CS284A Representations & Algorithms for Molecular Biology

Xiaohui S. Xie

University of California, Irvine

Today’s Goals

- Course information
- Challenges in computational biology
- Introduction to molecular biology

Course Information

- Lecture: TT 3:30-4:50pm in PSCB 220
- Grading
  - 30% Homework
  - 20% Scribe note
  - 50% Final project
- Exams
  - no final exams
- Course Prerequisites:
  - Programming skill (Perl/Python, Matlab/R)
  - Statistics, Calculus, basic knowledge of Biology

Course Goals

- Introduction to computational biology
  - Fundamental problems in computational biology
  - Statistical, algorithmic and machine learning techniques
  - Directions for future research in the field

- Final project:
  - Propose an innovative project
  - Design novel or implement previous algorithms to carry out the project
  - Write-up goals, approach and findings in a conference format
  - Present your project to your peers in a conference setting

References

- Recommended Textbooks:
  - R. Durbin, S. Eddy, A. Krogh and G. Mitchison.   Biological Sequence Analysis
  - P. Baldi and S. Brunak. Bioinformatics: the Machine Learning Approach

- Course Website:
  http://www.ics.uci.edu/~xhx/courses/CS284A/

Why computational biology?

Computational biology/Bioinformatics is the application of computational tools and techniques to biology (mostly molecular biology).

- 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!

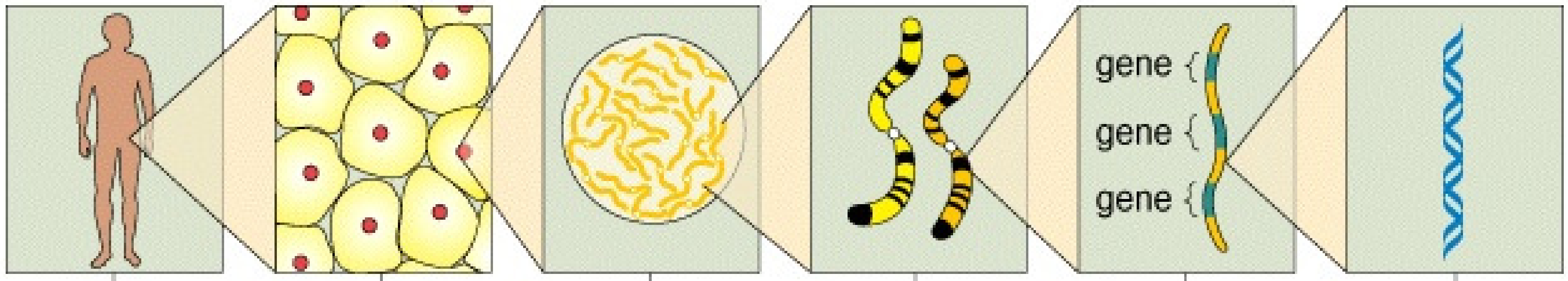
Four Aspects

- Biology
  - What’s the 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
  - Pattern finding (motif discovery, EM-algorithm)
  - Sequence alignment (Smith-Waterman, BLAST)
  - Models of sequences (HMM)
  - Gene discovery (HMM)
  - RNA folding (Stochastic context-free grammar SCFG)
- Algorithms for large-scale data analysis
  - Clustering algorithms (Hierarchical clustering, K-means)
  - Inference of networks (Regression, Bayesian networks)
  - Systems biology (Model and simulation)
- Evolutionary models
  - Phylogenetic trees
  - Comparative Genomics
- Protein world (if time allows)
  - Secondary & tertiary structure prediction

Introduction to Molecular Biology and Genomics



Organism (human)

A human body is made up of trillions of cells.

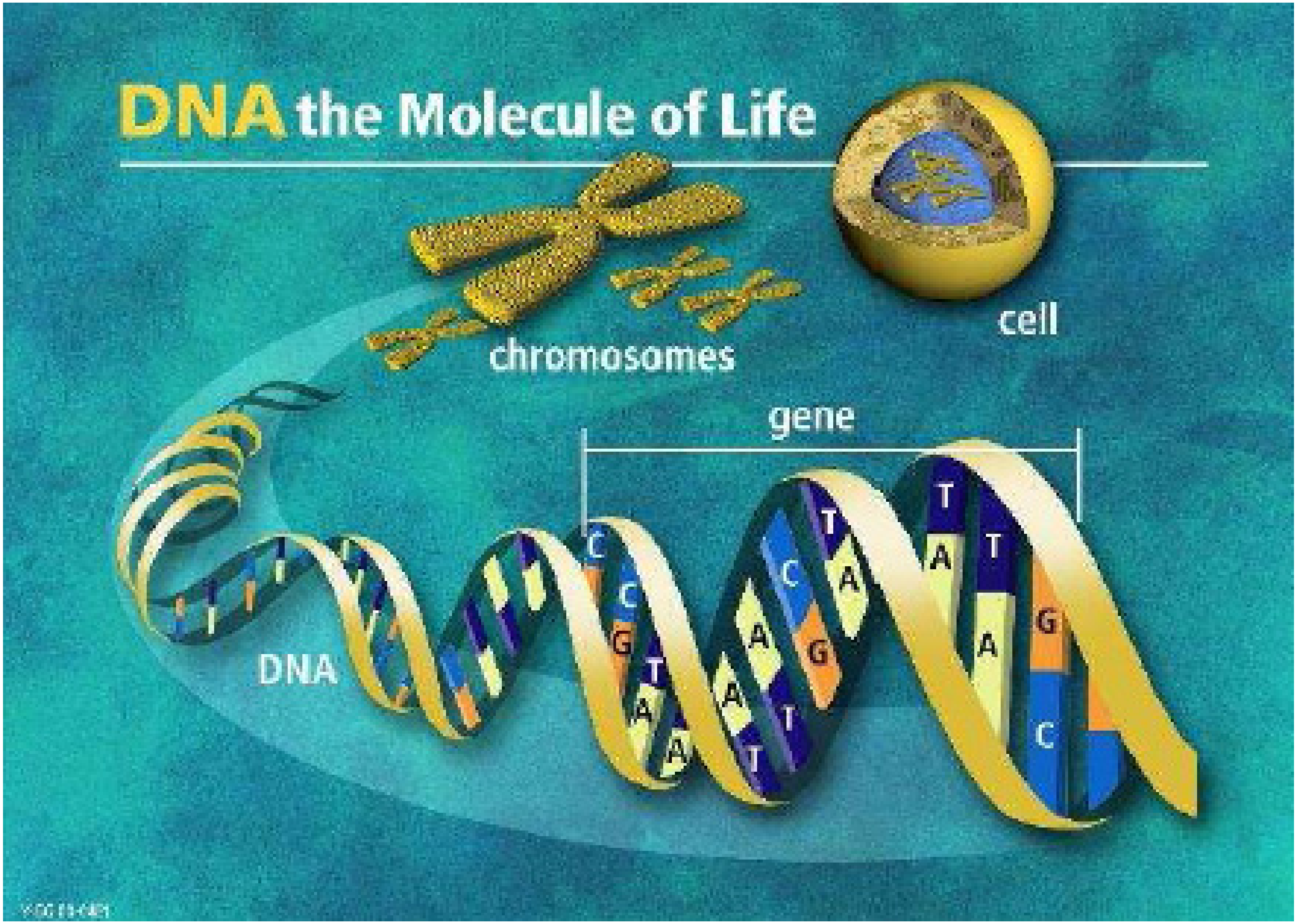
Each cell nucleus contains an identical complement of chromosomes.

One specific chromosome pair

Each chromosome is one long DNA molecule, and genes are functional regions of this DNA.

DNA is a double helix.

Different Life Forms Share a Common Genetic Framework


DNA the Molecule of Life
cell
chromosomes
gene
DNA
C
C
G
T
A
A
A
A
T
T
C
G
T
A
T
A
T
A
G
C

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)

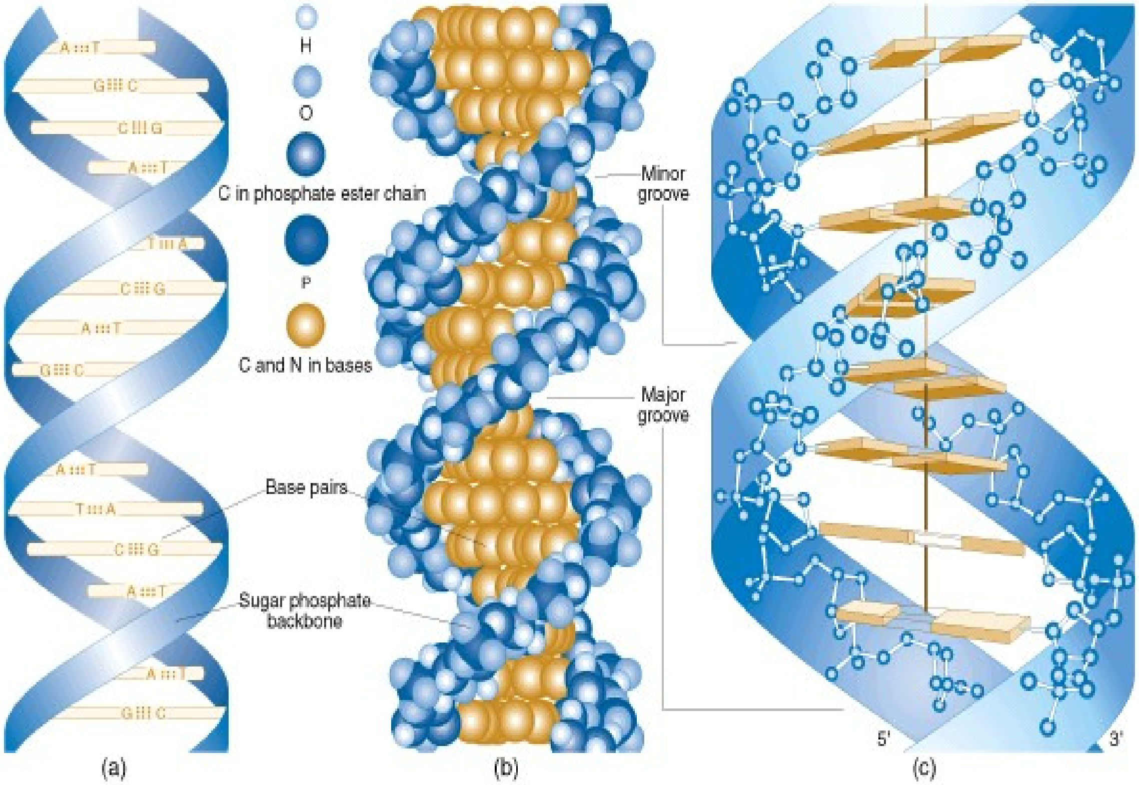
- DNA is digital information

- 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



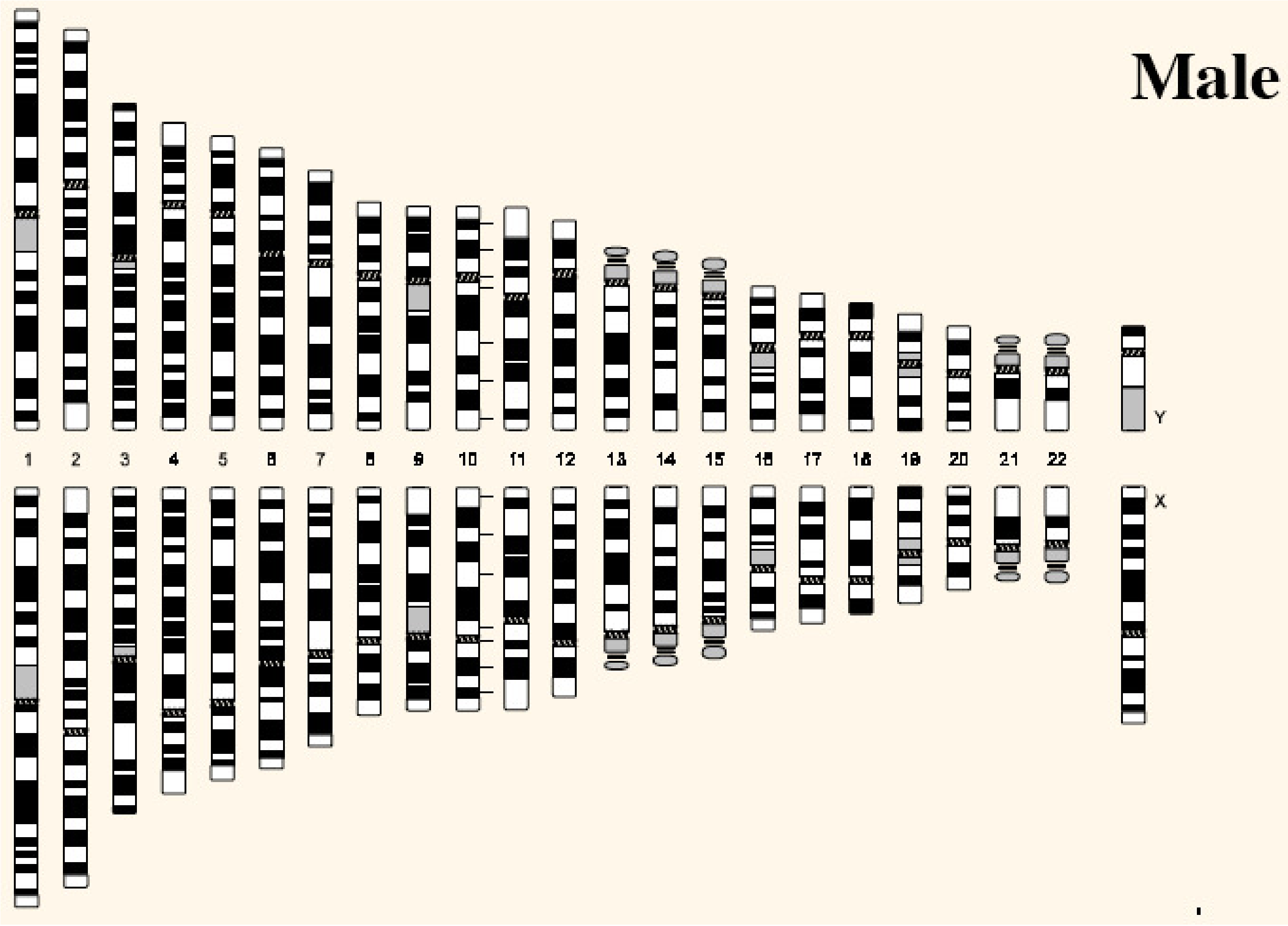
Genomes

- 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

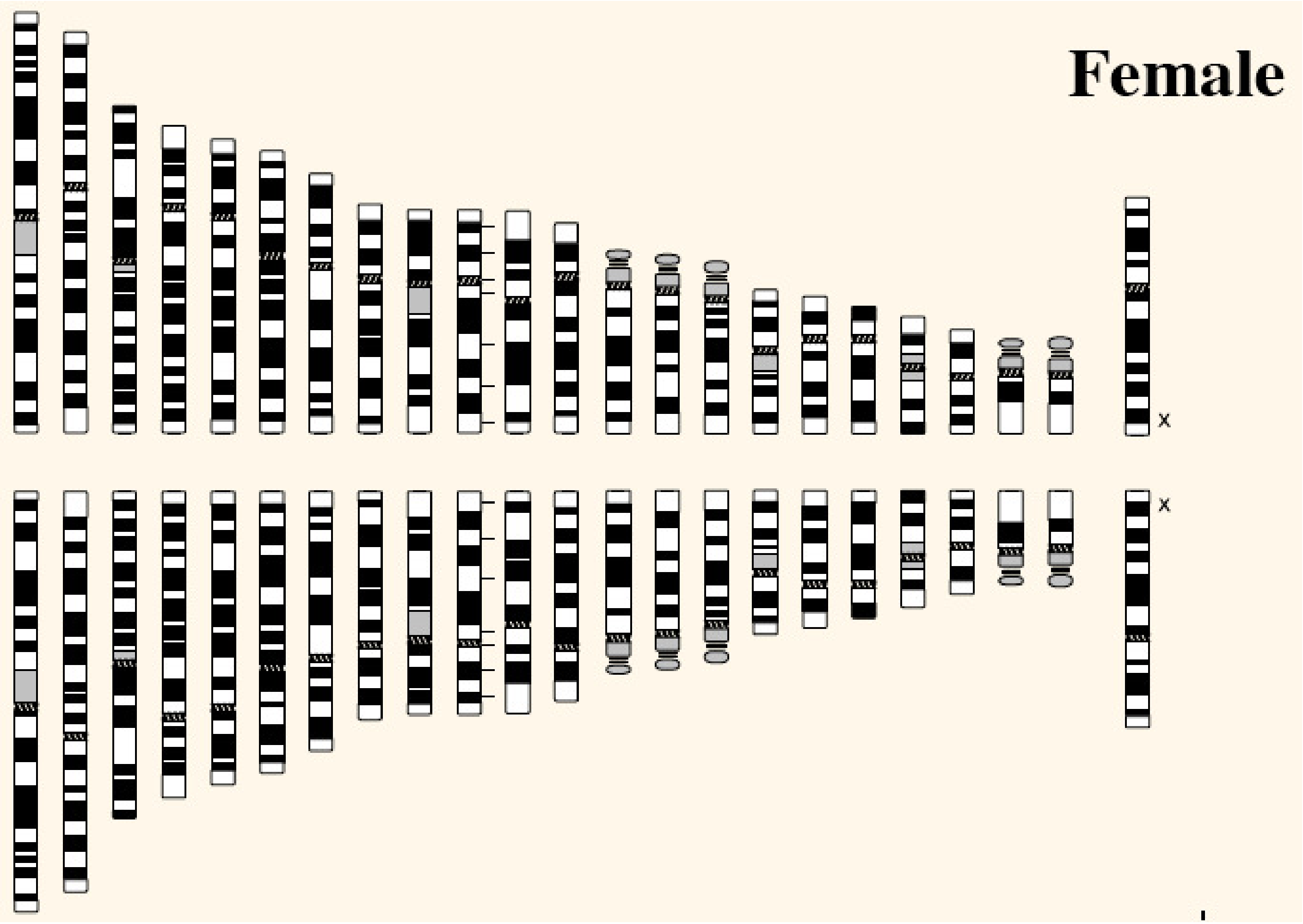
- Every cell (except sex cells and mature red blood cells) contains the complete genome of an organism

Human Genome (Male)



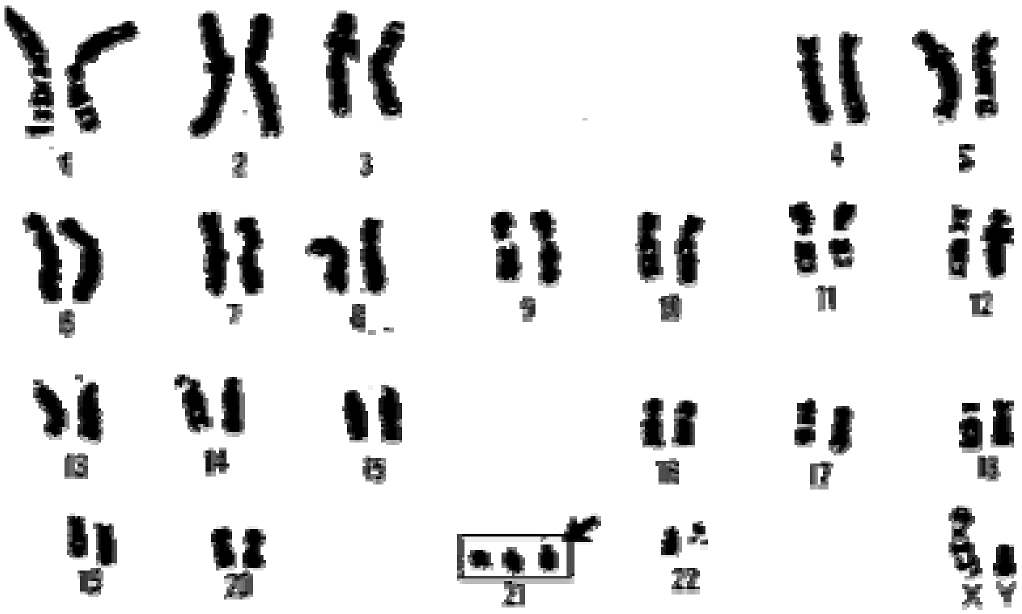
22 pairs of autosomes + sex chromosomes (XY)

Human Genome (Female)



22 pairs of autosomes + sex chromosomes (XX)

Human Chromosomes



Karyogram

Chromosomes

- DNA is packaged into individual *chromosomes* (along with proteins)
- *prokaryotes* (single-celled organisms lacking nuclei) have a single circular chromosome
- *eukaryotes* (organisms with nuclei) have a species-specific number of linear chromosomes
- DNA + associated chromosomal proteins = chromatin

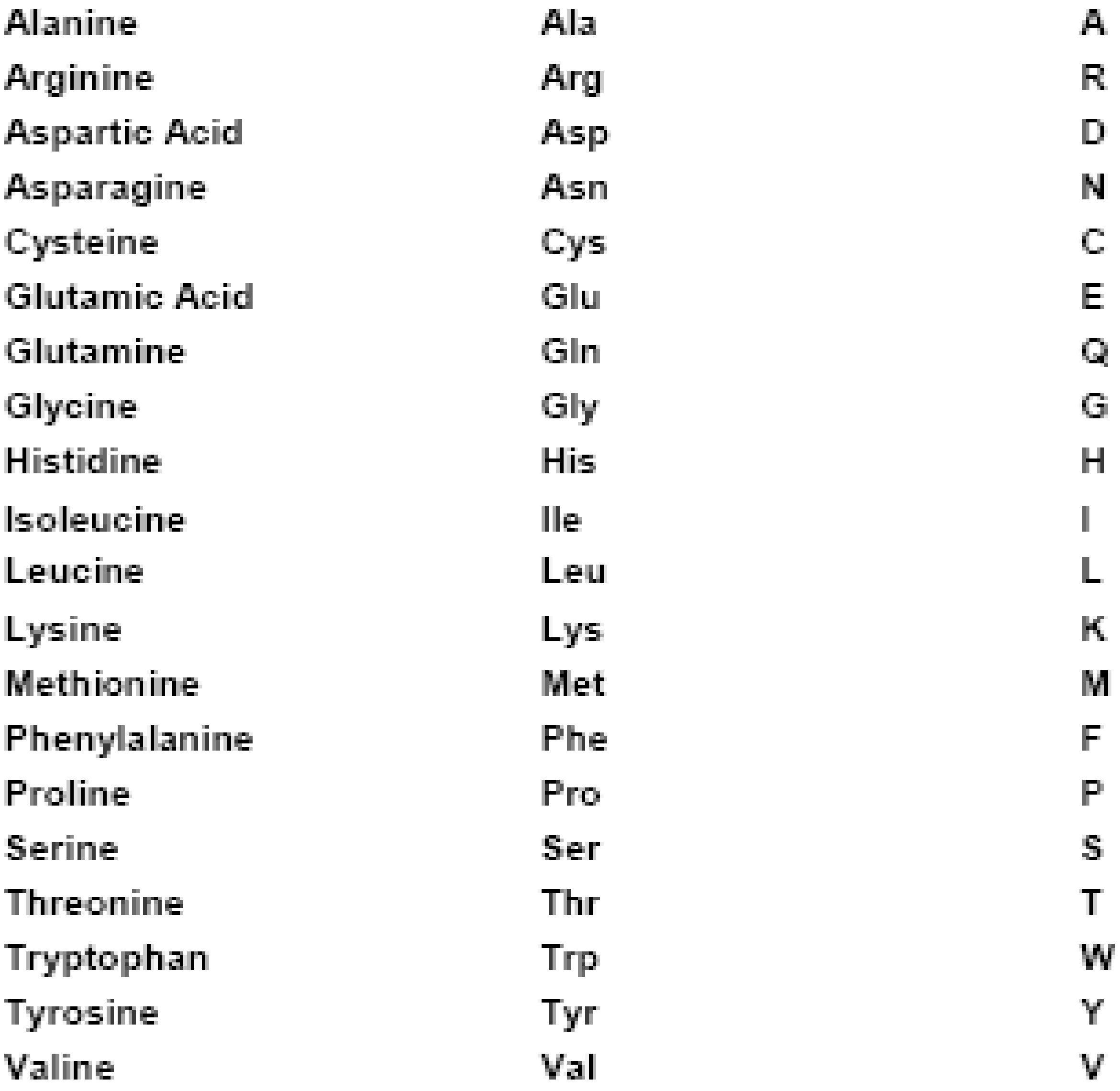
Proteins

- 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

Protein Functions

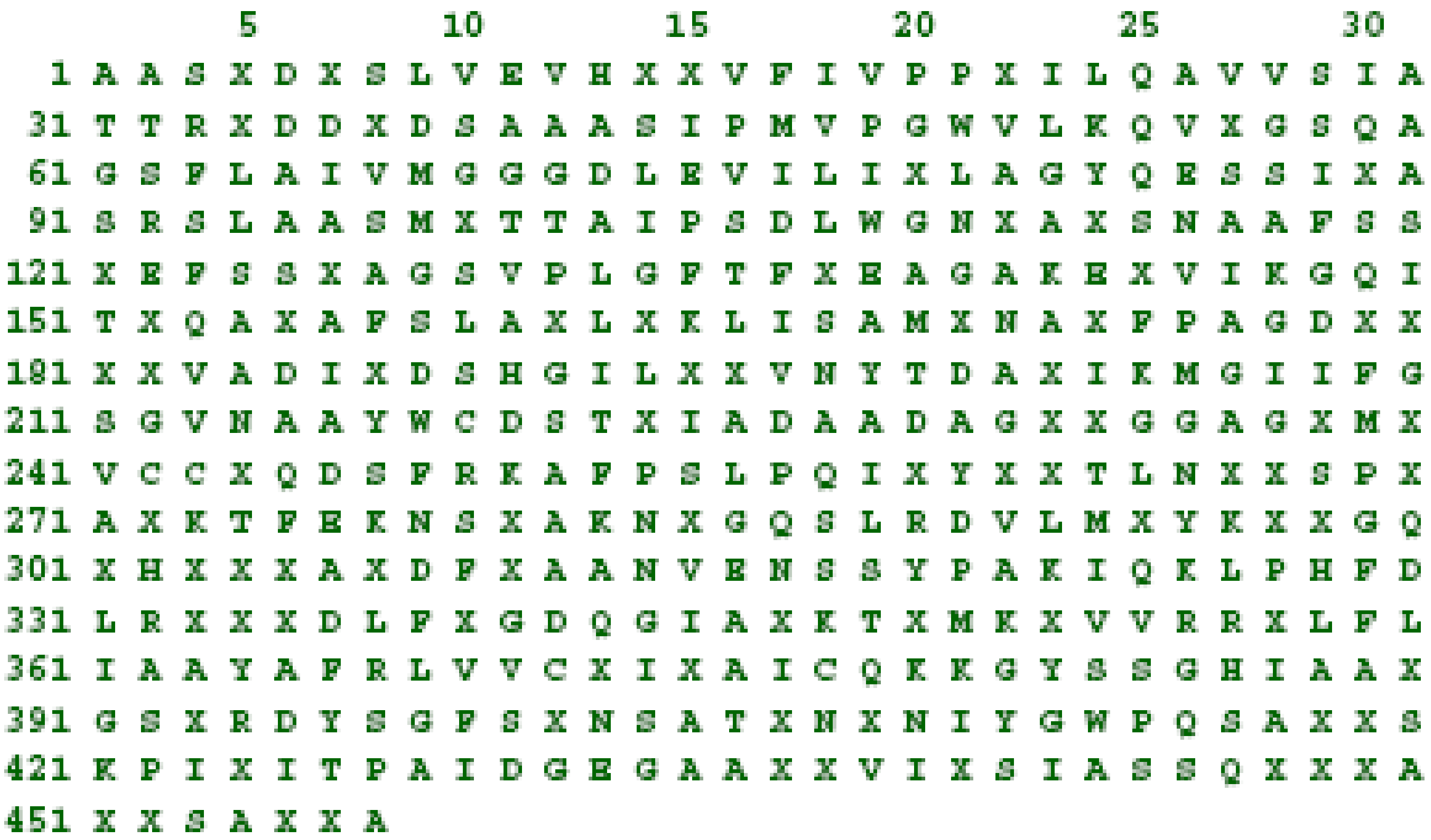
- structural support
- storage of amino acids
- transport of other substances
- coordination of an organism’s activities
- response of cell to chemical stimuli
- movement
- protection against disease
- selective acceleration of chemical reactions

Amino Acids



Alanine	Ala	A
Arginine	Arg	R
Aspartic Acid	Asp	D
Asparagine	Asn	N
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Amino Acid Sequence of Hexokinase

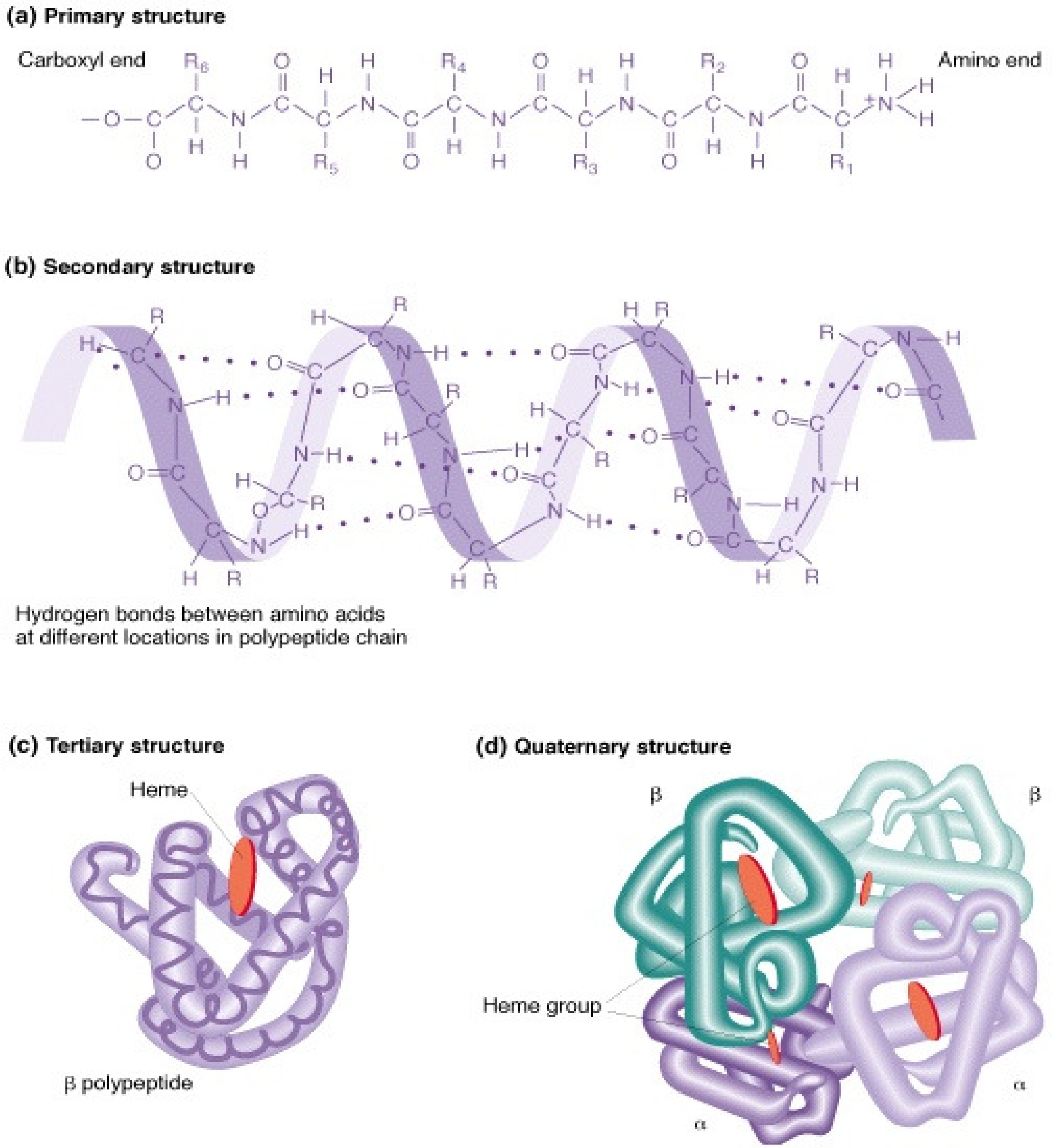


```
            5        10        15        20        25        30
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
121 X E F S S X A G S V P L G F T F X E A G A K E X V I K G Q I
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
```

Protein Structure

- Proteins are poly-peptides of 70-3000 amino-acids

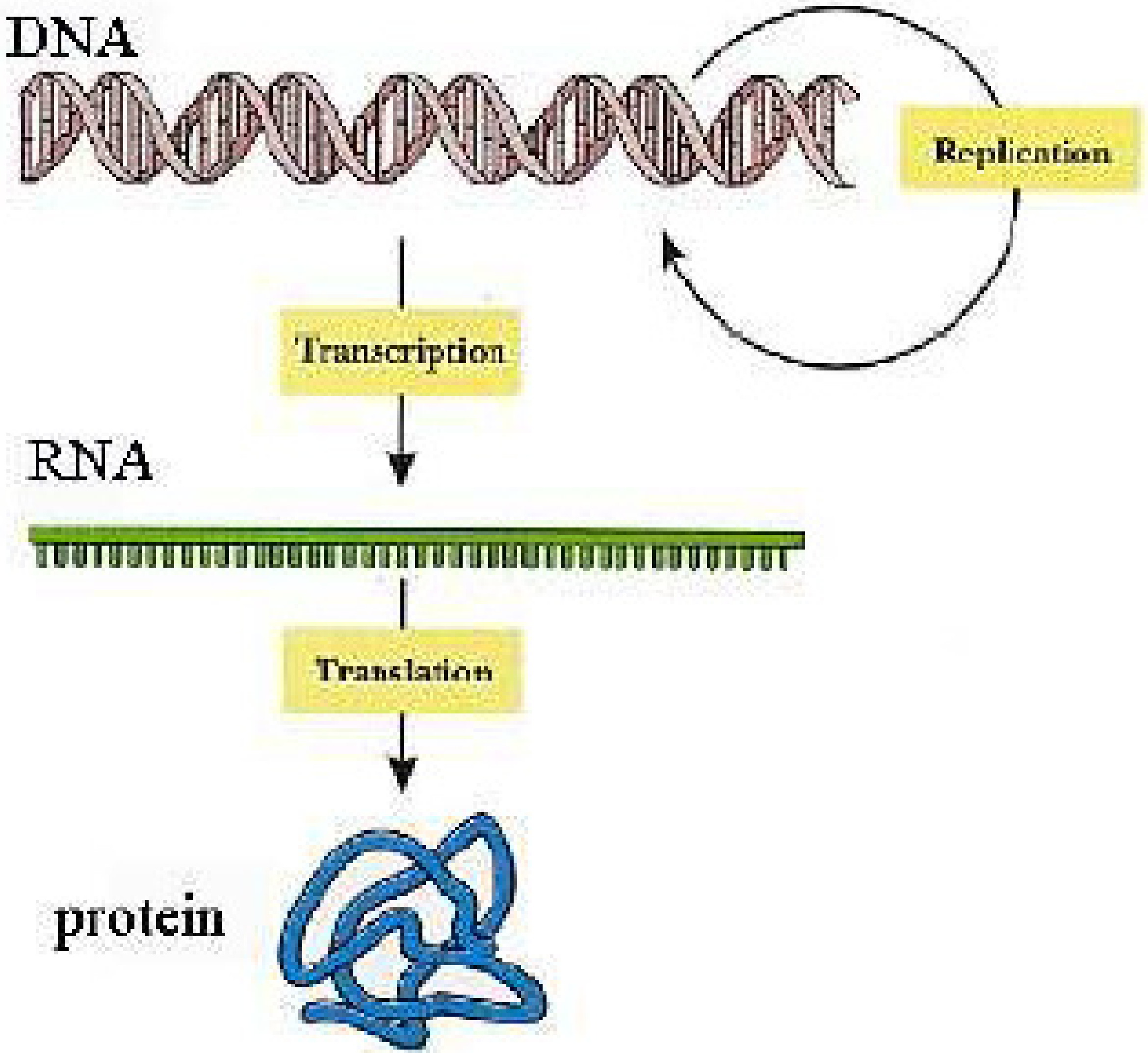
- This structure is (mostly) determined by the sequence of amino-acids that make up the protein



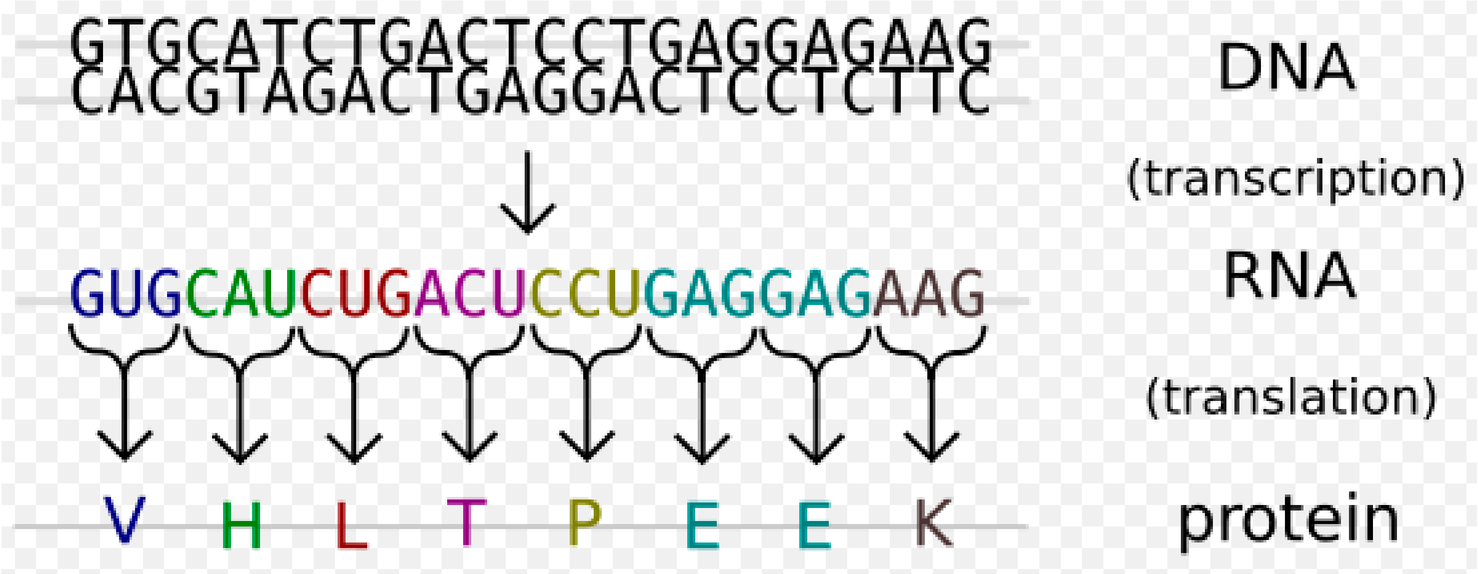
Genes

- Genes are the basic units of heredity
- A gene is a sequence of bases that carries the information required for constructing a particular protein (polypeptide really)
- Such a gene is said to *encode* a protein
- The human genome comprises ~22,000 genes
- Those genes encode >100,000 polypeptides
- RNA genes: microRNAs and other small RNAs

The Central Dogma



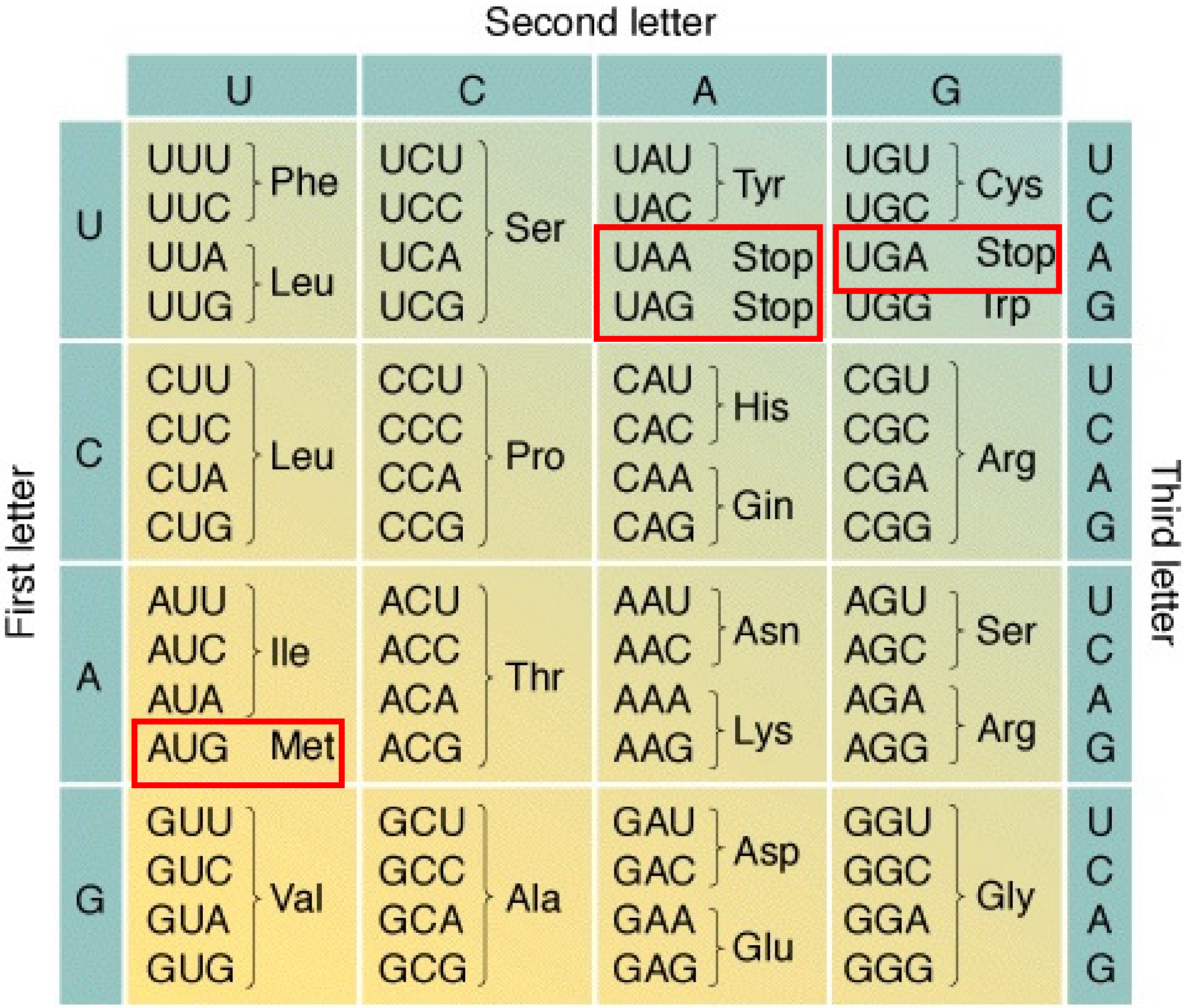
Genetic code: DNA -> mRNA -> protein



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



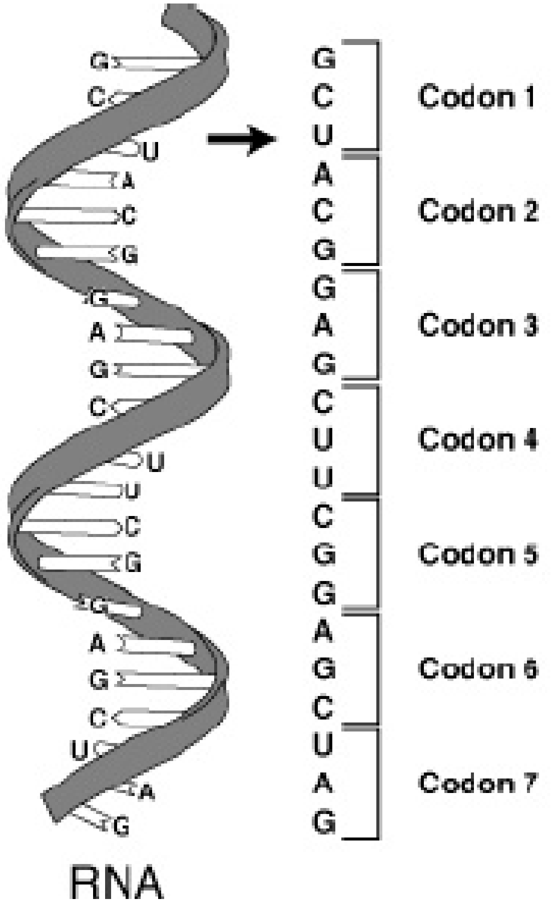
	Second letter								
First letter	U		C		A		G		Third letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	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

64 combinations: 20 amino acids + stop codon

Translation

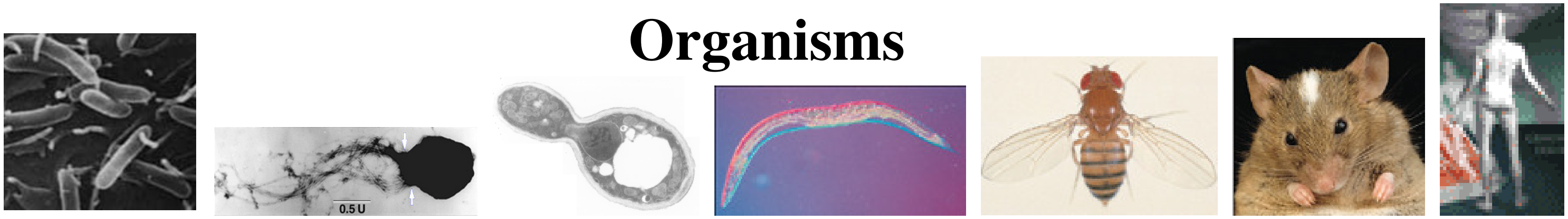
- *Ribosomes* are the machines that synthesize proteins from mRNA
- The grouping of codons is called the *reading frame*
- Translation begins with the *start codon*
- Translation ends with the *stop codon*

Codons and Reading Frames



Comparison of genome size

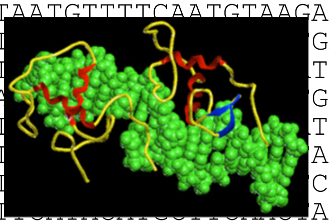
Organisms



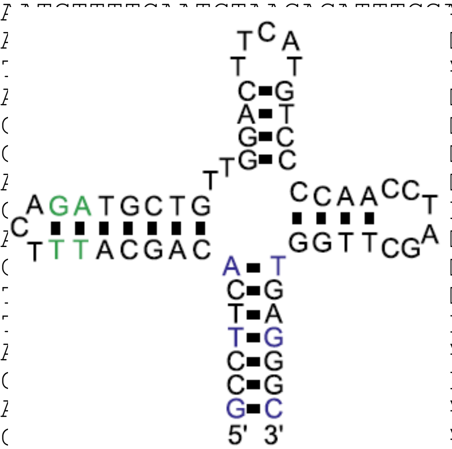
Genomes

	Haemophilus influenzae	Methannococcus jannaschii	Saccharomyces cerevisiae (baker's yeast)	Caenorhabditis elegans (nematode worm)	Drosophila Melanogaster (fruit fly)	Mus musculus (laboratory mouse)	Homo sapiens (man)
Genome (MB)	1.83	1.66	13	97	180	3200	3500
Number of genes	1709	1682	6241	18,424	13,500	~30,000	~30,000

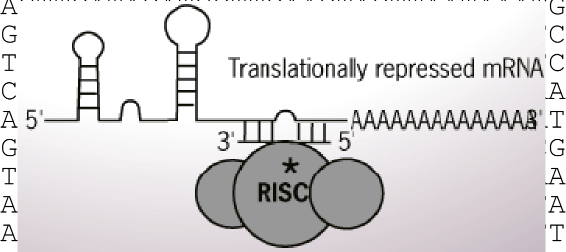
AGATTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATA
CATGCATGCTTCAATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTACTTAATAAATGATTGTATGATAATG



TTTTCAATG**TAAGAGATTTC**GAT GATAATGTTTTCTCCTTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAA
TTAATAAATGATTGTATGATAAT TTTCGATTATCCTTATAGTTCATACATGCATGCTTCAACTGAGATTTCGATTATCCTTATAGTTCATAC
ATGCATGCTTCAACTACTTAATA GTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTACTTAATAAATGATTGTATGATAATGTTTTCA
ATGTAAGAGATTTCGATTATCCT TGCTTCAACTACTTAATAAATGATTGTATGATA ATTATCCTTATA

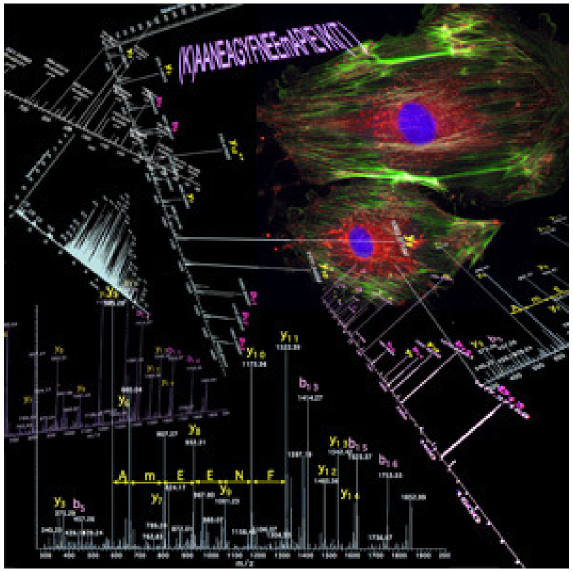


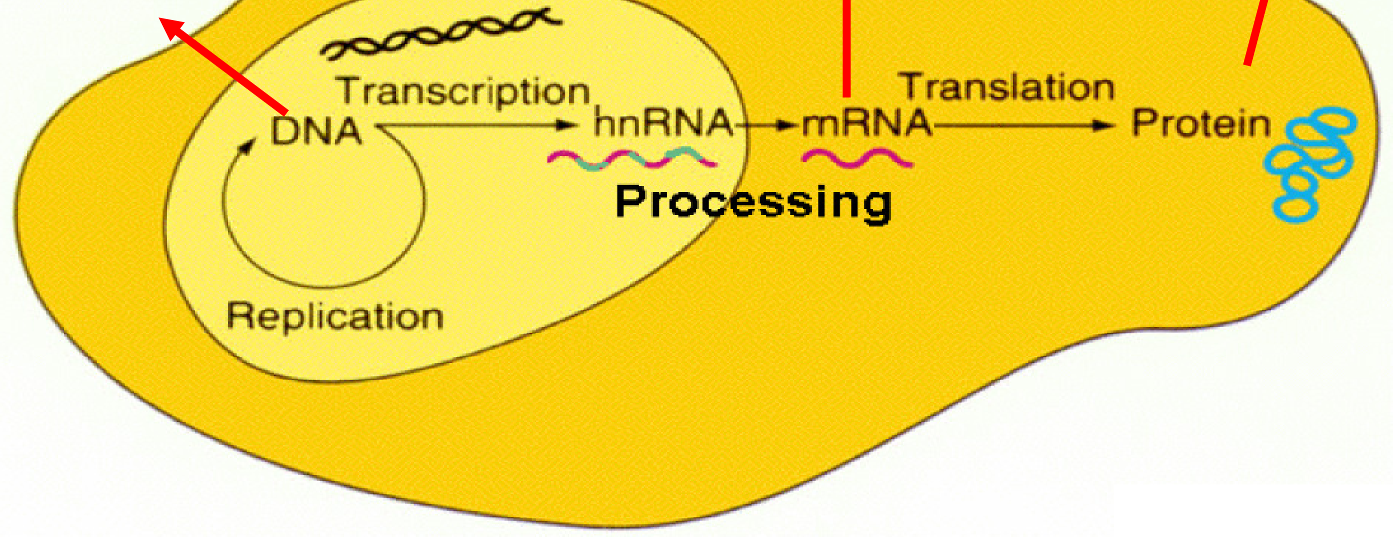
GTTCA**TACATGCATGCTT**CAACT ATGATAATGTTTTCTCCTTATCCTTATAGTTCA TGATTGTATGAT
AATGTTTTCAATGTAAGAGATTT CATACATGCTTCAACTACTTAATAAATGATTG ATTTCGATTATC
CTTATAGTTCATACATGCATAGT ACTTAATAAATGATTGTATGATAATGTTTTCA TAGTTCATACAT
GCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTAGTTCATAC TTGTATGATAAT
GTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTT**TCAATGTAA**GAGATTTCGATTATC TTAATAAATGAT
CAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACA TGTATGAATTTC
GATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAG CACATGCTTCAA
CTACTTAATAAATG**CAGATGCTGTTGGACTTCATGTCCCCAACCTAGCTTGGTGCACAGCATTT**ATTGTATGA TCATACATGCAT
AGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATC TACTTAATAAAT
GATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCAATGTAAGAGATTTCGATTATCCTTATAGT TAATAAATGATT
GTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGTATTGAATTTTCAAAAAT CAAAGAAGTTTA
ATAATCATATTACATGGCATTACCACCATATACATATCCATATCTAATCTTAC**TATA**TGTTGTGGAAATGTAA AAAACCTTCTCT
TTGGAACTTTCAGTAATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCC CTCCGTGCGTCC
TCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTA AGAGGAAAATT
GGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATTAC ATTAATCAGCGA
AGCGATGATTTTTGATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCA
TAAAAGTATCAACAAAAAAATTGTTAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATA**ATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA**
TTCTAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGGATTTTGTTGCTAGATCGCCT
GGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGATTTTGATATGCTTTGCGCCGTCAAAGTTTTGAACGATGAGA
TTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAATCTTTAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAG
ACTTTTTCAAGCAATTTGGTGCCTTGATGAACGAGTCTCATTCAGGTTGGTACGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTT
GTCAAATGGATCATATGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAAAAGAAGCC
CTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCAGCATTGGGCAGCTGTCTATATGAATTAG
TCAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACTTTAGCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTAT
AGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATA
TGCTTTCAACCGCTGCGTTTTGGATACCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGAAAATTTGC
GAAGTTCTTGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAG
ATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGGATACCTATTCTTGACATGATATGACTACCATTTTGTTA
TTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCTTGGCAAGTTGCCAACTGACGAGATGCAGTTTCCTACGCATAATAAGAATAGGAG
GGAATATGCAG**GAGAACGCCAGACAATCTATCATTACATTTAAGCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAA**
ATAATGTGGATTTGGAAAAAGAGTATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CAGCTCATTCTGGAAGAAATAGTGTTTCTTGTACAACCAGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGGATTGGGTACGGTTTCGTTGGTGCTTTTGTTGT
TTTGGCCTCTAGAGTTGGATCTGCTTATCATTTGTCATTCCCTATATCATCTAGAGCATCATTCGGTATTTTCTTCTCTTTATGGCCCGTTATTAACAGAGTCGTCATG
GCCATCGTTTGGTATAGTGTCCAAGCTTATATTGCGGCAACTCCCGTATCATTAATGCTGAAATCTATCTTTGGAAAAGATTTACAATGATTGTACGTGGGGCAGTTGA



CGTCTTATCATATGTCAAAGTCATTTGCGAAGTTCTTGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTTTATA TGA
TTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAACACAGGGACAAAATTCTTGATATG CT
TGACATGATAT**GACTACCAT**TTTGTTATTGTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGAT CC
TTATAGTTCATACATGCTTCAACTACTTAATAA**TGCACTGTA**TGATAATGTTTTCAATGTAAGAGATTTCGATTATC AT
GATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAA TA
GTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGATTACTTAATAAATGATTG GA
TTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTCATACATGCTTCAACTAC**TGTAAATAA**TTAAT AG
AGATTTCGATTATCCTTATAGTTCATACATGCATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA AT
AGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCA**AATAAAA**TGTAAGAGATTTCGATTA TG
ACATGATATGACTACCATTTTGTTATTGTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTCAATGTAAGAGATTTCGATTTGATTG
TATGATAATGTTTTCAATGTAAGAGATTTCGATTATCCTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATAATGTTTTGTATGATTTATATCG

Readout from the genome




Transcription
DNA
hnRNA
mRNA
Translation
Protein
Processing
Replication