COMP 555 Bioalgorithms

Fall 2014

Jan Prins

Lecture 1: Course Introduction

Read Chapter 1 and Chapter 3, Secns 3.1-3.7

Intended Audience

- COMP 555: Bioalgorithms
  - Suitable for both undergraduate and graduate students
  - CS majors who want to learn bioinformatics
  - Non CS majors from the biological sciences who are interested in algorithms used in bioinformatics.
  - Graduate students in the Computer Science department, the BCB curriculum or other departments with orientation to bioinformatics methods in research

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 Biologists
  - 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

Some details

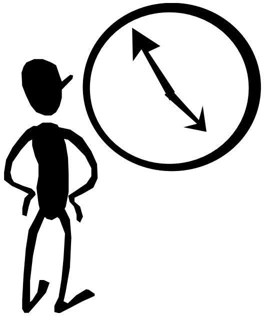
- Comp Sci undergraduates
  - COMP 555 does not substitute for the COMP 550 requirement, but both courses may be taken for credit.
- Comp Sci graduates
  - COMP 555 can count toward the Theory category in the distribution requirements
    - Subject to limitation on number of 500-level classes

Course Logistics

- Website:

  http://www.cs.unc.edu/~prins/Classes/555/

  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 23%)
  - Final Exam (worth 25%)
  - Class participation (worth 2%)



- Problem Sets
  - Roughly one every two weeks
  - Will include a short program to write

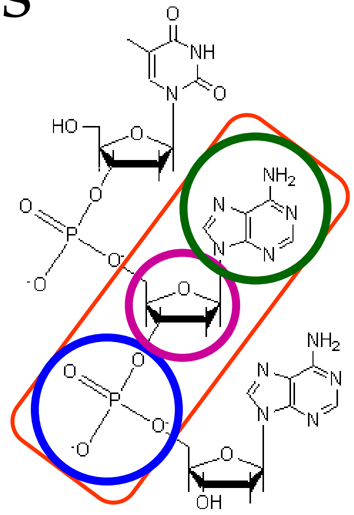
Today

- A few short examples of what we will study
  - Biology topic
    - Molecular biology
  - Algorithm topic
    - Finding the winning strategy for a simple game
  - Programming topic
    - Tiny program in python
  - Analysis topic
    - Asymptotic time complexity

Biology topic: Molecular biology

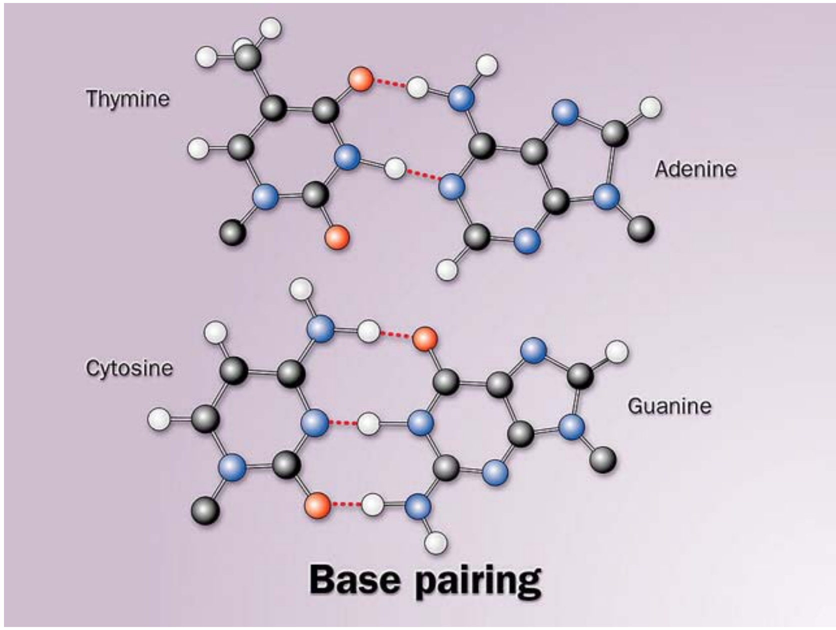
- The information stored in DNA organizes inanimate molecules into living organisms and orchestrates their lifelong function
- A long complementary chain of nucleotides

- Each nucleotide has 3 components
  - A phosphate group
  - A ribose sugar
  - One of four nitrogenous bases
- The sequence of nucleotides encodes information!



DNA Components

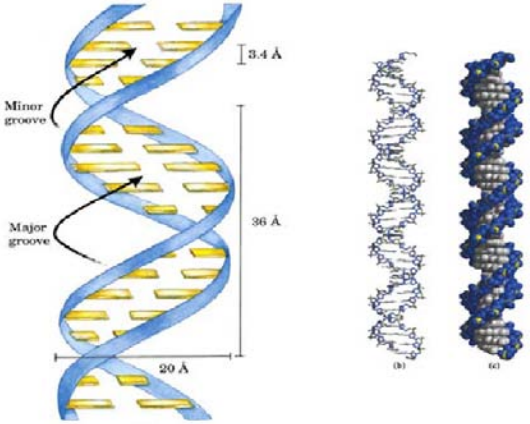
- DNA appears in all living organisms
- Different code sequences distinguish
  - Plants from animals
  - Species
  - Individuals
- A complete DNA sequence for an organism is called its *genome*
- Code sequences are composed of 4 bases (Adenine, Cytosine, Guanine, Thymine)
- Each base binds with another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

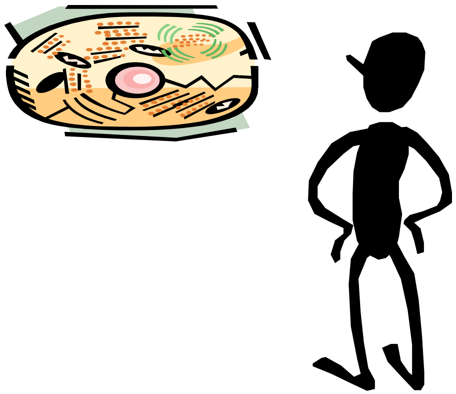
- Many more details are required to give a complete picture of DNA
  - Complementary strands are antiparallel and, thus, oriented (5′-bbbb-3′)
  - Not a simple twist, but has a major and minor grooves which are important for interacting with proteins



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

Biological Computing Machines

- DNA can be viewed as a "program" to
  - Collect raw materials and convert chemicals to energy
  - Perform specialized functions (neurons, muscle, retinal cones)
  - Protect and repair itself
  - Replicate itself, or duplicate entire organism



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

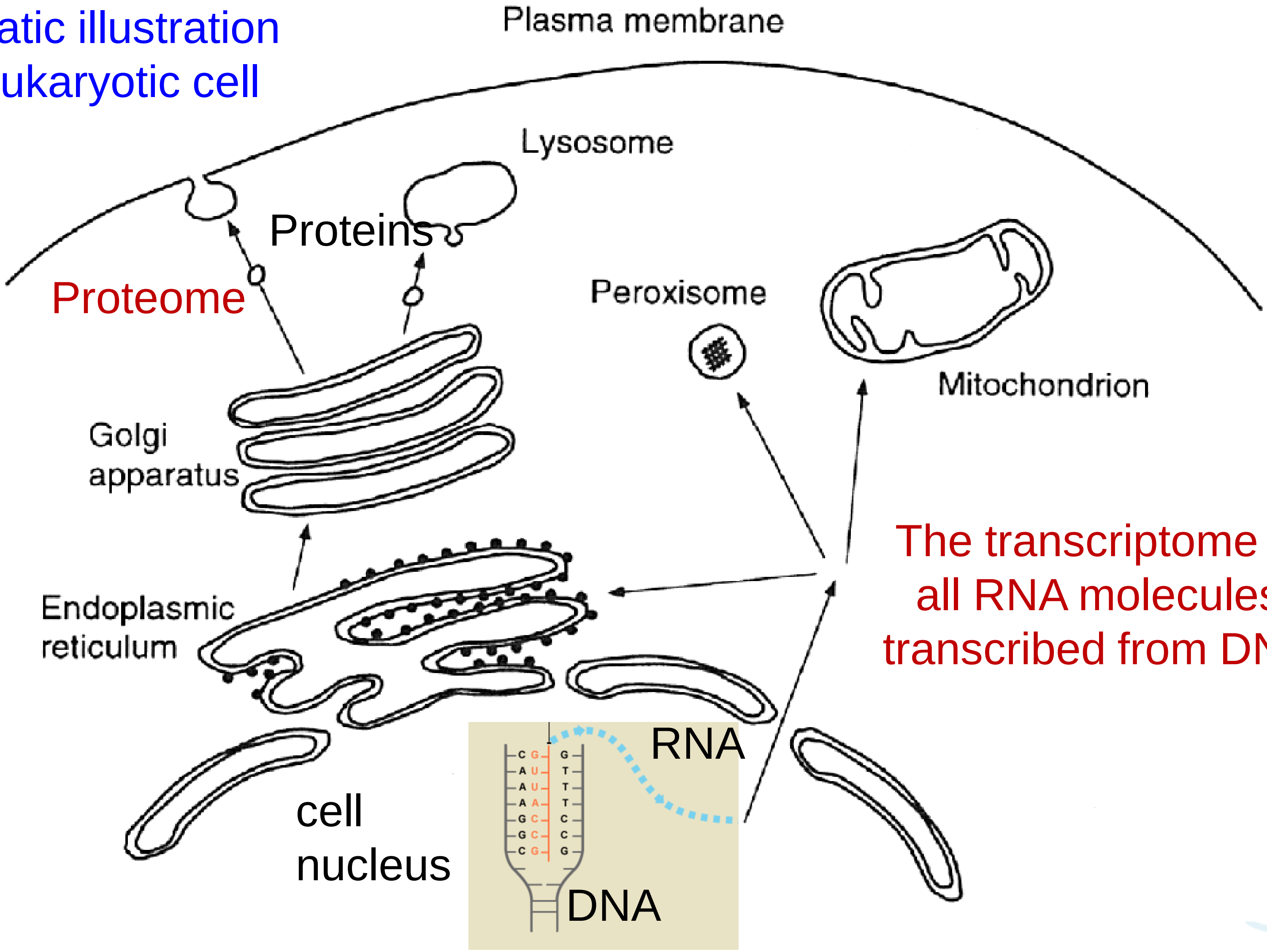
Genes are key parts of the Program

- A gene is a specific subsequences of DNA that controls specific functions of a cell
  - Genes are distributed throughout a genome
  - Not all DNA sequence sections contain genes

- Genes provide *instructions* for assembling proteins, which are the machinery within and beyond the cell
  - how are these instructions encoded?
  - how are the instructions carried out?

Genome, Transcriptome, Proteome

Schematic illustration of a eukaryotic cell



Proteins

Proteome

The transcriptome is all RNA molecules transcribed from DNA

RNA

cell nucleus

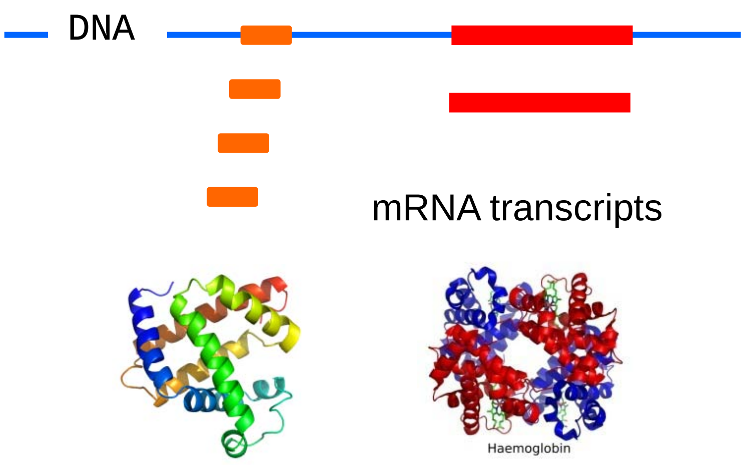
DNA

Genome

Execution of the gene “programs”

- Cells with the same genome may produce a different transcriptome … how?
  - Two main mechanisms controlled by external conditions

gene *expression*

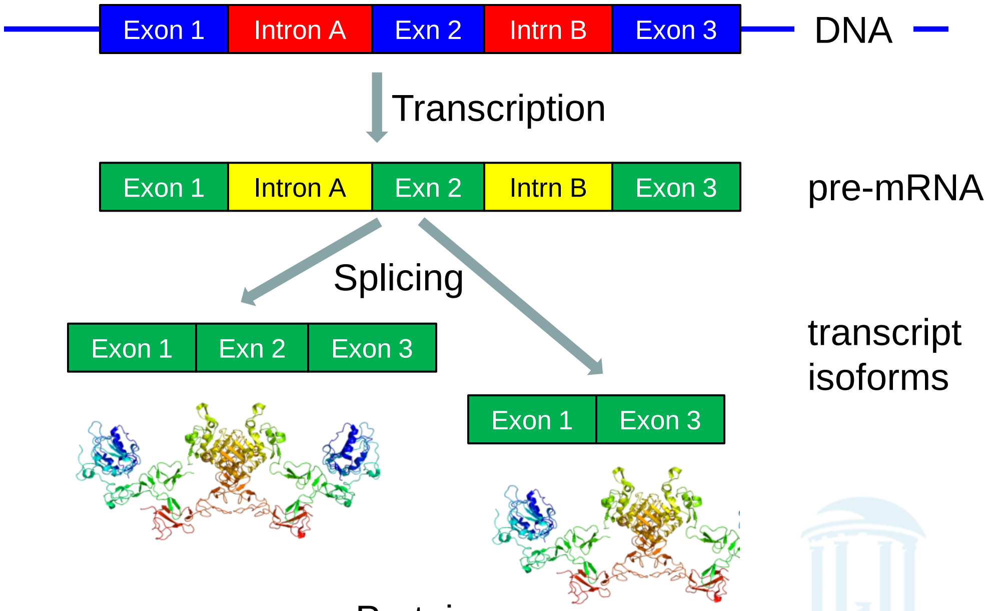


DNA

mRNA transcripts

Proteins

gene *splicing*



DNA

Exon 1	Intron A	Exn 2	Intrn B	Exon 3

Transcription

pre-mRNA

Exon 1	Intron A	Exn 2	Intrn B	Exon 3

Splicing

transcript isoforms

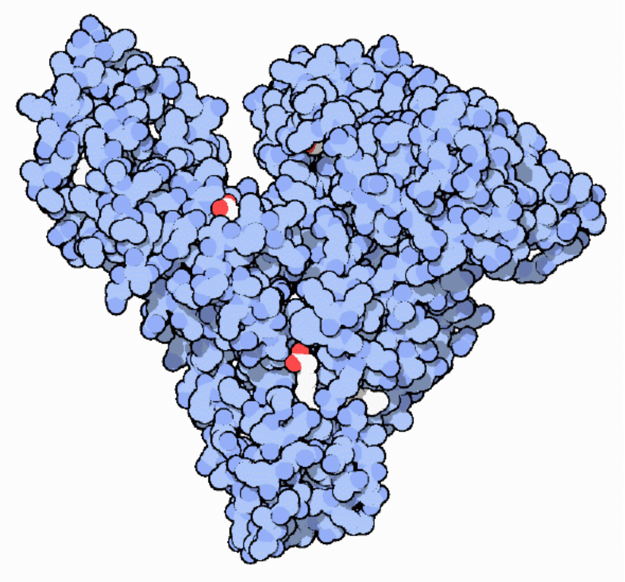
Exon 1	Exn 2	Exon 3

Exon 1	Exon 3

Proteins

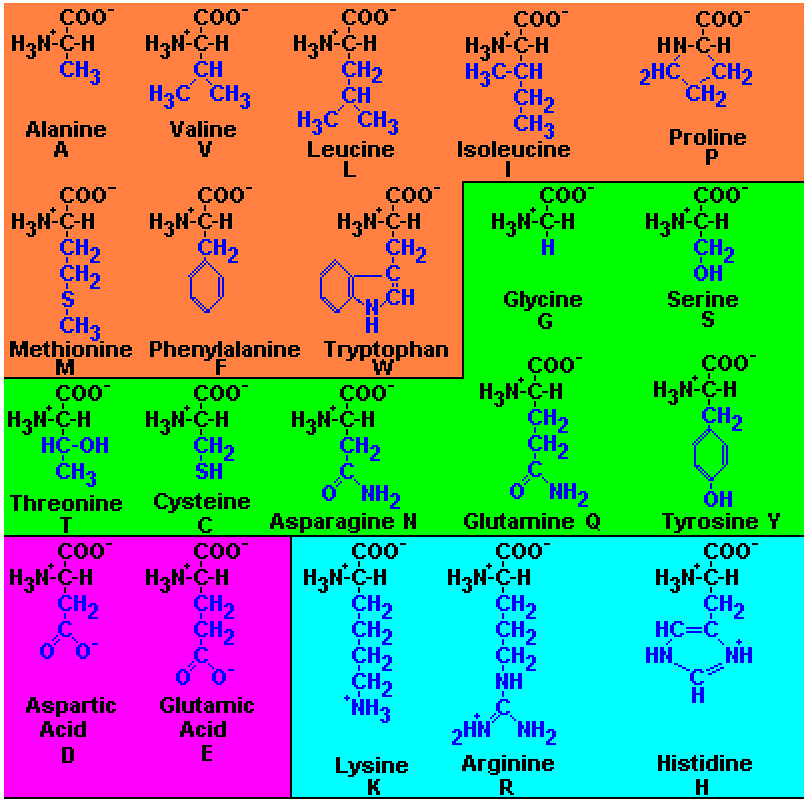
What are Proteins

- 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 with 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.

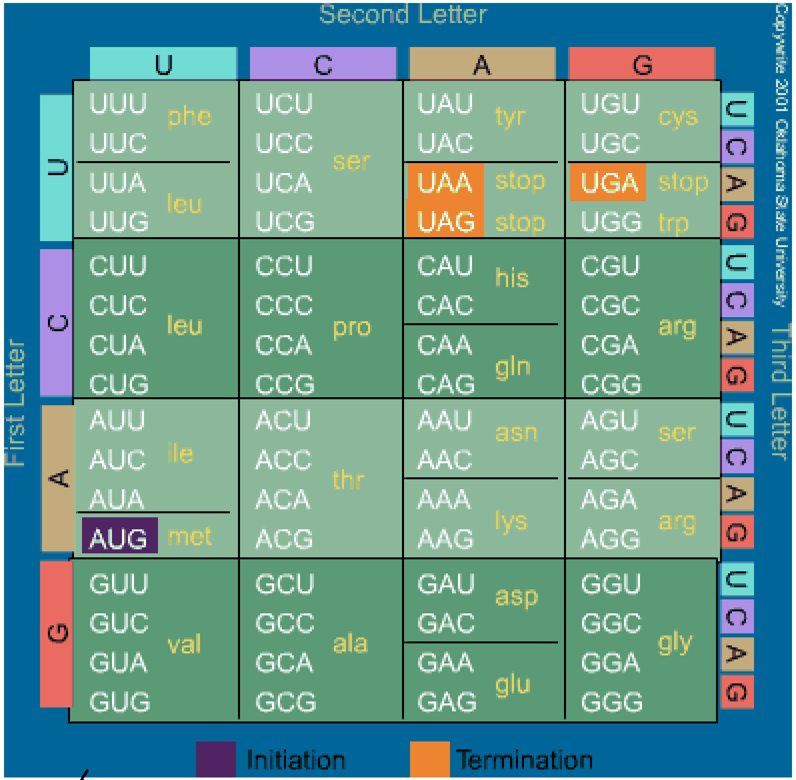
Encoding Protein Assembly

- 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

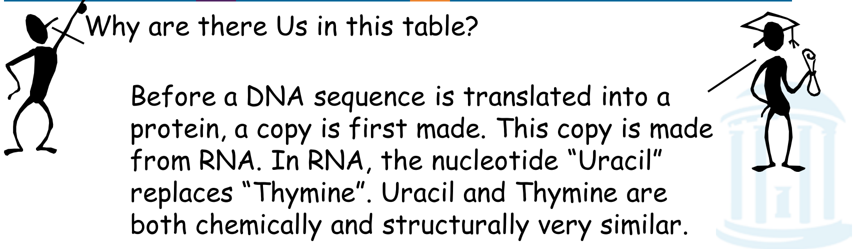
- Triplets of nucleotide bases determine the amino acid sequence of a protein
- All genes begin with a particular code, AUG, for the amino acid Methonine
- Three codes are used to indicate STOP, and thus end the transcription process for the gene
- Most amino acids have redundant encodings



	Second Letter								
First Letter	U		C		A		G		Third Letter
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	stop	UGA	stop	A
	UUG		UCG		UAG	stop	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

🟪 Initiation 🟧 Termination

Copywrite 2001 Oklahoma State University

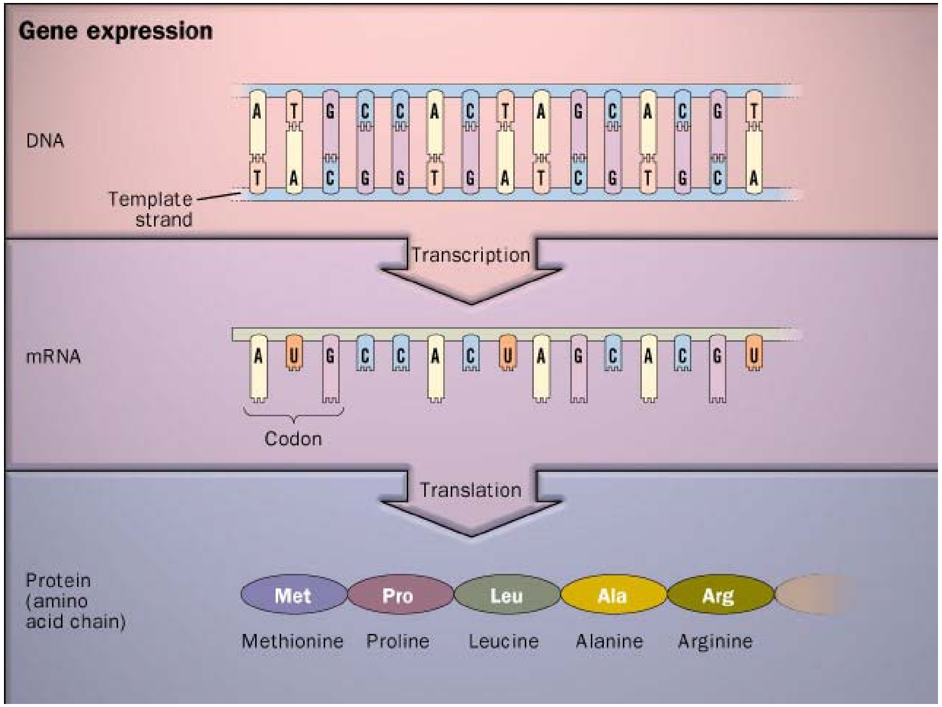


Why are there Us in this table?

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


Gene expression
DNA
A T G C C A C T A G C A C G T
T A C G G T G A T C G T G C A
Template strand
Transcription
mRNA
A U G C C A C U A G C A C G U
Codon
Translation
Protein (amino acid chain)
Met Pro Leu Ala Arg
Methionine Proline Leucine Alanine Arginine

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 *alleles,* 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.

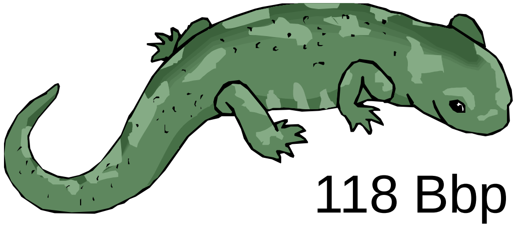
How Big is a Genome?

- The human genome is roughly 3 billion bases
  - A typical RNA transcript 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

670 Bbp



118 Bbp

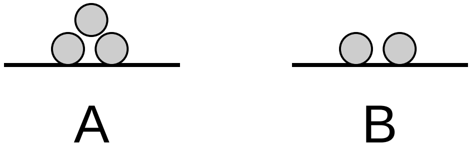
13 Bbp

Algorithm Topic

- Rocks: a simple game
  - Two piles of rocks A,B
  - Two players, alternating turns
  - A turn consists of
    - Take one rock from either pile

    <u>*OR*</u>

    - Take one rock from both piles
  - The game must end. Why?
  - Player that takes the last turn wins the game!
  - Rocks(*i*,*j*) = state of the game with *i* rocks in pile A, and *j* rocks in pile B
  - Can the player whose turn it is win Rocks(*i*,*j*)?



Solution Strategy

- Solve the simplest cases
- Reduce more complex cases to simpler cases

Game State	Player move	Player prospects
Rocks(0,0)	-	Loses
Rocks(0,1)		
Rocks(1,0)		

Programming Topic

- Sample python program

```
def total(v):
   “yields the sum of values in list v”
   n = len(v)
   sum = 0
   for i in xrange(0,n)
      sum += v[i]
   return sum
```

Analysis Topic

- Asymptotic time complexity of programs
  - What is the running time T(n) of the python program for a list of length n?
    - Graph it
  - How should we characterize the running time
    - Approximately, with accurate scaling in the limit

Expected Learning Outcomes

- Students completing this class should …
  - have an understanding of key algorithms used in bioinformatics and know their capabilities and limitations
  - have knowledge of algorithmic strategies that can be employed to design new algorithms for bioinformatics applications
  - be able to analyze the correctness and asymptotic time complexity of algorithms
  - have a working knowledge of concepts and terminology in molecular biology and understand the key challenges
  - be able to write bioinformatics programs using python and python bioinformatics libraries

For next time

- Text
  - Read Chapter 1 Introduction (6 pp)
  - Read Chapter 3 Molecular Biology Primer Secns 3.1 – 3.7 (10 pp)