CSCI2950-C
Genomes, Networks, and Cancer

Lecture 1: Sept. 10, 2009
http://cs.brown.edu/courses/csci2950-c/

Course Organization

- Seminar style
  1) Introductory lectures by me: 35-50% of class meetings.
  2) Student-led paper presentations and discussion. 50-65%
  3) Project. Groups of 1-3 students.
Course Organization

• Grading:
  – 30% paper reviews (~10)
  – 25% presentations (2-3)
  – 40% project.
    • 10% Written proposal.
    • 20% Final written report.
    • 10% Final presentation.
  – 5% participation

• CS requirements:
  PhD: Area B (Algorithms)
  ScM: Theory or Practice* *depending on project
  Significant programming*

Topics in Computational Biology

[Diagram of Genomes, Networks, Cancer]

**Computational biology**: develop computational techniques to solve biological problems. Ecology? Neuroscience? Botany?

**Computational molecular biology**: develop computational techniques to solve biological problems related to DNA, RNA, and/or proteins.
Computational Biology

Interdisciplinary
Biology
Computer Science
Mathematics
Statistics

Why choose just 1?

WSJ: Jan, 26, 2009.

Molecular Biology 101

Image adapted from: National Human Genome Research Institute.
Molecular Biology 101

Central Dogma

Three fundamental molecules.
1) DNA
   - Information storage.
2) RNA
   - Old view: messenger.
   - New view: Performs important functions.
3) Proteins
   - Perform most cellular functions (biochemistry, signaling, control, etc.)
Molecular Biology 101

Molecules of DNA, RNA, and protein are built from an “alphabet” of “characters.”

Computer Science!

DNA

Double helix
Two strands.

Each strand composed of sequence of covalently bonded nucleotides (bases).

Four possible nucleotides: A, C, T, G

DNA sequence: string from 4 character alphabet

..ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTGACTGAGACTGAGTTTA
TATATATATACGTCGTCGT
ACTGATGAACCTGAC
ACTGATTTAGATACCTGAC
TGATTTTTTTTTTTTTATT...
Computational Molecular Biology

For DNA, RNA, protein:

1. Measurement/Technology.
2. Comparison.
3. Prediction.

Measurement

Derive the DNA sequence of a human.

Which proteins are present in a cell at a given time?

No technology to answer such questions exactly.

Technology ↔ Algorithms
What molecules can we measure?

- Sequencing (expensive, high quality)
  Hybridization (inexpensive, lower quality)
- Sequencing (expensive, high quality)
  Hybridization (noisy, lower quality)
- Mass spectrometry (low quality)
  Hybridization (very low quality)

Computational Molecular Biology

For DNA, RNA, protein:

Comparison

What are the similarities/difference between the DNA sequences of human and chimpanzee?

[String comparison ↔ Algorithms]
Computational Molecular Biology

For DNA, RNA, protein:

Prediction
- How does the sequence of RNA or protein determine its 3D structure?
- Does a protein bind to DNA? To what sequence?
- Do two proteins bind/interact?

Course Topics

- Genomes
- Algorithms
- Machine Learning
- Data Mining/Analysis
- Networks
- Systems/Programming
- Cancer
Course Topics

- Genomes
  - The DNA sequence* of an individual.
- Networks
- Cancer

* diploid vs. haploid. Later…

DNA sequencing

How we obtain the sequence of nucleotides of a species

...ACGTGACTGAGGACCCTG
CGACTGAGACTGACTGAGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTGATTACAG
ACTGATTAGATAACCTGAC
TGATTAAAAAAATAATT...
DNA Sequencing

**Goal:**
Find the complete sequence of A, C, G, T's in DNA.
Human genome: String of 3 billion nucleotides.

**Challenge:**
There is no machine that takes long DNA as an input, and gives the complete sequence as output
Can only sequence ~500 letters at a time

Human Genome Sequenced
2000-2003, ???

*The New York Times*
“In the Genome Race, the Sequel Is Personal”
Shotgun Sequencing

Genomic segment

cut many times at random (shotgun)

Get one or two reads from each segment

~500 bp

Fragment Assembly

Cover region with ~7-fold redundancy

Overlap reads and extend to reconstruct the original genomic region
Human Genome Sequenced 2000-2003, ???

The New York Times
Sept. 3
“In the Genome Race, the Sequel Is Personal”

Now what?

1) Comparative Genomics
   • Genomes of other mammals
2) Disease genomics
   • HapMap
   • 1000 Genomes
3) Cancer genomics
   • The Cancer Genome Atlas
   • Tumor Sequencing Project

Comparative Genomics

Compare genome sequences of related species.

1. Sequence Alignment

   Sequences
   Evolve via substitutions
   Conservation suggests function

2. Phylogeny: defining ancestral relationships between species, genomes, genes.
Molecular Biology 101

Humans have 23 (pairs) of chromosomes.

History of Chromosome X

Comparative Genomic Architectures: Mouse vs Human Genome

- Humans and mice have similar genomes, but their genes are ordered differently
- ~245 rearrangements
  - Reversals
  - Fusions
  - Fissions
  - Translocation

Genome rearrangements

- What are the similarity blocks and how to find them?
- What is the architecture of the ancestral genome?
- What is the evolutionary scenario for transforming one genome into the other?
Reversals

Blocks represent conserved genes.

- In the course of evolution or in a clinical context, blocks 1,...,10 could be misread as 1, 2, 3, -8, -7, -6, -5, -4, 9, 10.
Reversals and Breakpoints

The reversion introduced two breakpoints (disruptions in order).

Sorting Permutations by Reversals
(Sankoff et al. 1990)

\[ \pi = \pi_1 \pi_2 \ldots \pi_n \] signed permutation

Reversal \( \rho(i,j) \) [inversion]

\[ \pi_1 \ldots \pi_{i-1} \pi_i \ldots \pi_j \pi_{i+1} \ldots \pi_n \]

**Problem:** Given \( \pi \), find a sequence of reversals \( \rho_1, \ldots, \rho_t \) with such that:

\[ \pi \cdot \rho_1 \cdot \rho_2 \ldots \rho_t = (1, 2, \ldots, n) \] and \( t \) is minimal.

**Solution:** Analysis of breakpoint graph

Polynomial time algorithms

- \( O(n^4) \) : Hannenhalli and Pevzner, 1995.
- \( O(n^2) \) : Kaplan, Shamir, Tarjan, 1997.
- \( O(n) \) [distance \( f \)] : Bader, Moret, and Yan, 2001.
- \( O(n^2) \) : Bergeron, 2001.
Course Topics

Graph algorithms:
- Eulerian and Hamiltonian paths.
- Cycle decompositions

Probability
- Maximum likelihood
- Markov chain Monte Carlo

* diploid vs. haploid. Later…

Course Topics

Network Alignment
Network Integration
Biological Interaction Networks

Many types:
• Protein-DNA (regulatory)
• Protein-metabolite (metabolic)
• Protein-protein (signaling)
• RNA-RNA (regulatory)
• Genetic interactions (gene knockouts)

Regulatory Networks
Cis-regulatory Network

Metabolic Networks

Nodes = reactants
Edges = reactions
labeled by enzyme (protein)
that catalyzes reaction
Protein-Protein Interaction (PPI) Network

- Proteins are nodes
- Interactions are edges
- Edges may have weights

Yeast PPI network
Computational Problems

1. Classifying Network Topology
   – Finding paths, cliques, dense subnetworks, etc.
2. Comparing Networks Across Species
3. Using networks to explain data
   – Dependencies revealed by network topology
4. Modeling dynamics of networks

Network Motifs

Subnetworks with more occurrences than expected by chance.

• How to find?
• How to assess statistical significance?
Network Alignment

The Network Alignment Problem

Given: \( k \) different interaction networks belonging to different species,

Find: **Conserved sub-networks** within these networks

Conserved defined by protein sequence similarity (node similarity) and interaction similarity (network topology similarity)
Protein Signaling Networks

Use machine learning methods (Bayesian networks, etc.) to derive network structure.

Course Topics

Graph algorithms:
- Path and clique finding.
- Graph isomorphism.
- Network flow.
- Finding heavy/dense.
- Random walks.

Statistics:
- Clustering.
- Hypothesis testing.
- Permutation tests.
- Enrichment.

Genomes
Networks
Cancer
Course Topics

Genomes

Networks

Cancer

Cancer Genomes
Cancer & Networks

Molecular Biology 101

Humans have 23 (pairs) of chromosomes.
Cell Division and Mutation

Cancer Genomes

Leukemia

Breast
Rearrangements in Tumors
Change gene structure, create novel fusion genes

• Gleevec (Novartis 2001) targets ABL-BCR fusion

Challenges

1) What is detailed organization of cancer genomes?
2) What genes affected?
3) What sequence of rearrangements produced this genome?
4) Can we create custom treatments for tumors based on mutational spectrum? (e.g. Gleevec)
The Cancer Genome Atlas (TCGA) Pilot Project

Charting a new course for prevention, diagnosis, and treatment of cancer

Tumor Genome → Phenotype
Gene Networks and Pathways

Integration of Multiple Data Sources
- ESP and copy number
- Mutation
- Expression (mRNA and miRNA)
- Binding (ChIP-chip)
- Pathways
- Epigenetics
Tumor Genome → Phenotype
Gene Networks and Pathways

Integration of Multiple Data Sources
• ESP and copy number
• Mutation
• Expression (mRNA and miRNA)
• Binding (ChIP-chip)
• Pathways
• Epigenetics

Significantly mutated pathways in lung adenocarcinomas.

Course Topics

Course goals

• Learn some cool applications of computer science techniques.
• Learn some new CS/mathematics.
• Read research papers.
  – Understand how to formulate a biological problem as a computational problem.
• Course project: Use your skills to solve a biological problem.
Reading

- Biology background.
- Probability background.
- See course website.