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

-- Chinese Proverb
Tonight

Admin

Why Comp Bio?

The world’s shortest Intro. to Mol. Bio.
Admin Stuff
Homework 0

Please do this ASAP
Course Mechanics & Grading

Web
http://courses.cs.washington.edu/courses/csep527/20au

Reading
In class discussion

Homeworks
reading blogs
paper exercises
programming

No exams, but possible oversized last homework in lieu of final

Check web for 1st, soon
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 \( \sim 3 \times 10^{12} \) 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>Acquisition</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>Storage</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>Analysis</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>Variant calling, ~2 trillion CPU hours</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Massive volumes</td>
<td></td>
<td>All-pairs genome alignments, ~10,000 trillion CPU hours</td>
<td></td>
</tr>
<tr>
<td>Distribution</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 Human Genome Project

1  gagcccgagcc cgggggacgg gcggcgggat agcgggacc cgggcccggcg gtcgctttca
61  ggggcacagc gcggcccgag ccggagcccc ggggcccccg aggggccgca agagccggcg gggcgcggtg
121  gcccgtcgcct atcatgcgtc gagaggctct gctggagatc gccctgggat ttaccgtgcct
181  ttttagcgtcc tacomcagcc atggggcgga cgcacatggg gaggcctggga acgtgaggag
241  aaccagagcc agtccgggca agagaagagg ccggtgagga cacgcgcgcgc taaaagggcc
301  caatgtcgtgt gkatcagcct ataatatgctt ctggtgcctt gataggaa cccctacctgg
361  cggaaaacctg tgtattgtcc cccatttgccgc catttctgtt ggggatggat tgttttccag
421  cgcacagctgcc tcatctgtca gatagctcct tcccgttgct ctacagatccat
481  acacacagtgc aatattcgct gtatgaatgg aggtagctgc agtgacgatc actgtctattg
541  ccagacatggg tacatagactt ctacgtgtgg aacactggtg tggtaaggt gctgtcctcag
601  tggaggaagg ttttgtggccc caaatcgcftc tgcactgcct tgcagatg gcggtgacccg
661  gtgtgacagtga gattacagag cgggcccactg ttttactgtg atcagacacc agatgtgcca
721  gggacaactc agcgggattg tctgcacaaa acagctctgc tgtgccacag tcggccggcgc
781  ctggggccac cccatgtgaga gggattgct ctggttcctgg ctgggtcacg gcgaccctgcgt
c91  tccacaatattc gcgcacggag ctgtgacgcag tggtaggaag tgcgcagcgc tttccbgccgt
961  acacacacttc aatgaagtgt cacaaaaatg tgaagatatt gatgaatgca gcaccattccc
1021  ...

...
The sea urchin *Strongylocentrotus purpuratus*
EndoMes model
Feb 19 2002

Endomes up to 20-24 hrs

Endo

Mes

Endoderm

Mesoderm

Veg1 endoderm

Endomes model
Feb 19 2002

Endomes up to 20-24 hrs

Endo

Mes

Endoderm

Mesoderm

Veg1 endoderm

16
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

Floords 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 2QQ, UK
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 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, …
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-
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
DNA, RNA, 3’, 5’, …

[Chemical structures of Ribose and Deoxyribose]

https://en.wikipedia.org/wiki/Ribose
Nucleotides

https://en.wikipedia.org/wiki/Nucleotide
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 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
Chromosomes

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

Most prokaryotes: just 1 chromosome

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 $\rightarrow$ RNA
# 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>Phe</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Ser</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</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>Thr</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>Ala</td>
</tr>
<tr>
<td>Arg</td>
<td>Arg</td>
</tr>
<tr>
<td>Asn</td>
<td>Asn</td>
</tr>
<tr>
<td>Asp</td>
<td>Asp</td>
</tr>
<tr>
<td>Cys</td>
<td>Cys</td>
</tr>
<tr>
<td>Glu</td>
<td>Glu</td>
</tr>
<tr>
<td>Gly</td>
<td>Gly</td>
</tr>
<tr>
<td>His</td>
<td>His</td>
</tr>
<tr>
<td>Ile</td>
<td>Ile</td>
</tr>
<tr>
<td>Leu</td>
<td>Leu</td>
</tr>
<tr>
<td>Lys</td>
<td>Lys</td>
</tr>
<tr>
<td>Met</td>
<td>Met</td>
</tr>
<tr>
<td>Phe</td>
<td>Phe</td>
</tr>
<tr>
<td>Pro</td>
<td>Pro</td>
</tr>
<tr>
<td>Ser</td>
<td>Ser</td>
</tr>
<tr>
<td>Thr</td>
<td>Thr</td>
</tr>
<tr>
<td>Trp</td>
<td>Trp</td>
</tr>
<tr>
<td>Tyr</td>
<td>Tyr</td>
</tr>
<tr>
<td>Val</td>
<td>Val</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
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 x 10⁹</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.
Homework #0, part 2

Meet your professor!

I’d like to schedule a 5-10 minute zoom with each of you over the next few days. Just chat, no nefarious agenda, ungraded.

Sign up via Google Doc linked from class web page.
Homework #1 (summary)

Read Hunter’s “bio for cs” primer;
Find & read another
Post a few sentences saying
  What you read (give me a link or citation)
  Critique it for your meeting your needs
  Who would it have been good for, if not you
See class web (coming soon) for full details
Bio Concept Summary

cells
DNA
base pairing
genome
replication, transcription, translation