CSCI2950-C
Topics in Computational Biology

Ben Raphael
Lecture 1: Sept. 10, 2010
http://cs.brown.edu/courses/csci2950-c/

Course Organization

Seminar style
1) Introductory lectures by me: 30-50% of class meetings.
2) Student-led paper presentations and discussion. 50-70%
3) Project. Groups of 1-3 students.
Course Organization

• Grading:
  Papers: 40%
  20% paper reviews (~15)
  20% presentations (2–3)
Projects: 50% project.
  • 10% Written proposal.
  • 10% Midterm report
  • 20% Final written report.
  • 10% Final presentation.
Participation: 10%

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

Topics in Computational Biology

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.
Three fundamental molecules.

1) **DNA**
   - Information storage.

2) **RNA**
   - *Old view*: Mostly a "messenger".
   - *New view*: Performs many 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!

algorithms on strings, trees, graphs

Molecules of DNA, RNA, and protein interact with each other in time and space

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

...ACGTGACTGAGGACCCTG
CGACTGAGACTGACTGTTG
CTAAGCTGACCTAAGCTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTGATACCTGAC
TGATTTTAAAAATATT...

Computational Molecular Biology

For DNA, RNA, protein:

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

For DNA, RNA, protein:

Measurement
Derive the DNA sequence of a human.
Which RNA molecules or proteins are present in a cell at a given time?

No technology to answer such questions exactly.

Measurement Technology + Algorithms ➔ Biological knowledge
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 sequences?

e.g. from human and chimpanzee?

[String comparison ↔ Algorithms]
Computational Molecular Biology

For DNA, RNA, protein:

Prediction

What differences in DNA determine specific traits? [Genetics]
Which mutations in DNA cause cancer?

Does a protein bind to DNA? To what sequence?

How does the sequence of RNA or protein determine its 3D structure?

Do two proteins bind/interact? ...

Course Topics

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

- Learn applications of computer science techniques to biology.
- 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.

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

---

DNA Sequencing

**Goal:**

Find the complete sequence of A, C, G, T’s in a DNA molecule.

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 <1000 letters at a time
Human Genome Sequenced
2000-2003

Shotgun Sequencing

getomic segment

cut many times at random (shotgun)

~500 bp
~500 bp

Get one or two reads from each segment
Fragment Assembly

Cover region with ~N-fold redundancy
Overlap reads and extend to reconstruct the original genomic region

First Topic: Algorithms to solve fragment assembly problem

Progress in Genome Sequencing

Old technology
~500 bp
~25-100 bp

New technologies
~500 bp
~25-100 bp

Human Genome Project
“Next-generation” DNA sequencing

Illumina
454
ABI SOLiD
Progress in Genome Sequencing

"Third-generation" DNA Sequencing

Strobe sequencing: Multiple reads per fragment

Comparison Genomics

Compare genome sequences of related species.

1. Sequence Alignment

Sequences: 
- Human: EFTPPVQAAYQKVAG
- Mouse: DFNPNVQAAYQKVAG

Evolve via substitutions
Conservation suggests function

2. Phylogeny: defining ancestral relationships between species, genomes, genes.
Humans have 23 (pairs) of chromosomes.

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 / segments of DNA sequence.

- In the course of evolution or cancer development, a mutation occurs causing blocks 1,…,10 to be reordered as 1, 2, 3, -8, -7, -6, -5, -4, 9, 10.
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

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

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

**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
- \( O(n^2) \): Hannenhalli and Pevzner, 1995.
- \( O(n^2) \): Kaplan, Shamir, Tarjan, 1997.
- \( O(n) \) [distance \( t \)]: Bader, Moret, and Yan, 2001.
- \( O(n^3) \): Bergeron, 2001.
Course Topics

Biology
DNA sequences: Alignment, Assembly
Rearrangements of Genomes

Computational techniques
Graph algorithms:
Eulerian and Hamiltonian paths.
Cycle decompositions
Probability
Maximum likelihood
Markov chain Monte Carlo

* diploid vs. haploid. Later...

Course Topics

Genomes
Networks
Cancer

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

A gene regulatory network

- INPUT signal A
  - receptor proteins
  - cascade of interacting kinase proteins or other molecules
  - active transcription factor A
- INPUT signal B
  - receptor proteins
  - inactive transcription factor B
  - inhibitory factor
  - active transcription factor B

OUTPUT
- mRNA
  - target gene
  - RNA polymerase

OUTPUT protein
  - cell functions
  - feedbacks
Cis-regulatory Network

Nodes = reactants
Edges = reactions labeled by enzyme (protein) that catalyzes reaction

Metabolic Networks
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?

Shen-Orr et al. 2002
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 subgraphs.
- Random walks.

Statistics:
- Clustering.
- Hypothesis testing: permutation tests.
- Enrichment.
Course Topics

Molecular Biology 101

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

- DNA Replication
- Mitosis
- Two Diploid Cells

Copy number

Single nucleotide change

Deletion
Duplication
Inversion
Translocation

Structural

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 are affected?
3) What sequence of rearrangements produced this genome?
4) Can we create custom treatments for tumors based on mutational spectrum? (e.g. Gleevec)
Multiple genomic measurements (DNA sequence, mRNA levels, etc.) from thousands of cancer samples.

Cancer Genome → Phenotype
Gene Networks and Pathways

Integration of Multiple Data Sources
- DNA sequence and somatic mutations (SNVs, copy number, rearrangements)
- Expression (mRNA and miRNA)
- Binding (ChIP-chip)
- Pathways
- Epigenetics
Cancer Genome → Phenotype

Significantly mutated pathways in lung adenocarcinomas.

Networks and Cancer

The Cancer Genome Atlas:
Mutations in ~200 brain cancer patients (TCGA, Nature 2008).

Are these results surprising?
Are there other “interesting” pathways?
Problem

Given:
1. Large-scale interaction network.
2. Mutation data from multiple cancer samples.

Find: Subnetworks mutated in a significant number of samples.

Mutated subnetworks: HotNet*

Alterations $\rightarrow$ heat diffusion

Extract “significantly hot” subnetworks

Two-stage multi-hypothesis test

HotNet Results

Subset of 23 mutated subnetworks significantly mutated in ~300 ovarian cancer samples.

Course Topics
Reading

• Biology background.
• Probability background.
• See course website.