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

-- Chinese Proverb
CSE 427, Wi '14: Computational Biology

Lecture: EEB 045 (seminars)  TuTh 12:00-1:20

Office Hours  Location  Phone
Instructor: Larry Ruzzo, ruzzo@cs  TBA  CSE 554 (206)
TA : Scott Lundberg, slund@cs  TBA  CSE 2xx

Course Email: cse427a_wi14@uw.edu. Staff announcements are posted to this list. Enrolled students are automatically subscribed. There is a student/staff Q&A about homework, lectures, etc. The instructor and TA are subscribed to this list. Enrolled students are automatically archived.

Discussion Board: Also feel free to use Catalyst. Check here for homework, etc.

Catalog Description: Algorithmic and analytical reasoning; a critical understanding of the underlying analysis of large-scale biological data sets such as DNA, RNA, and protein sequences or structures, expression and prediction of function. Hands-on experience with databases, analysis tools, and genome markers. Applications such as sequence alignment, BLAST, motif finding, and Markov models.

Prerequisites: CSE 312; CSE 337

Credits: 3

Learning Objectives: The ability to interpret and communicate large amounts of data such as the complete genome sequences of humans and other organisms is one of the landmark achievements of science. Understanding the vast 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 methods and treatments that arise in this interdisciplinary field. Students will learn enough of the basic concepts of molecular biology to understand the solutions to the computational problems presented in the rest of the course. They will learn how some of the computational methods they have studied in other courses can be applied to solve problems in modern molecular biology. An important component is to learn the nature and uses of some of the key public databases available for the solution of these problems, as well as publicly available computational resources and the algorithmic principles underlying them.

Grades: None.
Today

Admin

Why Comp Bio?

The world’s shortest Intro. to Mol. Bio.
Admin Stuff
Course Mechanics & Grading

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

Reading
In class discussion

Homeworks
  paper exercises & programming

No exams, but possible oversized last homework in lieu of final
Background & Motivation
Moore’s Law

Transistor count doubles approx every two years
Feature Size

0.7x every 2 years

45nm, 32nm, 22nm
Growth of GenBank (Base Pairs)

Excludes “short-read archive,” > 7 terabases by mid-2009 > 1 petabase by early 2013

SRA database growth

1,276,911,138,864,883 total bases
568,771,597,397,062 open access bases
Cost per Genome

National Human Genome Research Institute

genome.gov/sequencingcosts
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
The Human Genome Project

```
1  gagcccggcc  cgggggacgg  gcggcgggat  agcgggacc  cggcgcggcg  gtgcgcttca
61  gggcgtagcg  gcggccgcag  accgagcccc  gggcgcggca  agaggccgagc  gggagccggtg
121  gcggctcgcc  atcatgctcg  gagggcgctct  gctggagatc  gccctgggat  ttaccctgtct
181  tttagcgctcc  tacacgagcc  atggggcgga  cgccaaatgct  gaggctggga  acgtgaagga
241  aaccagagcc  agtcgggcca  agagaagagg  cggtgagagga  cacgcaccgc  ttgaaggacc
301  caatgtctgt  ggatcagctt  ataatgctta  ctgctggcct  ggtatggaa  ccttaaccttgg
361  cggaaatcag  tgtattgtct  ccattttcgg  gcatttcctg  gggatggat  tttgttctgg
421  gccaaatatg  tgcacttgcc  catctgccat  gatagctcct  tcctgtggct  ccagatccat
481  acaacactgc  aatattcgct  gtatgaatgg  aggtagctgc  agtgacgact  actgtctatg
541  ccagaaagga  tacatatagg  tctcactctg  acaacctgtt  tgtgaagagt  gtgtctctaa
601  tggaggaagg  tgtgtggccc  caaatcgatg  tgcactgcat  tacggatctt  ctggacccca
661  gtgtgaaaga  gattacagga  caggcccatg  tttactgtg  atcagcacc  agatgtgcca
721  gggacaacctc  agcgggtattg  tctgcacaaa  acagctctgc  tgtgccacag  tgcggccagagc
781  ctggggccac  cccctctgtg  tggctctgtg  ccacgttcac  ccctgcccgc  gtggctctcat
841  tccaaatatcc  gcacgctggag  tgtgtcaaca  acgctctgc  tgtgacccag  tggcacaagt
901  ctgtcagggga  gagataattgca  ttaatactgt  tgggcttttt  gaggctgaat  gcccctgctgg
961  acaacaaacctt  aatgaagttgt  cacaataattg  tgaagatattg  gatgaatgca  gcaccacattcc
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

Floode 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

Databases
  Integration of 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, …)

...
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 2QJ, UK
An Algorithm Example: ncRNAs

The “Central Dogma”:
DNA -> messenger RNA -> Protein

Last ~5 years:
100s – 1000s of examples of functionally important ncRNAs

Much harder to find than protein-coding genes

Main method - Covariance Models
≈ stochastic context free grammars

Main problem - Sloooow
$O(nm^4)$
“Rigorous Filtering” - Z. Weinberg

Convert CM to HMM
(aka: stochastic CFG to stochastic regular grammar)

Do it so HMM score always ≥ CM score

Optimize for most aggressive filtering subject to constraint that score bound maintained

A large convex optimization problem

Filter genome sequence with (fast) HMM, run (slow) CM only on sequences above desired CM threshold; guaranteed not to miss anything

Newer, more elaborate techniques pulling in key secondary structure features for better searching
(uses automata theory, dynamic programming, Dijkstra, more optimization stuff, …)

Details censored

But stay tuned...

Plenty of CS here
Results

Typically 200-fold speedup or more
Finding dozens to hundreds of new ncRNA genes in many families
The computational advance has enabled new biological discoveries

Newer, more elaborate techniques pulling in key secondary structure features for better searching (uses automata theory, dynamic programming, Dijkstra, more optimization stuff, …)
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
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 - much later
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
# 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>U</td>
<td>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td>U</td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>U</td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td>A</td>
<td>Ile</td>
<td>Thr</td>
</tr>
<tr>
<td>A</td>
<td>Met/Start</td>
<td>Thr</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
</tr>
</tbody>
</table>

- Ala : Alanine
- Arg : Arginine
- Asn : Asparagine
- Asp : Aspartic acid
- Cys : Cysteine
- Glu : Glutamic acid
- Flu : Glutamine
- 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>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>~25,000</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
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