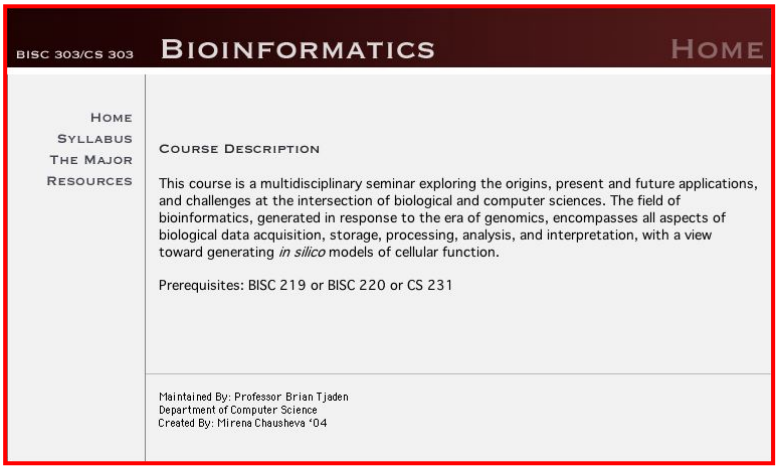
BISC/CS303: Bioinformatics

Spring 2008

Administrivia

Instructors:	Brian Tjaden and Brett Pellock
Meeting:	Wednesdays, 6:30 - 9:00pm
FirstClass Conference:	BISC/CS303-S08

Course Materials



http://cs.wellesley.edu/~cs303

Textbooks (optional, on reserve)

Bioinformatics and Functional Genomics
by Jonathan Pevzner
John Wiley & Sons, Inc., 2003

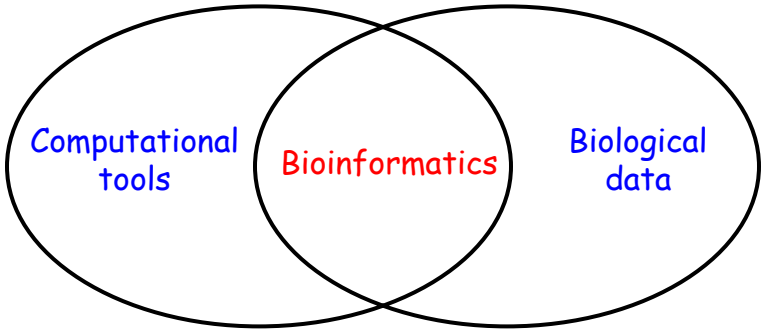
Functional Concepts of Bioinformatics
by D.E. Krane and M.L. Raymer
Benjamin Cummings, 2003

BISC/CS 303 overview/grading:

- Lecture
- Computer Lab and Milestones
- Final Project

- Grading:
- Milestones 60%
- Project 30%
- Presentation 10%

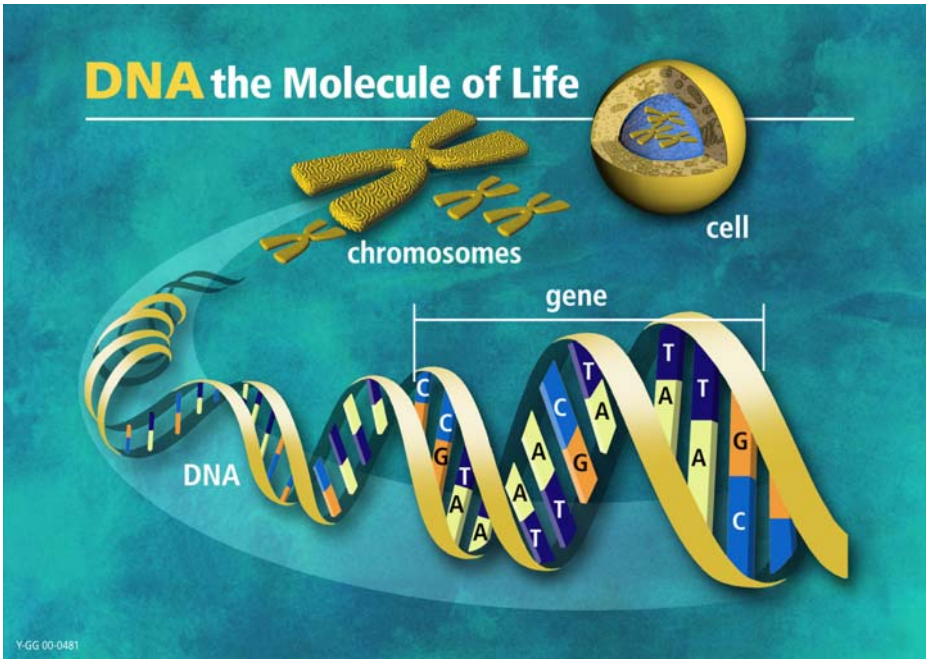
Bioinformatics is multidisciplinary



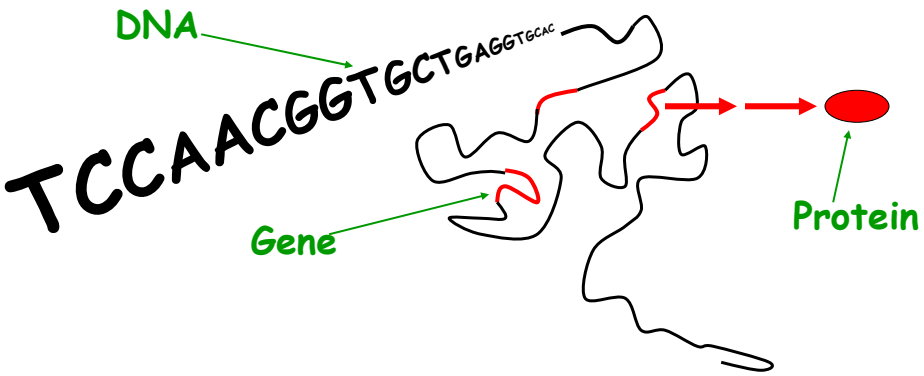
Focus is on how bioinformatic tools work

Today's class:

- Introduction to DNA
- Information flow
- The Genomics Era
- Lab exercise



DNA: simplified

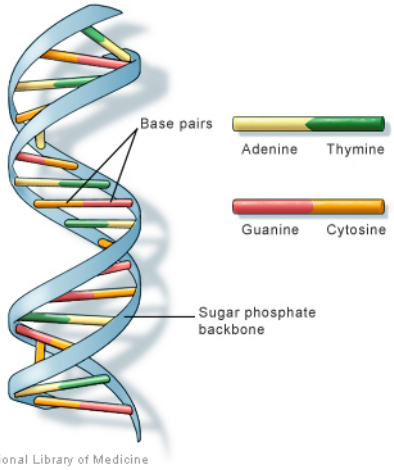


DNA: "program" for cell processes

Proteins: execute cell processes

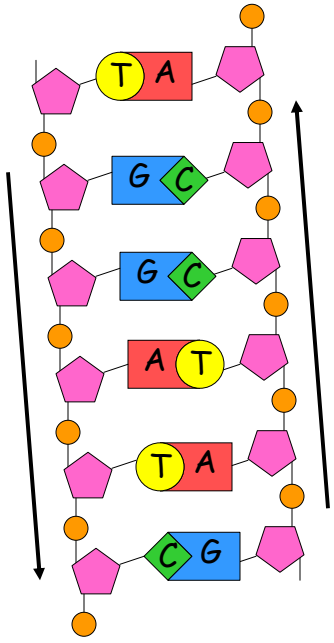
DNA Structure

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



DNA Structure

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

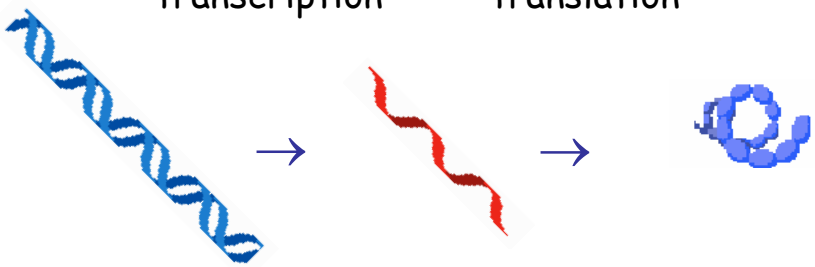


Information Flow

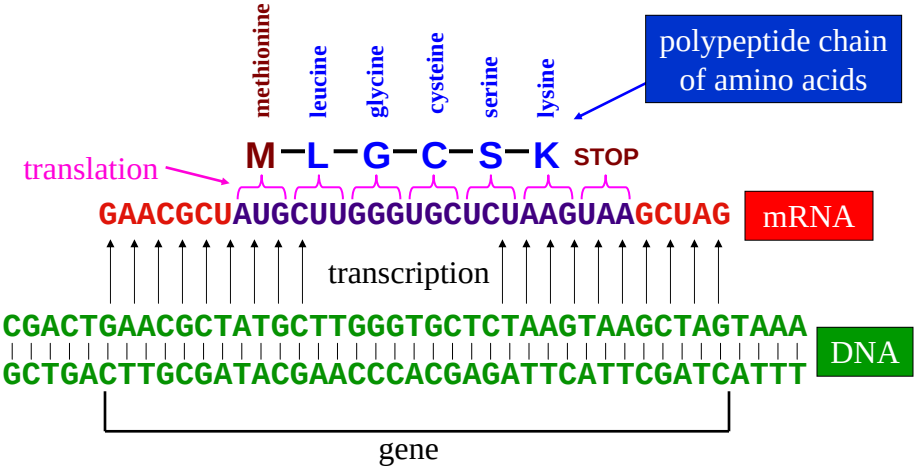
Nucleic acids | Amino acids

DNA → RNA → Protein

transcription translation



DNA → RNA → Protein



The Genetic Code

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

Second letter

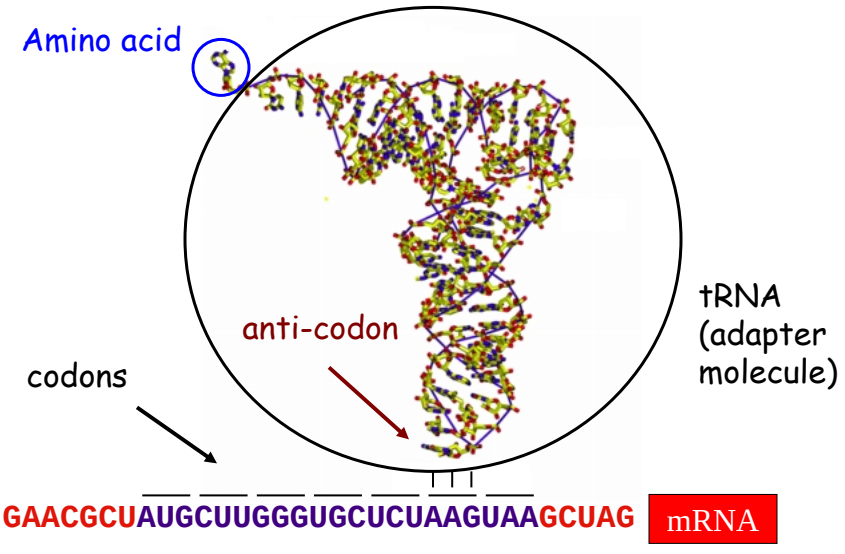
First letter	U	C	A	G	Third letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg,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

The Genetic Code

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

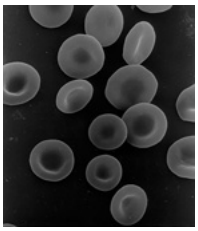
- 4 nucleotide bases, 20 amino acids
  - -1 base = 4 codons (4^1)
  - -2 bases = 16 codons (4^2)
  - -3 bases = 64 codons (4^3)

Commitment to information fidelity

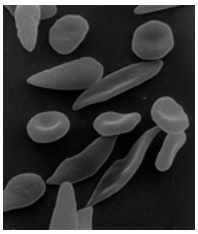


Mutations

- 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



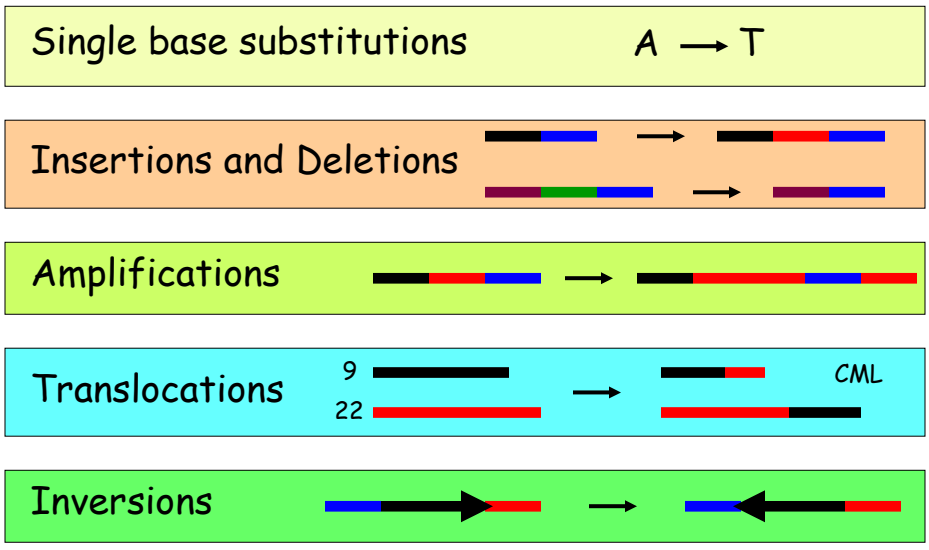
Normal RBCs



Sickle cell anemia

Ultraviolet (UV) light causes sunburn and DNA damage

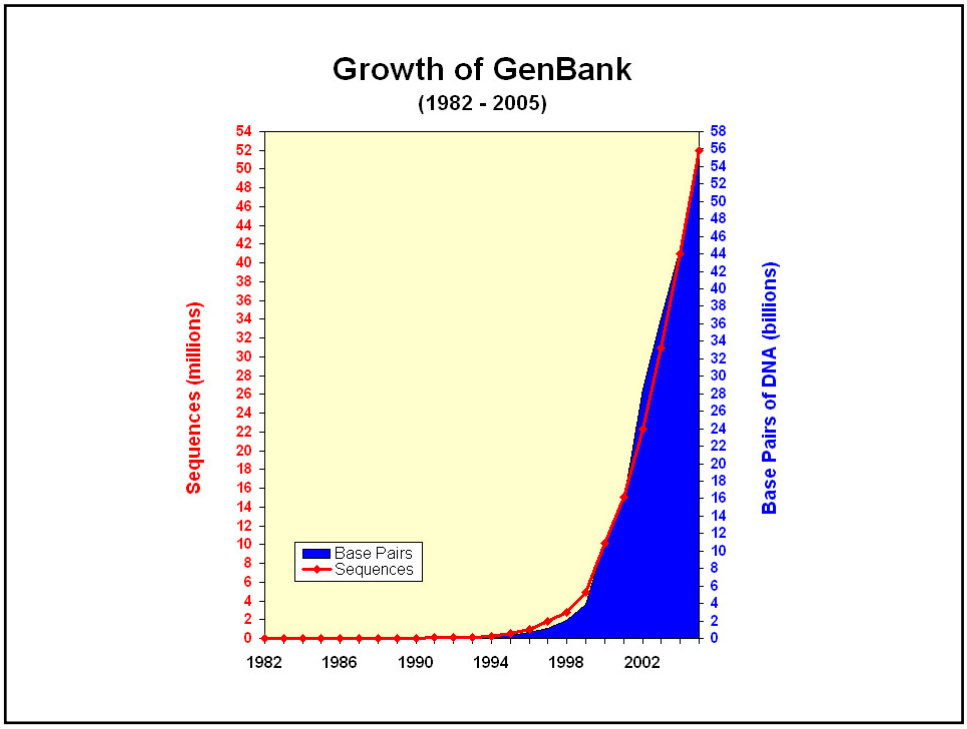
Types of Mutations

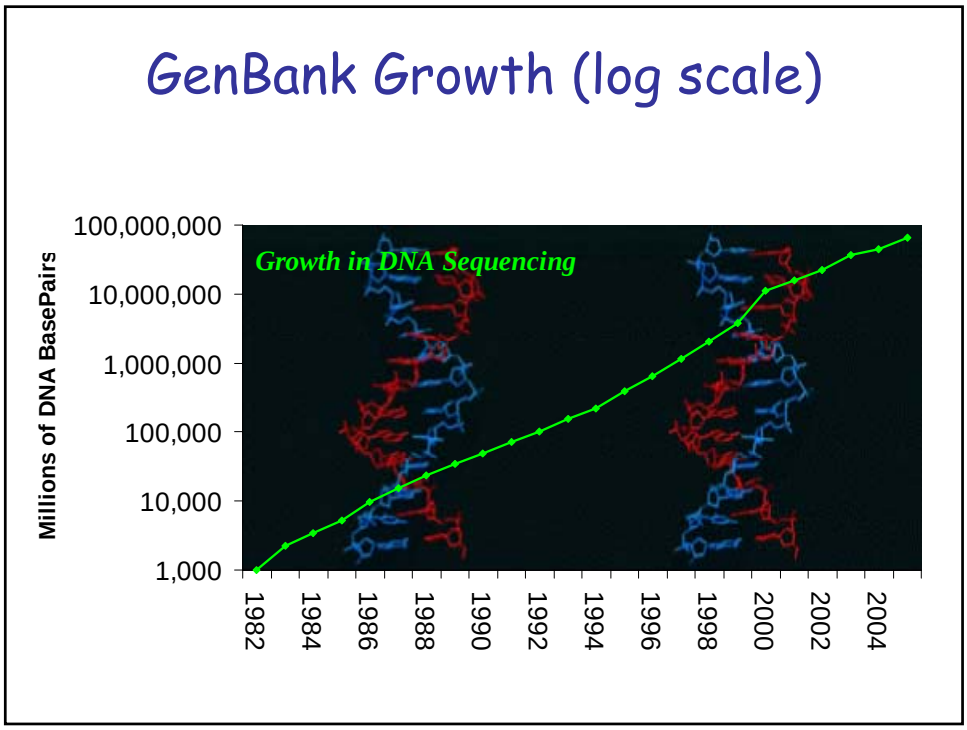


Sample Genomes

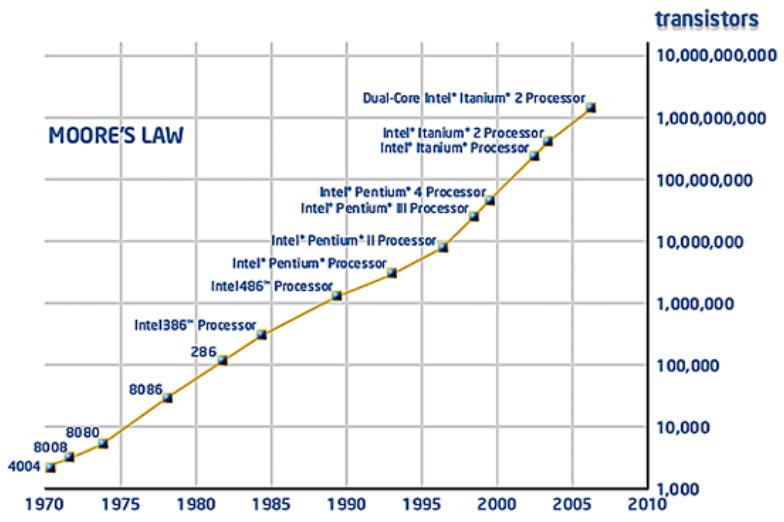
A *genome* is the total DNA in a cell

Species	Genome size	# of Genes
Epstein-Barr virus	172 Thousand	
Escherichia coli	4.6 Million	
Drosophila melanogaster	122 Million	
Homo sapiens	3.3 Billion	
Psilotum nudum (fern)	250 Billion	

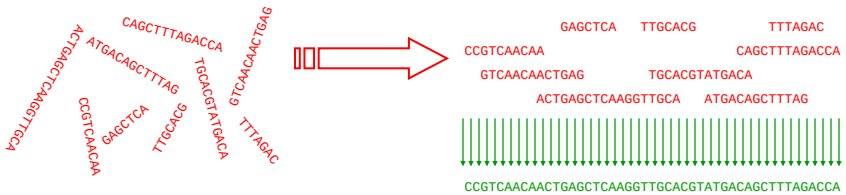

Growth of GenBank
(1982 - 2005)
Sequences (millions)
Base Pairs of DNA (billions)
Base Pairs
Sequences
1982
1986
1990
1994
1998
2002


GenBank Growth (log scale)
Millions of DNA BasePairs
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Growth in DNA Sequencing
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004

CPU Growth (log scale)



Sequencing a Genome



Comparative Genomics

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

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

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?

Recurring Themes

- Bioinformatic tools are often hypotheses-generating
- Identifying homologous regions between genomes can be very useful
- Properties of data guide choice of algorithm
- Determining statistical significance of results generated by bioinformatic tools