Lecture7_MolecularBio

September 3, 2018

1 Lecture 7: Molecular Biology Primer

CBIO (CSCI) 4835/6835: Introduction to Computational Biology

1.1 Overview and Objectives

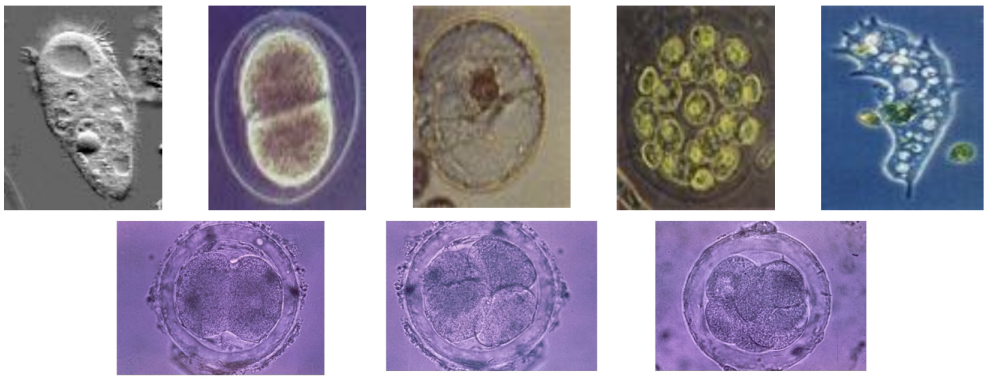
A respite from the "Python Firehose," we'll be starting our journey into applying the tools and techniques we've covered over the past couple weeks to answer biological questions. Here, we'll cover a basic introduction to molecular biology and the types of problems that crop up. By the end of this lecture, you should be able to:

- Recall the basic components for life and their types
- Discuss the three critical molecules and their roles in the "Central Dogma"
- Define genes, where they are found, and the process through which they are activated
- Explain the process through which DNA is sequenced
- Understand the role of bioinformatics and some common algorithms for analyzing sequence data

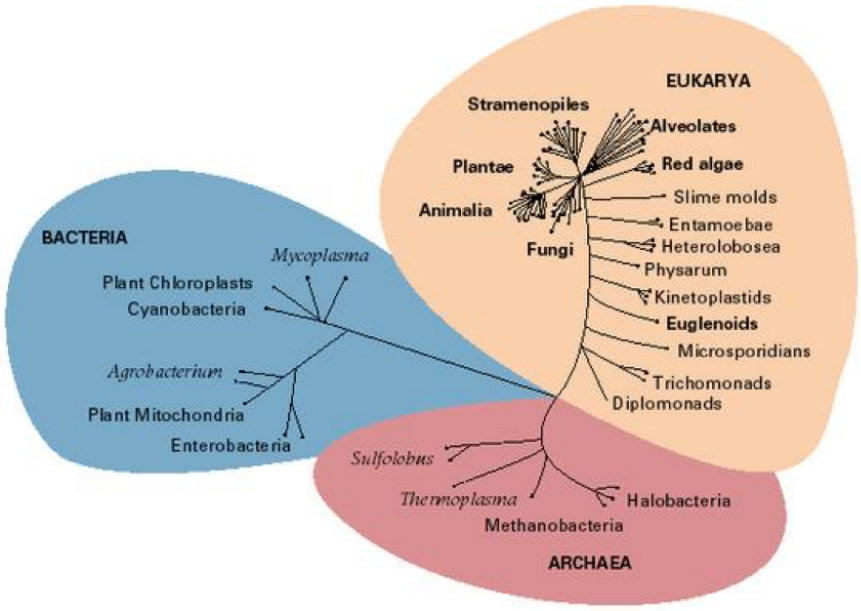
1.2 Part 1: Building Blocks of Life

1.2.1 Cells

- Fundamental unit of every living organism



cells



tree

- All cells evolved from a common ancestor roughly 3.5 billion years ago
- Cells fall into one of two main categories: **prokaryotes** and **eukaryotes**

Cells are the smallest functional unit of life; consequently, most cells share common structures…

…and functions.

1.2.2 Prokaryotic versus Eukaryotic

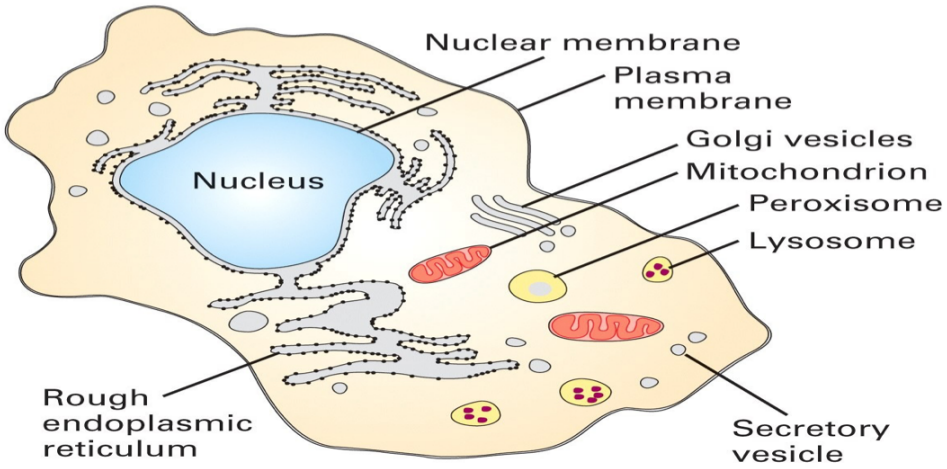
What are some of the main differences between prokaryotic and eukaryotic cells?

Prokaryotes

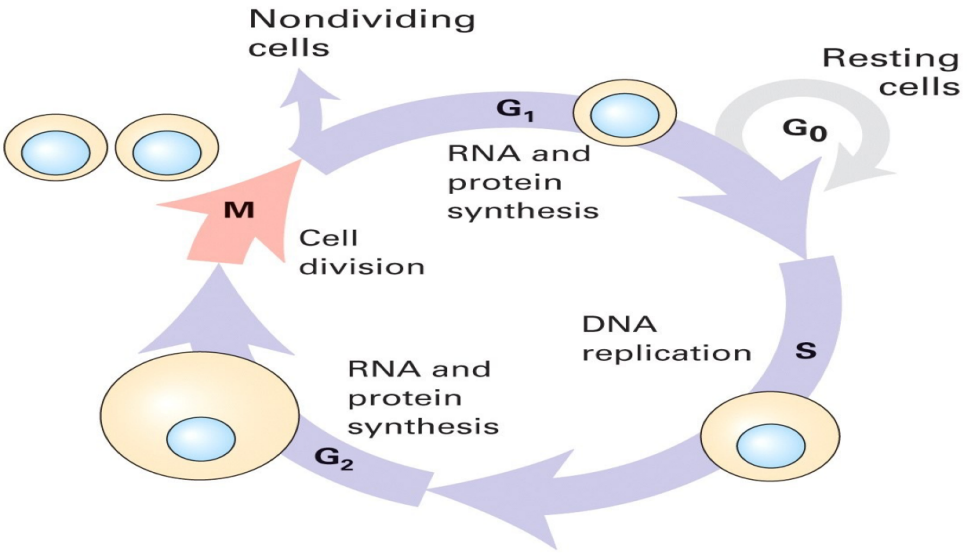
- Single-celled
- No nucleus or other organelles (lacks a membrane-bound nucleus)
- One circular piece of DNA
- No mRNA post-transcriptional modifications
- Only one kind of membrane: the plasma membrane
- Smallest known cells are bacteria (1000-2000 polypeptide species, 3×10^6 protein molecules
- *E. coli* genome has about 4×10^6 base pairs, 90% of which encode proteins

Eukaryotes

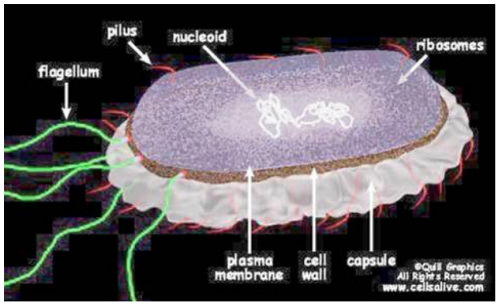
- Multi-celled (plants, animals, protists, fungi)
- Complex system of internal structures (organelles and vacules)
- Volume of cell is hundreds of times larger, e.g. HeLa cells (5,000-10,0000 polypeptide species, 5×10^9 protein molecules)
- Genome of yeast contains about 1.35×10^7 base pairs, only a small fraction of which encode proteins
- DNA contained in chromosomes, which are contained in membrane-bound nucleus

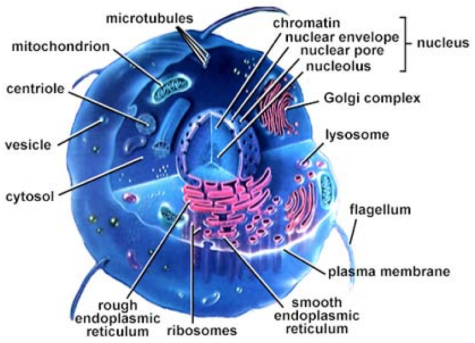


cell



cellcycle





proeu

1.2.3 Cell Signaling

Signaling pathways allow cells to coordinate complex behaviors and make decisions between them that would otherwise be impossible. Some of these decisions may include

- Synthesizing new materials
- Breaking down materials for "spare parts"
- Determining when to eat, replicate, or die

1.2.4 Three Critical Components

DNA

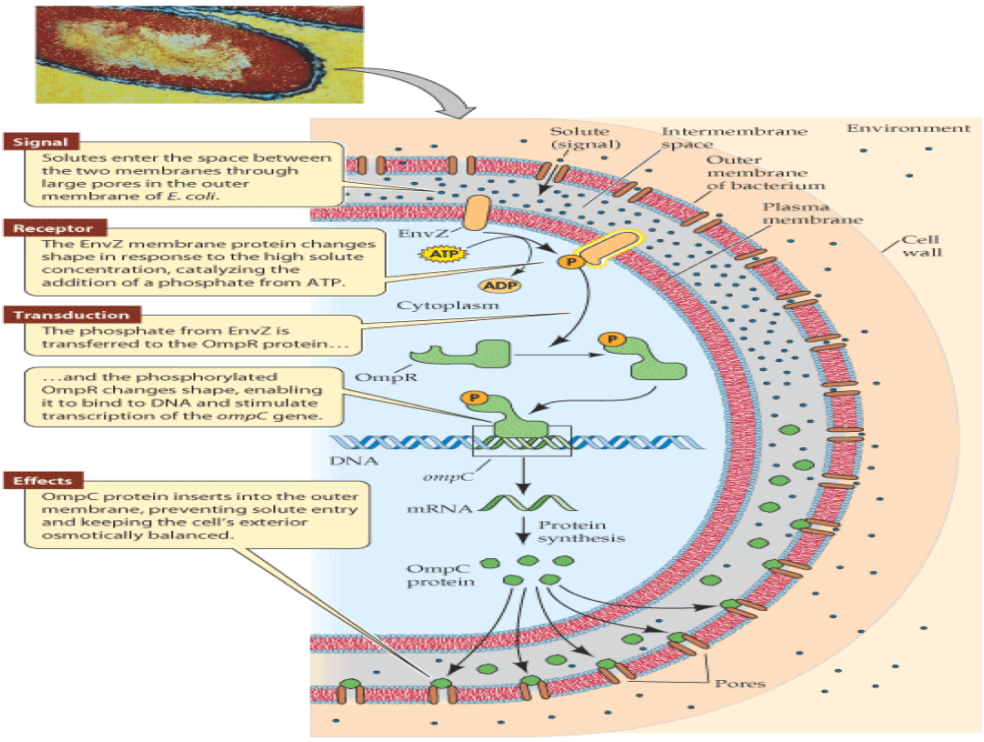
- Holds all the information for how the cell works. If the nucleus is a library, the DNA comprises the books.

RNA

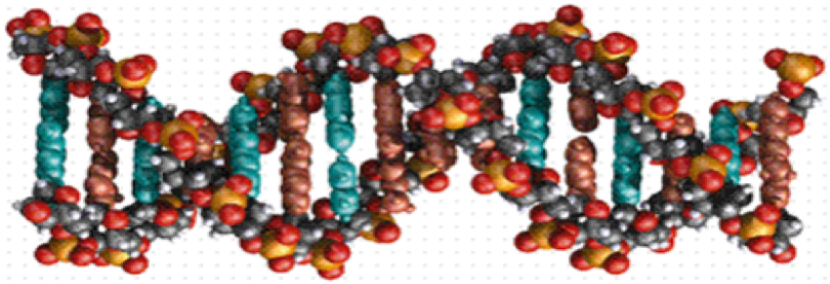
- Act to transfer short pieces of information to other parts of the cell.
- Provide templates for protein synthesis.
- Numerous other roles.

Proteins

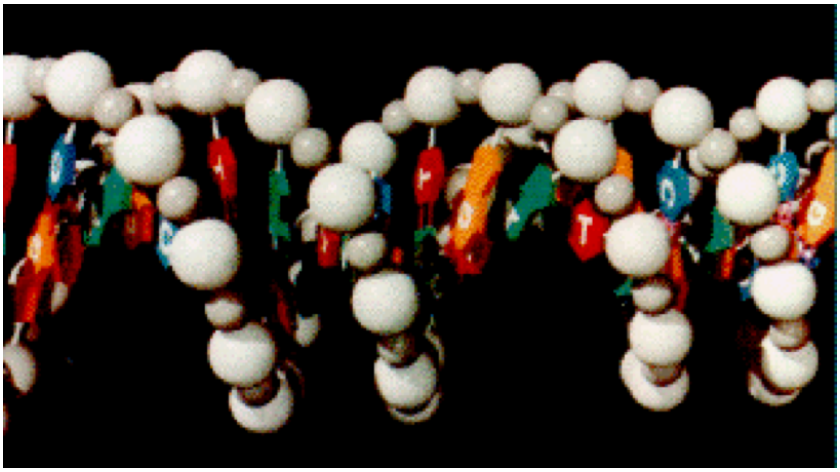
- The "workhorse", forming the body's major components (hair, skin, bones, muscles, etc).
- Form enzymes that send signals across cells and regulate gene activity.



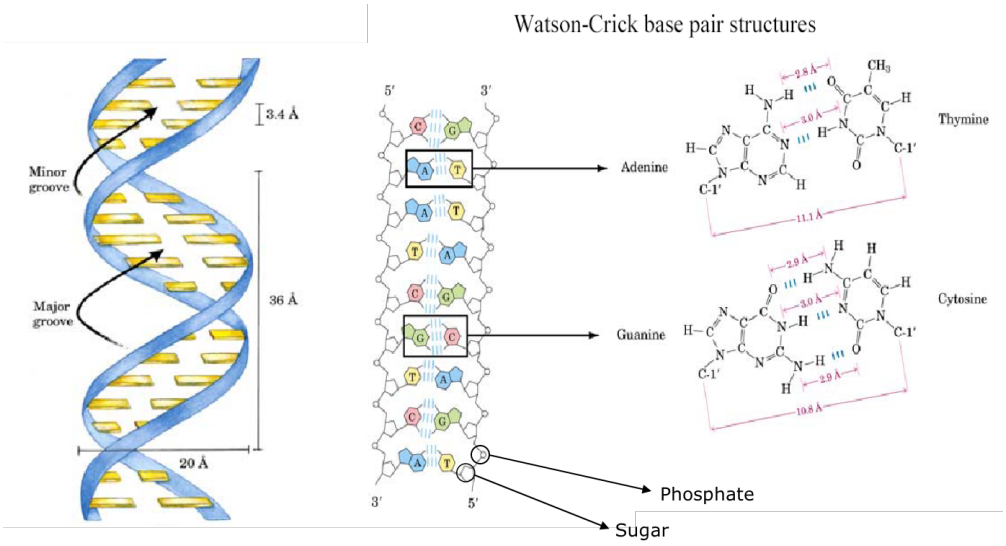
signaling



genes



dna



dnastruc

1.3 Part 2: Genetic Material

1.3.1 What is DNA?

Deoxy-ribonucleic acid

- DNA consists of a **nitrogenous base**, a **sugar**, and a **phosphate**
- Double helix struture–forward and backward direction, with 5′ and 3′ ends (why?)
- Always read (and transcribe) 5′ to 3′

1.3.2 Structure

- Main structure formed through linkage of base-pairs: Adenine (A), Thymine (T), Cytidine (C), and Guanine (G).
- A pairs with T, C pairs with G (which is stronger? why?).
- Backbone consists of the deoxyribose and phosphate groups

1.3.3 Replication

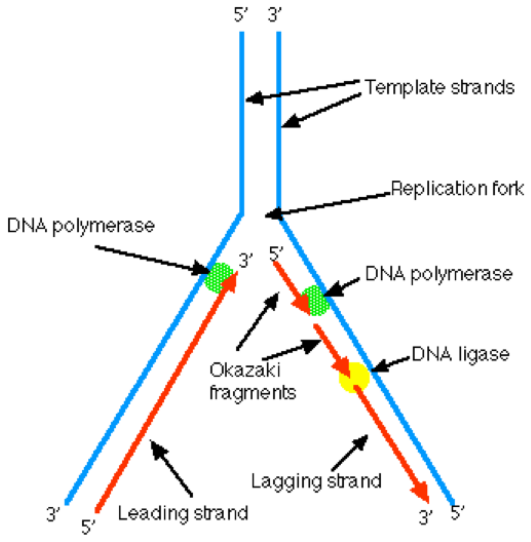
- Process through which DNA replicates, or clones, itself
- Splits the double strands and rebuilds each one
- Each daughter strand is an exact replica of the original (why?)

1.3.4 Coding for Traits

The *effects* of DNA were discovered long before the genetic structure itself.

Gregor Mendel discovered this effect in the 1860s while experimenting with pea plants. He was trying to answer the question: **Do traits come from a blend of both parents' traits, or from only one?**

Mendel discovered that genes were passed on to offspring by both parents in the form of either *dominant* or *recessive* traits.



replication

The dominant gene determined the *phenotype* of the offspring; unless it wasn't present, in which case the recessive gene would determine the phenotype. The gene itself is the *genotype*.

DNA and its building blocks were later discovered in the 1950s.

1.3.5 Mutations in Nucleotides

DNA is a sequence of nucleotides (A, G, T, and C). When one of these bases in the sequence change, this is known as a **mutation**.

Mutations can have one of three possible effects:

1. **Good**: The mutation results in a trait that enhances the organism's function (mutation in the sickle gene provides resistance to malaria)
2. **Bad**: The mutation causes a trait that is harmful, or even fatal, to the organism (Huntington's disease is a symptom of gene mutation that results in degeneration of the nervous system)
3. **Neutral**: The mutation can cause no discernible changes in the organism

1.4 Part 3: DNA to Proteins

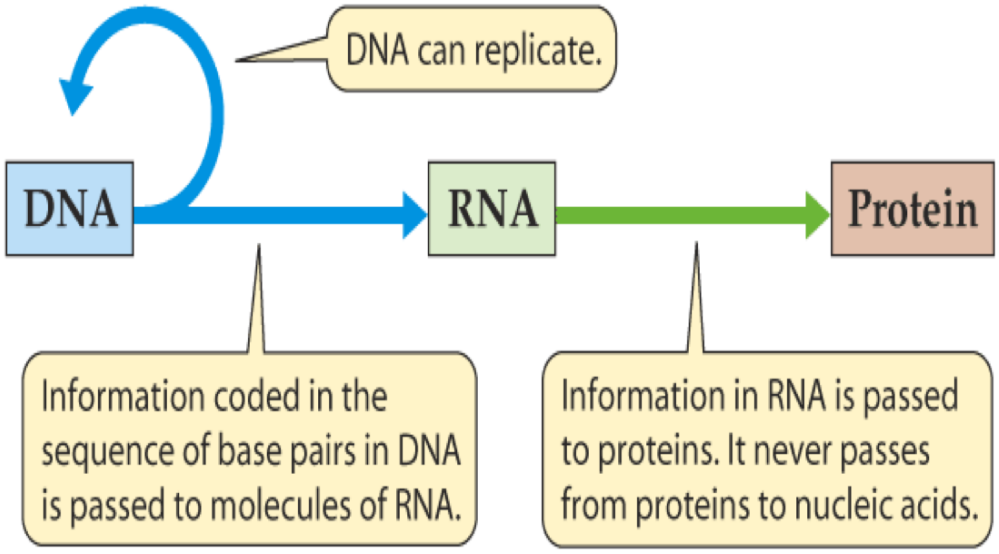
1.4.1 The Central Dogma

DNA -> RNA -> Protein

- Information for making proteins is stored in DNA
- Processes (transcription and translation) convert the coding regions of DNA to proteins
- By understanding this process and how it is regulated, we can make models and predictions of cells

1.4.2 RNA

Ribonucleic acid



centraldogma

- Similar, chemically speaking, to DNA (Uracil in lieu of Thyamine)
- Usually single-stranded
- Some forms of RNA form secondary structures by "pairing" with itself

Several "types" of RNA exist that are understood to perform specific functions.

- **mRNA**: the messenger RNA that encodes a gene from the DNA
- **tRNA**: delivers specific amino acids to ribosomes according to the sequences of the mRNA
- **rRNA**: ribosomal RNA that catalyzes the formation of peptide bonds between amino acids

1.4.3 Transcription

This is the process through which genes in the DNA are "converted" to RNA (specifically, mRNA).

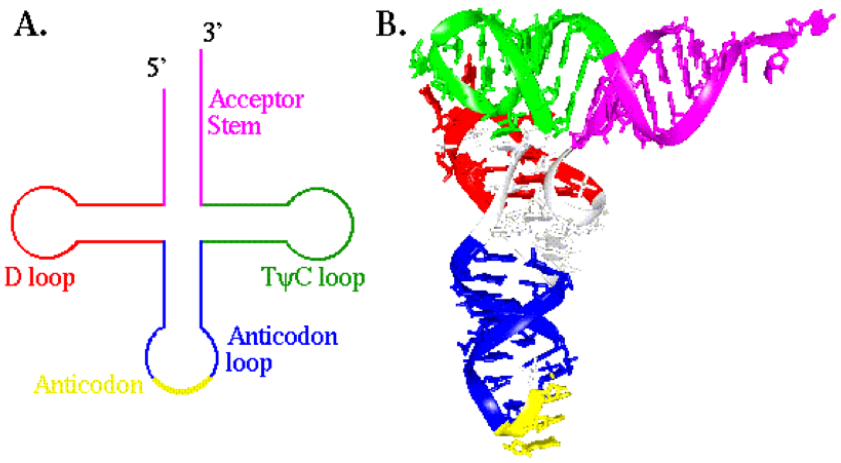
- Catalyzed by a transcriptase enzyme
- Transcribed by RNA polymerase
- Highly regulated process (promoter regions and motif finding, repressors and inhibitors)

1.4.4 Genes

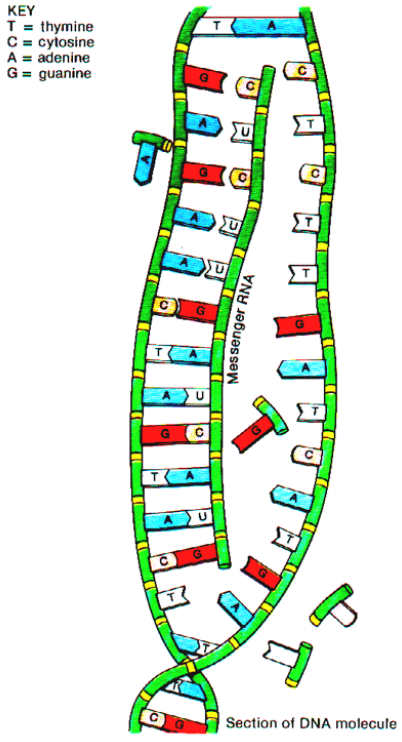
- Regulatory regions are often 50kb upstream of coding site
- **exons** are the protein-coding (or untranslated: UTR) regions [1 to 178 exons per gene, 8bp to 17kbp per exon]
- **introns** are splice acceptor (or junk coding) regions [1kb - 50kb per intron]

1.4.5 Splicing

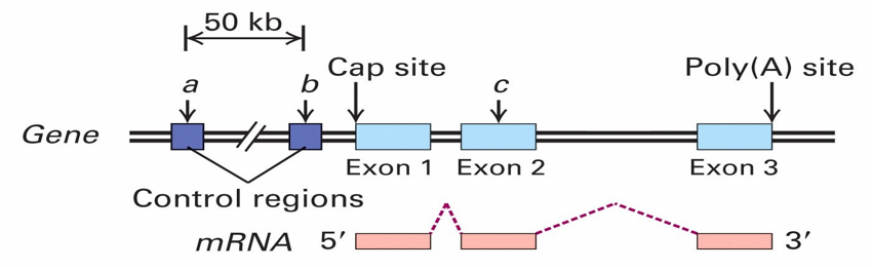
- Introns are removed from mRNA before the sequence is translated to proteins
- Alternate splicings can create different valid proteins from the same gene
- Post-transcriptional processing includes 5′ cap and poly-A tail on mRNA



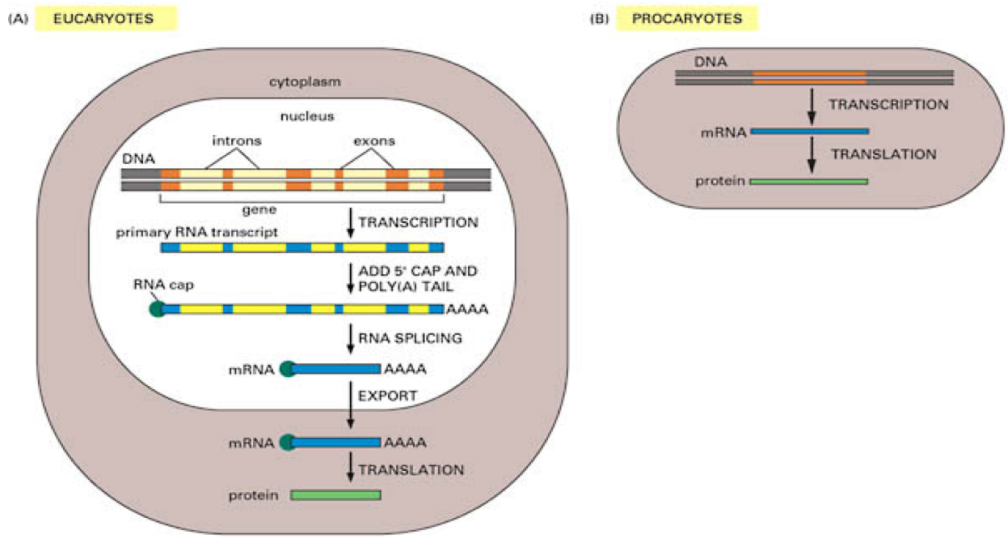
trna



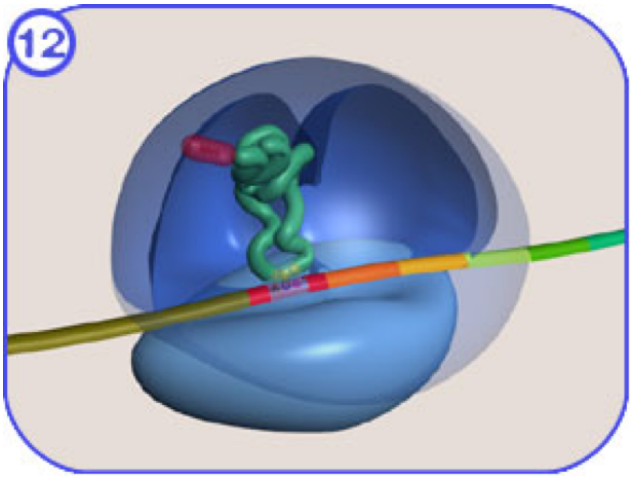
transcription



gene



splicing



translation

1.4.6 Translation

- Process through which mature mRNAs are translated into proteins
- Combination of mRNA, tRNA, and rRNA
- **How does the machinery know what amino acids to assemble?**

1.4.7 Codons

- Sequence of three consecutive bases
- Forms a language that specifies the amino acids
- 20* possible amino acids
- Always starts with a methionine and ends with a stop codon

1.4.8 Folding

The final step in protein synthesis is the folding of the assembled polypeptide into its primary configuration.

- Folding begins while the peptide is still being translated
- Occurs largely in endoplasmic reticulum and mitochondria
- α-helices or β-sheets
- "Molecular chaperones" can assist in protein folding
- Folding is understood to be a function of free energy
- Improper folding
- **Protein folding is an extremely hard problem** (we'll explore it later this semester)

1.5 Part 4: Analyzing DNA

Analyze a genome in four easy steps!

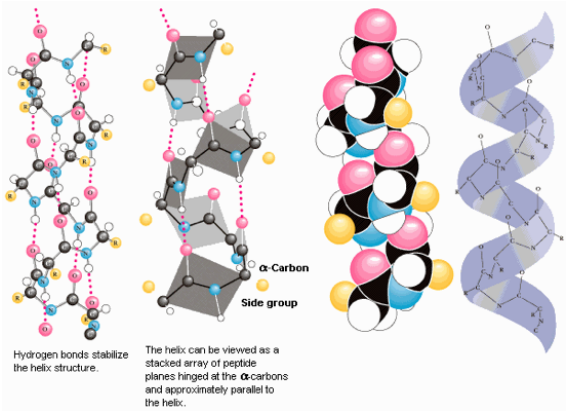
1. Copy the DNA many times.
2. Cut the DNA into small fragments.
3. Use specialized instruments to read the small fragments.

SECOND POSITION

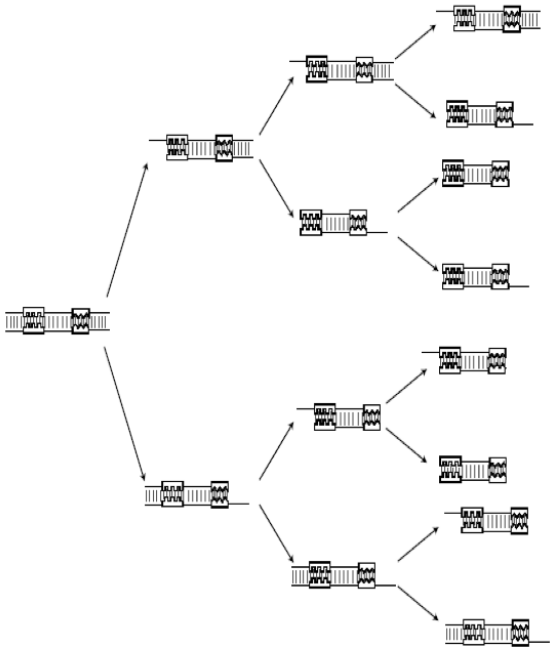
FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	* methionine		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

codons



folding



pcr

4. Assemble all the reads into a single coherent genome sequence.
5. Profit!

Steps 4 and 5 are where bioinformatics and computational biology play a role.

1.5.1 Copy the DNA

Why?

- DNA is very small; can't really "look" at it directly
- Need chemical techniques to detect the sequence
- These chemical techniques aren't terribly sensitive, so we need to make our DNA sample "bigger" by cloning it

Use **polymerase chain reaction (PCR)** to massively replicate the DNA.
PCR doubles the amount of DNA at every iteration (amount of DNA grows exponentially)

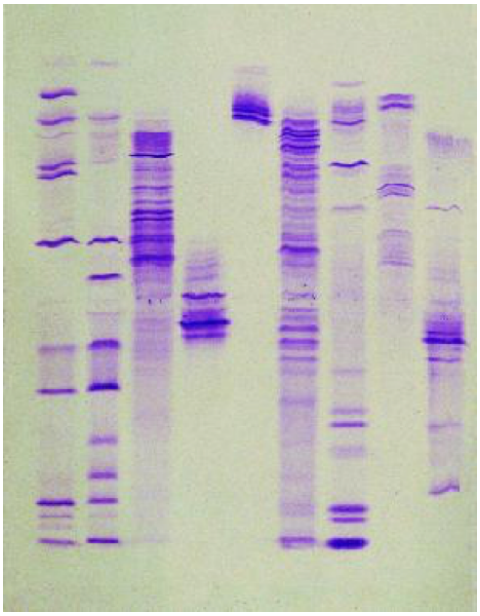
1.5.2 Cut up the DNA

Why?

- Sequencing machines can't read very many base pairs before becoming unreliable
- Shorter fragments = more reliable reads

Use various **restriction enzymes** to chop up the DNA at specific points.

- Discovered in the 1970s



electrophoresis

- Used by bacteria as a defense mechanism to break down the DNA of invading viruses; cut DNA into fragments
- Also used in sequencing: breaks down DNA into more manageable pieces
- Can then use standard purification techniques to identify single fragments and duplicate them

1.5.3 Sequence the fragments

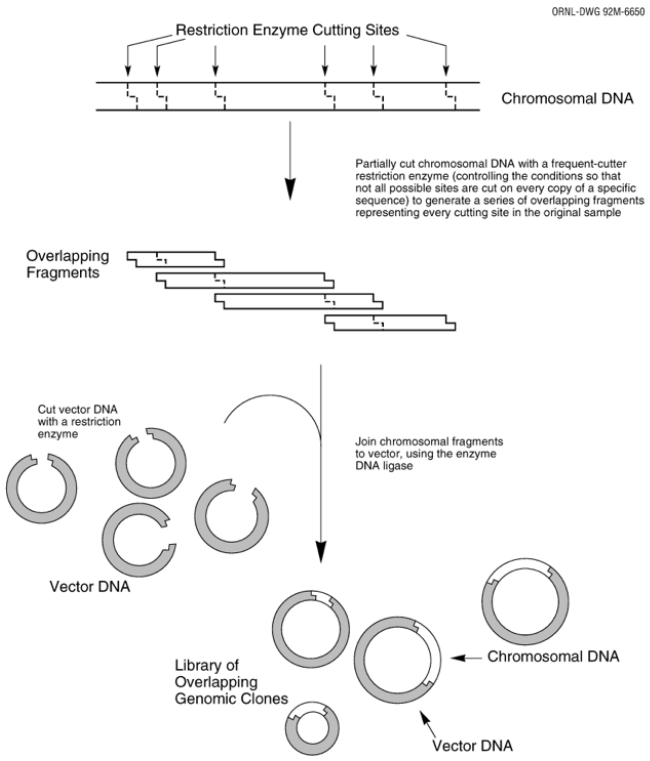
- This can be done through **gel electrophoresis**
- The phosphate backbone of DNA is highly negatively charged, therefore the DNA fragments will migrate in an electric field
- Size of DNA fragments (and, therefore, their identities) can be determined by how far the fragments migrate in the gel
- Direct sequencing can also be done using gene chips or next-gen sequencers

1.5.4 Assembling the Genome

- Solve the problem called the **shortest common superstring (SCS)**: fragments will have some overlap, so they need to be assembled in such a way that all fragments are used to create the shortest sequence possible
- **This is a very hard problem**

1.5.5 Genome Assembly Challenges

- DNA fragments may contain sequencing errors! (probability of replication error $\times$ huge number of replications)
- Two complements of DNA: need to take complementary strand into account when assembling reads



assembling

- Problem of repeats (50% of human DNA is just repeats)

1.6 Part 5: Role of Bioinformatics

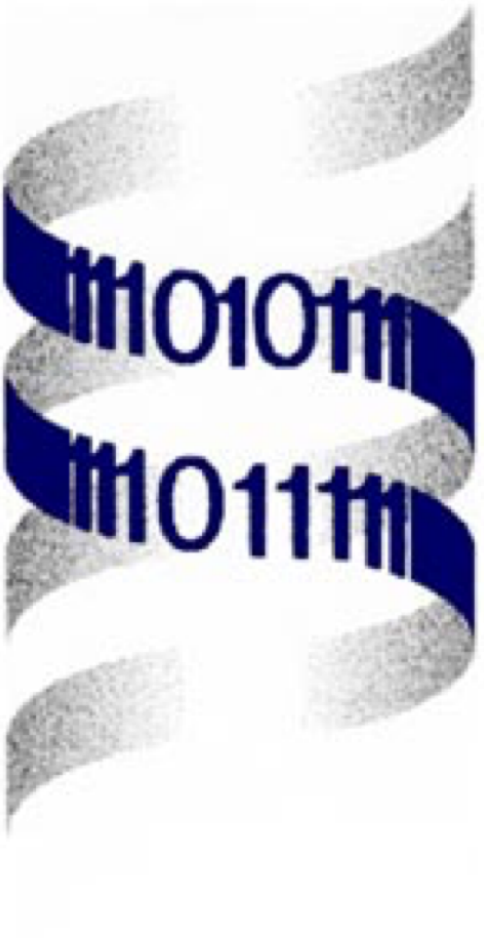
Bioinformatics or **Computational Biology** are generally defined as the analysis, prediction, and modeling of biological data with the help of computers.

in silico

- **Genomics**:
  - Fragment assembly of the DNA sequence
  - N-P complete problem (this means it's as hard a problem as they come)
  - Finding genes–identify open reading frames
- **Proteomics**:
  - Predict how proteins will fold, given their primary structure
    - Identify functional domains in protein sequence, and what the function is

1.6.1 Current Examples

- Sequence comparison
- Searching for similar sequences
- Queries to sequence databases



bioinformatics

1.6.2 BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

Allows researchers to compare query sequences (your sequence of interest) with entries in current biological databases

Great for predicting the function of unknown sequences using alignment to similar, known sequences

Emphasis is on **speed**: as a result, it does not search for an *optimal* result, but uses a heuristic to identify matches or close matches with high probability

1.6.3 PROSITE

Database of active sites in proteins

Similar to BLAST in that it takes a query sequence of an unknown protein and returns known active sites in proteins with similar amino acid sequences

1.6.4 Sequence Analysis

Critical component of computational genomics and proteomics.

- Finding open reading frames, RNA splice sites, conserved regions
- Amino acid propensities in proteins
- Protein secondary structure folding and active sites

Even further downstream, analyses of the primary structure can yield

- Phylogeny: finding relationships between species by tracking similarities between them
- Gene annotation (cooperative genomics) by comparing similar species
- Determination of regulatory networks
- Computational drug discovery

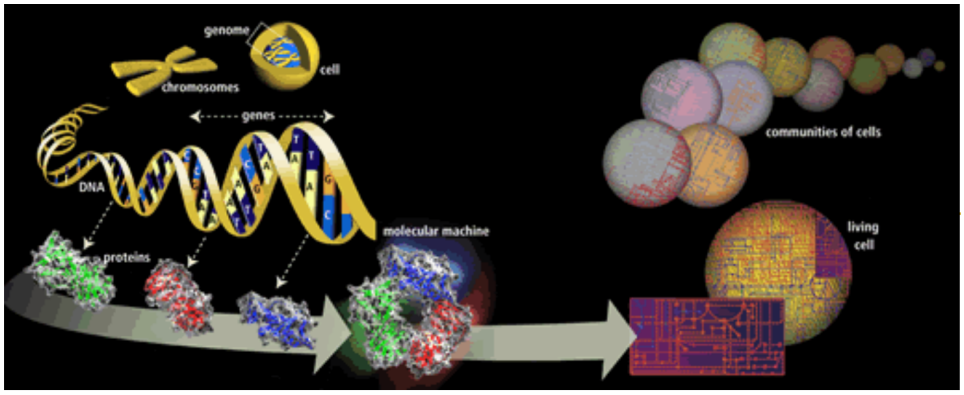
1.7 Administrivia

- **How is Assignment 2 going?** Due in a week and a half.
- **Will hopefully have Assignment 1 graded very soon.**

1.8 Additional Resources

1. Jones, Neil C. and Pevzner, Pavel A. *An Introduction to Bioinformatics Algorithms*, Chapter 3. 2004. ISBN-13: 978-0262101066
2. Based heavily on the modified slides of Dr. Phillip Compeau.

Molecular Biology Primer



Angela Brooks, Raymond Brown, Calvin Chen, Mike Daly, Hoa Dinh, Erinn Hama, Robert Hinman, Julio Ng, Michael Sneddon, Hoa Troung, Jerry Wang, Che Fung Yung

credits