

CS174 Bioinformatics

Instructor: Xiaohui Xie
University of California, Irvine

Today's Goals

- Course information
- Challenges in bioinformatics/computational biology
- Brief intro to molecular biology
- Python tutorial

Course Information

- Lecture: TT 3:30-4:50pm in DBH 1423
- TA: Lars Otten <lotten@ics.uci.edu>
- Grading
 - 40% Homework (4-5 problem sets)
 - 25% Mid-term quiz (in class)
 - 35% Final exam
- Office hours: TT after class
- Course Prerequisites:
 - Basic programming skills, we will teach Python
 - Statistics, Calculus, basic knowledge of biology

Course Goals

- Introduction to the growing field of **bioinformatics/computational biology**
 - Fundamental problems in computational biology
 - Statistical, algorithmic and machine learning techniques
 - Overall survey of the field

References

- Recommended Textbooks:
 - N.C. Jones and P.A. Pevzner. An Introduction to Bioinformatics Algorithms
 - R. Durbin, S. Eddy, A. Krogh and G. Mitchison. Biological Sequence Analysis
- Course Website:
<http://www.ics.uci.edu/~xhx/courses/CS174/>
where lectures, references and problem sets can be found.

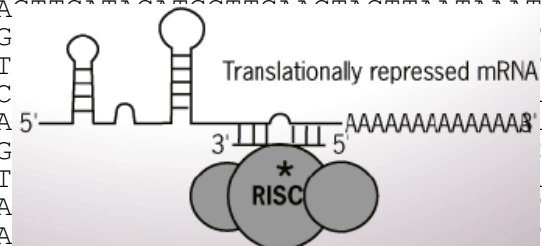
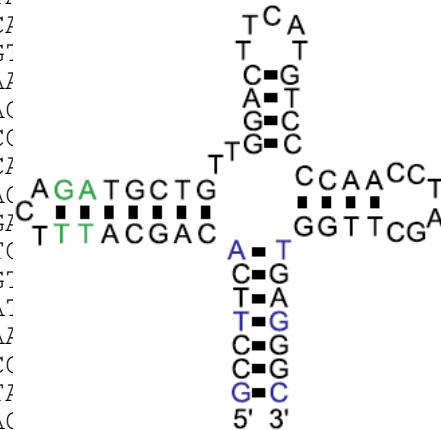
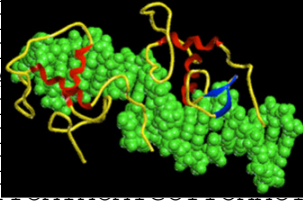
Why bioinformatics?

Bioinformatics = Biology + Information

AGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCAT
ACATGCATGCTTCAATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGTATGATAA
TGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCTCCTTATCCTTATAGTTCA
TACATGCTTCAACTACTTAATAAAATGATTGTATGATAAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCA
TGCTTCAACTGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCT
TTATAGTTACTTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGT
TGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCTCCTTATCCTTAT
AGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAA
TGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAAT
GTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCT
AGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTTCAATGTAAGAGATT
TCGATTATCCTTATAGTTCATATGCTTCAACTACTTAATAAAATGATCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTA
TAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGAATTTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTT
TCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTAT
AGTTCATACATGCATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTT
CAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATGCTTCAA
CTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGTATTGAATTTTCAAAAATCCTTACTTTTTTTTTTGGGA
TGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACATATCCATATCTAATCTTACTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTA
GCCTAAAAAACCTTCTCTTTGGAACCTTTCAGTAATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGA
AGACTCTCCTCCGTGCGTCTCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACTAGCTTTTA
TGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATTAGTTTTTTAGCCTTAT
TTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCACCTTTAACTAATACTTTCAACATTTTCAGTTTG
TATTACTTCTTATTCAAATGTCAATAAAGTATCAACAAAAAATTGTTAATATACCTCTATACTTTAACGTCAAGGAGAAAAAATAATGACTAAATCTCATTGAGA
AGAAGTGATTGTACCTGAGTTCAATTCTAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAA
ACCGGATTTTGTTGCTAGATCGCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGATTTTGATATGCTTTG
CGCCGTCAAAGTTTTGAACGATGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAATCTTTAAGAGTCTTGAAGGCTGTGAAATTAATGA
CTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATGAACGAGTCTCATTTCAGGTTGGTACGATAAACTTTACGAATGTTCTTGCCAG
AGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATCATATGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGTTCCAGGGGGGCCCAA
ATGGCAACATAGAAAAGGTAAAAGAAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAAC
CAGCATTTGGGCAGCTGTCTATATGAATTAGTCAAGTATACTTCTTTTTTTTTACTTTGTTTCAGAACAACTTCTCATTTTTTTTCTACTCATAACTTTAGCATCACAAAT
ACGCAATAATAACAGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCACA
AAGCTTAAACACAGGACAAAATTTCTTGATATGCTTTCAACCGCTGCGTTTTGGATACCTATTCTTGACATGTATGATAATGATATGACTACCATTTTGTTATTGTA
CGTGGGGCAGTTGACGTCTTATCATATGTCAAAGAAAATTTGCGAAGTTCTTGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTTATAGTTCATACATGCTTCA
ACTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTTCTTGATATGCTTTCAACCGCTGC
GTTTTGGATACCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCTTGGCAAGT
TGCCAACCTGACGAGATGCAGTTTCTACGCATAATAAGAATAGGAGGGAATATGCAGGAGAACGCCAGACAATCTATCATTACATTTAAGCGGCTCTTCAAAAAGATT
GAACTCTCGCCAACCTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAAGAGTATAAGTCATCTCAGAGTAATATACTACCGAAGTTTA
TGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGTAATACGCTGAAAAACCTCAATACAGCTCATTCTGGAAGAAATAGTGTTTTCTGTACAACCAGGACTTGAAGC
CCGTCGAAAAAGAAAGCGGGTTTTGGGATTTGGGTACGGTTTTCGTTGGTGCTTTTGTTGTTTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTTCATTCCCTATATCA
TCTAGAGCATCATTCGGTATTTTCTTCTCTTTATGGCCCGTTATTAACAGAGTCGTATGGCCATCGTTTTGGTATAGTGTCCAAGCTTATATTGCGGCAACTCCCGTA
TCATTAATGCTGAAATCTATCTTTGGAAAAGATTACAAATGATTG
CAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTC
CTTTAAACACAGGGACAAAATTTCTTGATATGCTTTCAACCGCTGC
TGCTTCAACTACTTAATAAAATGATTGTATGATAATGTTTTCAATG
ATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAAATAAA
TTATAGTTCATACATGCTTCAACTACTTAATAAAATGATTGTATGA
ATGTTTTCAATGTAAGATTACTTAATAAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAAAT
GATTTCATACATGCTTCAACTACTGTAAATAAATAAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTTCGATTATCCTTATAGTTCATACATGCATAGTTCA
TACATGCTTCAACTACTT

The human genome is 400,000 longer
than the sequence shown here.

TGC
AAA
ACA
ATG
TCC
ATA



Why bioinformatics?

- Lots of data
- Pattern finding, rule discovery
- Allowing analytic and predictive methodologies that support and enhance lab work
- Informatics infrastructure (data storage, retrieval)
- Data visualization
- Life itself is a computer!

Genome as a computer program

Computer Program	Genome
Data Method Class Encapsulation Virus	Gene Regulatory code Modularity Histone code Virus
Deterministic Precise	Stochastic Redundant

1997

DOE forms Joint Genome Institute



NCHGR becomes NHGRI



E. coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the U.S. published



RIKEN Genomic Sciences Center (Japan) established

Roundworm (*C. elegans*) genome sequenced

SNP initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999

Full-scale human sequencing begins



Sequence of first human chromosome (chromosome 22) completed



2000

Draft version of human genome sequence completed

President Clinton and Prime Minister Blair support free access to genome information

Fruit fly (*D. melanogaster*) genome sequenced

Mustard cress (*A. thaliana*) genome sequenced



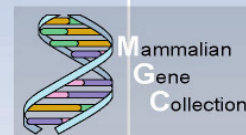
Executive order bans genetic discrimination in U.S. federal workplace

2001

Draft version of human genome sequence published



10,000 full-length human cDNAs sequenced



2002

Draft version of mouse genome sequence completed and published



Draft version of rat genome sequence completed

Draft version of rice genome sequence completed and published

2003

Finished version of human genome sequence completed

HGP ends with all goals achieved

to be continued..

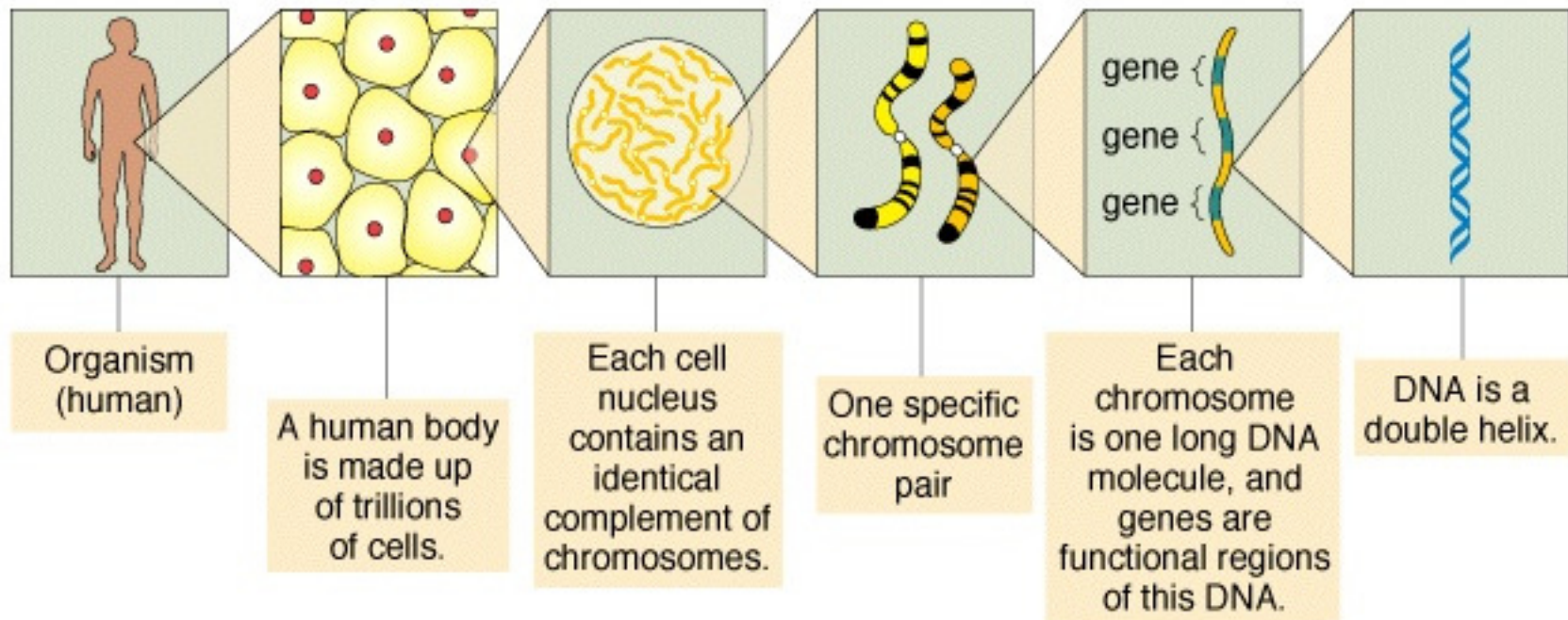
Four Aspects

- Biology
 - What's the underlying problem?
- Algorithm
 - How to solve the problem efficiently?
- Learning
 - How to model biology systems and learn from observed data?
- Statistics
 - How to differentiate true phenomena from artifacts?

Topics to be covered

- DNA/RNA/Protein sequence analysis
 - Gene discovery
 - Pattern finding (motif discovery, EM-algorithm)
 - Sequence alignment (Smith-Waterman, BLAST)
 - Models of sequences (HMM)
 - RNA folding (Stochastic context-free grammar SCFG)
- Algorithms for large-scale data analysis
 - Clustering algorithms (Hierarchical clustering, K-means)
 - Inferring gene networks (Regression, Bayesian networks)
- Evolutionary models
 - Phylogenetic trees
 - Comparative Genomics
- Protein world (if time allows)
 - Secondary & tertiary structure prediction

Introduction to Molecular Biology and Genomics



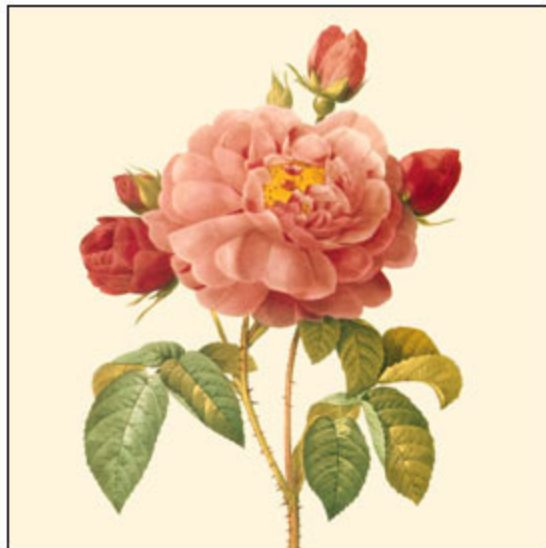
Different Life Forms Share a Common Genetic Framework



(A)



(B)

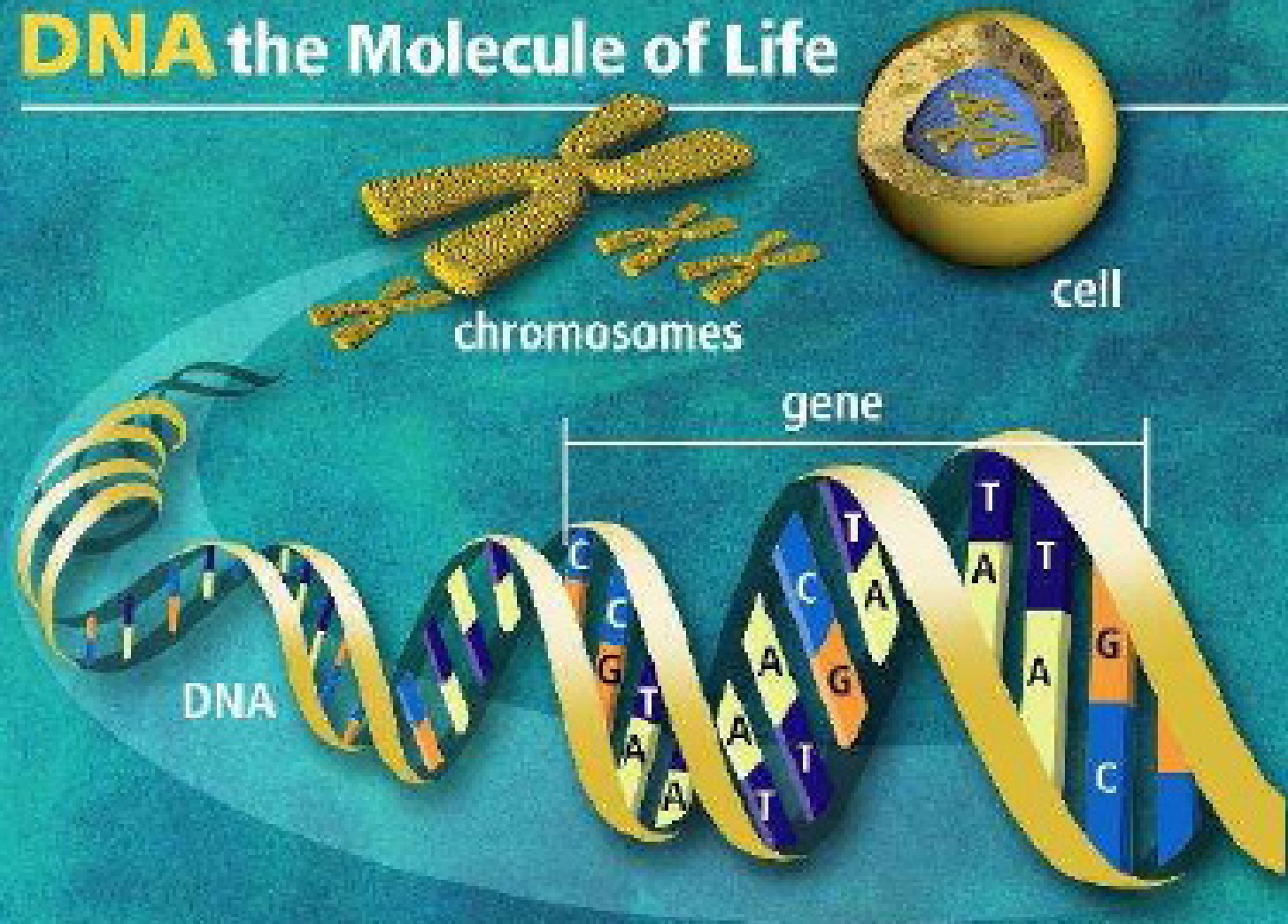


(C)



(D)

DNA the Molecule of Life



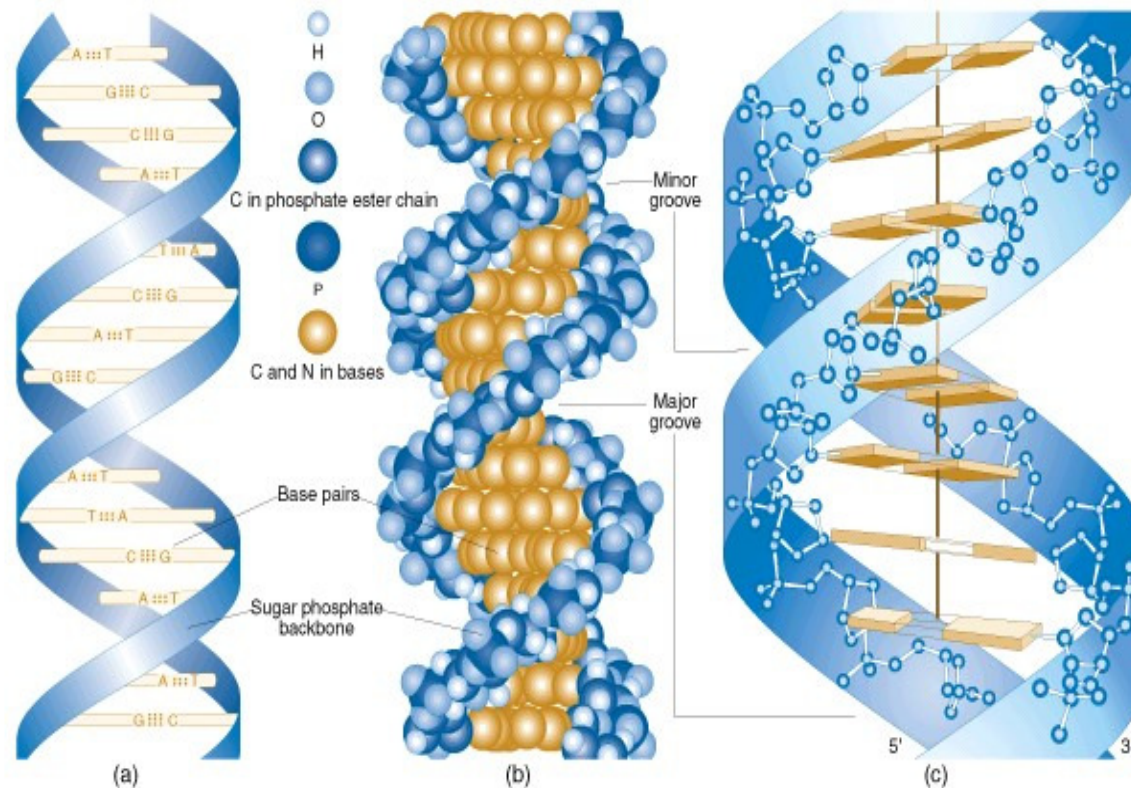
Deoxyribonucleic acid (DNA)

- can be thought of as the “blueprint” for an organism
- composed of small molecules called *nucleotides*
 - four different nucleotides distinguished by the four *bases*: adenine (**A**), cytosine (**C**), guanine (**G**) and thymine (**T**)
- is a *polymer*: large molecule consisting of similar units (nucleotides in this case)
- DNA is digital information
- a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

AGCGGTTAAGGCTGATATGCGCTTTAA
TCGCCAATTCCGACTATACGCGAAATT

The Double Helix

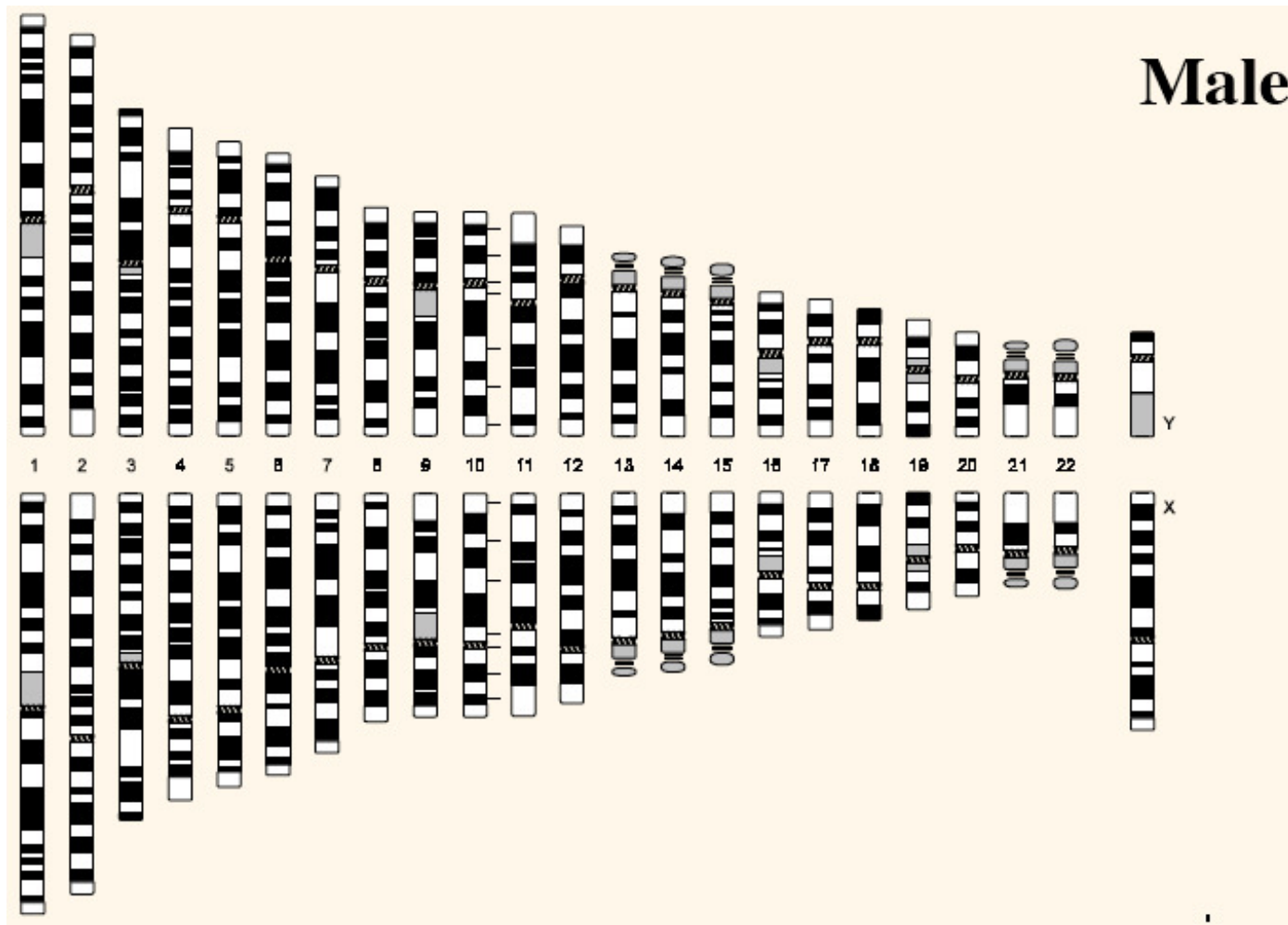
DNA molecules usually consist of two strands arranged in the famous double helix



Genomes

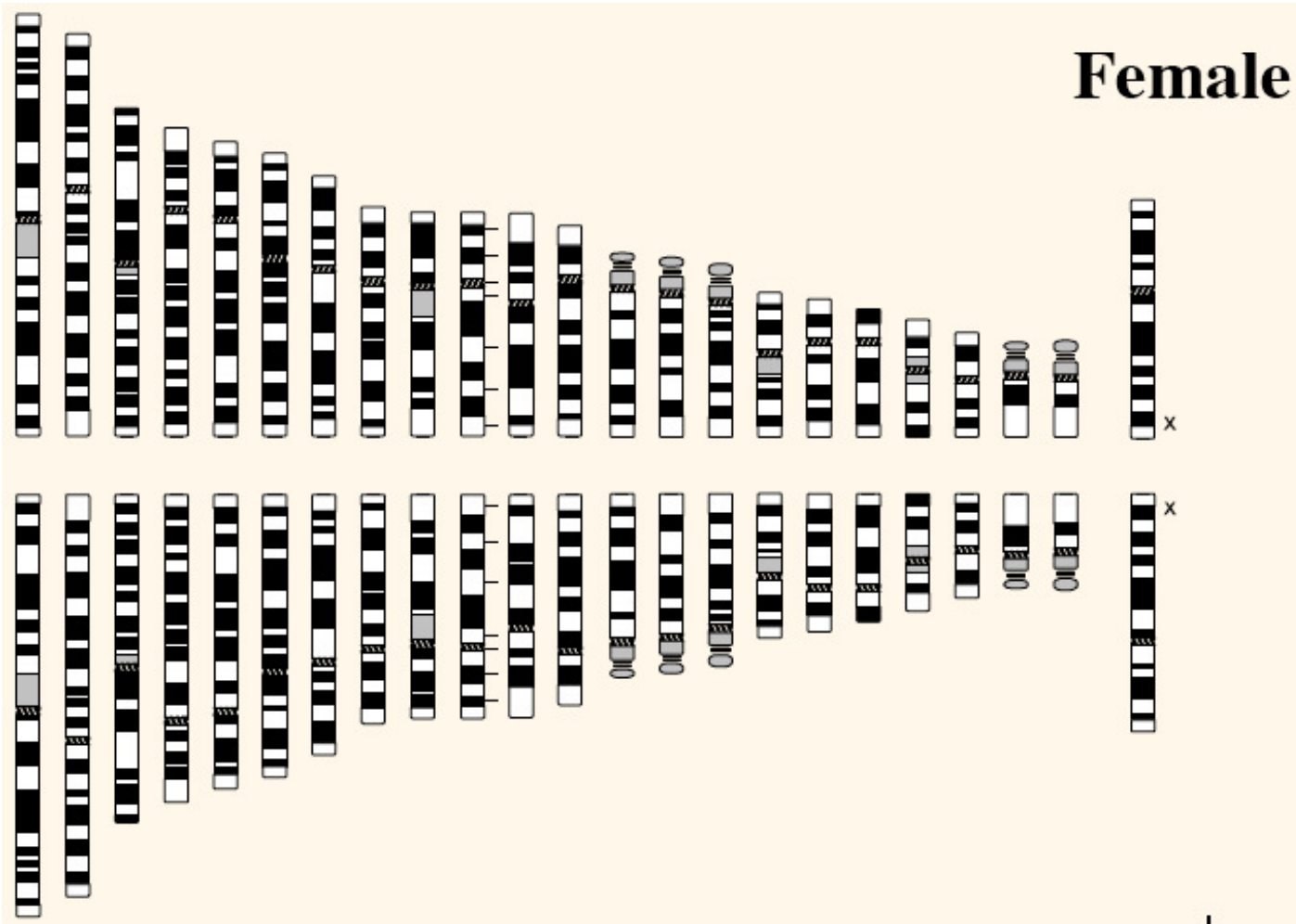
- The term *genome* refers to the complete complement of DNA for a given species
- The human genome consists of 46 chromosomes
 - Male: 22 pairs of autosomes + XY
 - Female: 22 pairs of autosomes + XX
- Every cell (except sex cells and mature red blood cells) contains the complete genome of an organism

Human Genome (Male)



22 pairs of autosomes + sex chromosomes (XY)

Human Genome (Female)



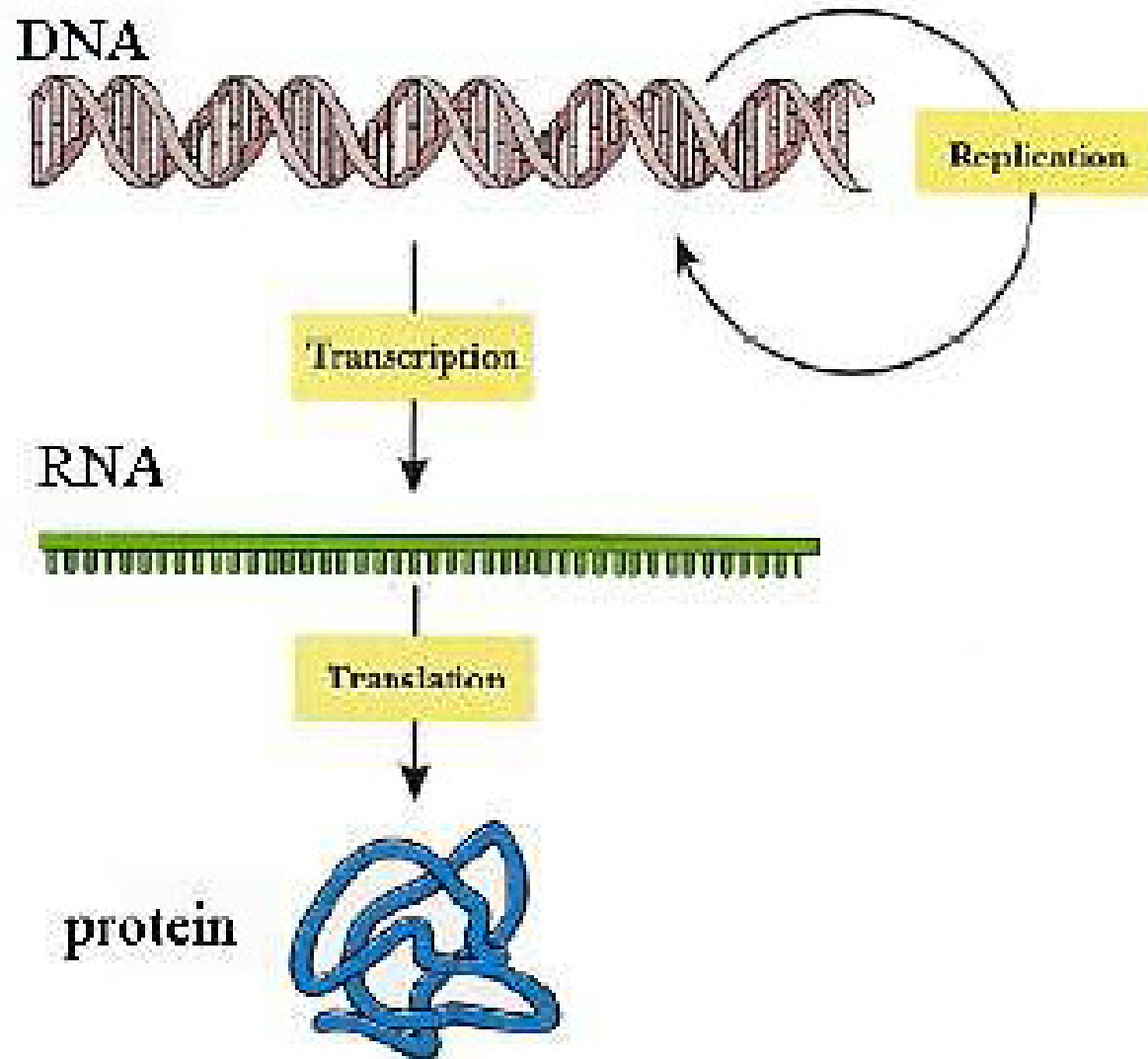
22 pairs of autosomes + sex chromosomes (XX)

Human Chromosomes



Karyogram

The Central Dogma



RNA

- RNA is like DNA except:
 - backbone is a little different
 - usually single stranded
 - the base uracil (**U**) is used in place of thymine (T)
- A strand of RNA can be thought of as a string composed of the four letters: A, C, G, **U**

The Genetic Code

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

64 combinations: 20 amino acids + stop codon

Proteins

- Proteins are molecules composed of one or more *polypeptides*
- A polypeptide is a polymer composed of *amino acids*
- Cells build their proteins from **20** different amino acids
- A polypeptide can be thought of as a string composed from a 20-character alphabet

Readout from the genome

