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
Course Mechanics & Grading

Reading
In class discussion
Lecture scribes
Homeworks
reading
paper exercises
programming
Project
No exams

Background & Motivation

Source: http://www.intel.com/research/silicon/mooreslaw.htm

The Human Genome Project

```plaintext
1 gagccggccc gggggggcgg ccgggagcgc ggccggcggc ggcggggggt cgggtgcttca
gggccggcc cgggggacgg gcggcgggat agcgggaccc cggcgcggcg gtgcgcttca
61 ggccgggagc ggcccggcag atcagctgct ggggggctt gctgggagct acgtgctggc
121 cggccggccc cggggggcac cggccgggac ggcggggcgg gcggcgggat gggctgcttca
gacggggggt cgggggggtt cgggtgcttca
gtgcgcttca cgggtgcttca
181 tgtggcctgc tcaagcgagc atggggcggc cggccgggac ggcggggggt cgggtgcttca
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
121 cggccggccc cggggggcac cggccgggac ggcggggcgg gcggcgggat gggctgcttca
gacggggggt cgggggggtt cgggtgcttca
gtgcgcttca cgggtgcttca
61 ggccgggagc ggcccggcag atcagctgct ggggggctt gctgggagct acgtgctggc
181 tgtggcctgc tcaagcgagc atggggcggc cggccgggac ggcggggggt cgggtgcttca
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
cgggggacgg gcggcgggat agcgggaccc cggcgcggcg
gggccggcc
```

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

"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
Databases
- Integration of disparate, overlapping data sources
- Distributed genome annotation in face of shifting underlying genomic coordinates
AI/NLP/Text Mining
Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…

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

Algorithms

An Algorithm Example: ncRNAs

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

Last ~5 years: many examples of functionally important ncRNAs
- 175 -> 350 families just in last 6 mo.
Much harder to find than protein-coding genes

Main method - Covariance Models (based on stochastic context free grammars)
Main problem - Slooow … 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 CM score subject to constraint that score bound maintained
A large convex optimization problem
Filter genome sequence (run 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
Has enabled discovery of many new families

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

Course Focus & Goals

Sequence analysis, maybe some microarrays
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

More Admin

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

DNA

Discovered 1869
Role as carrier of genetic information - much later
The Double Helix - Watson & Crick 1953
Complementarity
A $\leftrightarrow$ T, C $\leftrightarrow$ G

Visualizations:
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)
Genotype vs phenotype
Mendel
Each individual two copies of each gene
Each parent contributes one (randomly)
Independent assortment
**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 have just 1 chromosome

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 → RNA

Codons & The Genetic Code

Translation: mRNA → Protein

Watson, Gilman, Witkowski, & Zoller, 1992
Ribosomes

Gene Structure

Transcribed 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>MimiVirus</td>
<td>1,200,000</td>
<td>1,260</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 10^8</td>
<td>~25,000</td>
</tr>
</tbody>
</table>

Genome Surprises

Humans have < 1/3 as many genes as expected
But perhaps more proteins than expected, due to *alternative splicing, alt start, alt polyA*
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
90% of DNA is transcribed (< 2% coding)
Complex, subtle “epigenetic” information
… and much more …
Read one of the many intro surveys or books for much more info.

Homework #1 (partial)
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 for more details