Introduction to Biological Sequences

Background: What is DNA?

- Deoxyribonucleic acid
- Blueprint that carries genetic information from one generation to the next
- Resides in cell nucleus
- DNA contains *genes*
- Each gene is responsible for the production of a particular *protein*
- A strand of DNA is a *chromosome*
- Set of chromosomes carried by an organism is a *genome*

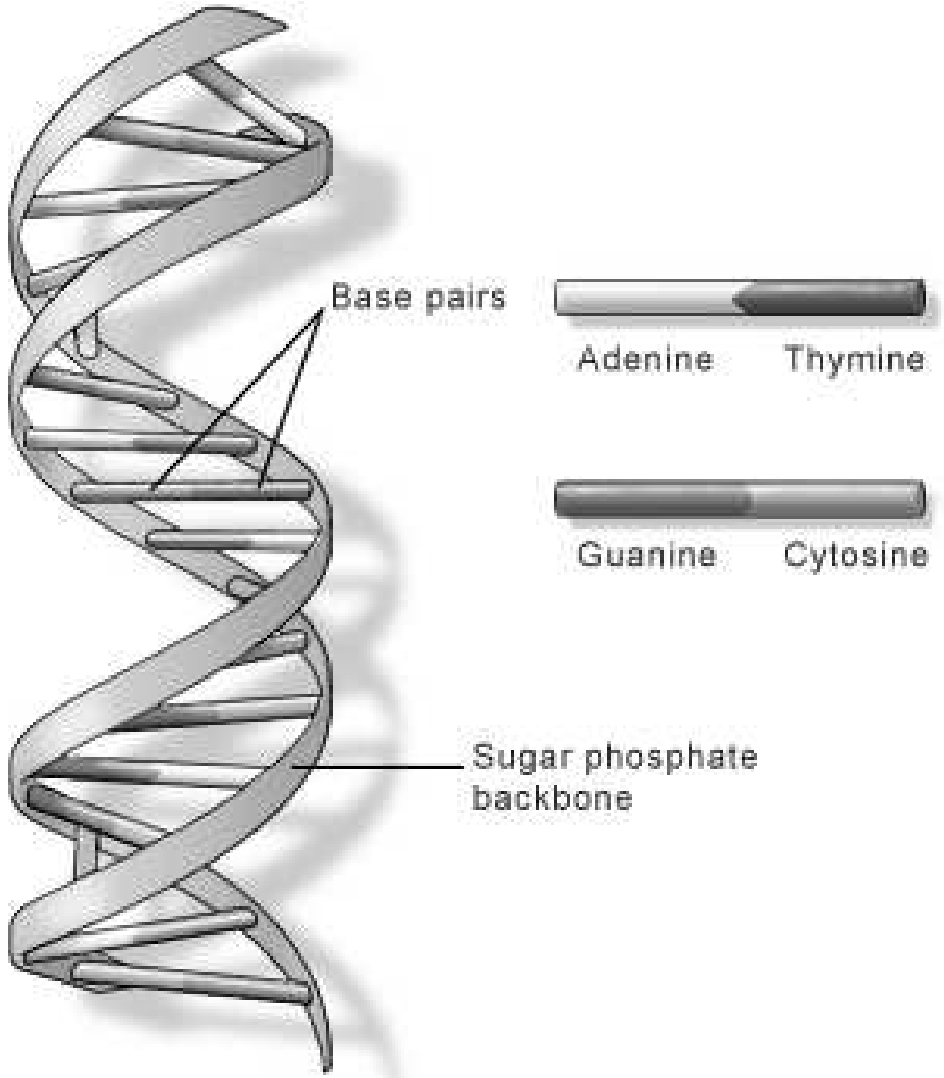
DNA Structure

- Double helix
- DNA consists of four nucleotides: adenine, cyostine, guanine, thymine (A, C, G, T)
- 5 prime (beginning), 3 prime (end)
- Base Pairs: A-T, C-G
- Ordering of these pairs is a *sequence*
- An example:

G A A T T C

C T T A A G

DNA Structure



U.S. National Library of Medicine

Example Genome Sizes

Species	Number of base pairs	Number of genes
E. coli	4,600,000	3200
Fruit fly	180,000,000	13,600
Chicken	1,000,000,000	23,000
Mouse	2,500,000,000	30,000
Corn	2,500,000,000	59,000
Human	3,000,000,000	25,000-30,000
Grasshopper	180,000,000,000	?
Amoeba	670,000,000,000	?

Why so many base pairs?

- “Junk” DNA (or non-coding DNA)
- Portions of DNA sequence for which no function identified
- 98.5% of human genome
- May serve functions that are not yet understood

Why is DNA useful?

- Every living thing on earth uses DNA to store and transmit information
- Catalogs all the different functions performed in an organism
- Identify similarities among organisms
- Identify inherited traits
- Solving crimes

DNA replication

- DNA molecule splits, each half gets copied
- Why is DNA double stranded?
  - More stable
    - Replication consists of half-new, half old
  - Allows for error-correction
    - If a base is damaged, can correct
    - Like a RAID

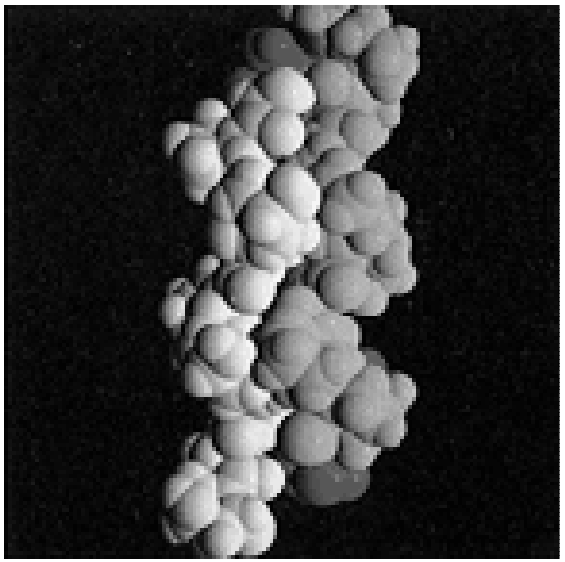
RNA

- DNA is blueprint, doesn't do much on its own
- Transcription – copy DNA into RNA
  - Only genes get transcribed
  - Promoter – DNA sequence that enables gene to be transcribed
  - Exon – protein coding sequence of gene
  - Intron – sections of DNA that are spliced out after transcription
- Messenger RNA (mRNA) moves out of cell nucleus to provide building plans for *proteins*
- RNA alphabet: Uracil (U) instead of thymin (T)

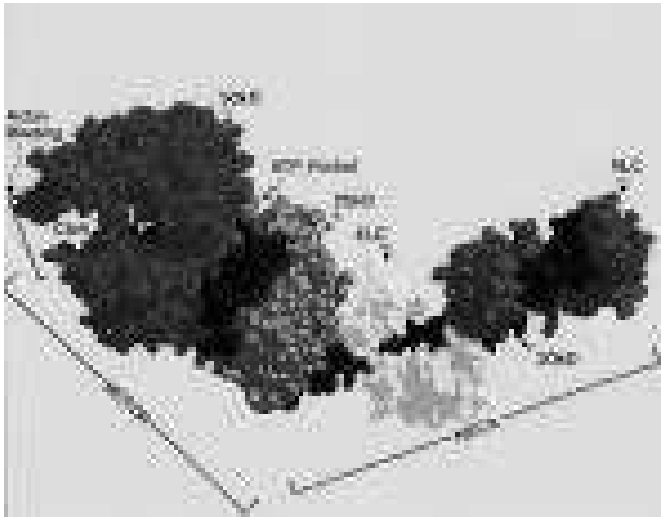
Proteins

- Buildings” and “machines” inside cell
- Composed of amino acids
- Analogy: DNA and RNA send and store information, proteins make things happen
- Can fold into shapes dependent on their amino acid sequences

Protein Examples



Collagen – shaped like a rod to be used as structural support



Myosin – has a hook that can be used as a motor
Found in muscles

Protein alphabet

- 20 symbols (20 amino acids)
- Genetic code – Ribosome reads mRNA and writes protein, 3 at a time
- Each 3 letter RNA sequence translates to one of 20 symbols
- Degenerate – 4^3 = 64 combinations – but only 20 symbols
- Redundancy

An Example

- Codon translation
- Position is important!

What can we do with sequences?

- Biological sequences show complex patterns of similarity to each other
- Organisms also show similarities
- Sequences can change over time due to different forces
  - mutation
  - natural selection
  - genetic drift
- Similarity searches among sequences can help identify these occurrences

Mutation

- Change in a DNA sequence
- Every time cell divides, DNA is duplicated
- Replication process isn't perfect
  - 1 error in every 300 million letters
  - ~10 mutations per genome replication
- Change in DNA leads to change in RNA, which *may* change protein
- Mutations may also insert or remove nucleotides

Natural Selection

- Assumptions
  - There must be variation within a population
  - The variation must be heritable
  - Must be differential reproduction based on variation
- Variations at DNA level explain differences within population
- Mutation is constantly happening

Genetic Drift

- Not just natural selection
- Frequencies can change by purely random processes
- Suppose 10 individuals, 5 C and 5 T
- Next generation, sampling error may cause shift C = 0.6, T = 0.4
- Over time, C may increase and T may decrease

Molecular Clocks

- In sequences of related organisms, some positions change at different rates
- If you know rate of change, can determine how long ago two sequences diverged
- Simple example:
  - Given a protein sequence from cats and dogs, 10 differences between them
  - Cats and dogs had common ancestor 50 million years ago
  - Cats and humans have 12 differences in sequence
  - Conclusion: cats and humans shared ancestor ~60 million years ago

Homology

- We know which organisms exist today
- But don’t know what existed 100 million years ago
- What was last common ancestor of humans, chimpanzees, and gorillas?
- Phylogenetics – study of relationships between organisms
- Two sequences are *homologous* if they share a common ancestor

Tree of Life

- Recall the taxonomy hierarchy from relational lecture
- Five taxonomic kingdoms: animals, plants, fungi, monera, protista
  - Based on what you see
- Using DNA sequences, new classification:
  - Prokaryotes (bacteria and archaea)
  - Eukaryotes – have a nucleus (e.g., birds, trees)

Why is similarity useful?

- Identify which species are most closely related
- Before sequences- morphological data
  - Hair, teeth, limbs, fins, hearts, livers, eyes, etc.
- Sometimes morphological data misleading!

An exercise

- How do we measure similarity?
- Exercise: measure the similarity of a sequence in three different animals
  - Bone Morphogenetic Protein 7 gene (BMP7)
  - Represent signals that induce bone growth

Results

- Animals are (in no particular order):
  - Rabbit
  - Pig
  - Sheep
- Any idea which is which?