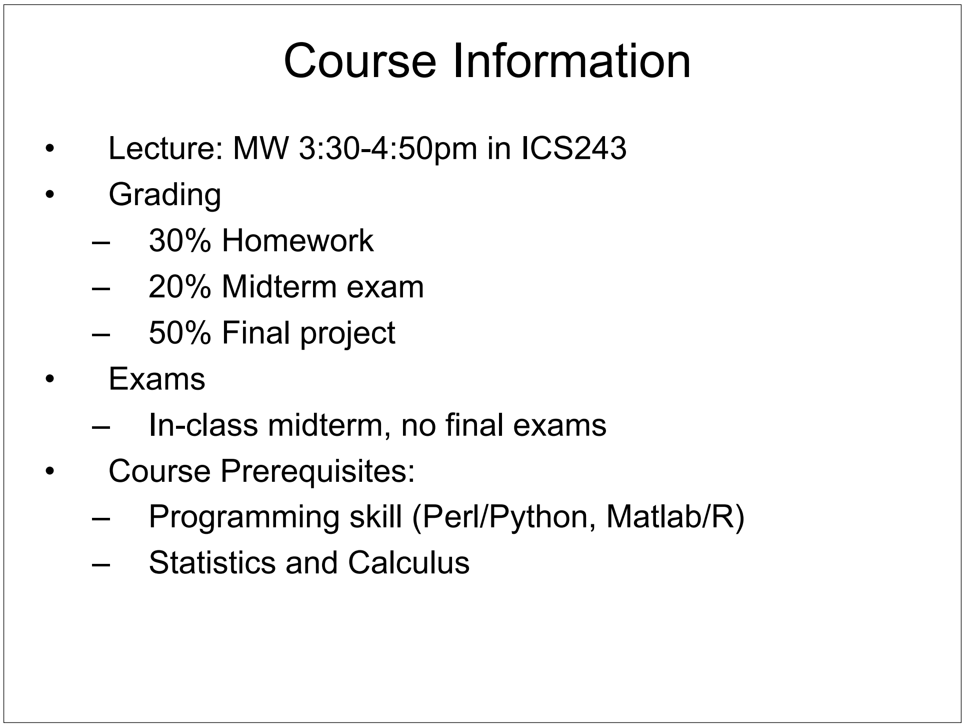
# CS284A Introduction to Computational Biology and Bioinformatics

Xiaohui S. Xie
University of California, Irvine

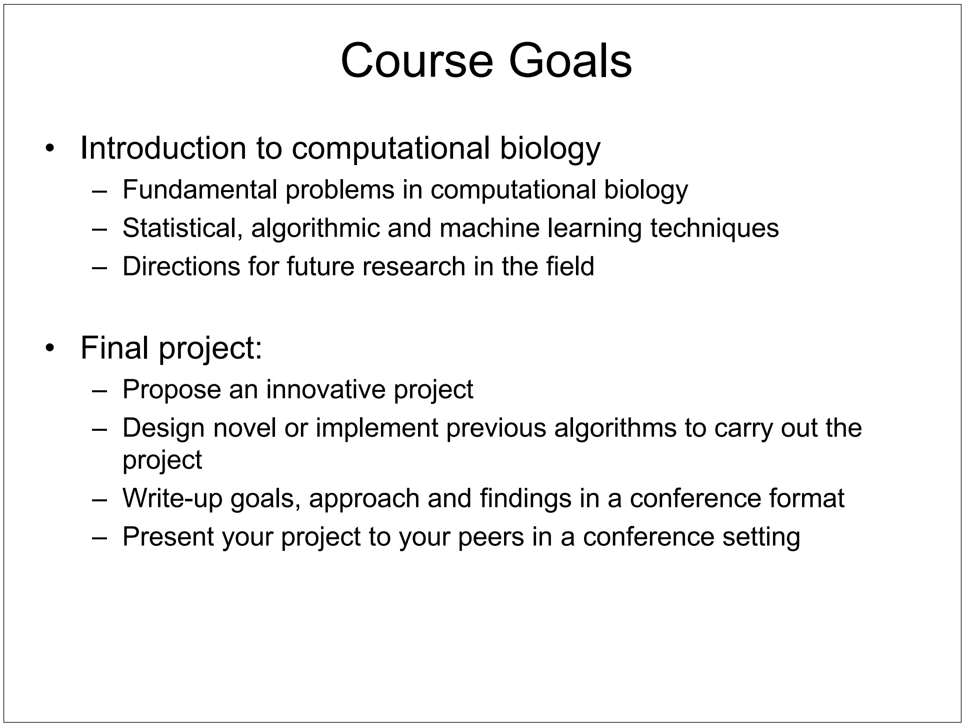
## Today’s Goals

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



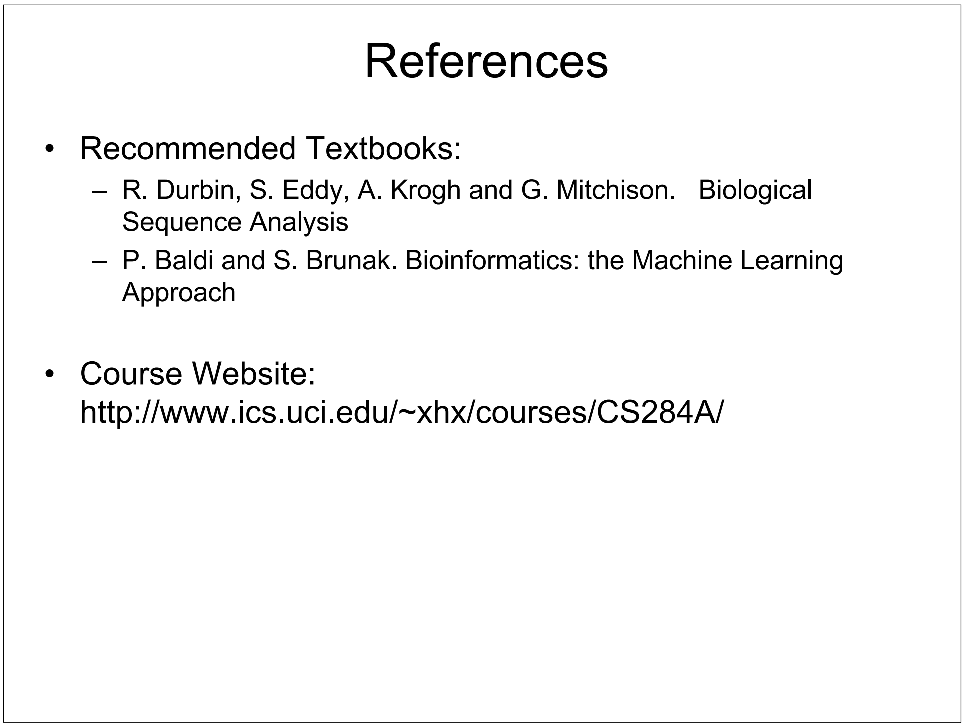
# Course Information

- Lecture: MW 3:30-4:50pm in ICS243
- Grading
  - 30% Homework
  - 20% Midterm exam
  - 50% Final project
- Exams
  - In-class midterm, no final exams
- Course Prerequisites:
  - Programming skill (Perl/Python, Matlab/R)
  - Statistics and Calculus



# Course Goals

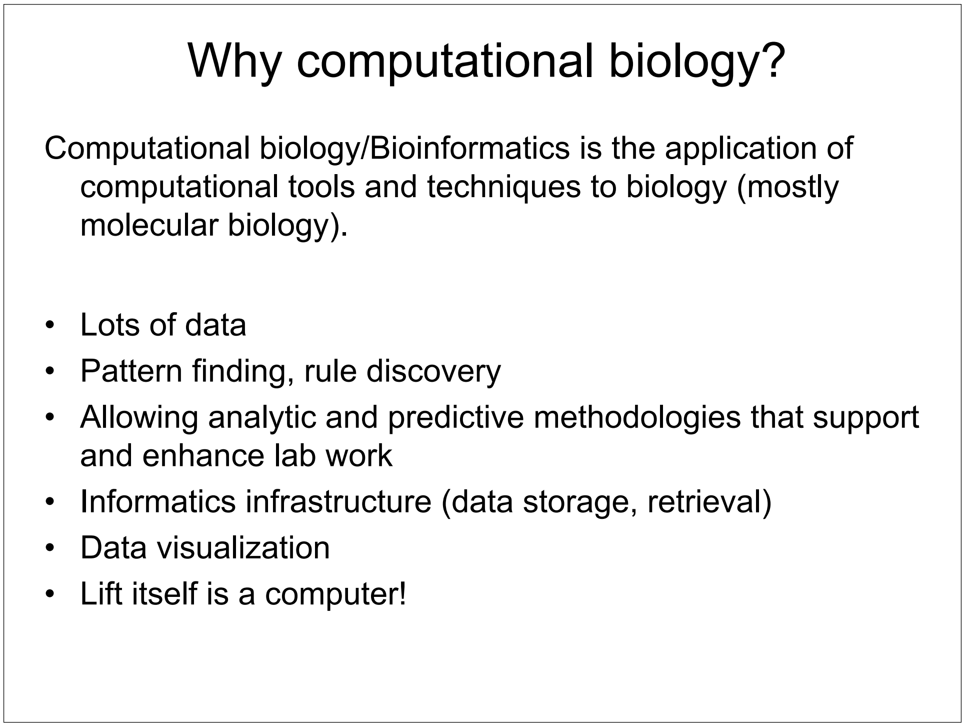
- Introduction to computational biology
  - Fundamental problems in computational biology
  - Statistical, algorithmic and machine learning techniques
  - Directions for future research in the field
- Final project:
  - Propose an innovative project
  - Design novel or implement previous algorithms to 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

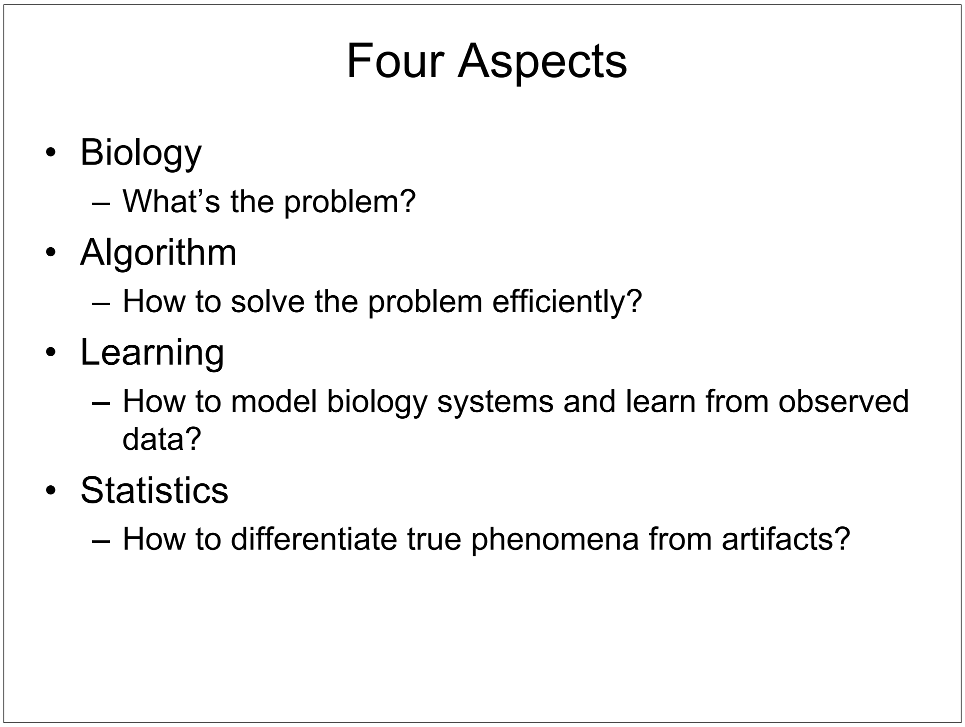
- Course Website: http://www.ics.uci.edu/~xhx/courses/CS284A/



# Why computational biology?

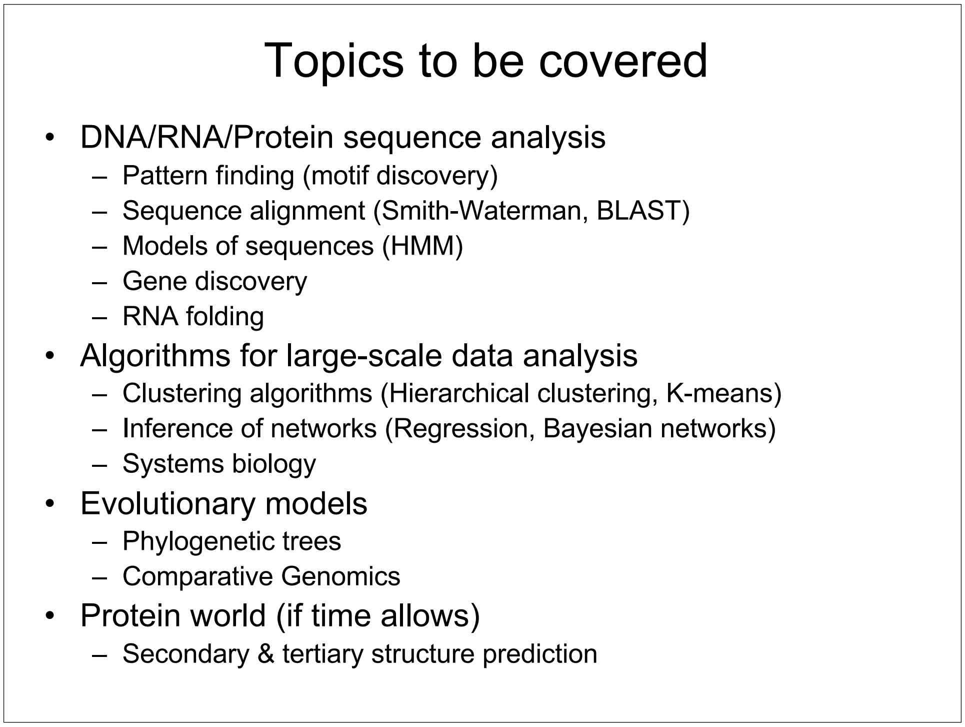
Computational biology/Bioinformatics is the application of computational tools and techniques to biology (mostly molecular biology).

- 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
- Lift 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?

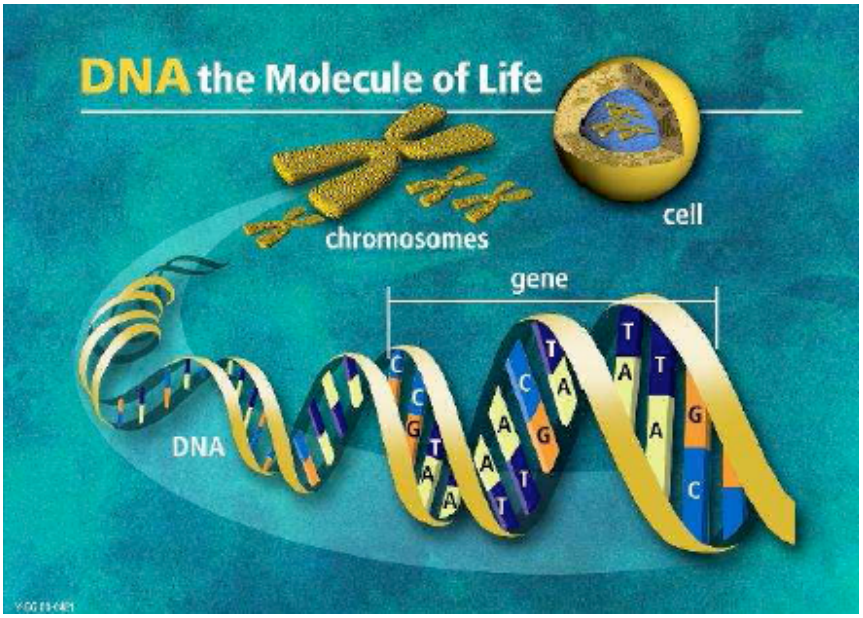


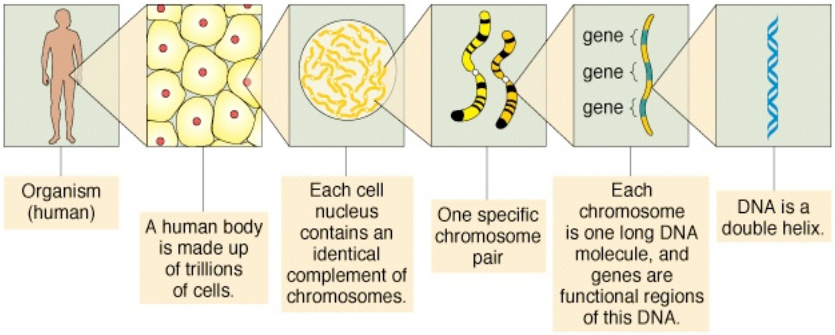
# Topics to be covered

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

# Introduction to Molecular Biology and Genomics

Slides from Mark Cravens





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.

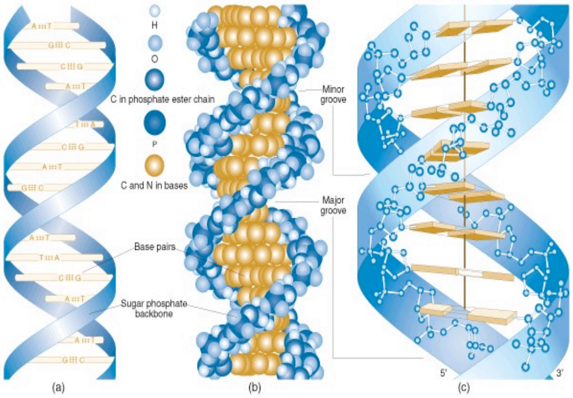
# Deoxyribonucleic acid (DNA)

- 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
```

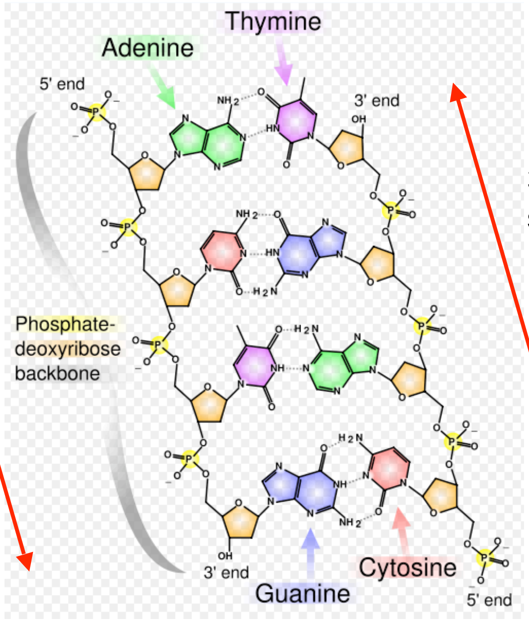
# The Double Helix

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



# Watson-Crick Base Pairs

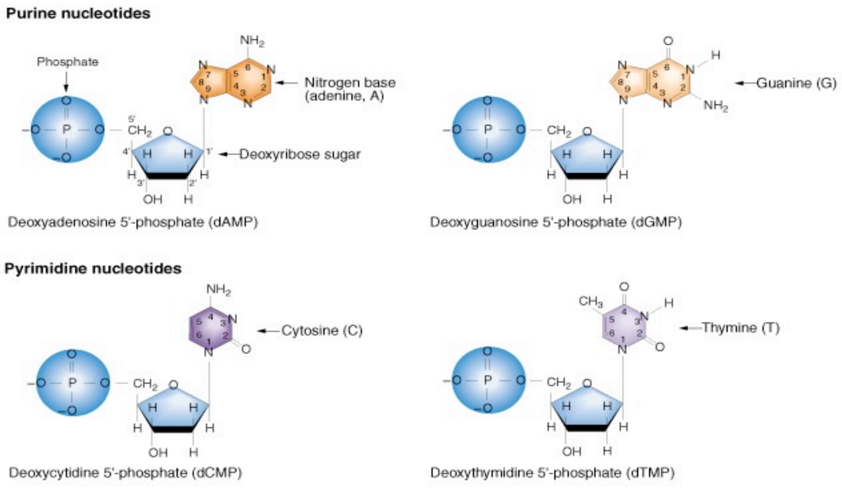
- A bonds to T
- C bonds to G



5'-3' strand

3'-5' strand

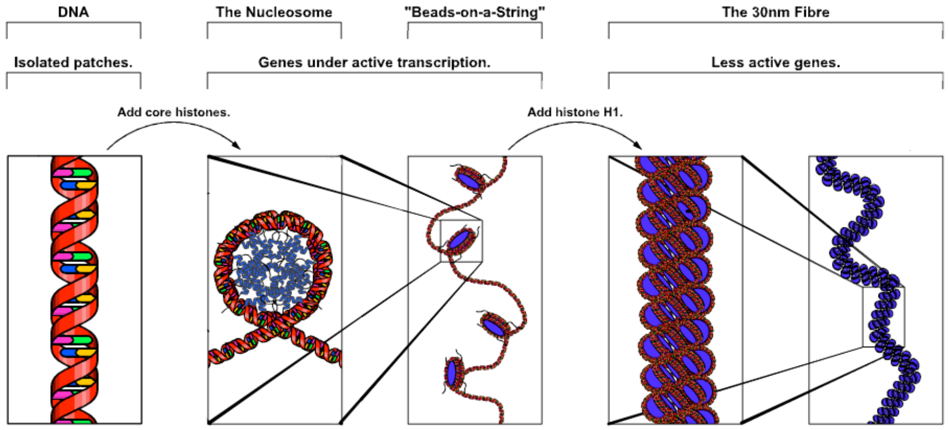
# Four nucleotides


Purine nucleotides
Phosphate
NH2
Nitrogen base (adenine, A)
5'
CH2
4'
1'
3'
2'
Deoxyribose sugar
OH
H
Deoxyadenosine 5'-phosphate (dAMP)
Guanine (G)
NH2
Deoxyguanosine 5'-phosphate (dGMP)
Pyrimidine nucleotides
NH2
Cytosine (C)
Deoxycytidine 5'-phosphate (dCMP)
CH3
Thymine (T)
Deoxythymidine 5'-phosphate (dTMP)

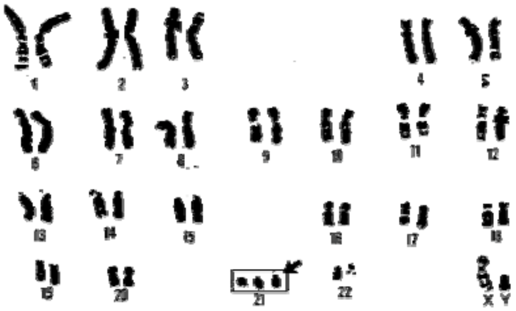
# Chromosomes

- 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

## DNA organization



## Human Chromosomes



Karyogram of a human male

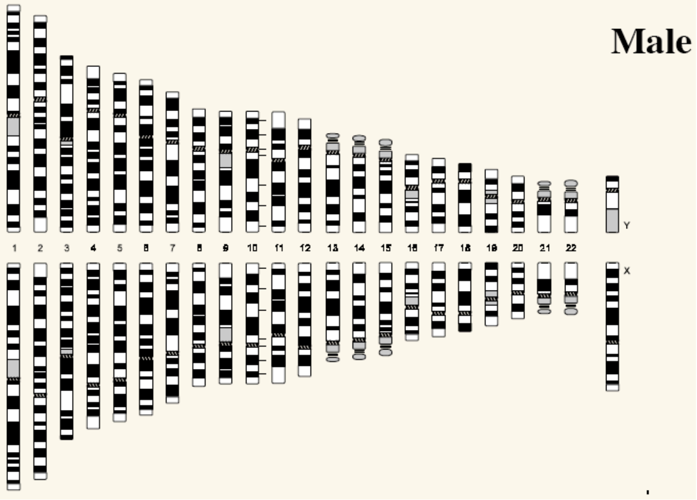
# Genomes

- 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

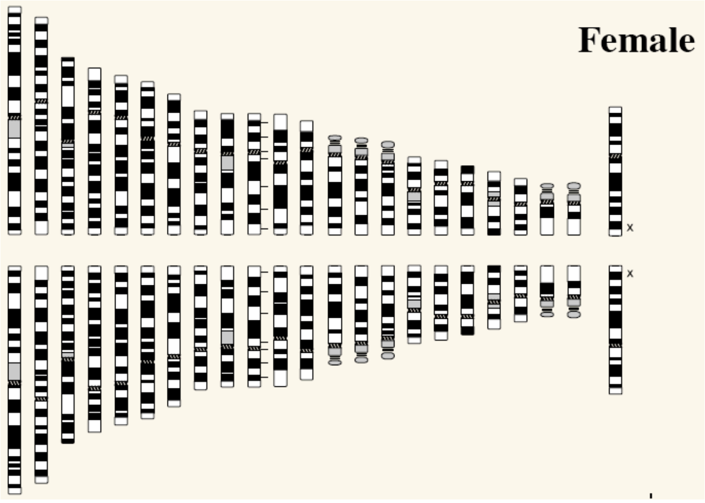
- Every cell (except sex cells and mature red blood cells) contains the complete genome of an organism

# Human Genome (Male)



22 pairs of autosomes + sex chromosomes (XY)

## Human Genome (Female)



22 pairs of autosomes + sex chromosomes (XX)

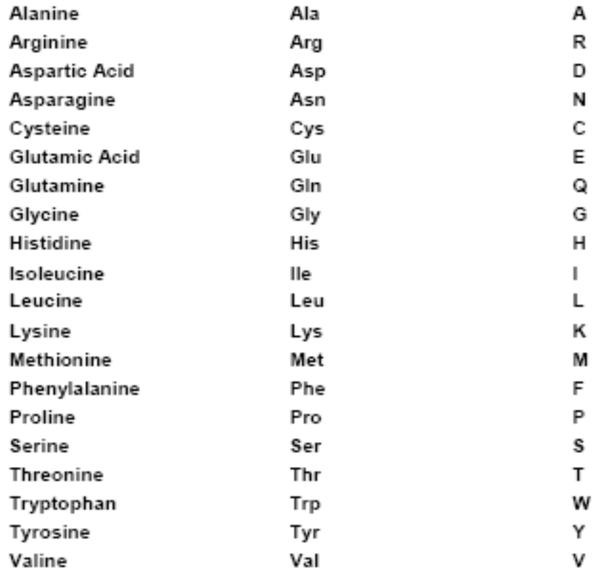
## Proteins

- 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

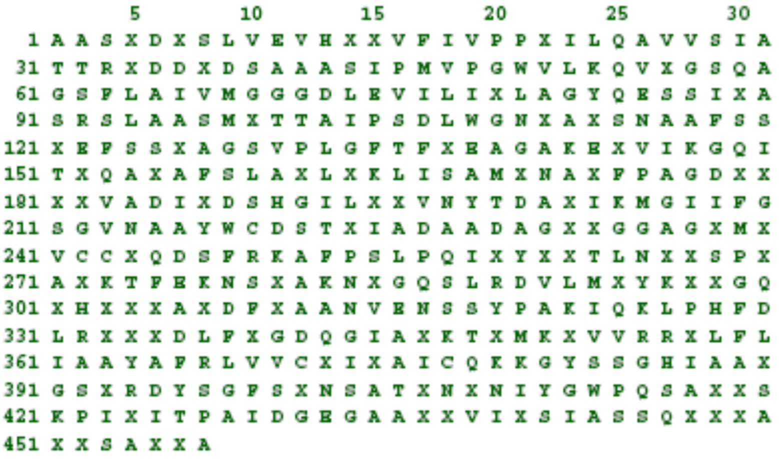
- structural support
- storage of amino acids
- transport of other substances
- coordination of an organism’s activities
- response of cell to chemical stimuli
- movement
- protection against disease
- selective acceleration of chemical reactions

# 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

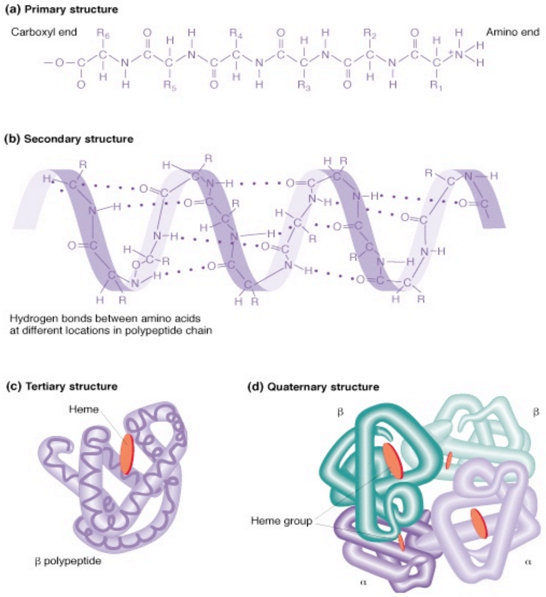


```
            5          10          15          20          25          30
  1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
 31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
 61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
 91 S R S L A A S M X T T A I P 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



(a) Primary structure

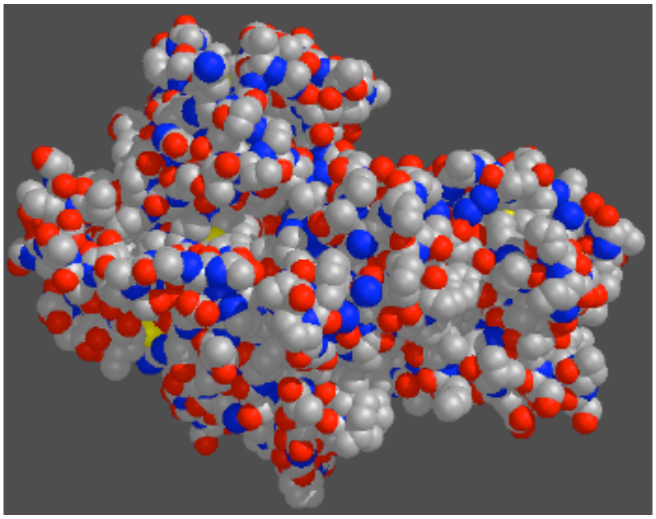
(b) Secondary structure

Hydrogen bonds between amino acids at different locations in polypeptide chain

(c) Tertiary structure

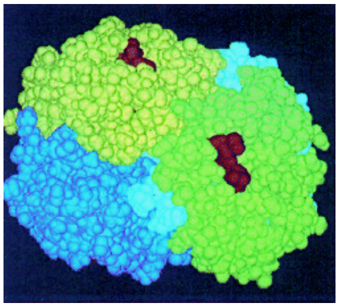
(d) Quaternary structure

# Space-Filling Model of Hexokinase



# Hemoglobin

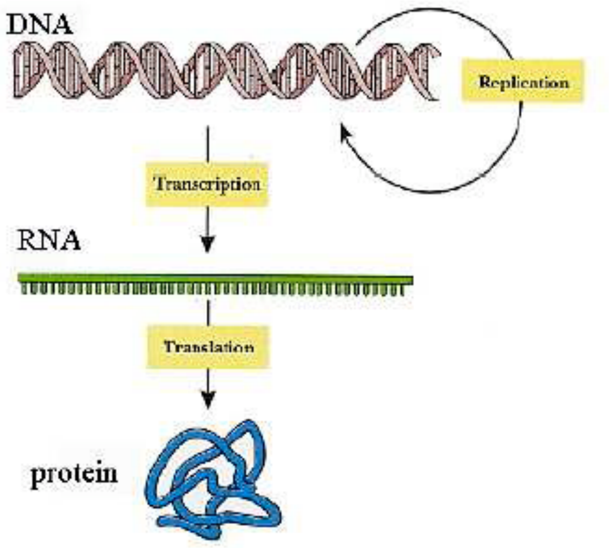
- protein built from 4 polypeptides
- responsible for carrying oxygen in red blood cells



## 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

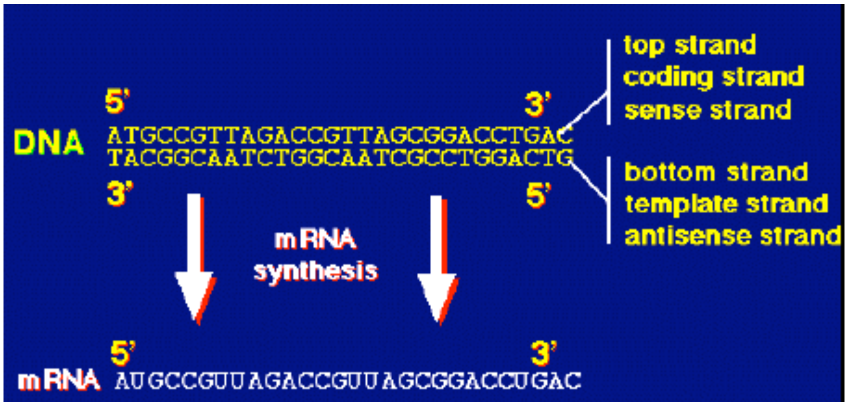
## The Central Dogma



# 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

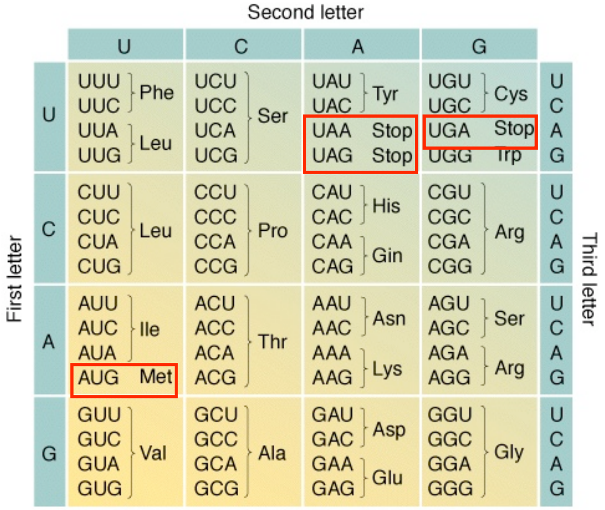
# Transcription



# Transcription

- *RNA polymerase* is the enzyme that builds an RNA strand from a gene
- RNA that is transcribed from a gene is called *messenger RNA (mRNA)*

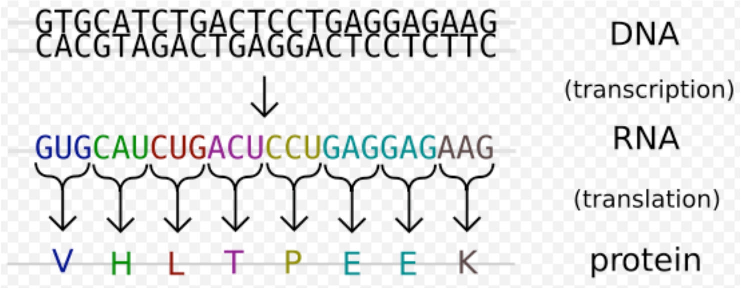
# The Genetic Code



| | 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

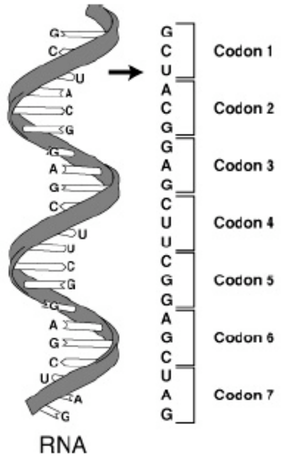
# Genetic code: DNA -> mRNA -> protein



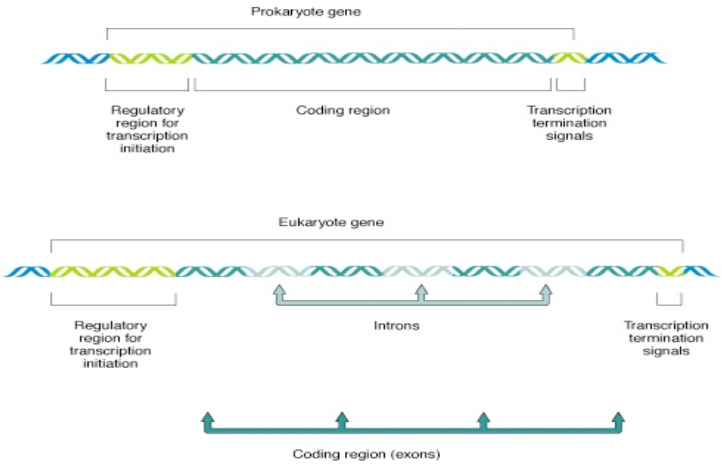
# Translation

- *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*

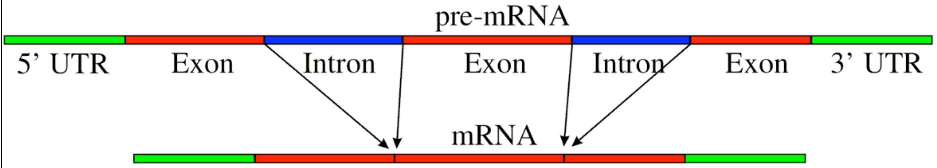
# Codons and Reading Frames



## Genes include both coding regions as well as control regions


Prokaryote gene
Regulatory region for transcription initiation
Coding region
Transcription termination signals
Eukaryote gene
Regulatory region for transcription initiation
Introns
Transcription termination signals
Coding region (exons)

## RNA Splicing: pre mRNA --> mRNA

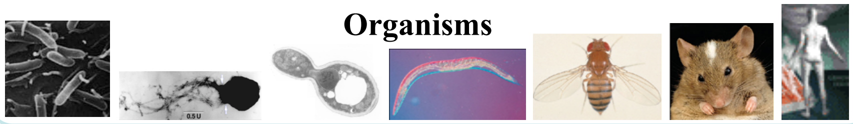


## Different Life Forms Share a Common Genetic Framework

©1998 GARLAND PUBLISHING

# Comparison of genome size

## Organisms



## Genomes

| | Haemophilus<br>influenzae | Methannococcus<br>jannaschii | Saccharomyces<br>cerevisiae<br>(baker's yeast) | Caenorhabditis<br>elegans<br>(nematode worm) | Drosophila<br>Melanogaster<br>(fruit fly) | Mus<br>musculus<br>(laboratory mouse) | Homo<br>sapiens<br>(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 |

# Sequenced Genomes

Science 1995 Jul 28;269(5223):496-512
Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Fleischmann RD et al.

Science 1996 Aug 23;273(5278):1058-73
Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Bult CJ et al.

Science 1996 Oct 25;274(5287):546, 563-7
Life with 6000 genes. Goffeau A et al.

Science 1998 Dec 11;282(5396):2012-8; errata in Science 1999 Jan; 283(5398):35 and 1999 Mar 26;283(5410):2103 and 1999 Sep 3;285(5433):1493
Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium.

Science 2000 Mar 24;287(5461):2185-95
The genome sequence of Drosophila melanogaster. Adams MD et al.

Feb, 2001 Human Genome in both *Nature* and *Science*

Science 2002 Aug 23;297: 1301-1310
Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes Aparicio S. et al.

Nature 2002 Dec 5; 420:520-62
Initial sequencing and comparative analysis of the mouse genome. Waterston et al.

Nature 2004 Apr 5; 428:493-512 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Gibbs et al.

Nature 2005 Sep 1; 437:69-87 Initial sequence of the chimpanzee genome and comparison with the human genome

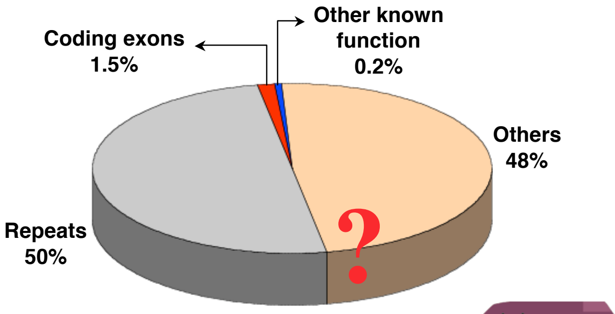
# Genes

The DNA strings include:

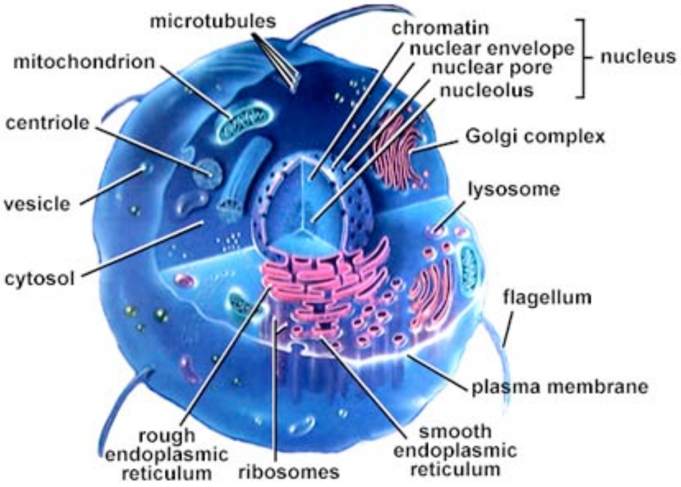
- <u>Coding regions</u> ("genes")
  - *E. coli* has ~4,000 genes
  - Yeast has ~6,000 genes
  - C. Elegans has ~18,000 genes
  - Humans have ~30,000 genes
- <u>Control regions</u>
  - These typically are adjacent to the genes
  - They determine when a gene should be "expressed"
- <u>"Junk" DNA</u> (better to be called DNA with unknown function)

# 98% of the human genome unknown

Human Genome ~3Gb

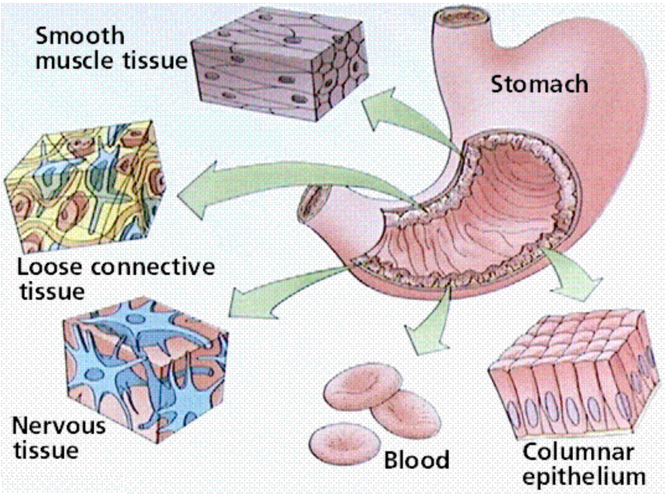


## The Cell

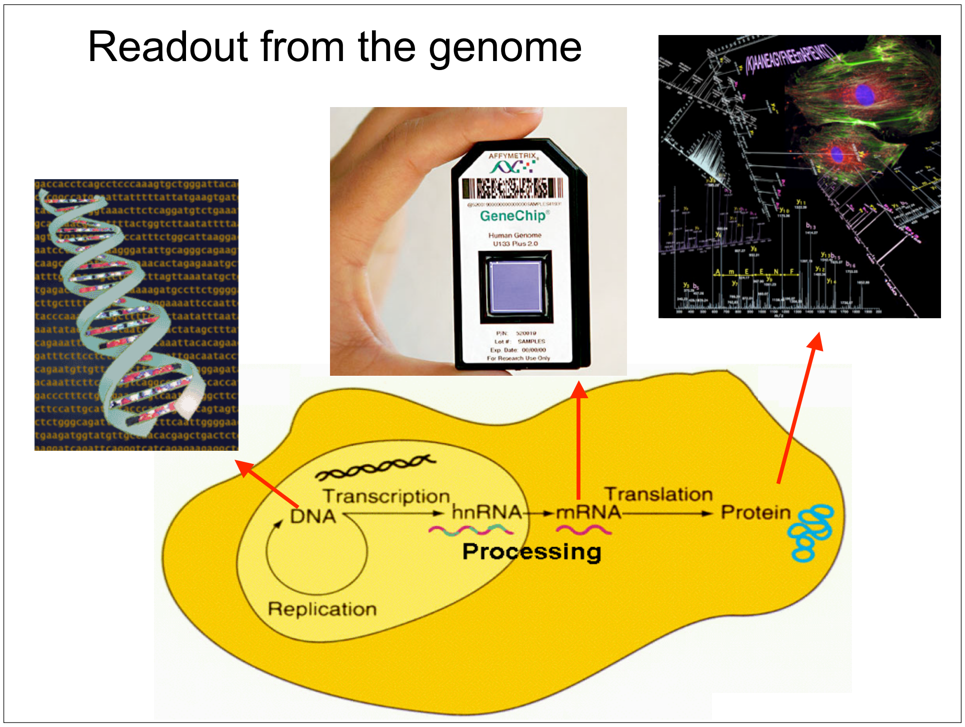


All cells of an organism contain the same DNA content (and the same genes) yet there is a variety of cell types.

## Example: Tissues in Stomach



How is this variety encoded and expressed ?



# Readout from the genome