CS313 Computational Biology

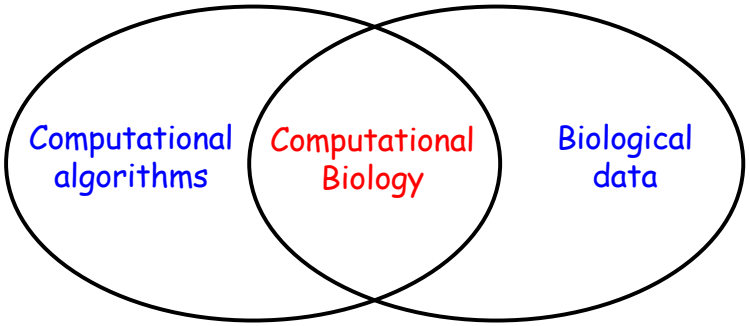
Course Information

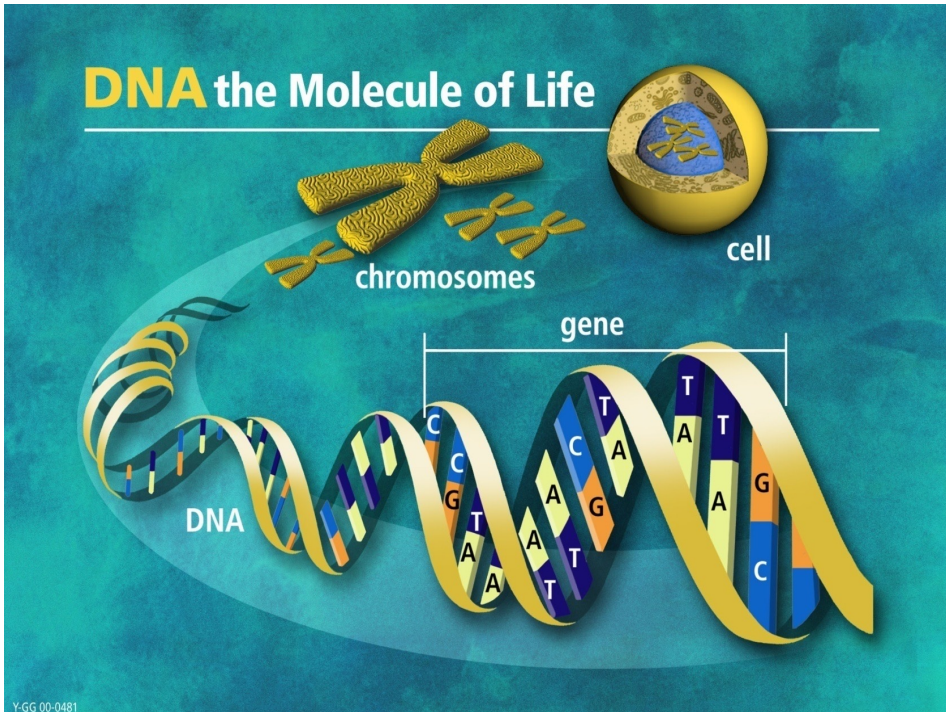
Instructor: Brian Tjaden

Pronouns: He, Him, His

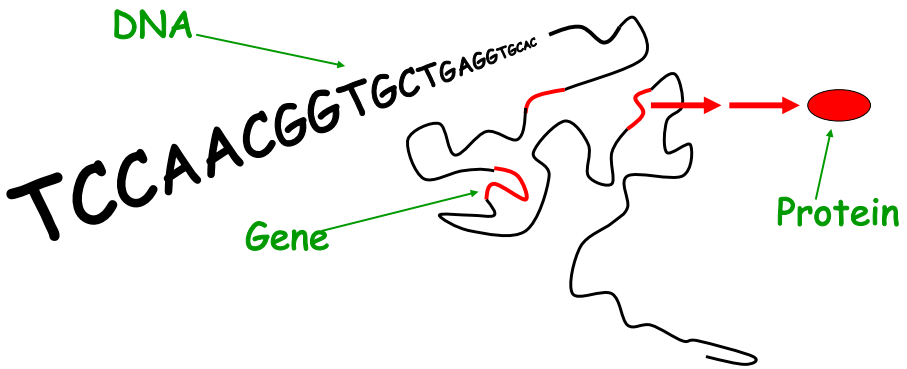
Course Materials: http://cs.wellesley.edu/~cs313

Computational Biology is Multidisciplinary





DNA: simplified

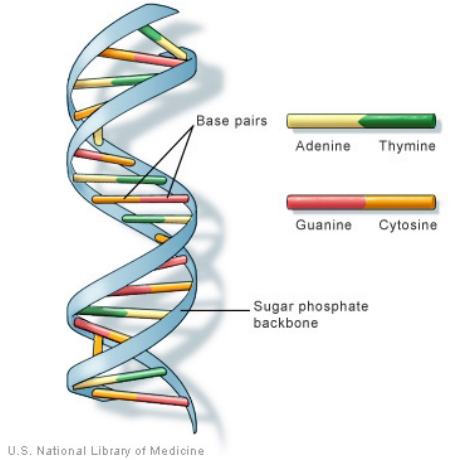


DNA: "program" for cell processes

Proteins: execute cell processes

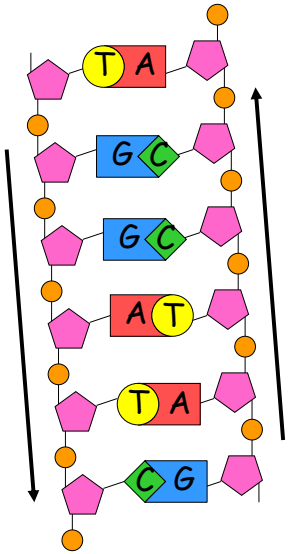
DNA Structure

- Double helix
- Deoxyribose (sugar) - phosphate backbone
- Four bases – A, T, G, C
- Base pairing

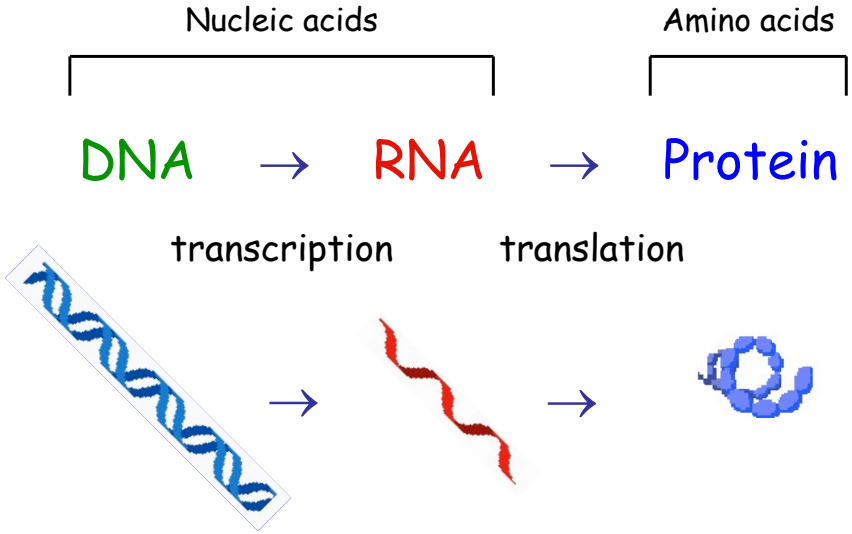


DNA Structure

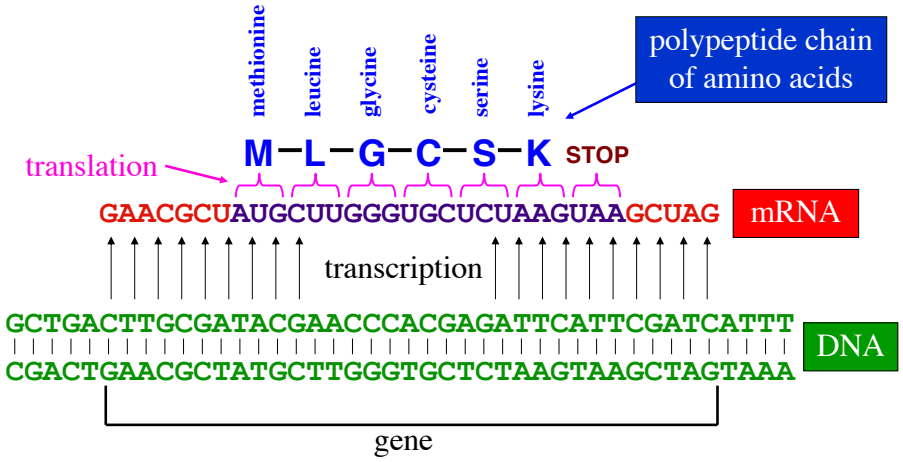
- Information polarity (anti-parallel strands)
- Either strand can function as a template (complementary strands)



Information Flow



DNA → RNA → Protein



The Genetic Code

- 61 amino acid codons
- 1 start codon (Met)
- 3 stop codons

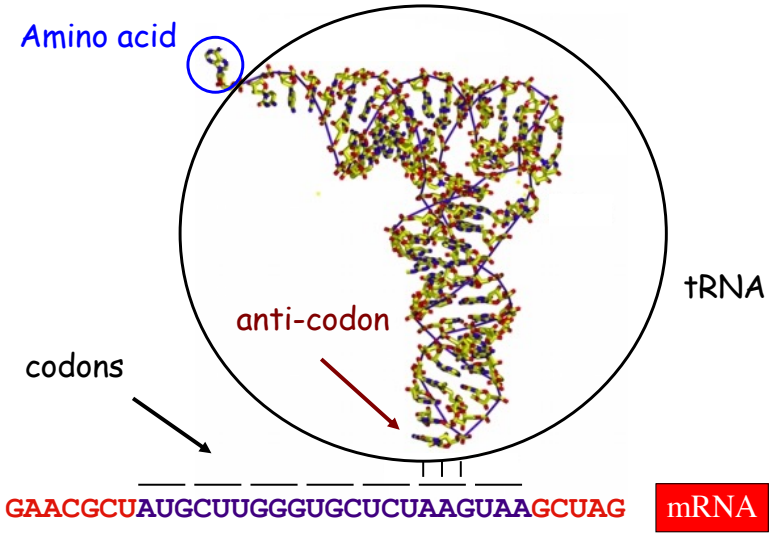
First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

The Genetic Code

- Triplet code
- Non-overlapping codons
- Start and stop codons
- Degeneracy

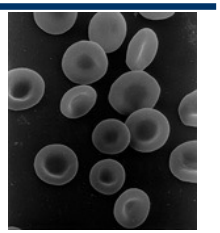
4 nucleotides, 20 amino acids

tRNAs (transfer RNAs)

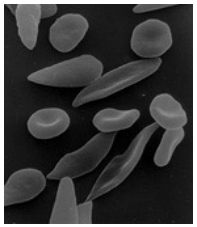


Mutations

- Changes in DNA occur, despite cell's best efforts
- Spontaneous events, copying errors, enviromental factors
- Mutations might change gene function
- Can be harmful, neutral, or beneficial



Normal RBCs



Sickle cell anemia

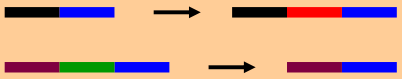
```
ATGGTGCATC TGACTCCTGA GGAGAAGTCT GCCGTTACTG CCCTGTGGGG CAAGGTGAAC GTGGATGAAG TTGGTGGTGA
GGCCCTGGGC AGGCTGCTGG TGGTCTACCC TTGGACCCAG AGGTTCTTTG AGTCCTTTGG GGATCTGTCC ACTCCTGATG
CTGTTATGGG CAACCCTAAG GTGAAGGCTC ATGGCAAGAA AGTGCTCGGT GCCTTTAGTG ATGGCCTGGC TCACCTGGAC
AACCTCAAGG GCACCTTTGC CACACTGAGT GAGCTGCACT GTGACAAGCT GCACGTGGAT CCTGAGAACT TCAGGCTCCT
GGGCAACGTG CTGGTCTGTG TGCTGGCCCA TCACTTTGGC AAAGAATTCA CCCCACCAGT GCAGGCTGCC TATCAGAAAG
TGGTGGCTGG TGTGGCTAAT GCCCTGGCCC ACAAGTATCA CTAA
```


Ultroviolet (UV) Light Causes Sunburn and DNA Damage

Types of Mutations

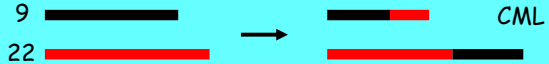
Single base substitutions A → T

Insertions and Deletions

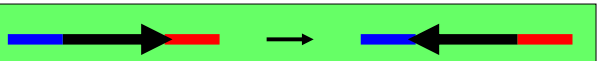


Amplifications

Translocations



Inversions

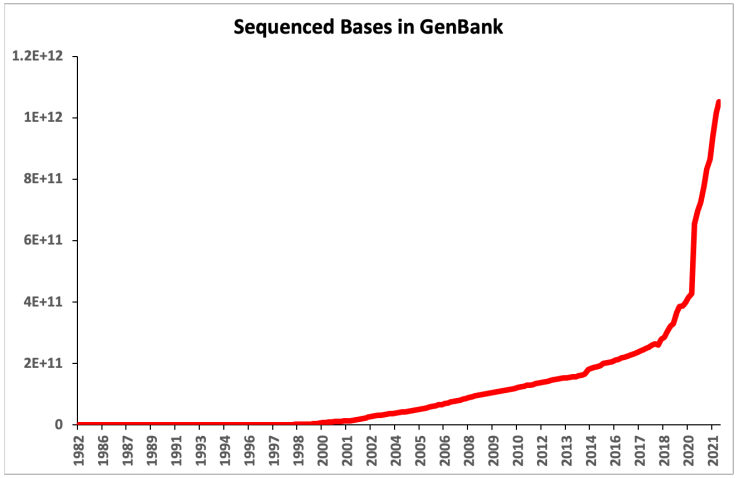


Sample Genomes

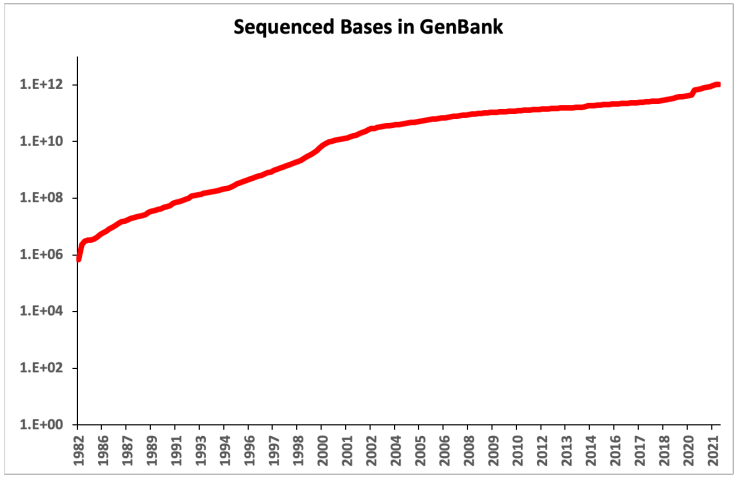
A genome is the total DNA in a cell

Species	Genome Size	# of Genes
SARS-CoV-2	30 thousand	
Epstein-Barr virus	172 thousand	
Escherichia coli	4.6 million	
Drosophila melanogaster	122 million	
Homo sapiens	3.3 billion	
Paris japonica	150 billion	

GenBank Growth



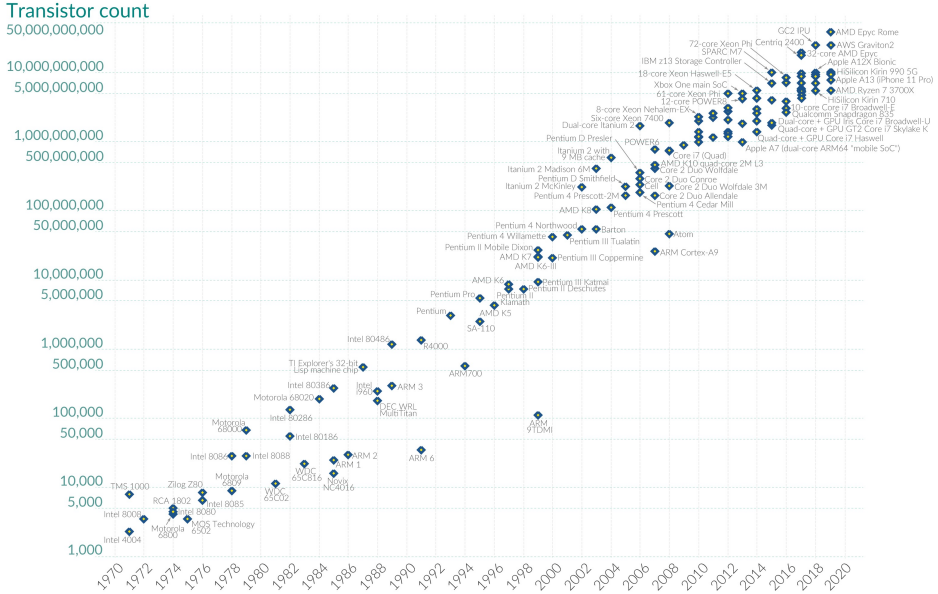
GenBank Growth (log scale)



Moore's Law: The number of transistors on microchips doubles every two years

Our World in Data

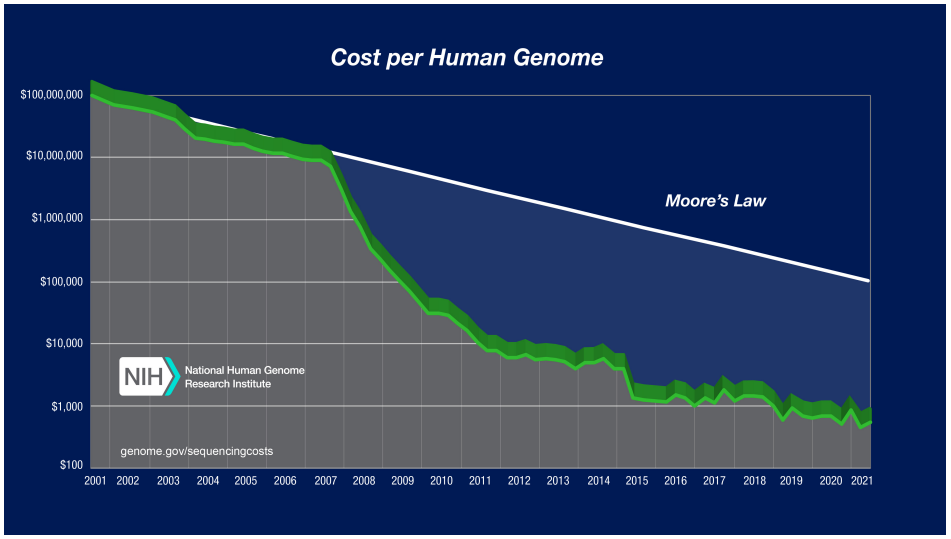
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



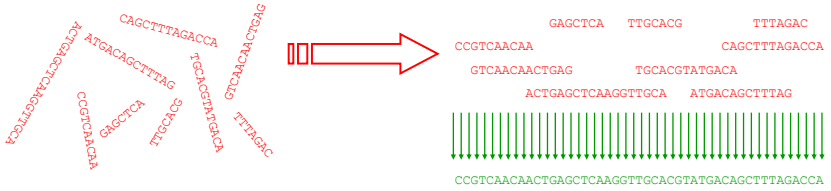
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world's largest problems.

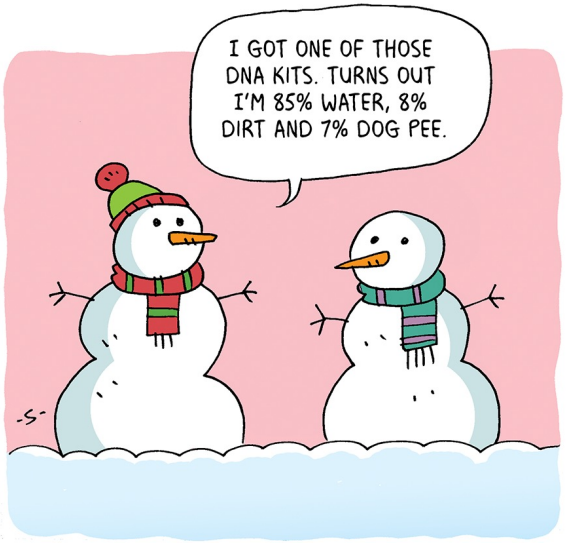
Cost of Sequencing a Human Genome



Sequencing a Genome



Is sequencing genome the end or beginning?



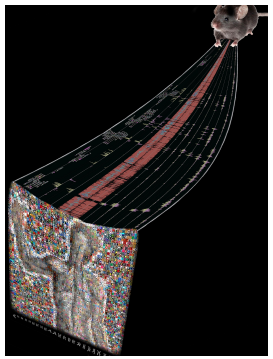
Open Questions

- Which regions of DNA have biologic function? (What are the genes?)
- What are their functions?
- When and how are genes turned on and off?
- How do genes and their products (proteins) interact with each other?
- What are the implications to health and medicine?

in other words...
How does the cell's DNA "program" work?

Comparative Genomics

Comparative genomics involves understanding the relationships between the genomes of different species.



- Many genes have unknown functions
- Which genes are present (conserved vs unique)?
- How are genes arranged in the genome
- Infer function of genes by sequence similarity – homology to known genes

Human and banana: 55%
Human and human: 99.9%
Human and chimpanzee: 96%
Human and mouse: 85%
Human and bacteria: ??%

Recurring Themes

- Bioinformatic tools are often hypotheses-generating
- Determining statistical significance of results generated by bioinformatic tools is useful
- Properties of data guide choice of algorithm
- Some problems are solved exactly or optimally. Other problems are addressed using a *heuristic* approach.
- Many computational approaches are improved by incorporating additional biological insights into their underlying method or model.
- Recent advances have allowed scientists to gather large amounts of, often heterogeneous, data. One of the roles of bioinformatic tools is efficient analysis of large data sets with the aim of extracting new biological insights.