CS612 - Algorithms in Bioinformatics

Introduction

February 3, 2025

Contact Information

- Instructor: Nurit Haspel
- `http://www.cs.umb.edu/~nurith`
- nurith@cs.umb.edu or nurit.haspel@umb.edu
- Phone – 617-287-6414.
- Office – M03-201-04
- Office hours – Mo We 2:30–3:50 or by appointment.
- Course schedule: Mo We 4:00–5:15 at Y-04-4140.

Course Description

- Introduction to Molecular Biology.
- Protein structure – basic concepts.
- Structural representation and storage of protein molecules.
- Protein folding and docking.
- Geometric conformational search algorithms.
- Introduction to classification and machine learning methods.
- Bio-molecular simulations and Molecular dynamics.
- Introduction to systems biology, networks if time permits.
- Course website: `http://www.cs.umb.edu/~nurith/cs612`.
- Syllabus.

Course Requirements

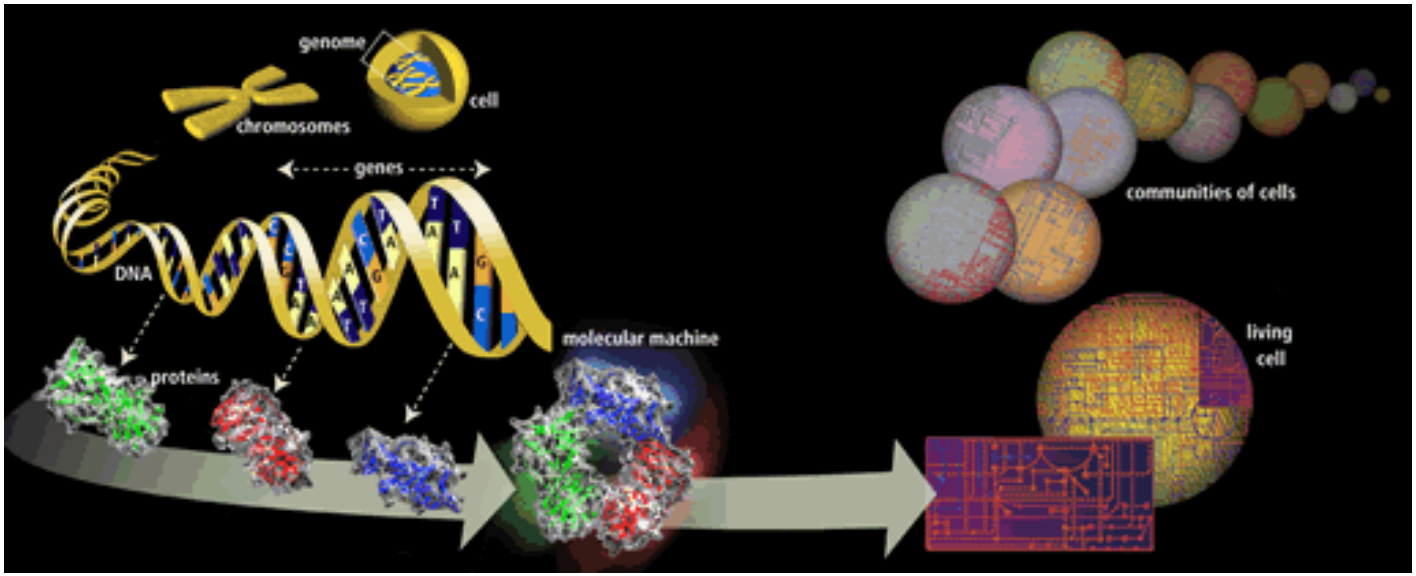
- Prerequisite: CS210, MATH260 or equivalent.
- Knowledge in biology is not required – all the concepts will be taught. It is an advantage, though!
- Homework/programming assignments – 4-5 during the course (70% total).
- You may consult with your friends, but the final work should be individual.
- I strongly prefer typed homework. If handwritten – make it CLEAR.
- Term projects + presentation – 30%.

Course Requirements (Cont.)

- The homework due date is strict. No late assignments will be accepted without a good and documented reason.
- Final project will include individual programming/research project and class presentation.
- Choice should be made by the middle of February and coordinated with me.
- Your final grade should be at least C (60%) to pass.
- Recommended text books: See syllabus.

General Stuff

- The course material will be available online and updated regularly with class notes and assignments.
- Attendance is required (barring valid medical or other emergency situations). You are responsible for updating yourselves if you miss a class.
- Don't be afraid to ask questions in or out of class. I won't think you are stupid and it won't lower your grade.
- Don't hesitate to send me e-mails. I expect e-mails. It won't lower your grade.
- See syllabus for AI policy.



`bioalgorithms.info`, `cnx.org`, and instructors across the country. Special thanks to Dr. Lydia Kavraki, Rice and Dr. Amarda Shehu, GMU

Critical Events in Molecular Biology

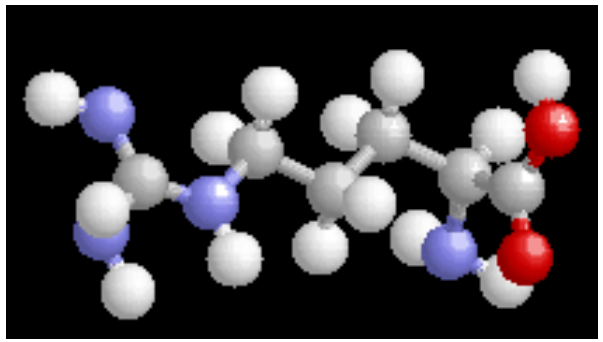
1665 – Rise of microscope: Robert Hooke discovered organisms are made of cells.

1869 – Discovery of DNA: Johan Friedrich Miescher discovered DNA and named it nuclein.

1865 – Rise of Genetics: Gregor Mendel discovered that an organism has two alternative hereditary units for a given trait: dominant vs. recessive.

1881 – Edward Zacharias showed that chromosomes are composed of nucleins.

1899 – Richard Altmann renamed nuclein to nucleic acid.

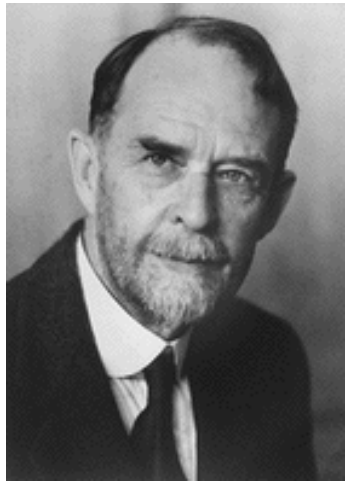


1900 – Chemical structures of all twenty classic amino acids had been identified.

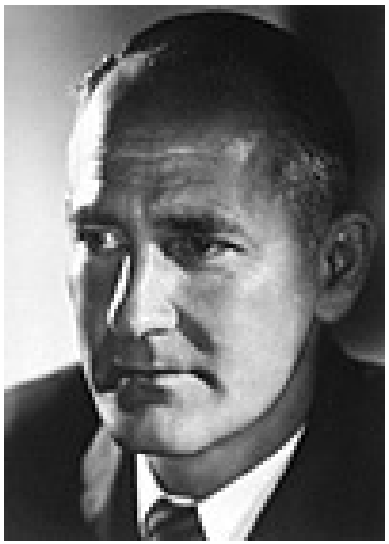
Critical Events in Molecular Biology

1902 – Emil Fischer won the Nobel Prize for showing that amino acids link to form proteins.
Postulated: protein properties are defined by amino-acid composition and arrangement
This postulate is now fact

1911 – Thomas Morgan discovered genes on chromosomes are discrete units of heredity
Pheobus Lerene discovered RNA



1941 – George Beadle and Edward Tatum identify that genes make proteins



1952 – Alfred Hershey and Martha Chase make genes from DNA
1952–1953 – James Watson and Francis Crick deduce the double helical structure of DNA
Watson and Crick "used" the X-ray diffraction data produced by Rosalind Franklin to deduce the structure of DNA

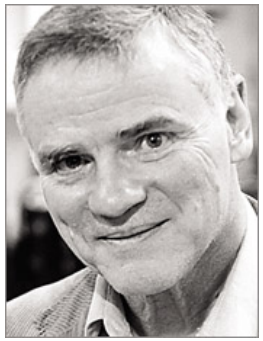
Critical Events in Molecular Biology

1977 – Philip Sharp and Richard Roberts show pre-mRNA is processed by excision of introns and splicing together of exons

1995 – John Craig Venter sequences first bacterial genomes – automated sequencing

1996 – First eukaryotic genome, yeast, sequenced

1986 – Leroy Hood developed automated sequencing mechanism – Human Genome Initiative announced



1990 – Congress launches Human Genome Project

2000 – Complete sequence of euchromatic portion of Drosophila Melanogaster genome

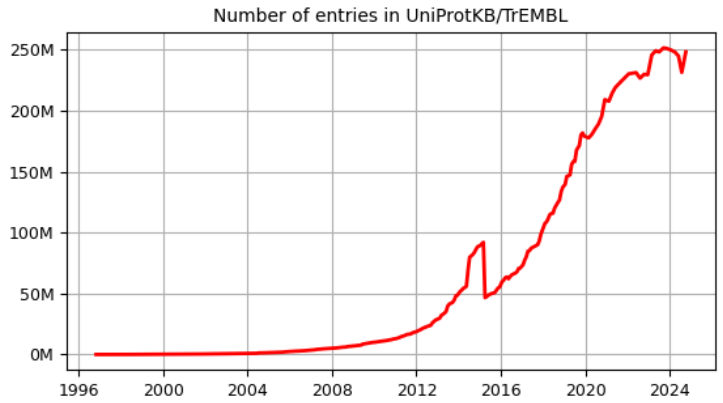
2001 – First draft of human genome published

2003 – Mouse genome sequenced

2007 – James Watson's genome sequenced

Molecular Biology as Information Science

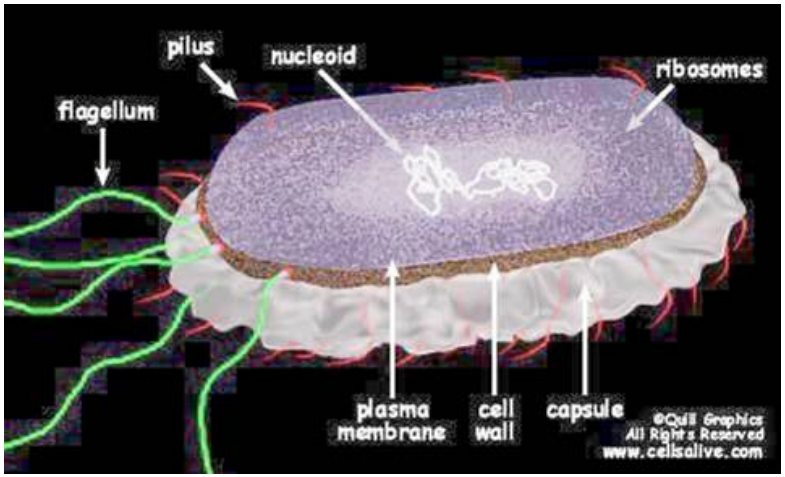
- $> 38,000$ genomes fully sequenced, $> 484,000$ permanent draft, mostly bacterial (2025)
- $254,254,987$ sequences (Nov. 2024), $572,619$ reviewed.
- What do we do with them?
  - Compare them to find what is common and different among organisms (Comparative Genomics)
  - Find out how and which genes encode for which proteins
  - Identify changes that lead to disease
  - Associate structural and functional information with new gene sequences

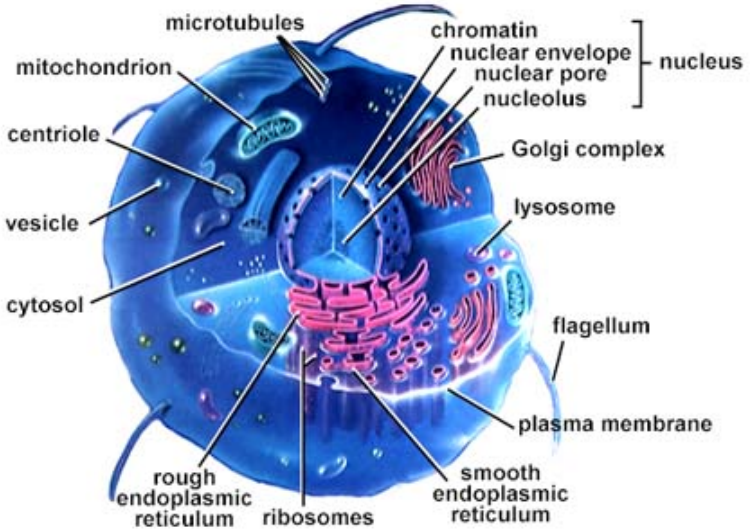


source: `http://www.uniprot.org`

- JGI (Genomes Online Database) `https://gold.jgi.doe.gov/`
- Most of the sequences do not have a solved structure
- Experiments lagging behind
- Way too much data for computer scientists to sit around doing nothing
- Recently – AlphaFold and Large Language Models filling the gap

Life as we Know it: The Cell



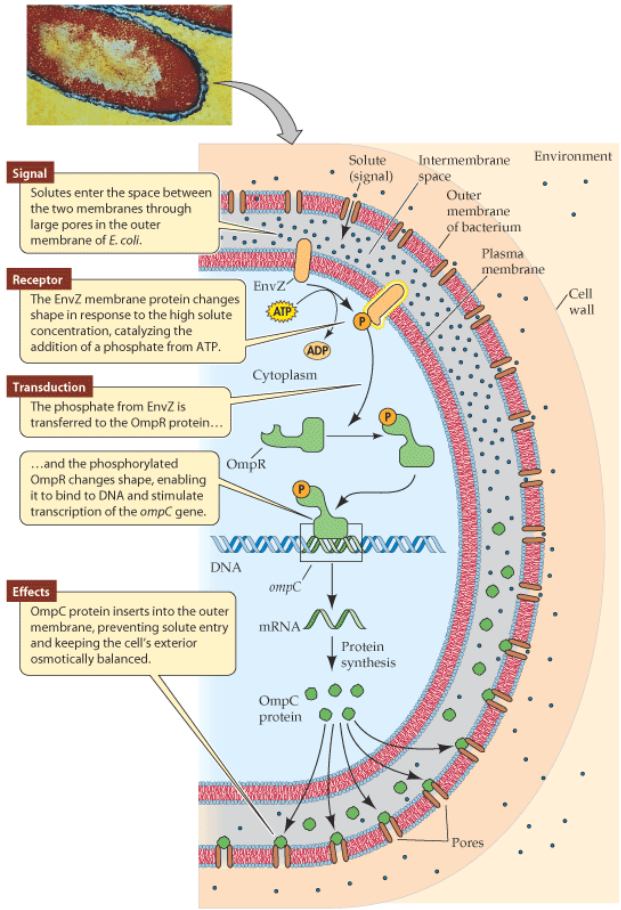


Prokaryote cell	**Eukaryote cell**
Single Cell	Single or multi cell
No nucleus	Nucleus
No organelles	Organelles
One piece of circular DNA	Chromosomes
No mRNA post transcriptional modification	Exon–intron splicing

Signaling, Control, and Gene Activity

Cells make decisions through complex networks of chemical reactions, called pathways

- Synthesize new materials
- Break material down for spare parts
- Signal to eat, die, or divide
- Signal to one another to communicate

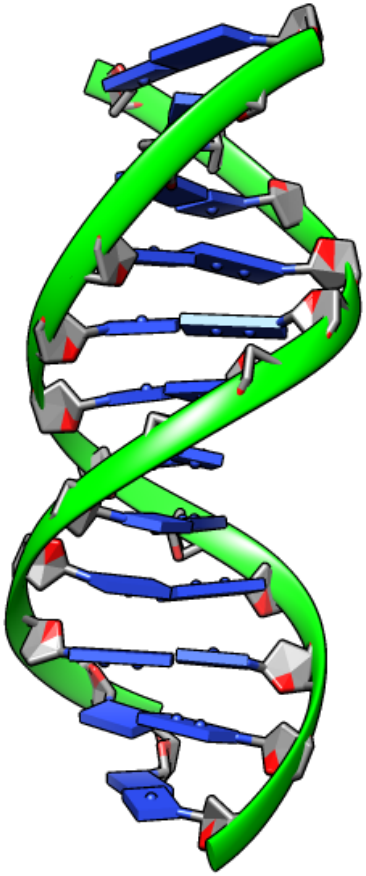


Large-scale studies of gene regulatory, protein, metabolic networks in cells are conducted in systems biology

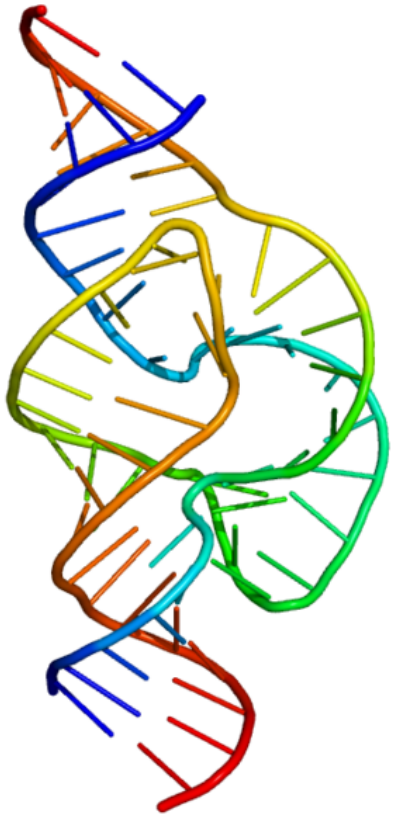
Cell Information and Machinery

- A cell stores all the information needed to replicate itself
- The human genome is around 3 billion base pairs long
- Almost every cell in the human body contains the same set of genes
- What differentiates cells in your body?
  - Not all genes are expressed at the same time in the same way in all cells
- A cell is a machinery
- It collects and manufactures its own components
- It carries out its own replication
- It kicks the start of its new offspring

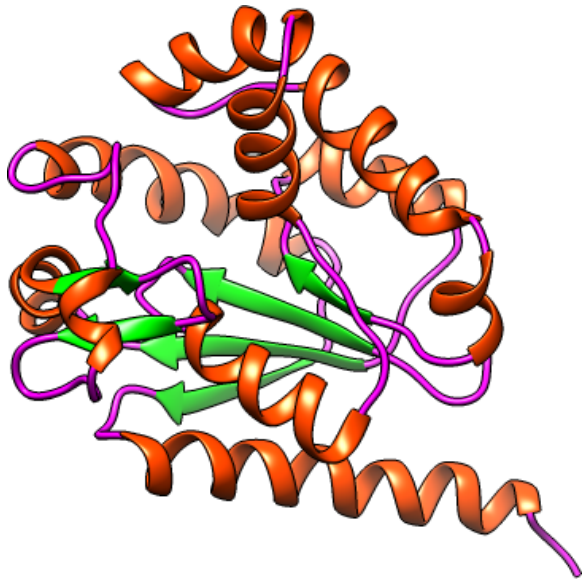
The Three Life-Critical Molecules



DNA



RNA



Protein

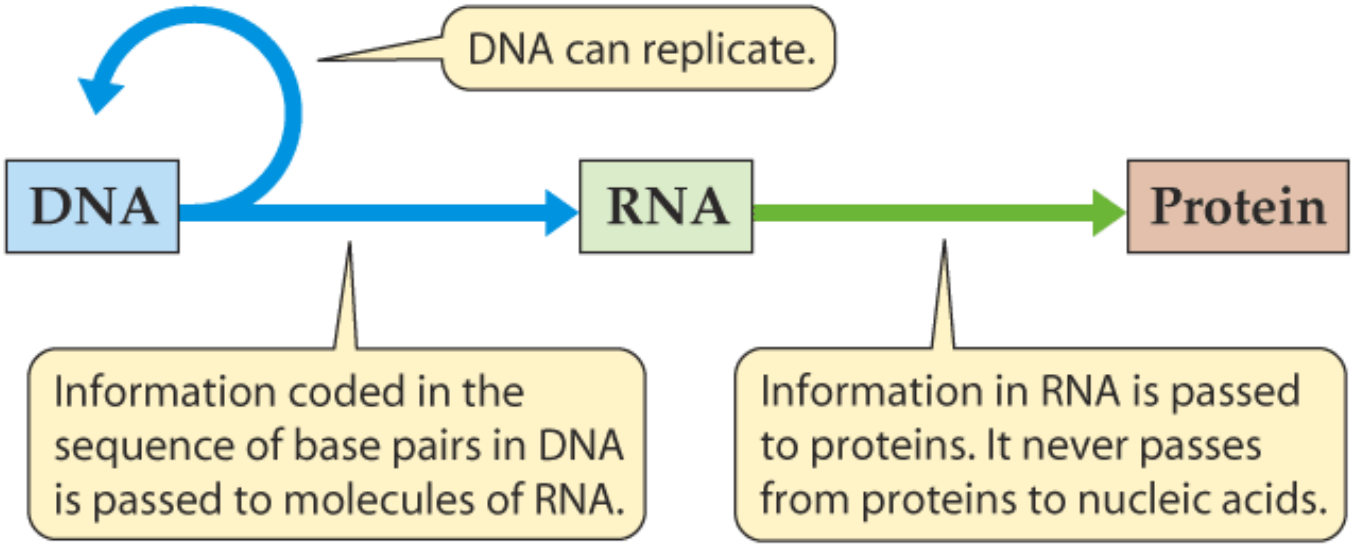
Overview Before Heading to Central Dogma

- Nucleus = Library.
- Chromosomes = Bookshelves.
- Genes = Books.
- Books represent the information (DNA) that every cell needs so it can grow and carry out its own set of functions.
- Genome – an organism's genetic material.
- Gene – discrete unit of hereditary information located on chromosomes and consisting of DNA.
- Genotype – genetic makeup of an organism.
- Phenotype – physical expressed trait of an organism.
- Nucleic acids – biological molecules (RNA and DNA) that allow organisms to reproduce.
- Proteins – complex molecules made up of smaller subunits called amino acids.

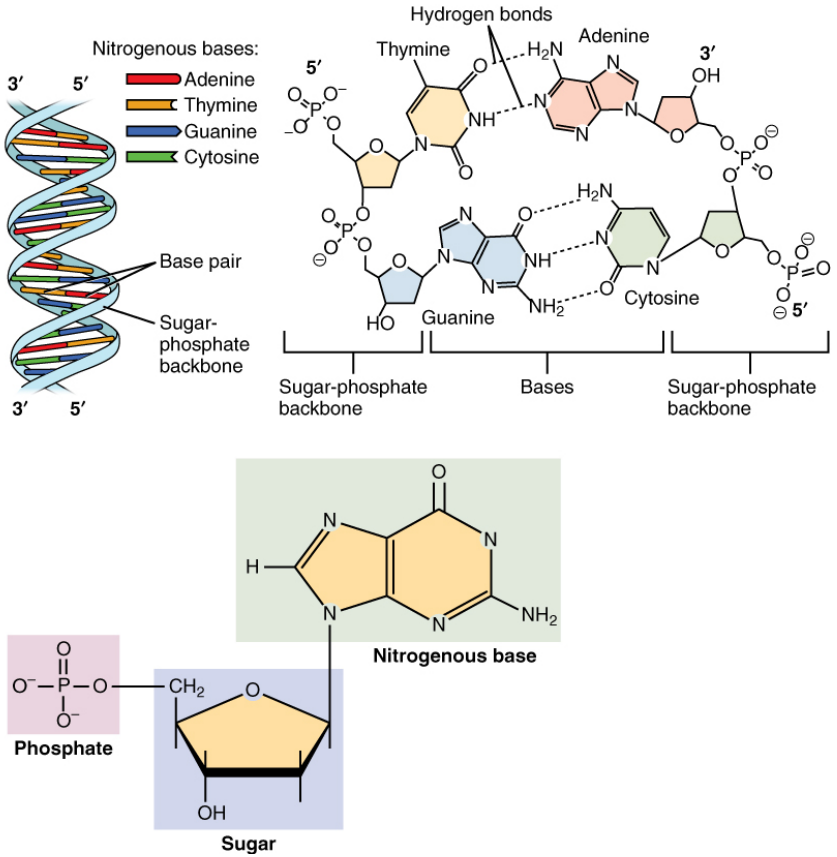
Central Dogma of Molecular Biology

- Information flows from DNA through RNA to Synthesize Proteins in cells
- DNA
  - Holds information on how the cell works
- RNA
  - Acts to transfer short pieces of information to different parts of the cell
  - Provides templates to synthesize into proteins
- Proteins
  - Can be enzymes that send signals to other cells and regulate gene activity
  - Form the body's major components (hair, skin, etc.)
  - Often referred to as the workhorses of the cell

Central Dogma of Molecular Biology

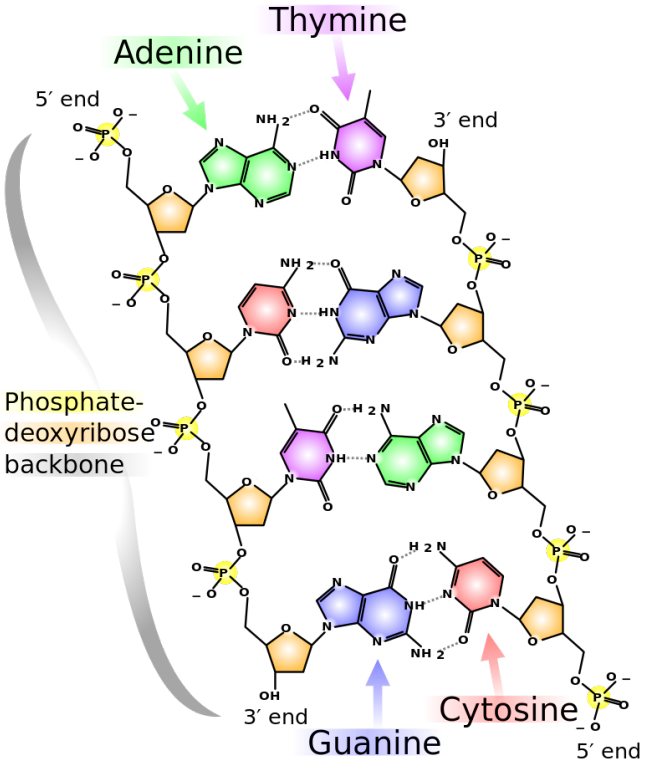


DNA: Sequence and Structure



The four fundamental units of DNA are: Adenine (A), Guanine (G), Thymine (T), and Cytosine (C) They pair up on complementary strands: A-T and C-G. Like a four-letter alphabet.

DNA: Sequence and Structure

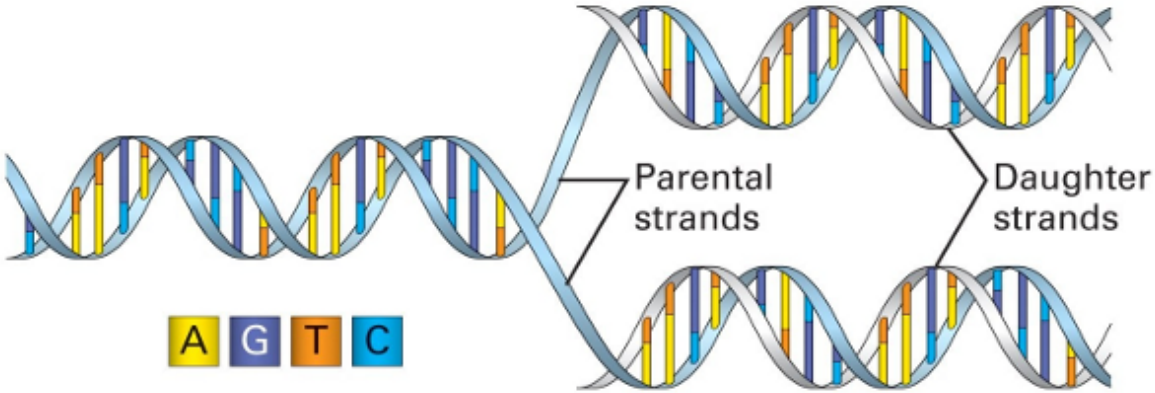


The double helix structure is composed of: sugar molecules phosphate groups bases (A, C, G, T)

Base pairs form hydrogen bonds 2 bonds link A to T 3 bonds link C to G

DNA always reads from 5' end to to 3' end for transcription and replication 5' ATTTAGGCC 3' 3' TAAATCCGG 5

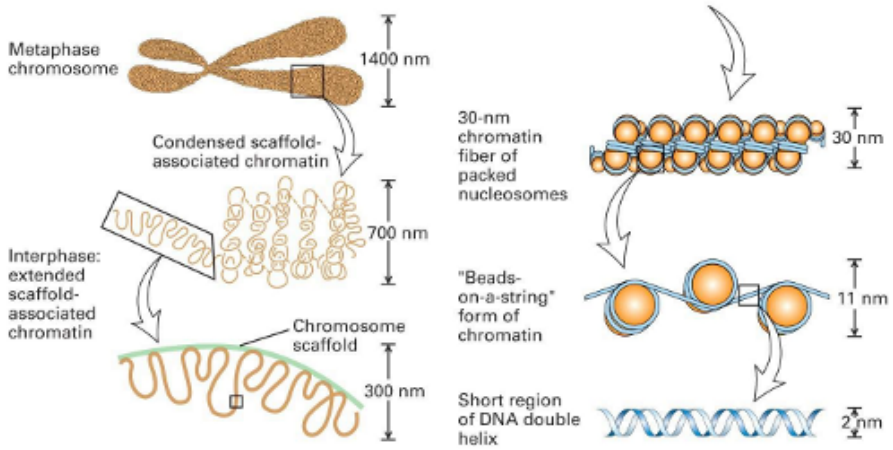
DNA: Sequence and Structure



How DNA copies itself:

`https://www.youtube.com/watch?v=2_-jSoSaaTA`

DNA: Sequence and Superstructure



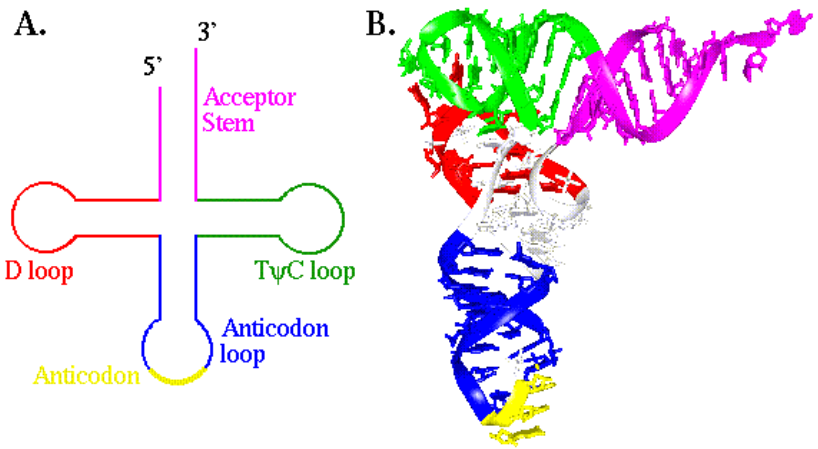
Lodish et al. Molecular Biology of the Cell (5th ed.). W.H. Freeman & Co., 2003.

DNA in living cells is highly compact and structured.

Transcription factors and RNA polymerase need access to DNA.

Transcription is dependent on the structural state – sequence alone does not tell the whole story.

RNA: Sequence and Structure



RNA is chemically similar to DNA, but T(hymine) is replaced with U(racil) and the ribose instead of deoxy-ribose.

Some forms of RNA can fold to create secondary structures – has implication for function

DNA and RNA can pair with each other

There are several forms of RNA:

1. mRNA (messenger RNA) – carries a gene's information out of nucleus
2. tRNA (transfer RNA) – transfer's mRNA's information onto a protein chain of amino acids
3. rRNA (ribosomal RNA) – part of the ribosome, where proteins are synthesized

Transctiption: From DNA to mRNA

Transcription refers to the process of copying a piece of the DNA onto mRNA

Catalyzed by transcriptase enzyme

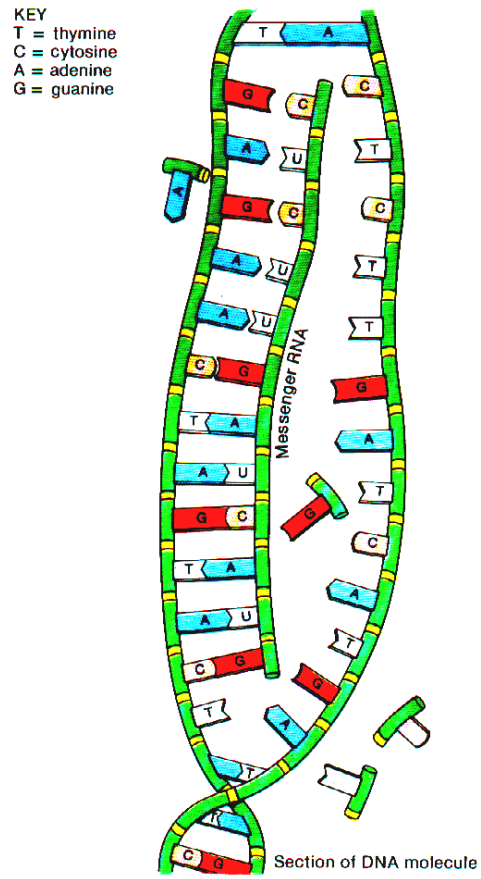
The enzyme recognizes a promoter region to begin transcription

About 50 base pairs can be transcribed per second in bacteria – multiple transcriptions can occur

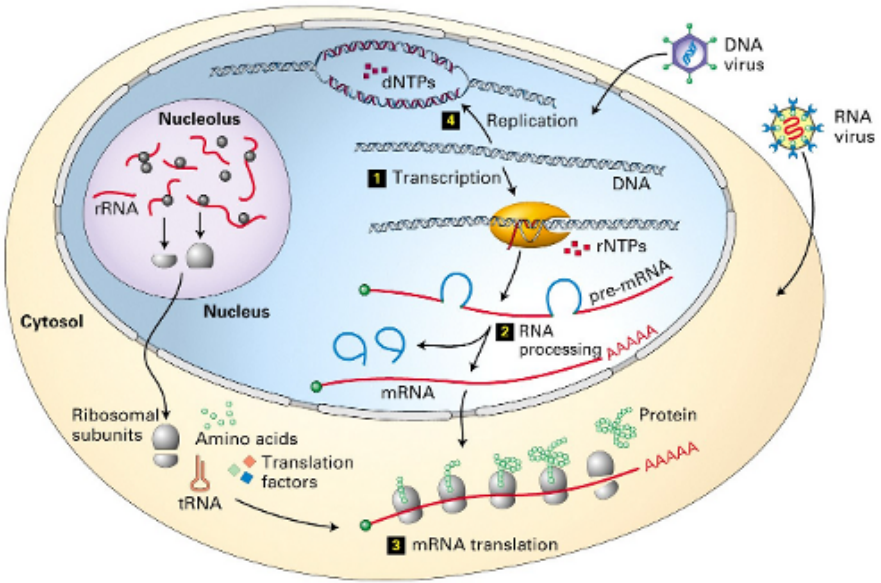
The process of how the enzyme finds the promoter regions is partially understood and related to the problem of motif finding in bioinformatics

Repressor and inhibitor enzymes act in various ways to stop transcription. This makes the regulation of gene transcription difficult to understand, model, or control

Question: How does splicing complicate the picture?



From DNA and RNA to Proteins



DNA contains introns and exons.
Introns ("junk DNA") – not fully understood, probably not junk. Exons read to transfer information to mRNA $\rightarrow$ Transcription (RNA synthesis)

mRNA goes to the ribosome tRNAs line up to link amino acids in a chain $\rightarrow$ Translation – Protein Synthesis
`https://youtu.be/sntF8XEjl7Q?si=BKiTupso_ExKb34f&t=220`

Translation – Codons and Amino Acids

DNA: strings of four letter codes A, T, C, G

RNA: strings of four letter codes A, U, C, G (uracyl replaces thymine).

Proteins: strings of twenty letter codes. The letters are the fundamental building blocks called amino acids.

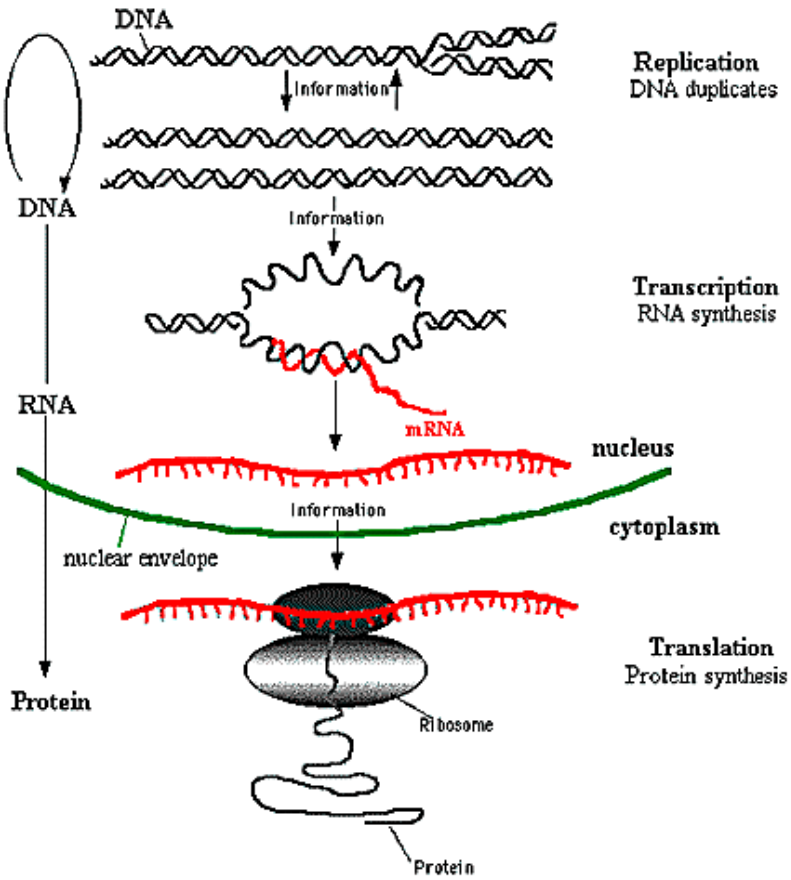
Each amino acid is coded by 3 nucleotides called codon

There are 20 amino acids, Many codons code for the same amino acid. Some codons indicate when to stop reading the genetic information

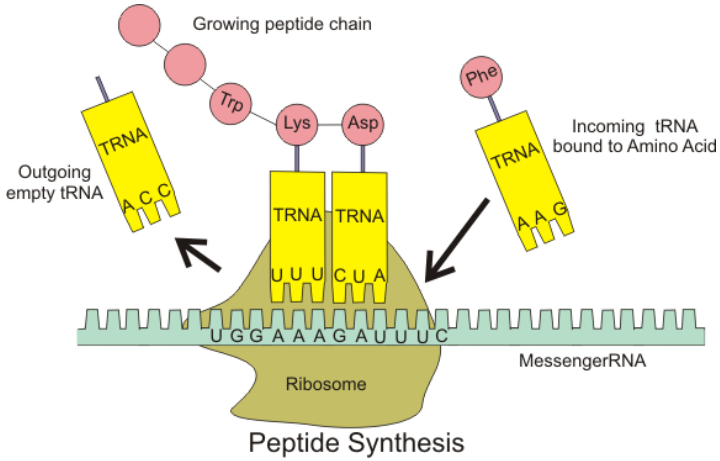
The Universal Genetic Code

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

From mRNA to Proteins



The Central Dogma of Molecular Biology



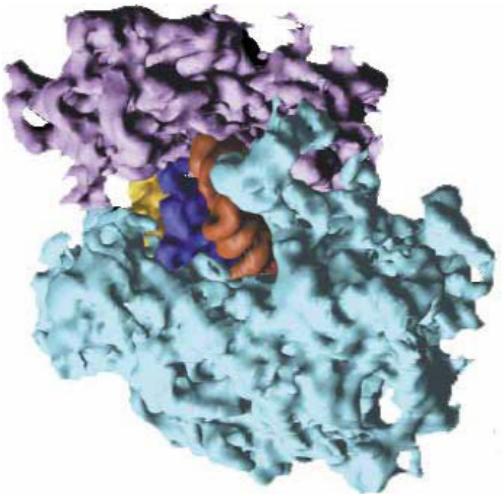
mRNA travels to the ribosome

Each codon of mRNA determines what tRNA to line itself up in order

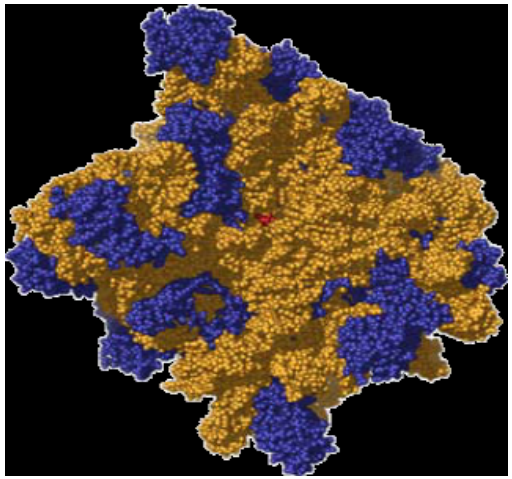
Each tRNA then transfers its aminoacyl group to the growing chain of amino acids

Example: the tRNA with the anticodon AAG corresponds with the codon UUC of mRNA and attaches Phenylalanine (PHE, F) onto the growing protein chain

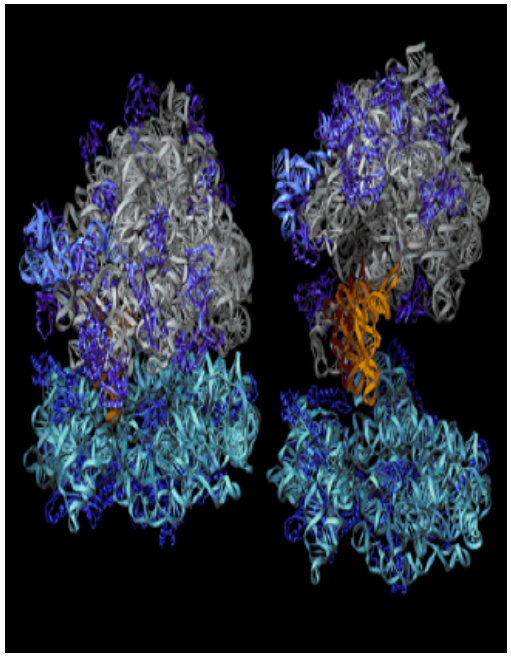
Ribosome - Protein Assembly



Bacterial ribosome (nigms.nih.gov)

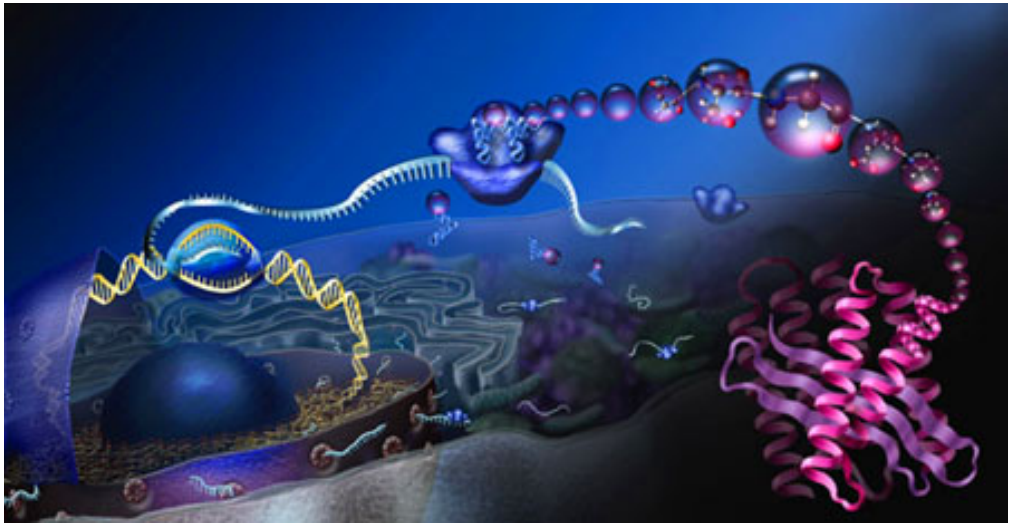


50S large subunit of the ribosome. Proteins are in blue, RNA in orange, and active sites are shown in red (wikipedia.org).



Closed depiction (above) and "cracked open" depiction (below) exposing the tRNAs. [adapted from M Yusupovet al. (2001) Science]

Proteins: Workhorses of the Cell



In the ribosome:
the chain of amino acids folds (arranges itself in three dimension) into a unique structure where the protein is functional
This is known as protein folding.

Deviations from this structure can be lethal and give rise to disease.
This is known as misfolding.

Proteins do all essential work in cells:
a. build cellular structures
b. digest nutrients
c. execute metabolic functions
d. mediate information flow within a cell and among cell communities
e. work together with other proteins or nucleic acids as molecular machines

Understanding (computing) how the chain of amino acids determines what structure(s) a protein assumes in cells is an important problem in computational (structural) biology: The protein folding problem (A variant: The structure prediction problem).

Basic Summary of What We Know

- DNA, RNA, and Proteins are specified linearly (linear strings of characters)
- DNA and RNA are constructed from nucleic acids (nucleotides)
  - Strings written in a four-letter alphabet (C, G, A, T/U)
- Proteins are constructed from amino acids
  - Strings written in a twenty-letter alphabet
  - These strings fold into complex 3D structures

Sequence Bioinformatics

patterns about the sequence can reveal insight into transcription, translation, and function of synthesized proteins.

Genomic sequences represent a written language of 4-letter alphabet

DNA decoding techniques not very different than those for decoding an ancient language

Structural Bioinformatics

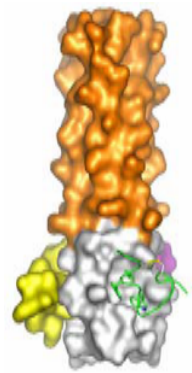
When sequence decoding reaches limit

When structure reveals further information - relevance of DNA, RNA, Protein structure.

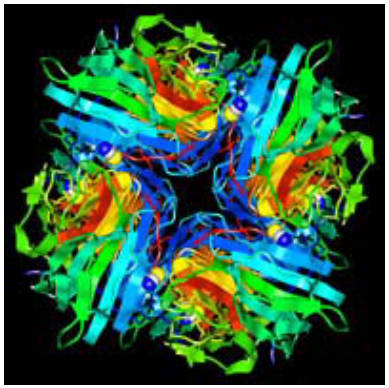
When understanding how molecules fit to create machines requires more than sequence information

Structure to Function

- Structure determines reactions in cells
- Structures of proteins that are complementary fit with one another
- Problem: Using structure (and possibly sequence) infer active (infer) sites
- Sites where proteins interact with other molecules
- Problem: Using structure (and possibly sequence) infer the structure and function of an amino-acid chain synthesized from a decoded gene sequence



Inhibitors (green, yellow, purple) bind to (block) an HIV protein mimic in three "pockets" that are essential to the virus' ability to enter cells. (bnl.gov)



Left (anl.gov) and right (nigms.nih.gov) structures suggest that the proteins are involved in DNA binding or transfer

Biological Databases

- An increasing amount of biological and sequence data is freely available in online databases
- Large amounts of data also pose an interesting computational problem of how to store them

An always improving, changing, increasing list of biological databases:

- NCBI – `https://www.ncbi.nlm.nih.gov/`
  - contains many subdatabases
  - nucleotide sequence database the is most prominent
- Protein Data Bank `http://www.rcsb.org`
  - contains protein structures
- Uniprot – `http://www.uniprot.org/`
  - contains annotated protein sequences
- Prosite – `http://kr.expasy.org/prosite`
  - Database of motifs of protein active sites

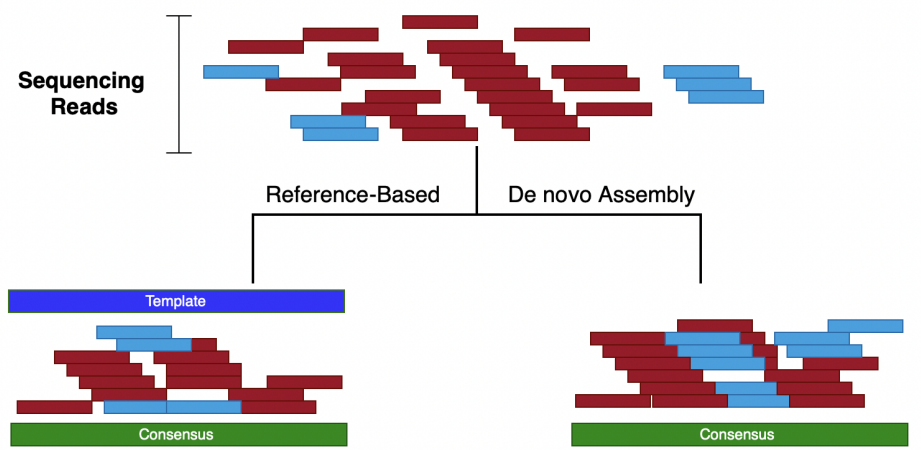
Employing Databases for Sequence Analysis

- Analyze biological sequences for patterns
  - RNA splice sites – what are they and why?
  - Open reading frames (ORFs) – can be used to generate proteins?
  - Amino acid propensities in a protein – why?
  - Conserved regions in proteins – possible active sites
  - Conserved regions in DNA and RNA – possible protein (Transcription Factors) binding sites
- Predict from sequence
  - Protein and RNA topology and 3D structure.
  - Protein binding/active sites
  - Protein Function

Fundamental sequence question: How are genomes assembled, mapped, annotated?

Genome Assembly

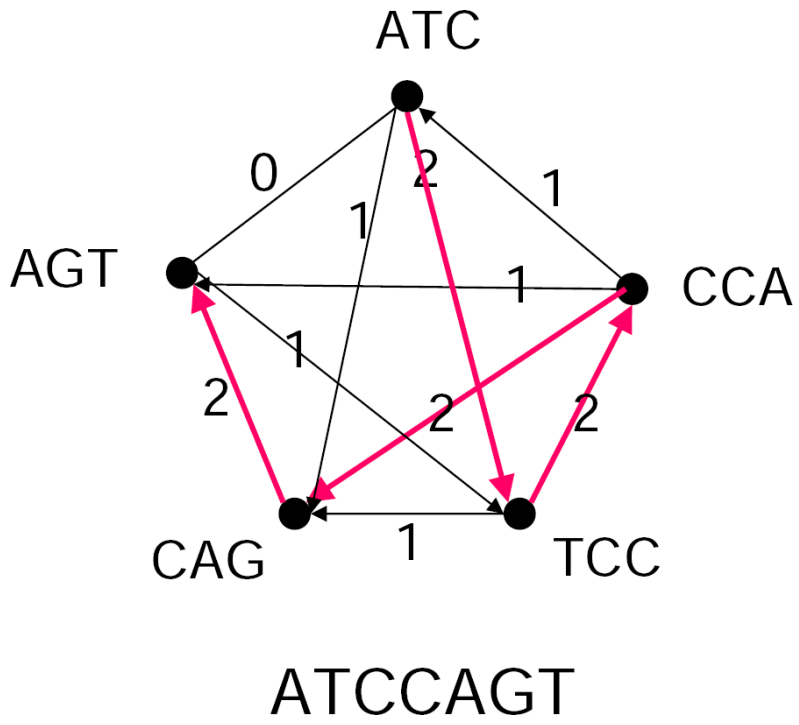
- Length of sequenced genome fragments is limited
- Genome is fragmented
  - Enzymes splice/cut
- Fragments then need to be taken and put back together in right order
- Either de-novo (From scratch) or reference based
- Not easy to do



By Erekevan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=116777958

Genome Assembly Using De Bruijn Graphs

- Shortest Common Superstring Problem (SCS)
- Needed because fragments overlap
- Fit overlapping sequences together to find the shortest sequence that includes all fragment sequences
- Fragments may contain sequencing errors
- Two complements of DNA
  - Need to take into account both 3' and 5'
- Repeat problem:
  - 50% of human DNA is just repeats



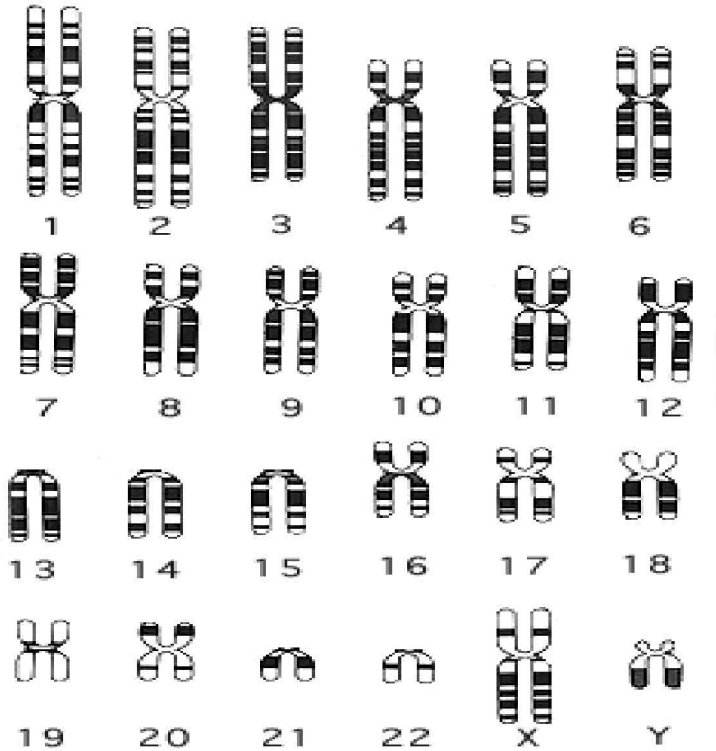
Genome Assembled – Now What

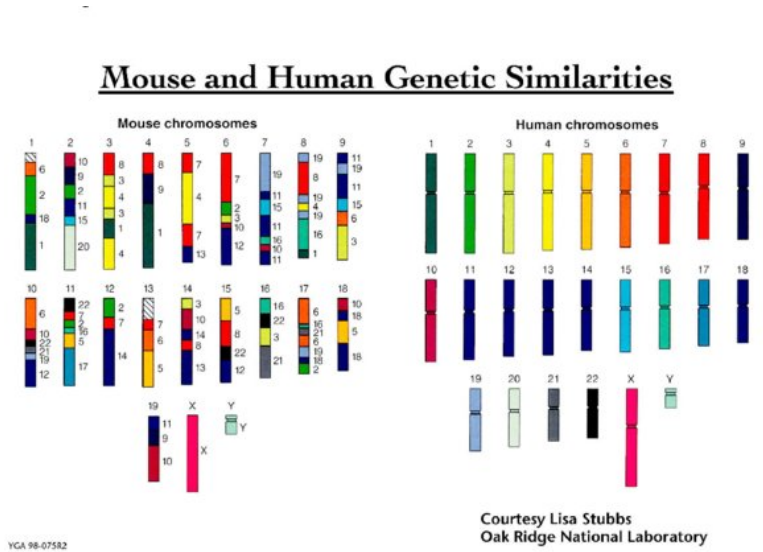
- Tracing Phylogeny
  - Find (evolutionary) family relationships by tracking similarities between species
- Gene Annotation/Finding (comparative genomics)
  - Comparison of Similar Species
- Determining Regulatory Networks
  - The variables that control the body's response to stimuli
- Proteomics
  - From DNA sequence to a folded protein with known function
  - The main part of this course will focus heavily on computational proteomics

Gene Finding

- Identifying protein-encoding regions (exons) from "junk" DNA (introns)
- Ab-initio predicting methods – Hidden Markov Models, Machine learning approaches...
- Comparative genomics.

Human Chromosomes



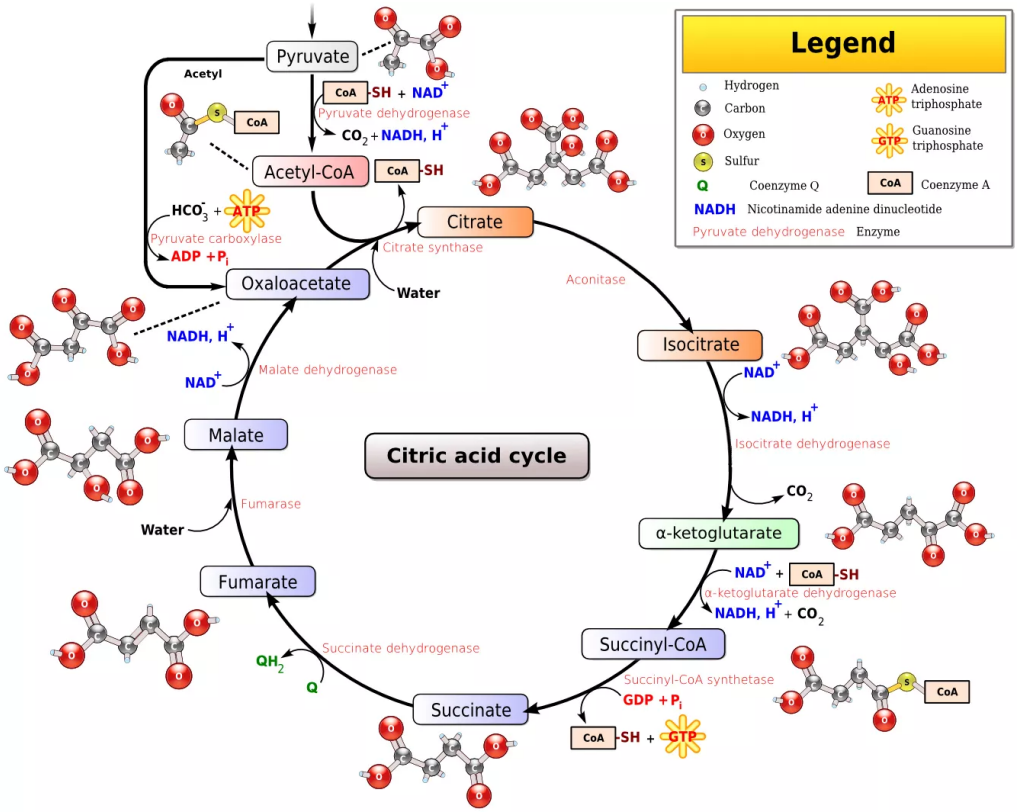


More Than Sequence and Structure

Nodes: Metabolites
Edges: Biochemical reaction

Analysis of this metabolic network often employs tools from graph theory in computer science

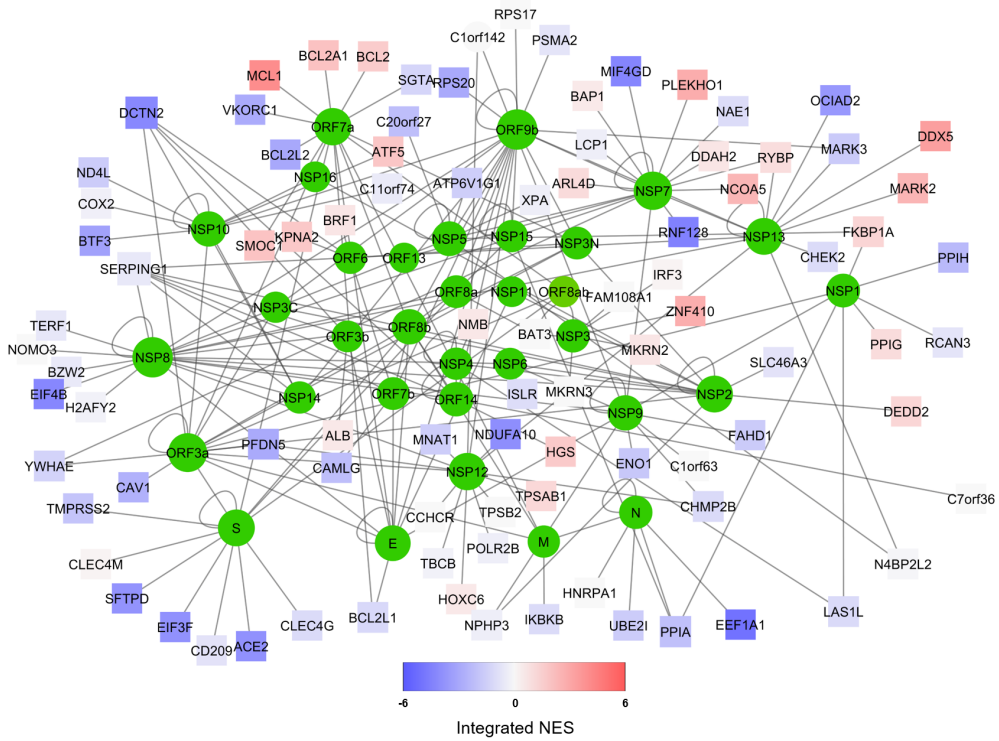
What are some questions we can ask about a metabolic network?



Citric acid cycle

Protein Interaction Networks

- Analysis of protein interaction networks also employs tools from graph theory in computer science
- What are some questions we can ask about p2p networks?
- e.g.: how could we predict the function of a gene through such a network?



Nodes: Proteins. Edges: Interactions

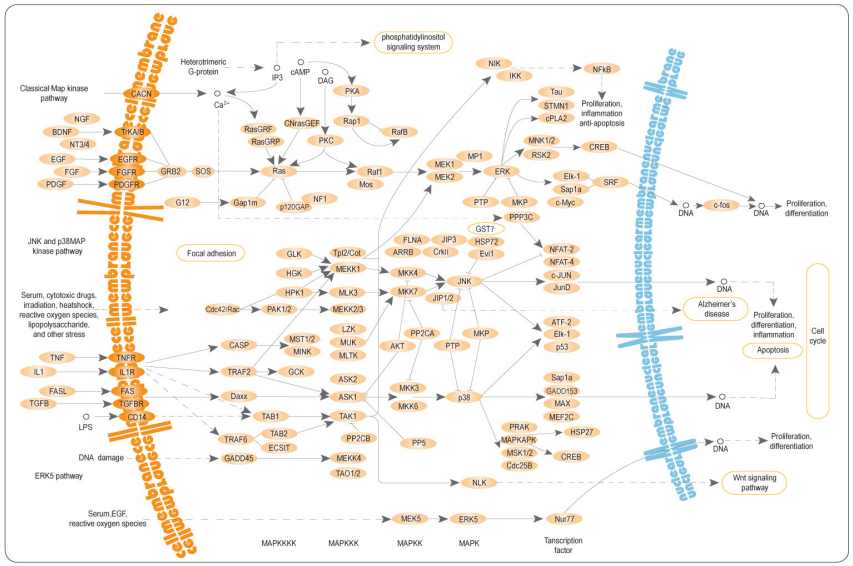
Master Regulator Analysis of the SARS-CoV-2/Human Interactome
https://www.mdpi.com/2077-0383/9/4/982

Signaling Networks

Nodes: Different molecules: Proteins or neurotransmitters
Edges: Activation or deactivation

Analysis of signaling networks also employs tools from graph theory in computer science

What are some questions we can ask about p2p networks? e.g.: what does a path in the network mean?



MAPK pathway, from Wikipedia

Mining/Analyzing, Modeling, Predicting

- Modeling biological processes (such as what?) allows us to test whether we fully understand the process and whether we know all the variables that control it
  - We build models of proteins to explore the sequence-structure-function relationship
  - Models of molecular interactions to test whether molecules interact to achieve a biological function in cells
  - Models of gene regulatory networks to understand how genes interact with one another
  - We build (systems biology) models of entire cells
- We use information from biologists, chemists, physicists, computer scientists to build such models
- We then use computers to simulate the behavior or properties of molecules or cells being modeled over time and space
- If the behavior or properties are different from those "seen" in the wet lab, we correct our model – this improves our (theoretical and computational) understanding
- If the model is correct and we observe additional properties – we have crossed into the discovery and prediction contribution of computational biology

The complexity of the models and the simulations requires fast, efficient, accurate computer algorithms and powerful machines

What do Bioinformaticians Do?

- Mining for information:
  - Let scientists discover the biology of cells.
  - Encourage the deposition of gene sequences, protein sequences, protein structures decoded, resolved by experimentalists in databases.
  - Organize and cross-link databases so information can be quickly extracted and cross-referenced.
  - Conduct fast large-throughout searches in sequence databases.
  - Compare sequences of existing and novel genes, proteins to infer knowledge about structure and function.
- ab initio modeling:
  - Apply principles of physics to fold chains into 3D structures.
- Combination of the two:
  - Statistical information from the databases with ab initio computing
- Study large patterns in interaction networks
- ...

The Future of Computational Biology

- A significant part of our time is spent in improving the accuracy of our modeling
- The rest of that time is spent in extending the speed, accuracy, and applicability of our algorithms in order to make meaningful predictions through computers
- Bioinformatics and Computational Biology are still in infant stages
  - There are a lot of things we do not understand
  - A lot of questions to be resolved
  - The main one revolves around control: how can we control the behavior of a molecule, a group of molecules, and then a cell?
- These questions are often asked in the context of biomedical research or bioengineering, where our focus is on building effective therapeutics (to improve the health of society) or novel functional materials (to improve the living conditions of our citizens)

Sources Cited

- Ernst Mayr, "What evolution is".
- Neil C. Jones, Pavel A. Pevzner, "An Introduction to Bioinformatics Algorithms".
- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell. New York: Garland Science. 2002.
- Mount, Ellis, Barbara A. List. Milestones in Science & Technology. Phoenix: The Oryx Press. 1994.
- Voet, Donald, Judith Voet, Charlotte Pratt. Fundamentals of Biochemistry. New Jersey: John Wiley & Sons, Inc. 2002.
- Campbell, Neil. Biology, Third Edition. The Benjamin/Cummings Publishing Company, Inc., 1993.
- Snustad, Peter and Simmons, Michael. Principles of Genetics. John Wiley & Sons, Inc, 2003.
- Bioinformatics workbook `https://bioinformaticsworkbook.org/`