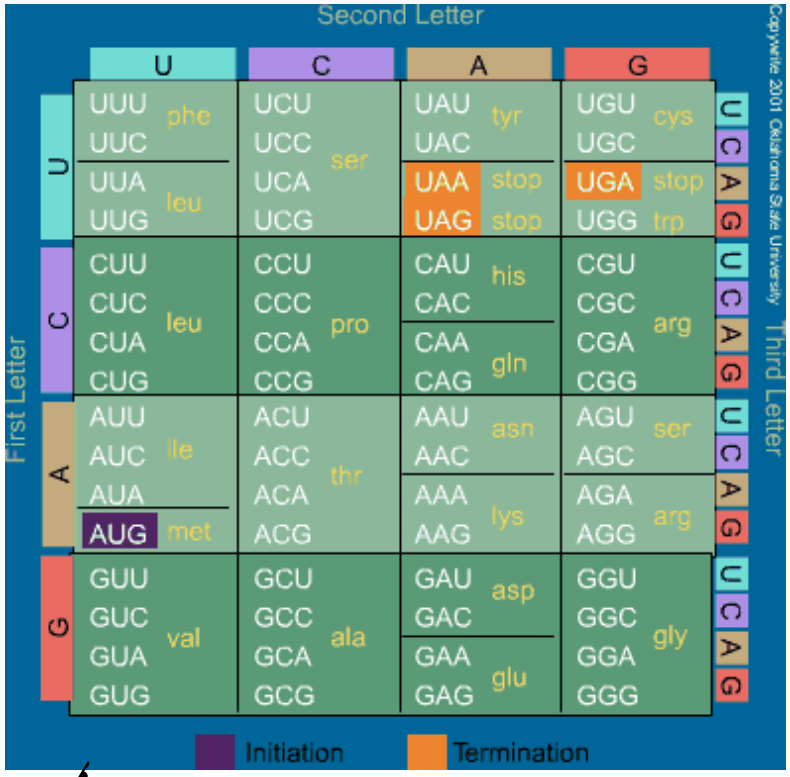
Encoding Protein Assembly

- Each DNA base can be one of 4 values (G,C,A,T)
- Proteins are polymer chains of amino acids ranging in length from tens to millions
- There are 20 amino acids
- How do you encode variable length chains of 20 amino acids using only 4 bases?
- Do you need other codings?

Clearly, we can't encode 20 different amino acids using only one base. How many encodings are possible using a pair of bases? How many with three?

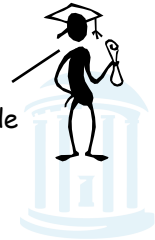
Codons

- Triplets of nucleotide bases determine the amino acid sequence of a protein
- All genes begin with a particular code, AUG, for the amino acid Methonine
- Three codes are used to indicate STOP, and thus end the transcription process for the gene
- Most amino acids have redundant encodings



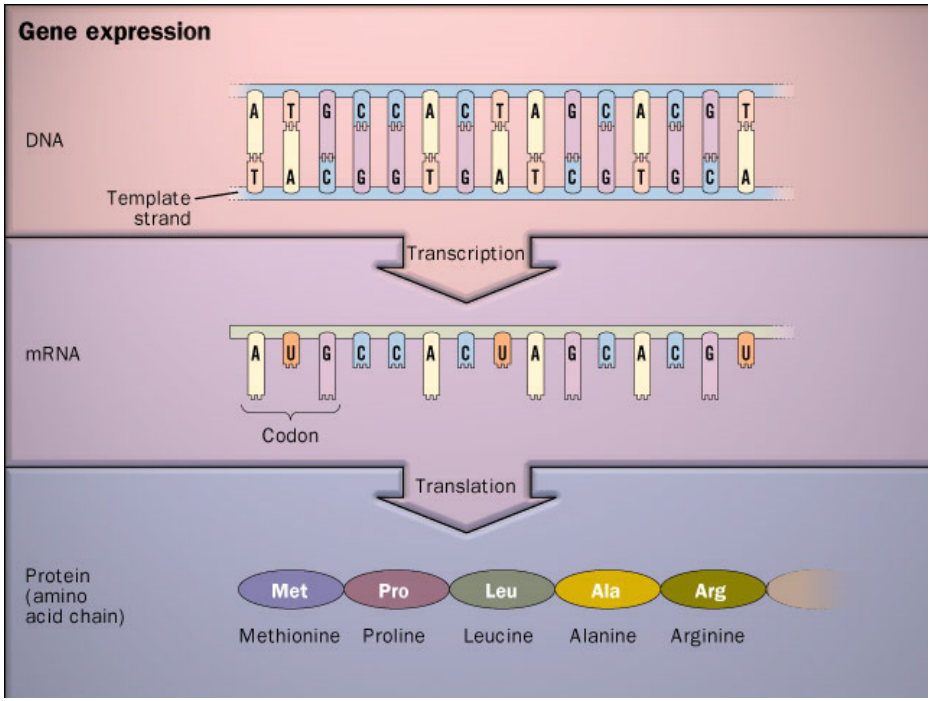
Why are there Us in this table?

Before a DNA sequence is translated into a protein, a copy is first made. This copy is made from RNA. In RNA, the nucleotide "Uracil" replaces "Thymine". Uracil and Thymine are both chemically and structurally very similar.



From Genes to Proteins

- The central dogma of molecular biology is that information encoded by the bases of DNA are transcribed by RNA and then converted into proteins



Is the Code Perfect?

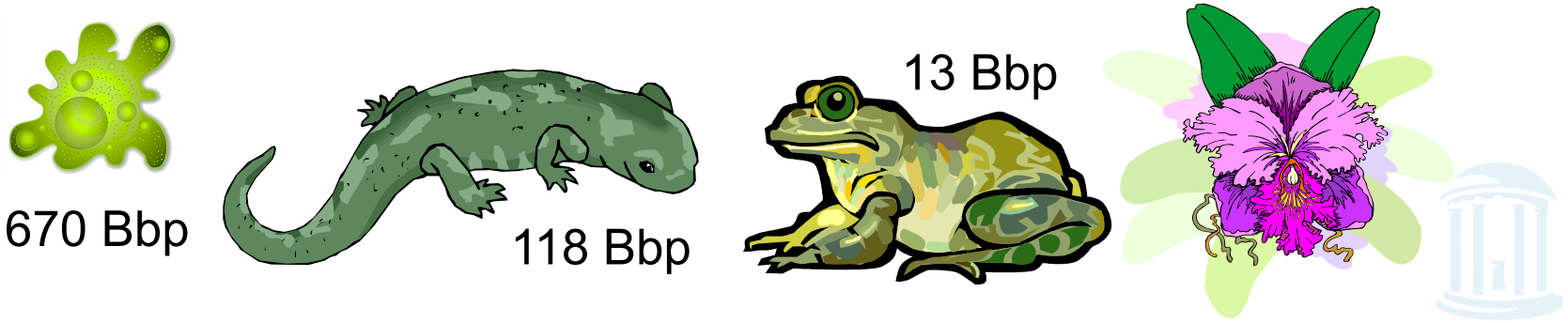
- Proteins are generally unaffected by small variations in their code sequence, particularly changes to a small number of bases
- Minor variations in genes, called *allels,* are responsible for individual variations (blood-type, hair color, etc.)
- Errors in translation (the substitution for one amino acid for the one encoded by the gene), occur at roughly 0.1% of all residues. This means that a single large protein will have at least one incorrect amino acid somewhere! Many of these will still function, in part because the substituted residue will often be adequate. Still, is a bit curious that this level of error is acceptable.
- In eukaryotes (humans, plants) gene sequences are not contiguous. They are broken into codon carrying segments called *exons* separated by seemingly meaningless base sequences called *introns*.

How Big is a Genome?

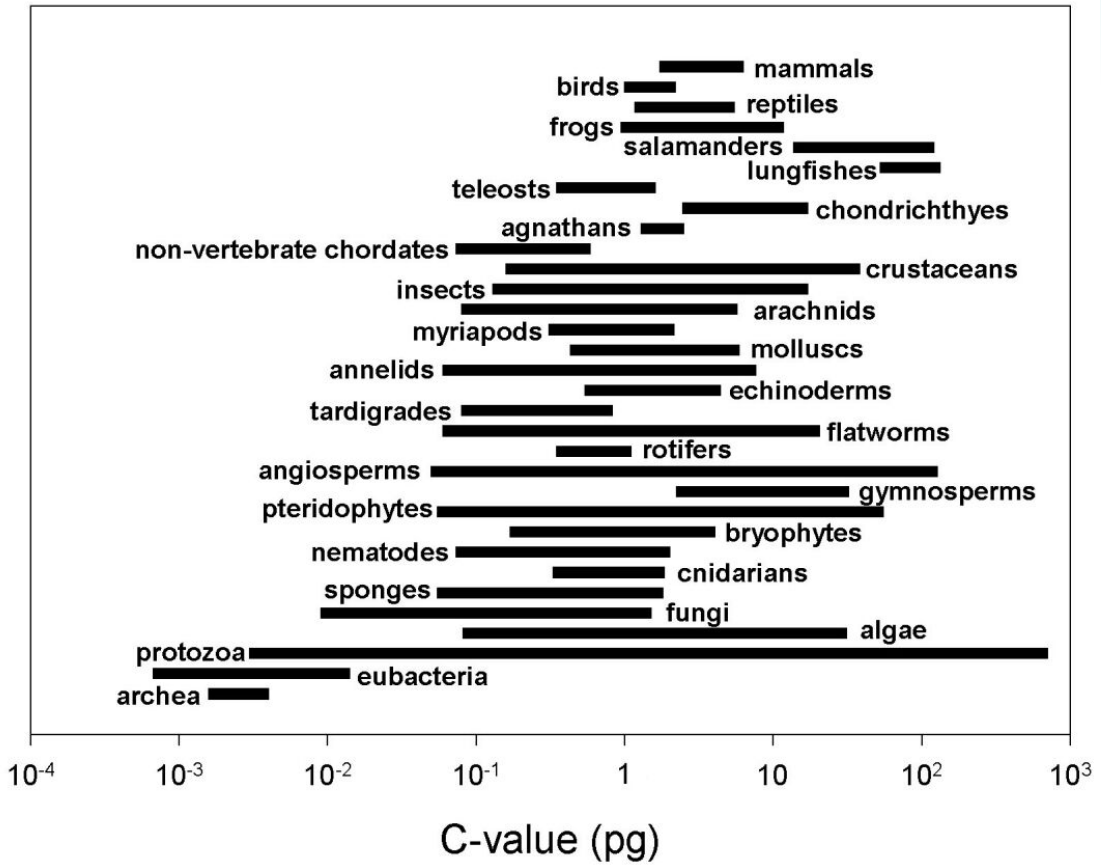
- The human genome is roughly 3 billion bases
  - A typical gene is roughly 3000 bases
  - The largest known human gene (dystrophin) has 2.4 million bases
  - The estimated number of human genes is roughly 30K
  - The genome is nearly identical for every human (99.9%)
  - Human DNA is 98% identical to chimpanzee DNA.
  - The functions are unknown for more than 50% of discovered genes.
  - Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.

Is Bigger Better?

- The genome size of a species is constant
- Large variations can occur across species lines
- Not strictly correlated with organism complexity
- Genome lengths can vary as much as 100 fold between similar species
- Length and variability are more of an indications of a phylum's susceptibility to mutation



Genome Variation



- Length and variability are more of an indications of a phylum's susceptibility to mutation than complexity

Summary

- Information resolves uncertainty
- Heredity provides evidence that information is passed on by biological systems
- Biological information is stored as DNA
- Genes are segments of DNA sequence that encode assembly instructions for proteins
- The central dogma of molecular biology is that DNA sequences are transcribed by RNA polymerases into mRNAs that are then translated by ribosomes into proteins.
- A genome's length is not a good indicator of its information content

Next Time

- Assembling Biological Puzzles
- Duplicating DNA using PCR
- DNA sequencing
- Restriction Enzymes
- Gel Electrophoresis
- Blotting and Hybridization