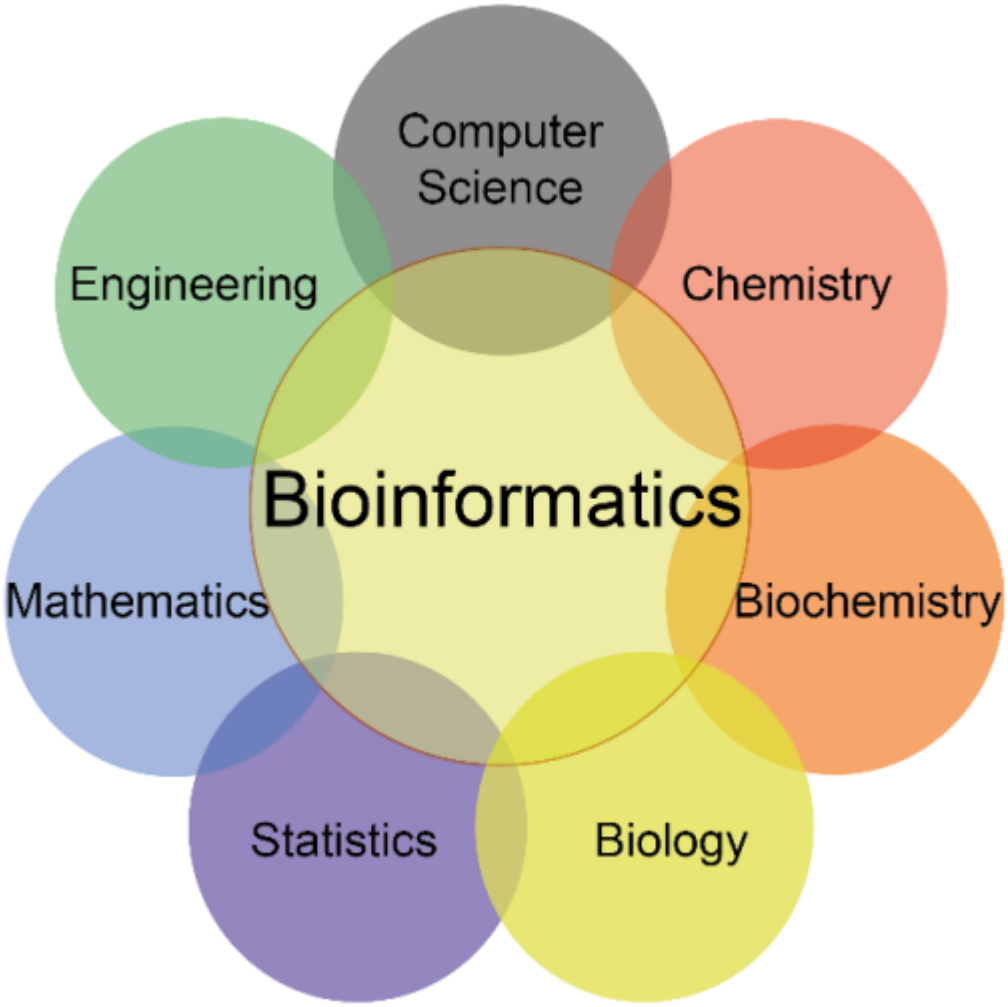
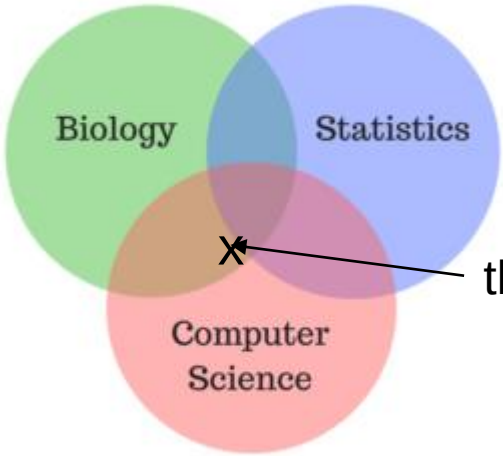
CS 342: Bioinformatics
Lecture 1

Catie Welsh

Bioinformatics is Interdisciplinary!





Biology 101

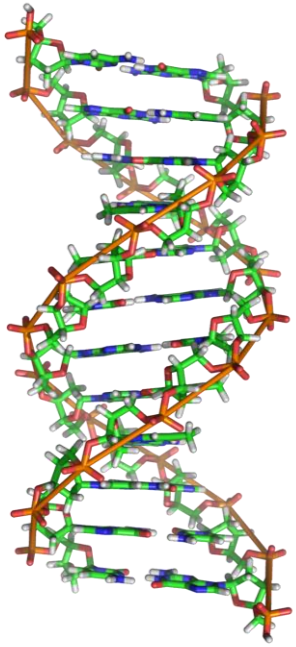
Actually…just the cliff notes
Actually…we'll just focus on one branch of biology…

Molecular biology is a branch of science concerning **biological** activity at the **molecular** level.

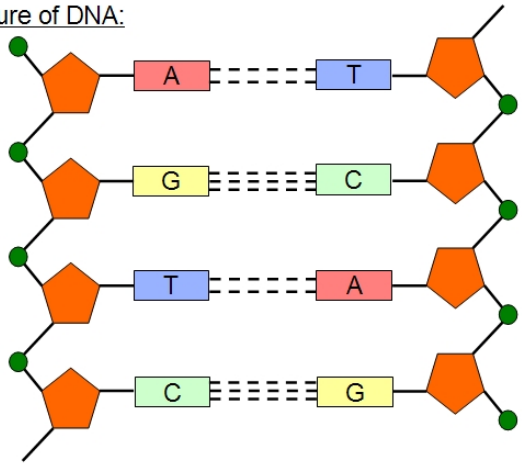
- **DNA**
- **RNA**
- **Protein**

DNA

Each strand composed of sequence of covalently bonded **nucleotides** (**bases**).



Structure of DNA:



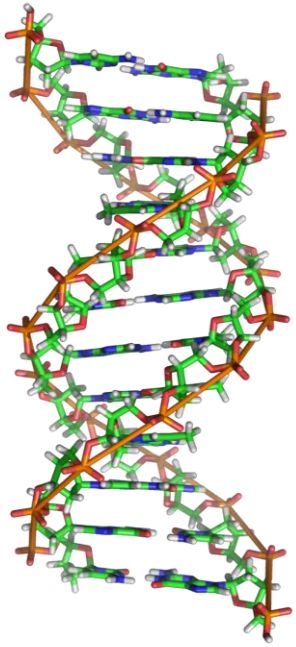
Four nucleotides

A (adenine)
C (cytosine)
T (thymine)
G (guanine)

A ←→ T, C ←→G Watson-Crick base-pairing

DNA

Each strand composed of sequence of covalently bonded **nucleotides** (**bases**).



5’ …ACGTGACTGAGGACCGTG… 3’

… | | | | | | | | | | | | | | | | | | …

3’ …TGCACTGACTCCTGGCAC… 5’

Pair of strings from 4 character alphabet

Read: 5’ ⟶ 3’

5’ …ACGTGACTGAGGACCGTG

CGACTGAGACTGACTGGGT

CTAGCTAGACTACGTTTTA

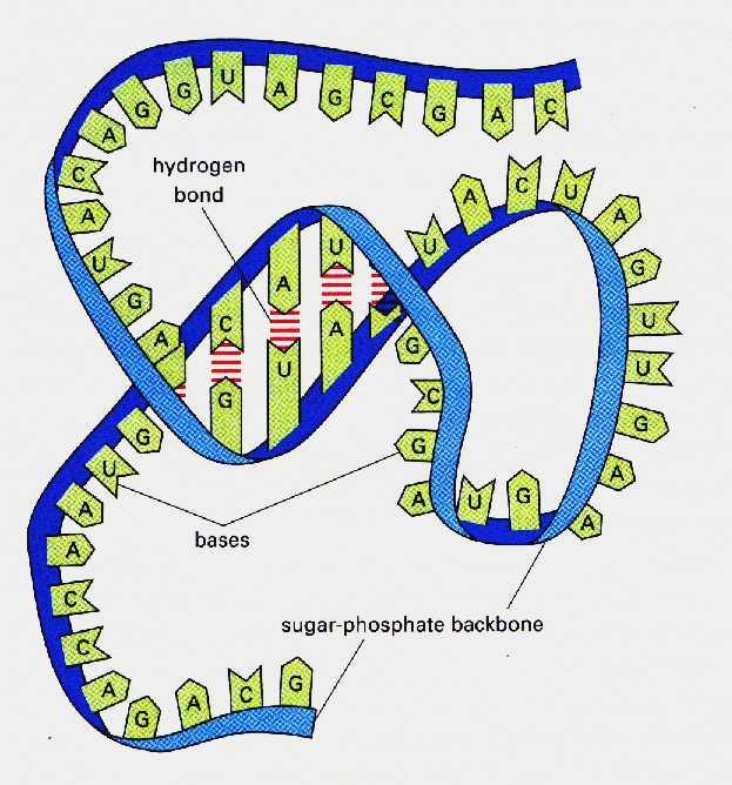
TATATATATACGTCGTCGT

ACTGATGACTAGATTACAG

TGATTTTAAAAAAATATT… 3’

Single **string** from 4 character alphabet

RNA

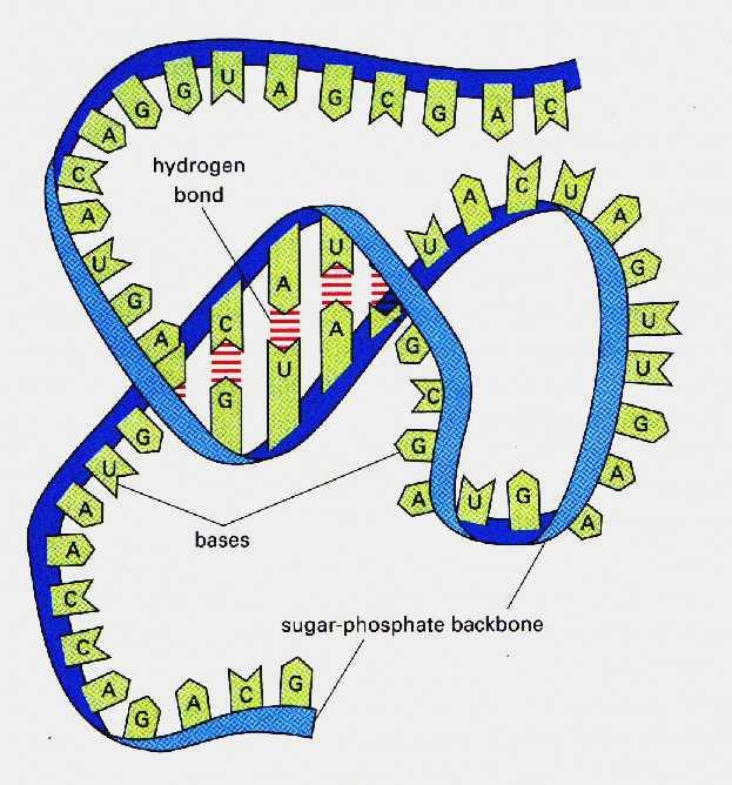


- **Single-stranded**
  - A (adenine)
  - C (cytosine)
  - U (uracil)
  - G (guanine)
- Can fold into **structures** due to base complementarity.

  A ←→ U, C ←→G
- Comes in many flavors:

  mRNA, rRNA, tRNA, tmRNA, snRNA, snoRNA, scaRNA, aRNA, asRNA, piwiRNA, etc.

RNA



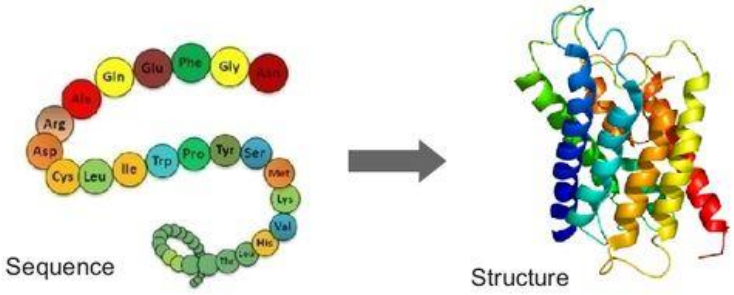
...ACGUGACUGAGGACCGUG...

A single string from 4 character alphabet

Protein

Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
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

- A large molecule consisting of a long chain of amino acids.
- Folds into 3D structures to perform various functions in cells



Protein

A single string from a 20 character alphabet

… DTIGDWNSPSFFGIQLV
SSVHTTLWYRENAFPVLGG
FSWLSWFNWHNMGYYYPVY
HIGYPMIRCGTHLVPMQFA
FQSIARSFALVHWNAPMVL
KINPHERQDPVFWPCLYYS
VDIRSMHIGYPMIRCYQA…

Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
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

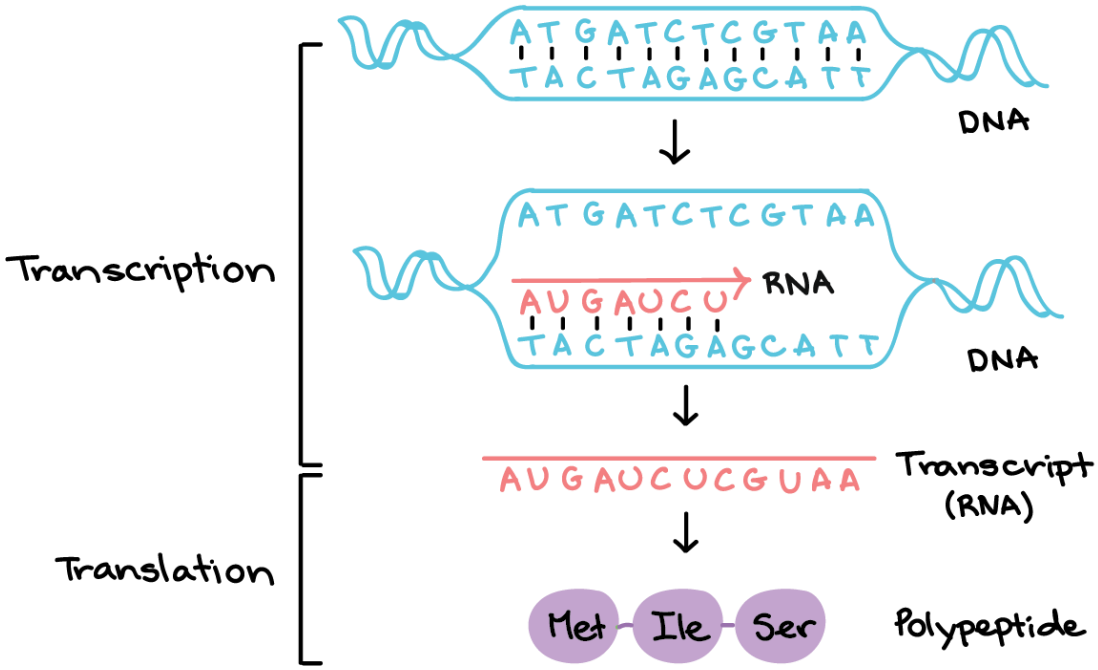
Molecule	Representation
DNA	String from a 4 character alphabet ({A,C,G,T})
RNA	String from a 4 character alphabet ({A,C,G,U})
Protein	String from a 20 character alphabet ({A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V})

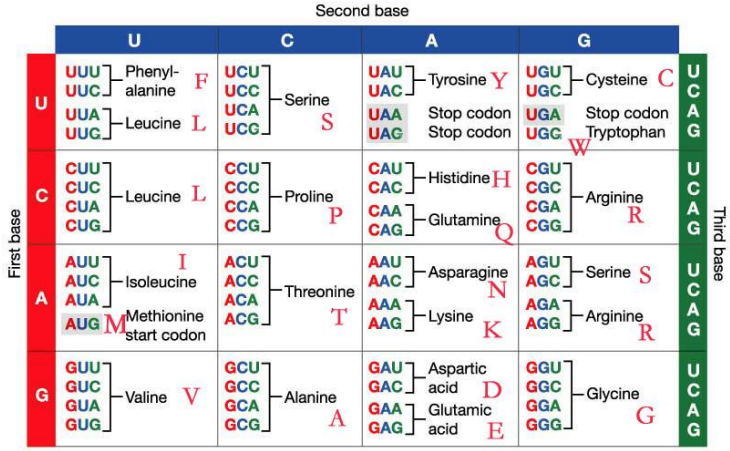
Molecule	Representation	Function
DNA	String from a 4 character alphabet ({A,C,G,T})	Information storage
RNA	String from a 4 character alphabet ({A,C,G,U})	Old: Messenger, New: Many
Protein	String from a 20 character alphabet ({A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V})	Perform cellular functions (biochemistry, signaling, etc.)

Central Dogma: DNA makes RNA makes Protein

The process by which cells “read” the genome

DNA → RNA → Protein (The Central Dogma)





https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription

Molecule	Representation	Function
DNA	String from a 4 character alphabet ({A,C,G,T})	Information storage
RNA	String from a 4 character alphabet ({A,C,G,U})	Old: Messenger, New: Many
Protein	String from a 20 character alphabet ({A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V})	Perform cellular functions (biochemistry, signaling, etc.)

This class will focus on **Algorithms** on strings, trees, and graphs designed for analyzing DNA, RNA and Proteins

Activity

Come up with an "algorithmic" problem (specify input and output) that deals with strings. Be as specific as possible. Be creative!

Input: A DNA string $D = d_1d_2d_3\ldots d_n$

Output: The corresponding RNA sequence $R = r_1r_2r_3\ldots r_n$ where r_i corresponds to the base d_i after transcription.

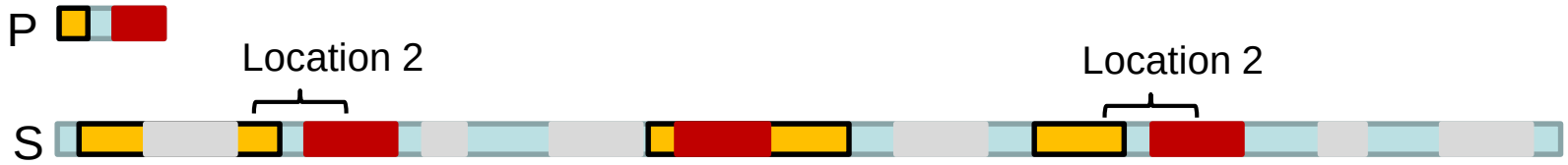
Input: Two DNA strings D_1 and D_2

Output: True if D_1 appears somewhere in D_2

Topic: Pattern Matching

Input: String S and a Pattern P

Output: Find the location of anywhere that P appears exactly in S

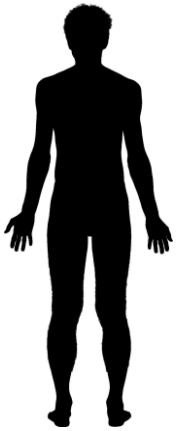


Key Question: How do we do this quickly?

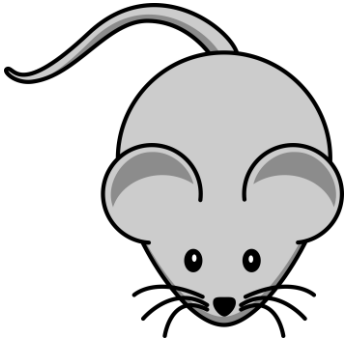
LOTS of biologically motivated applications!
(e.g. finding start and stop codons)

Topic: Sequence Alignment

Question: How do we compare two genomes?



vs.



Human Genome:

…ACTCGACTGAGAGGATTTCGAGCATGA…

≈3.2 x 10^9 bp

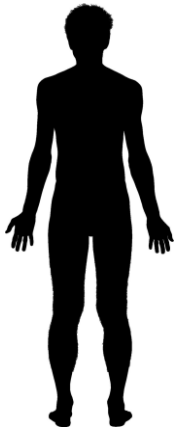
Mouse Genome:

…ACTCAACTGAGATTCGAGCTTCAATGA…

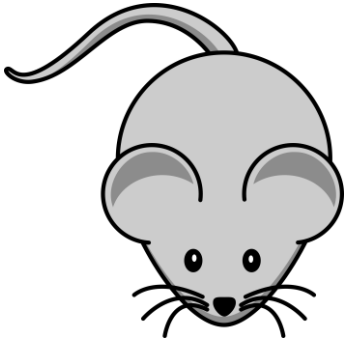
≈2.8 x 10^9 bp

Topic: Sequence Alignment

Question: How do we compare two ~~genomes~~ genes?



vs.



Human Gene:

`ACTCGACTGAGAGGATTTCGAGCATGA`

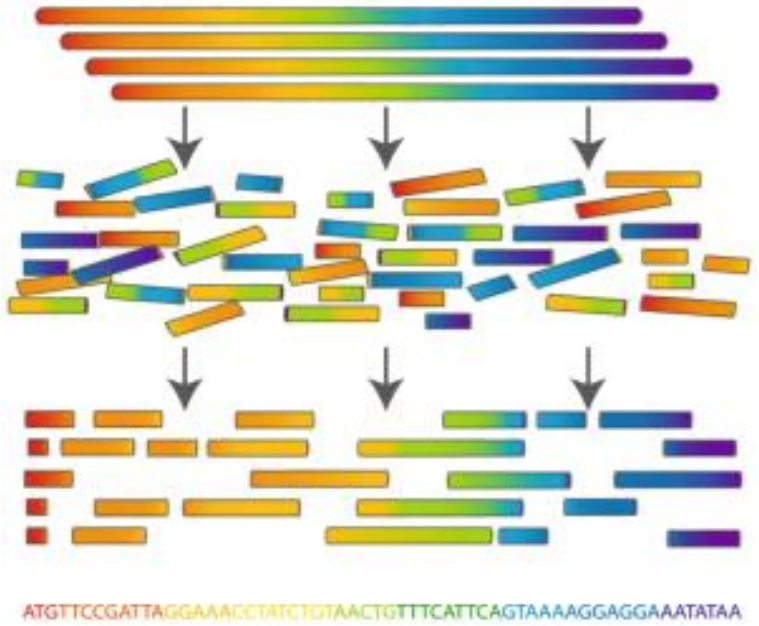
≈10,000 – 15,000 bp

Mouse Genome:

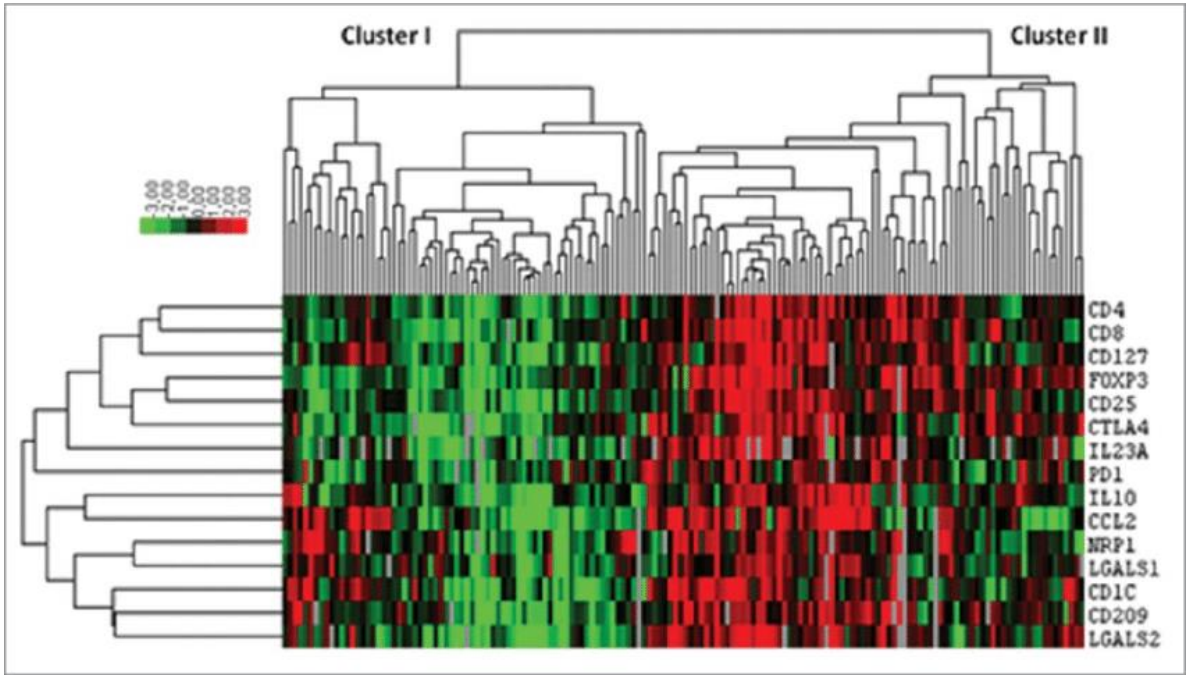
`ACTCAACTGAGATTCGAGCTTCAATG`

Topic 3: Genome Assembly

Question: Given a bunch of sequences from a genome, how do we reconstruct the original genome? (This is necessary because of how DNA sequencing works.)

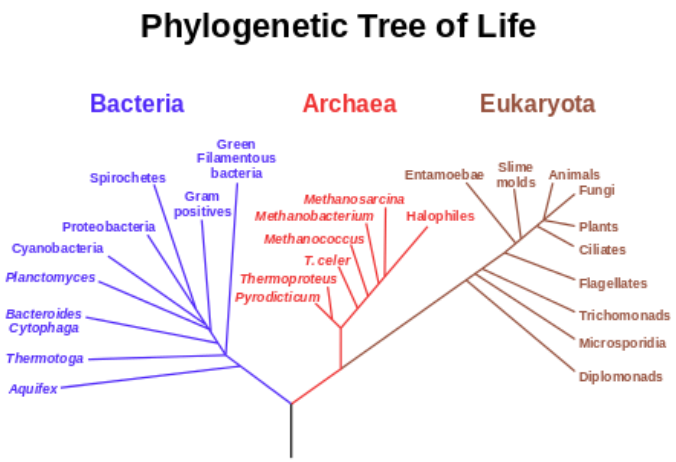


Topic: Clustering



Question: Given a dataset, can you break it into sub groups where items in a group are similar to each other, but different from items in other groups?

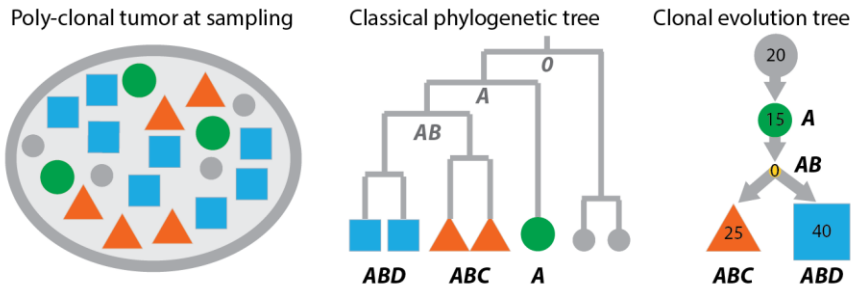
Topic: Phylogenetics



https://en.wikipedia.org/wiki/Phylogenetic_tree

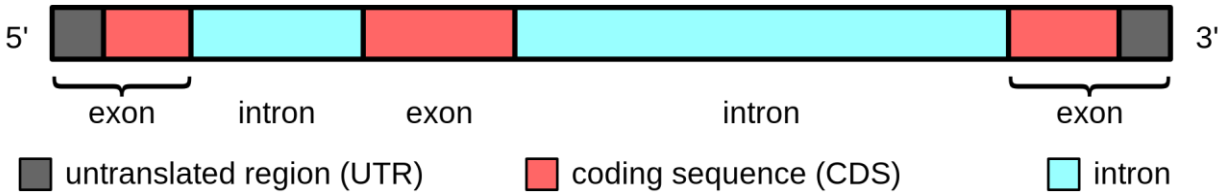
Question: How can we reconstruct the evolutionary history of different species?

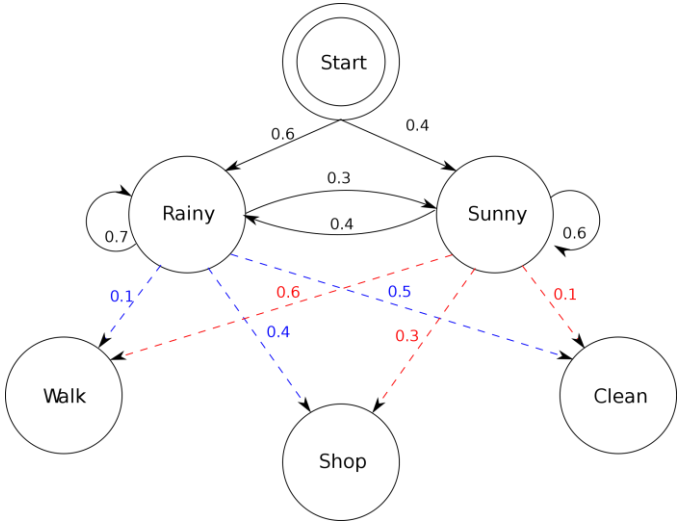
Question: Can we recover how a tumor has evolved overtime?



https://scientificbsides.wordpress.com/2014/06/09/inferring-tumour-evolution-2-comparison-to-classical-phylogenetics/

Topic: Gene Finding and HMMs





Question: Not all DNA "codes" for proteins (is a gene). How do you find the portion of DNA that is a gene?

Course Logistics