CSCI2950-C
Topics in Computational Biology
Lecture 1

Ben Raphael
September 12, 2011
http://cs.brown.edu/courses/csci2950-c/

Topics in Computational Biology

Biology

Computer Science
Mathematics & Statistics

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.
Technology Transforming Biology

Until late 20th Century

Hypothesis Generation and Validation

21st Century and Beyond

Algorithms

Hypothesis Generation and Validation

High throughput technologies

Human Genome Sequenced

2000-2003

June 26, 2000
Biology 101

Image adapted from: National Human Genome Research Institute.

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.)
DNA

Double helix
Two strands.

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

Four possible nucleotides: A, C, T, G

Bases (nucleotides) in each strand are complementary:
A \leftrightarrow C, \ T \leftrightarrow G
Watson-Crick base-pairing

DNA sequence: string from 4 character alphabet
RNA

• RNA: single stranded, 4 letter alphabet: A, C, U, G (U = T)
• Can fold into structures due to base complementarity.
• Many classes of RNA
  – mRNA (messenger: transcription
    • DNA → RNA → protein)
  – tRNA (transfer RNA: translation
    • mRNA → protein
  – miRNA
  – snoRNA
  – piwiRNA
  – Etc...

Protein

• String of amino acids: 20 letter alphabet
• Folds into 3D structures to perform various functions in cells
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

Computational Molecular Biology

For DNA, RNA, protein:

1. Measurement/Technology.
2. Comparison.
3. Prediction.
Computational Molecular Biology

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

Genome Sequencing and Comparison

DNA sequencing
… ACGTATTAGG …
Computational Molecular Biology

For DNA, RNA, protein:

Comparison
What are the similarities/difference between sequences?
e.g. from human and chimpanzee?
[String comparison ↔ Algorithms]

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?

...
Why Study Computational Biology?

Interdisciplinary Biology Computer Science Mathematics Statistics = FUN!

Why choose just 1?

Course goals

- Learn applications of computer science, mathematics, and statistics to biology.
- Learn some new CS/mathematics.
- Learn to read and critique research papers.
  - Understand how to formulate a biological problem as a computational problem.
- Undertake a research project
  - Course project: Use your skills to solve a biological problem.
- Write a proposal (NIH style)
Course Organization

Seminar style
1) Introductory lectures by me: 30-50% of class meetings.
2) Student-led paper presentations and discussion.
   50-70%
3) One computer assignment: next-gen sequencing data.
4) Project. Groups of 1-3 students.
   Proposal: NIH style
   Initial proposal
   Midterm report
   Final report
   Project presentation

Course Organization

See handout

• Prerequisites
• Grading
• Department Requirements
Topics in Computational Biology

Genomes

Networks

Cancer

Genome Sequencing and Comparison

Species

DNA sequencing

... ACGTATTAGG ...

Baseline information

Cost of genome sequencing compared with von Neumann's law for computers

Cost of computing (von Neumann's law)

$ per million DNA bases

Source: Broad Institute

"Next-generation" DNA sequencing
DNA Replication and Mutation

- DNA Replication
- Mitosis
- Two diploid cells
- Deletion
- Duplication
- Inversion
- Translocation

Species

Comparative genomics
Differences between species?

Genetic

Personal genomics
Genetic basis for traits of individuals?

Somatic

Cancer genomics
Which somatic mutations lead to cancer?
Course Topics

1. **Measurement** of genomes

- **Human genome:** ~3 billion letters
- **Next-generation DNA sequencing**
  - **Genome and Mutations**
  - 10-100’s million reads:
    - 30-1000 letters

---

Course Topics

- **Genome**
  - Millions -billions nucleotides
- **Next-generation DNA sequencing**
  - 10-100’s million reads
  - Reads: 30-1000 nucleotides

1. **“Resequencing”** algorithms (also RNA-Seq, ChIP-Seq, *-Seq)
2. **De novo assembly algorithms**
Course Topics

2. Interpretation of genomes

- Personal genomics: Genetic basis for traits of individuals?
- Cancer genetics: Which somatic mutations lead to cancer?

![Gene Algorithms: Identify functional variants](chart)

3. Beyond the genome: epigenetics

- Human genome: ~3 billion letters
- Next-generation DNA sequencing
- Algorithms

- Genome Modifications
  - DNA Methylation
  - Packaging in chromatin
Molecular Biology 101

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.)

What molecules can we measure?

- Sequencing (expensive, high quality)
- Hybridization (inexpensive, lower quality)
- Mass spectrometry (low quality)
- Hybridization (very low quality!)
DNA sequencing

How we obtain the sequence of nucleotides of a species

...ACGTGACTGAGGACCGTG
CGACTGAGACTGGACTG
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCG
ACTGATGACTAGATTACAG
ACTGATTAGATACCTGAC
TGATTTTTAAAAATATT...

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
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 ~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

ABI SOLiD

Baseline Information

<table>
<thead>
<tr>
<th>Cost of Genome Sequencing (compared to human's laws for computers)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
</tr>
</tbody>
</table>

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

ABI SOLiD

Baseline Information

<table>
<thead>
<tr>
<th>Cost of Genome Sequencing (compared to human's laws for computers)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
</tr>
</tbody>
</table>
“Third-generation” DNA Sequencing

**Strobe sequencing**: Multiple reads per fragment

![Diagram showing 3-strobe subread advancement](image)

**Single Molecule Real Time (SMRT)** Sequencing from Pacific Biosciences

Ritz, Bashir, and Raphael (2010). *Bioinformatics.*

---

**Course Topics**

2. **Interpretation** of genomes

- Personal genomics
  - Genetic basis for traits of individuals?

- Cancer genetics
  - Which somatic mutations lead to cancer?

**Identify functional variants**

---

- Algori...
Somatic Mutations and Cancer

Clonal Theory (Nowell 1976)

Founder cell

Passenger mutations

Driver mutation

Time (cell divisions)

Sequence genome

“typical tumor”: ~10 driver mutations
100’s – 1000’s of passenger mutations

Cancer Genomes

Leukemia

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

Gleevec (Novartis 2001) targets ABL-BCR fusion

Functional Mutations

Next-generation DNA sequencing

Recurrent mutations/mutated genes \(\rightarrow\) functional

91 glioblastoma multiforme samples
Recurrent Mutations

Cancer is a disease of “pathways”.

What pathways are altered/mutated?

[Hanahan and Weinberg, Cell 2000]

Biological Interaction Networks

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

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

Yeast PPI network


Problem

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

**Find:** Subnetworks mutated in a significant number of samples
Somatic Mutations and Cancer

Clonal Theory (Nowell 1976)

- Driver mutation
- Passenger mutations

"typical tumor": ~10 driver mutations
100’s – 1000’s of passenger mutations

Founder cell

Time (cell divisions)

Sequence genome

Given: cross-sectional data from many individuals.
Goal: Infer order of mutations.

Course Topics

2. Interpretation of genomes

Cases

Controls

Personal genomics

Genetic basis for traits of individuals?

Gene

gene

Algorithms

Identify functional variants
Course Topics

3. Beyond the genome: epigenetics

- Human genome: ~3 billion letters
- Next-generation DNA sequencing
- Algorithms
- Genome Modifications
- DNA Methylation
- Packaging in chromatin

Course Topics

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

- Biology background.
- Probability background.
- DNA sequencing technologies
- See course website.