CS 342: Bioinformatics
Lecture 1
Catie Welsh
Bioinformatics is Interdisciplinary!

- Mathematics
- Biology
- Computer Science
- Chemistry
- Biochemistry
- Engineering
- Statistics

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

A $\leftrightarrow$ T,
C $\leftrightarrow$ G

Watson-Crick base-pairing

Four nucleotides
A (adenine)
C (cytosine)
T (thymine)
G (guanine)
Each strand composed of sequence of covalently bonded nucleotides (bases).

Pair of strings from 4 character alphabet

Read: 5’ → 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.
  \[ A \leftrightarrow U, \quad C \leftrightarrow G \]

- Comes in many flavors:
  - mRNA, rRNA, tRNA, tmRNA, snRNA, snoRNA, scaRNA, aRNA, asRNA, piwiRNA, etc.
RNA

...ACGUGACUGAGGACCACGUG...

A single string from 4 character alphabet
Protein

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

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-Letters</th>
<th>1-Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>
Protein

A single string from a 20 character alphabet

... DTIGDWNPSFFGIQLV SSVHTTLYWRENAPFV VDLG ...
<table>
<thead>
<tr>
<th>Molecule</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>String from a 4 character alphabet ({A,C,G,T})</td>
</tr>
<tr>
<td>RNA</td>
<td>String from a 4 character alphabet ({A,C,G,U})</td>
</tr>
</tbody>
</table>
### Molecule Representation Function

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Representation</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>String from a 4 character alphabet ({A,C,G,T})</td>
<td>Information storage</td>
</tr>
<tr>
<td>RNA</td>
<td>String from a 4 character alphabet ({A,C,G,U})</td>
<td>Old: Messenger, New: Many</td>
</tr>
</tbody>
</table>

**Central Dogma:** DNA makes RNA makes Protein

The process by which cells “read” the genome
DNA $\rightarrow$ RNA $\rightarrow$ Protein (The Central Dogma)

DNA: ATGATCTCGTAAA

Transcription:

RNA: ATGATCTCGTAAA

Translation:

Transcript (RNA): AUGAUCUCGUAA

Met - Ile - Ser - Polypeptide

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Representation</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>String from a 4 character alphabet ({A,C,G,T})</td>
<td>Information storage</td>
</tr>
<tr>
<td>RNA</td>
<td>String from a 4 character alphabet ({A,C,G,U})</td>
<td>Old: Messenger, New: Many</td>
</tr>
</tbody>
</table>

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 $D = d_1d_2d_3…d_n$

**Output:** The corresponding RNA sequence $R = r_1r_2r_3…r_n$ 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)
**Question**: How do we compare two genomes?

- **Human Genome**: ...ACTCGACTGAGAGGATTTTCGAGCATGA...
  ≈3.2 x 10^9 bp

- **Mouse Genome**: ...ACTCAACTGAGATTCGAGCTTCAATGA...
  ≈2.8 x 10^9 bp

**Comparison**:

- Human Genome: ≈3.2 x 10^9 bp
- Mouse Genome: ≈2.8 x 10^9 bp

**Note**: The comparison is illustrative and does not reflect the exact number of bases.
**Topic:** Sequence Alignment

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

Human Gene:
ACTCGACTGAGAGGATTTCGAGCATGA
≈10,000 – 15,000 bp

Mouse Genome:
ACTCAACTGAGATTCGAGCTTCAATG
≈10,000 – 15,000 bp
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 subgroups where items in a group are similar to each other, but different from items in other groups?
**Phylogenetics**

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

**Question:** Can we recover how a tumor has evolved over time?
Question: Not all DNA “codes” for proteins (is a gene). How do you find the portion of DNA that is a gene?