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
CSE 427, Au '17: Computational Biology

Lecture: EEB 045
TuTh 1030-1150

Instructor: Larry Ruzzo, ruzzo@cs
Office Hours: TBA
Location: CSE 554
Phone: (206) 54

TA: Katie Doroschak, kdoroschak@cs
TA: Daniel Jones, djeones@cs
TA: David Wadden, dwaddene@cs

Course Email: cse427a.au17@uw.edu. Staff announcements will be posted here. You can post student/staff Q&A about homework, lectures, etc. The instructor and TAs are subscribed to this list. Enrolled students are as well. You should change their default subscription options. Messages are automatically archived.

Discussion Board: Also feel free to use Catalyst to ask and answer homework, etc.

Catalog Description: Algorithmic and analytical underpinnings of large-scale biological data sets such as DNA, RNA, and protein sequences or structures, expression and regulatory networks. Hands-on experience with databases, analysis tools, and genome markers. Applications such as sequence alignment, BLAST, and Markov models.

Prerequisites: CSE 312, CSE 377

Credits: 3

Learning Objectives:
- Understanding the complete genome sequences of humans and other organisms is one of the landmark achievements of science. Understand manipulation of such volume of data is a problem that will challenge scientists for decades to come, and the nature and scope of the problem means that scientists will play a vital role. The primary objective of the course is for students to understand the variety of computational problems that arise in this interdisciplinary field. Students will learn enough of the basic concepts of molecular biology to understand and solve the computational problems presented in the rest of the course. They will learn how some of the computational methods they have learned in other courses can be applied to solve problems in modern molecular biology. An important component is to learn the nature and structure of the key public databases available for the solution of these problems, as well as publicly available computational analysis tools and dynamic principles underlying them.
- Homework, possibly including a small project: 90%; class participation: 10%.

Books: None.

This page is protected by copyright. Portions of the CSE 427 Web may be reprinted or adapted for academic nonprofit purposes, providing the source is accurately quoted and duly credited. The CSE 427 Web © 1993-2017, the Authors and the Department of Computer Science and Engineering, University of Washington.
Course Mechanics & Grading

Web:
   https://courses.cs.washington.edu/courses/cse427

Reading
In class discussion
Homeworks
   paper exercises & programming
No exams, but probable 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”

13.3 peta-bases

Short Read Archive Growth

Modern DNA Sequencing

A table-top box the size of your oven (but costs a bit more … ;-) can generate ~100 billion BP of DNA seq/day; i.e. = 2008 genbank, = 30x your genome
PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens¹, Skylar Y. Lee¹, Faraz Faghri², Roy H. Campbell², Chengxiang Zhai³, Miles J. Efron⁴, Ravishankar Iyer¹, Michael C. Schatz⁵*, Saurabh Sinha³*, Gene E. Robinson⁵*
Fig 1. Growth of DNA sequencing.

http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195
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></td>
<td>Real-time processing</td>
<td>Metadata analysis</td>
<td></td>
<td>Variant calling, ~2 trillion CPU hours</td>
</tr>
<tr>
<td></td>
<td>Massive volumes</td>
<td></td>
<td></td>
<td>All-pairs genome alignments, ~10,000 trillion CPU hours</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>


http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195
The Human Genome Project

1  gagcccgccg ccaggacgggc gcggcgagat agcgggaccc cggcgcggcg gtgcgcttca
61  gggcgcagcg gcggccgcag accgagcccc gggcgcggca agaggccgcgg ggcagcggctg
121  gcggctcgcc atcatgctgct gagggcgtct gctgagagatc gcccctgggat ttagcggtct
181  tttaggcgtcc tacacgagcc atggggcggga cgcccaatgtt gaggctggga acgtgagaaga
241  aaccagagcc agtcggggcca agagaagagg cgggtggagga cagcgcggcgtt taaaagggacg
301  caatgtctgt ggtcagctgg aataagctta cttggtccct ggtggaaaa cctttacctgg
361  cggaaaaatcag tgttattgct ccatttgccg gcattcctgtt ggggatggat tttggttcgag
421  gccaaatatg tgcaactggtgc atctgttggc gataagctct tccgttgggt ccagattcctg
481  acaacaactgc aatattcgct gtatgaatgg aggtagctgc agtgacgatc actgtctatg
541  ccaagaaagga tacatagggg ctcattgtttg acaacctgttt ttgaaaggtg gctgctctgaa
601  tggagaggaag tgtgtgggcc caaattggatt gtagcatcact taccaggatta cttgacaacca
661  gtgtgtgaaaga gattacagga caggcccactg tttactgtg atccagactg aagatctgaca
721  gccgcactgc agcggcagttg tctgcacatag ccagctctggt cttgaccacac tggcgagaagc
781  ctggggccag caacctttggt tgtgtctgtgc ccctgtgagg cctggtccgc gttgccctgat
841  tccaaatatag cgcagactgc actttcgggg tttgtacgta atgctgttcgct ccagacgcaag
901  ctggcaggag ggaattgctt ataatgctta tgggtctttt gactgtgcaat gctgctgggttgc
961  acacaactctt aatgaagtgt cacaataatg tgaagatatt gatgaatgca gcaccattccc
1021 ...
The sea urchin *Strongylocentrotus purpuratus*
Goals

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

“High-Throughput BioTech”

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

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

Flooding of data "Grand Challenge" problems
What’s all the fuss?

The human genome is “finished”…
But that’s only the beginning
Explosive growth in 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

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
More Admin
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.
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
A *VERY* Quick Intro To Molecular Biology
The Genome

The hereditary info present in every cell
DNA molecule -- a long sequence of 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
Genetics - the study of heredity

A *gene* – classically, an abstract heritable attribute existing in variant forms (*alleles*)
  ABO blood type – 1 gene, 3 alleles

**Mendel**
  Each individual 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
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 “higher” 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 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 $\rightarrow$ RNA

RNA polymerase

DNA sense strand

DNA antisense strand

RNA strand

5' $\rightarrow$ 3'

3' $\rightarrow$ 5'
# Codons & The Genetic Code

<table>
<thead>
<tr>
<th>First Base</th>
<th>Second Base</th>
<th>Third Base</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Met/Start</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
</tr>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
</tr>
</tbody>
</table>

- Ala : Alanine
- Arg : Arginine
- Asn : Asparagine
- Asp : Aspartic acid
- Cys : Cysteine
- 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 → Protein

Watson, Gilman, Witkowski, & Zoller, 1992
Ribosomes

Watson, Gilman, Witkowski, & Zoller, 1992
## Genome Sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Base Pairs</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<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>~20,000</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 unexpectedly many proteins, due to *alternative processing*
Protein-wise, all mammals are just about the same
But more individual variation than expected
And many more *non-coding* RNAs -- more than protein-coding genes, by some estimates
Many other non-coding regions are highly conserved, e.g., across all vertebrates
Subset of DNA being transcribed is » 2% coding
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