CSE 427
Computational Biology

http://courses.cs.washington.edu/courses/cse427

Larry Ruzzo
Autumn 2021
He who asks is a fool for five minutes, but he who does not ask remains a fool forever.

-- Chinese Proverb
Today

Admin

Why Comp Bio?

The world’s shortest Intro. to Mol. Bio.
Admin Stuff
Lecture: CSE2 G10, TuTh 11:30-12:50

Instructor: Larry Ruzzo, ruzzo@cs TBA
TA: Cailin Winston, cailinw@cs TBA
TA: Zoey Shi, shiz27@cs TBA

Office Hours Location: TBA

Course Email: cse427a_au21@uw.edu

Course Description:
Algorithmic and analytic techniques underlying analysis of large-scale biological data, such as DNA, RNA, and protein sequences or structures, expression and proteomic profiling. Hands-on experience with databases, analysis tools, and genome markers. Applications such as sequence alignment, BLAST, phylogenetics, and Markov models.

Prerequisites: CSE 312; CSE 332
Credits: 3

Learning Objectives: The availability of the complete genome sequences of humans and other organisms is one of the landmark achievements of science. Understanding this enormous volume of data is a problem that will challenge scientists for decades to come, and the nature and scope of the problem means that computer scientists will play a vital role. The primary objective of the course is for students to understand the variety of computational problems and solutions that arise in this interdisciplinary field. Students will
Course Mechanics & Grading

Hybrid in-class/livestream format
Web:

http://courses.cs.washington.edu/courses/cse427/21au

Reading
In class discussion
Homeworks: paper exercises & programming
No exams: maybe oversized last homework in lieu of final
Background & Motivation
Moore’s Law

Transistor count doubles approx every two years
Growth of GenBank (Base Pairs)


Excludes “short-read archive”
44.7 peta-bases

Short Read Archive Growth
Modern DNA Sequencing

A box the size of a double oven (but costs a bit more … ;-) can generate
\(~3 \times 10^{12}\) BP of DNA seq/day; i.e.,

1\textsuperscript{st} 30 yrs of genbank

1000 x your genome
Big Data: Astronomical or Genomical?


Table 1. Four domains of Big Data in 2025.

In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

<table>
<thead>
<tr>
<th>Data Phase</th>
<th>Astronomy</th>
<th>Twitter</th>
<th>YouTube</th>
<th>Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acquisition</strong></td>
<td>25 zetta-bytes/year</td>
<td>0.5–15 billion tweets/year</td>
<td>500–900 million hours/year</td>
<td>1 zetta-bases/year</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>1 EB/year</td>
<td>1–17 PB/year</td>
<td>1–2 EB/year</td>
<td>2–40 EB/year</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td>In situ data reduction</td>
<td>Topic and sentiment mining</td>
<td>Limited requirements</td>
<td>Heterogeneous data and analysis</td>
</tr>
<tr>
<td>Real-time processing</td>
<td>Metadata analysis</td>
<td>Variant calling, ~2 trillion CPU hours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Massive volumes</td>
<td></td>
<td></td>
<td>All-pairs genome alignments, ~10,000 trillion CPU hours</td>
<td></td>
</tr>
<tr>
<td><strong>Distribution</strong></td>
<td>Dedicated lines from antennae to server (600 TB/s)</td>
<td>Small units of distribution</td>
<td>Major component of modern user’s bandwidth (10 MB/s)</td>
<td>Many small (10 MB/s) and fewer massive (10 TB/s) data movements</td>
</tr>
</tbody>
</table>
The sea urchin *Strongylocentrotus purpuratus*
Goals

Basic biology
Drug discovery, validation & development
Disease diagnosis/prognosis/treatment
Individualized/precision medicine

…
“High-Throughput BioTech”

Sensors
- DNA / RNA sequencing
- Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction

Controls
- Cloning
- Gene knock out/knock in
- CRISPR

Floods of data

“Grand Challenge” problems
What’s all the fuss?

The human genome is “finished”…
Even if it were, that’s only the beginning
Explosive growth in biological data is revolutionizing biology & medicine

“All pre-genomic lab techniques are obsolete”
(and computation and mathematics are crucial to post-genomic analysis)
CS Points of Contact & Opportunities

Scientific visualization
  Gene expression patterns, development, immune response, ...

Databases
  Integration of complex, disparate, overlapping data sources
  Distributed genome annotation in face of shifting underlying genomic coordinates, individual variation, ...

AI/NLP/Text Mining
  Information extraction from text with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models, ...

Machine learning
  System level synthesis of cell behavior from low-level heterogeneous data (DNA seq, gene expression, protein interaction, mass spec, …)

Algorithms
  ...

...
Computers in biology: Then & now

Sequence alignment by word processor

D. Ross Boswell

Department of Haematological Medicine, University of Cambridge School of Clinical Medicine, Addenbrooke’s Hospital, Hills Road, Cambridge CB2 2QH, UK

ACGGGTAA
ACGGGTAA
More Admin
Course Focus & Goals

Mainly sequence analysis
Algorithms for alignment, search, & discovery
  Specific sequences, general types (“genes”, etc.)
  Single sequence and comparative analysis
Techniques: HMMs, EM, MLE, Gibbs, Viterbi…
Enough bio to motivate these problems
  including very light intro to modern biotech supporting them
Math/stats/cs underpinnings thereof
Applied to real data
Why Take This Course?

IT and Genomics are, and probably will remain, the 2 most explosively transformative technologies of your lifetimes.

Even if you don’t choose to work at that interface, having some knowledge of it will be valuable.

Hopefully, you will learn useful alg, ML, stats techniques and ideas for how to apply them in novel domains.
A VERY Quick Intro To Molecular Biology
The Genome

The hereditary info present in every cell
DNA molecule -- a long sequence of \textit{nucleotides} (A, C, T, G)
Human genome -- about $3 \times 10^9$ nucleotides
The genome project -- extract & interpret genomic information, apply to genetics of disease, better understand evolution, …
The Double Helix

As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b) is an uncoiled fragment of a polymer made up of four re- called deoxyribonucleotides.
DNA

Discovered 1869
Role as carrier of genetic information – 1940’s
4 “bases”:
  adenine (A), cytosine (C), guanine (G), thymine (T)
The Double Helix - Watson & Crick (& Franklin) 1953
Complementarity
  A ↔ T  C ↔ G

Visualization:
  http://www.rcsb.org/pdb/explore.do?structureId=123D
A gene -- classically, an abstract heritable attribute existing in variant forms (*alleles*)

ABO blood type—1 gene, 3 alleles

Mendel

Each individual has two copies of each gene
Each parent contributes one (randomly)
Independent assortment (approx, but useful)

*Genotype vs phenotype*

I.e., genes vs their outward manifestation
AA or AO genotype →“type A” phenotype
Cells

Chemicals inside a sac - a fatty layer called the *plasma membrane*

*Prokaryotes* (bacteria, archaea) - little recognizable substructure

*Eukaryotes* (all multicellular organisms, and many single celled ones, like yeast) - genetic material in nucleus, other organelles for other specialized functions, like photosynthesis
Chromosomes

1 pair of (complementary) DNA molecules (+ protein wrapper)

Most prokaryotes: just 1 chromosome

Most Eukaryotes - all cells have same number of chromosomes, e.g. fruit flies 8, humans & bats 46, rhinoceros 84, …
Mitosis/Meiosis

Most eukaryotes are diploid - have homologous pairs of chromosomes, one maternal, other paternal (exception: sex chromosomes)

*Mitosis* - cell division, duplicate each chromosome, 1 copy to each daughter cell

*Meiosis* - 2 specialized divisions form 4 haploid gametes (egg/sperm)

Recombination/crossover -- exchange maternal/paternal segments
Proteins

Chain of amino acids, of 20 kinds

Proteins: the major functional elements in cells

- Structural/mechanical
- Enzymes (catalyze chemical reactions)
- Receptors (for hormones, other signaling molecules, odorants, …)
- Transcription factors

... 

3-D Structure is crucial: the protein folding problem
The “Central Dogma”

Genes encode proteins
DNA transcribed into messenger RNA
mRNA translated into proteins
Triplet code (codons)
Transcription: DNA → RNA

RNA polymerase
## Codons & The Genetic Code

<table>
<thead>
<tr>
<th>First Base</th>
<th>Second Base</th>
<th>Third Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td></td>
<td>Phe</td>
<td>Tyr</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Stop</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Stop</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>His</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>His</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Gln</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Gln</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td>Ile</td>
<td>Lys</td>
</tr>
<tr>
<td></td>
<td>Met/Start</td>
<td>Thr</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
</tr>
</tbody>
</table>

### Codon Correspondences
- **Ala**: Alanine
- **Arg**: Arginine
- **Asn**: Asparagine
- **Asp**: Aspartic acid
- **Cys**: Cysteine
- **Gln**: Glutamine
- **Glu**: Glutamic acid
- **Gly**: Glycine
- **His**: Histidine
- **Ile**: Isoleucine
- **Leu**: Leucine
- **Lys**: Lysine
- **Met**: Methionine
- **Phe**: Phenylalanine
- **Pro**: Proline
- **Ser**: Serine
- **Thr**: Threonine
- **Trp**: Tryptophane
- **Tyr**: Tyrosine
- **Val**: Valine
Translation: mRNA $\rightarrow$ Protein
Ribosomes

Watson, Gilman, Witkowski, & Zoller, 1992
Gene Structure

mRNA built 5’ to 3’
Promoter region and transcription factor binding sites (usually) precede 5’ end
Transcribed region includes 5’ and 3’ untranslated regions
In eukaryotes, most genes also include *introns*, spliced out before export from nucleus, hence before translation
## Genome Sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Bases</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-CoV-2</td>
<td>29,903</td>
<td>12</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>580,073</td>
<td>483</td>
</tr>
<tr>
<td>Pandora Virus</td>
<td>2,900,000</td>
<td>2,500</td>
</tr>
<tr>
<td>E. coli</td>
<td>4,639,221</td>
<td>4,290</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>12,495,682</td>
<td>5,726</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>95,500,000</td>
<td>19,820</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>115,409,949</td>
<td>25,498</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>122,653,977</td>
<td>13,472</td>
</tr>
<tr>
<td>Humans</td>
<td>$3.3 \times 10^9$</td>
<td>~21,000</td>
</tr>
<tr>
<td>Amoeba dubia</td>
<td>~ 200 x human</td>
<td></td>
</tr>
</tbody>
</table>
DNA content (picograms)

Genome Surprises

Humans have < 1/3 as many genes as expected

But perhaps more proteins than expected, due to alternative splicing, alt start, alt end

Protein-wise, all mammals are just about the same

But more individual variation than expected

Many other non-coding regions are highly conserved, e.g., across all vertebrates

Subset of DNA being transcribed is >> 2% coding, giving many non-coding RNAs -- more than protein-coding genes, by some estimates

Complex, subtle “epigenetic” information
... and much more ...

Read one of the many intro surveys or books for much more info.
Bio Concept Summary

cells
DNA
base pairing
genome
replication, transcription, translation