Course Staff

Instructor:

- Mohammed El-Kibir (melkebir)
- Office hours: Tuesdays, 3:15-4:15pm

Developing combinatorial algorithms for studying all stages of cancer progression.
Course Information

Course website:
• www.el-kebir.net/teaching/cs598MEB
• Slides are password protected: p!neapplep3n

Piazza: (please sign up)
• https://piazza.com/illinois/spring2019/cs598meb

Description:
• This course focuses on recent algorithmic methods in cancer genomics, including somatic variant calling, phylogeny inference and identification of driver mutations. Students will study the underlying principles of these methods and the application of these methods to cancer genomics data.
Course Objectives

Learn:
• Learn underlying ideas of common algorithms in cancer genomics.
• Learn to translate a biological problem into a computational problem.
• Learn to read and critique scientific papers.
• Learn to propose and conduct independent research.
• Learn to present key ideas of a paper to other people.
• Learn to ask critical questions.

Not learn:
• Will not learn to run popular cancer genomics packages.
• Will not learn how to program.
Grading

• Class participation (20%)
  • Peer reviews
  • Asking questions

• Paper presentation (30%)

• Course project (50%)
  • Proposal
  • Report/paper
  • Presentation
Tentative Course Schedule

Introductory lectures (Jan 15 to Feb 28)
• Molecular biology and cancer biology
• Fundamental algorithms in computational biology
• Algorithms in computational genomics

Paper presentations (Mar 5 to Mar 29)
• Student presentation of research/survey paper

Course projects (Apr 2 to Apr 30)
• Proposal presentation
• Final presentation + report
Paper Presentation

• Each student will present a paper picked by the student. The goal of the presentation is to facilitate a discussion, focusing on:
  • Presenting the biological problem and corresponding computational problem
  • How did the authors solve the problem?
  • Did they manage to answer the original biological question?
  • How can we improve the results? What are future directions?

• The remaining students are required to answer questions that will help read the paper critically, as well as write a short peer review.
Course Project

• 1-2 students per project
• First write a proposal, which will receive feedback from instructor and fellow students
• Then, conduct research and write a paper
• Pick venue (conference/journal) and use LaTeX style for your paper
• Students will anonymously peer review submitted papers using EasyChair (if time permits)
Lecture Outline

• Primer on Molecular Biology
• Primer on Computational Biology
• Primer on Cancer Biology
• Tumor Phylogeny Inference

Reading

• “Biology for Computer Scientists” by Lawrence Hunter
  (http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)
Molecular Biology is the field of biology that studies the composition, structure and interactions of cellular molecules – such as nucleic acids and proteins – that carry out the biological processes essential for the cell's functions and maintenance.

<table>
<thead>
<tr>
<th>Cellular molecules:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DNA</td>
</tr>
<tr>
<td>2. RNA</td>
</tr>
<tr>
<td>3. Protein</td>
</tr>
</tbody>
</table>

https://www.nature.com/subjects/molecular-biology
DNA

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

**Four nucleotides:**
- A (adenine)
- C (cytosine)
- T (thymine)
- G (guanine)

**Watson-Crick base-pairing**
- A $\leftrightarrow$ T
- C $\leftrightarrow$ G
DNA

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

5’ …ACGTGACTGAGGACCGTG…
   …||||||||||||||||||…
   …TGCACTGACTCCTGGCAC…
3’

Pair of strings from 4 character alphabet

5’ …ACGTGACTGAGGACCGTG
   CGACTGAGACTGACTGGGT
   CTAGCTAGACTACGTTTTA
   TATATATATACGTCGTCGT
   ACTGATGACTAGATTACAG
   TGATTTTTTTTTTTTTTATT…
3’

Single string from 4 character alphabet
RNA

- Single-stranded
  - A (adenine)
  - C (cytosine)
  - U (uracil)
  - G (guanine)

- Can fold into structures due to base complementarity.
  \[ A \leftrightarrow U, \quad C \leftrightarrow G \]

- Comes in many flavors:
  - mRNA, rRNA, tRNA, tmRNA, snRNA, snoRNA, scaRNA, aRNA, asRNA, piwiRNA, etc.
Protein

• String of amino acids: 20 letter alphabet

...DTIGDWNSSPSFFGIQLVSSVHT TLWYRENAFPVLLGGFSWLSWFNW HNMGYYYPVYHIGYPMIRCGLHL VPMQFQFQSIARSFALVHYNAPM VLKINPHERQDPVFWPCLLYSVD IRSMHIGYPMIRCYQA...
Protein

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

Three fundamental molecules:

1. **DNA**
   Information storage.

2. **RNA**
   Old view: Mostly a “messenger”.
   New view: Performs many important functions.

3. **Protein**
   Perform most cellular functions (biochemistry, signaling, control, etc.)
Central Dogma of Molecular Biology

DNA → RNA → Protein: The process by which cells “read” the genome

First proposed by Francis Crick in 1956.
Transcription and Translation


Transcription and Translation


http://bioinfo.bisr.res.in/project/crat/pictures/codon.jpg
Lecture Outline

• Primer on Molecular Biology
• Primer on Computational Biology
• Primer on Cancer Biology
• Tumor Phylogeny Inference

Reading

• “Biology for Computer Scientists” by Lawrence Hunter
  (http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)
Computational biology and bioinformatics is an interdisciplinary field that develops and applies computational methods to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology.

https://www.nature.com/subjects/computational-biology-and-bioinformatics
Technology and Bioinformatics are Transforming Biology

Until late 20\textsuperscript{th} Century

Hypothesis Generation and Validation

21\textsuperscript{th} Century and Beyond

Algorithms

Hypothesis Generation and Validation

High throughput technologies
A Deluge of Data

Cost per Genome

What happened here?

Moore's Law

Log Scale

NIH National Human Genome Research Institute

genome.gov/sequencingcosts

November, 2017
A Deluge of Data
Question: What does it mean that we can sequence a genome?

No technology exists that can sequence a complete (human) genome from end to end!

Genome
Millions - billions nucleotides

Next-generation DNA sequencing

10-100’s million noisy reads
Reads: 30-1000 nucleotides

Making sense of this data absolutely requires the use and development of algorithms!
Why Study Computational Biology?

Interdisciplinary
Biology
Computer Science
Mathematics
Statistics
= FUN!

Why choose just 1?

Best Jobs
1. Actuary
2. Audiologist
3. Mathematician
4. Statistician
5. Biomedical Engineer
6. Data Scientist
7. Dental Hygienist
8. Software Engineer
9. Occupational Therapist
10. Computer Systems Analyst

Worst Jobs
200. Newspaper reporter
199. Lumberjack
198. Enlisted Military Personnel
197. Cook
196. Broadcaster
195. Photojournalist
194. Corrections Officer
193. Taxi Driver
192. Firefighter
191. Mail Carrier

“I can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.”
Computational Biology: Sequence Alignment

**Question:** How do we compare two genes/genomes?

Human Genome:
...ACTCGACTGAGAGGATTTCGAGCATGA...
≈3.2 x 10^9 bp

Mouse Genome:
...ACTCAACTGAGATTCGAGCCTTCAATGA...
≈2.8 x 10^9 bp
Computational Biology: Genome Assembly

Question: How do we put all the pieces back together?
Computational Biology: Phylogenetics

Phylogenetic Tree of Life

https://en.wikipedia.org/wiki/Phylogenetic_tree

Question: Can we reconstruct the evolutionary history of different species?

https://scientificbsides.wordpress.com/2014/06/09/infering-tumour-evolution-2-comparison-to-classical-phylogenetics/

Question: Can we recover how a tumor has evolved overtime?
Computational Biology: Pattern Matching

Question: How do we start to make sense of all these sequences?

Suffix Trees

Motif Finding

 Burrows Wheeler Transform

Computational Biology is Computer Science

1. Sequence alignment
   ‘How do we compare two genes/genomes?’
   **Dynamic programming**: edit distance

2. Genome assembly
   ‘How do we put all the pieces back together?’
   **Graphs**: de Bruijn graph, Eulerian and Hamiltonian paths

3. Phylogenetics
   ‘What is the evolutionary history of different sequences?’
   **Trees and distances**: distance matrices, neighbor joining, hierarchical clustering, Sankoff/Fitch algorithms, perfect phylogeny and compatibility

4. Pattern matching
   ‘How do we start to make sense out of all these sequences?’
   **Suffix trees/arrays**: Burrows-Wheeler transform, Hidden Markov Models (HMMs)
### Problem $\Pi$ with instance $X$ and solution set $\Pi(X)$:

- **Decision problem:**
  - Is $\Pi(X) = \emptyset$?

- **Optimization problem:**
  - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.

- **Counting problem:**
  - Compute $|\Pi(X)|$.

- **Sampling problem:**
  - Sample uniformly from $\Pi(X)$.

- **Enumeration problem:**
  - Enumerate all solutions in $\Pi(X)$

### Algorithms:

Set of instructions for solving problem.

- **Exact**
- **Heuristic**
Key Challenge in Computational Biology

- Formulating a combinatorial problem
- Interpreting solutions and validating the algorithm
- Analyzing complexity & combinatorial structure
- Designing an algorithm

Translating a biological problem into a computational biology
Lecture Outline

• Primer on Molecular Biology
• Primer on Computational Biology
• Primer on Cancer Biology
• Tumor Phylogeny Inference

Reading

• “Biology for Computer Scientists” by Lawrence Hunter (http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)
Cancer Statistics: Incidence and Mortality

The Burden of Cancer in the United States

• In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed in the United States and 609,640 people will die from the disease.

• The number of new cases of cancer (cancer incidence) is 439.2 per 100,000 men and women per year (based on 2011–2015 cases).

• The number of cancer deaths (cancer mortality) is 163.5 per 100,000 men and women per year (based on 2011–2015 deaths).

• Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes (based on 2013–2015 data).

Source: Surveillance, Epidemiology, and End Results (SEER) Program
Cancer Statistics: Incidence and Mortality

The Burden of Cancer in the United States

• In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed in the United States and 609,640 people will die from the disease.

• The number of new cases of cancer (cancer incidence) is 439.2 per 100,000 men and women per year (based on 2011–2015 cases).

• The number of cancer deaths (cancer mortality) is 163.5 per 100,000 men and women per year (based on 2011–2015 deaths).

• Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes (based on 2013–2015 data).

Source: Surveillance, Epidemiology, and End Results (SEER) Program
90% of cancer patients die of metastasis [Gupta, G. P. & Massagué, Cell, 2006]
Figure 1. Acquired Capabilities of Cancer
We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.
Hallmarks of Cancer

Inter-tumor heterogeneity: Every tumor is different!

Figure 1. Acquired Capabilities of Cancer
We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.
### Hallmarks of Cancer

<table>
<thead>
<tr>
<th>Component</th>
<th>Acquired Capability</th>
<th>Example of Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-sufficiency in growth signals</td>
<td>Activate H-Ras oncogene</td>
</tr>
<tr>
<td></td>
<td>Insensitivity to anti-growth signals</td>
<td>Lose retinoblastoma suppressor</td>
</tr>
<tr>
<td></td>
<td>Evading apoptosis</td>
<td>Produce IGF survival factors</td>
</tr>
<tr>
<td></td>
<td>Limitless replicative potential</td>
<td>Turn on telomerase</td>
</tr>
<tr>
<td></td>
<td>Sustained angiogenesis</td>
<td>Produce VEGF inducer</td>
</tr>
<tr>
<td></td>
<td>Tissue invasion &amp; metastasis</td>
<td>Inactivate E-cadherin</td>
</tr>
</tbody>
</table>

**Figure 1. Acquired Capabilities of Cancer**

We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.
Cancer is Caused by Somatic Mutations
Cancer is Caused by Somatic Mutations

**Question:** Why is there inter-tumor heterogeneity?
Tumorigenesis: Cell Mutation

Founder tumor cell with somatic mutation: (e.g. BRAF V600E)
Clonal Evolution Theory of Cancer
[Nowell, 1976]

**Clone** is a group of cells with the same mutations { , }
Tumorigenesis: Cell Mutation, Division

**Clonal Evolution Theory of Cancer**
[Nowell, 1976]

A clone is a group of cells with the same mutations. In the image, a group of cells is shown splitting to form new clones.
Tumorigenesis: Cell Mutation, Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Intra-Tumor Heterogeneity
Tumorigenesis: Cell Mutation, Division & Migration

Clonal Evolution Theory of Cancer
[Nowell, 1976]

Intra-Tumor Heterogeneity

Phylogenetic Tree $T$

Question: Why are tumor phylogenies important?
Phylogenies are Key to Understanding Cancer

Identify targets for treatment

Understand metastatic development

Recognize common patterns of tumor evolution across patients
Phylogenies are Key to Understanding Cancer

Identify targets for treatment
Understand metastatic development
Recognize common patterns of tumor evolution across patients

These downstream analyses critically rely on accurate tumor phylogeny inference
Phylogenies are Key to Understanding Cancer

These downstream analyses **critically rely** on accurate tumor phylogeny inference

**Key challenge in phylogenetics:**
Accurate phylogeny inference from data at present time
Lecture Outline

• Primer on Molecular Biology
• Primer on Computational Biology
• Primer on Cancer Biology
• Tumor Phylogeny Inference

Reading

• “Biology for Computer Scientists” by Lawrence Hunter
  (http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)
Character-Based Tree Reconstruction

• Characters may be morphological features
  • Shape of beak \{generalist, insect catching, \ldots\}
  • Number of legs \{2,3,4, \ldots\}
  • Hibernation \{yes, no\}

• Character may be nucleotides/amino acids
  • \{A, T, C, G\}
  • 20 amino acids

• Values of a character are called states
  • We assume discrete states
Character-Based Phylogeny Reconstruction

**Input characters** → **Output optimal tree**

**Question:** What is optimal?

**Want:** Optimization criterion
Character-Based Phylogeny Reconstruction

**Input characters** → **Output optimal tree**

**Question**: What is optimal?

**Want**: Optimization criterion

**Question**: How to optimize this criterion?

**Want**: Algorithm
Character-Based Phylogeny Reconstruction: Input

<table>
<thead>
<tr>
<th>Characters / states</th>
<th>State 1</th>
<th>State 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth</td>
<td>Smile</td>
<td>Frown</td>
</tr>
<tr>
<td>Eyebrows</td>
<td>Normal</td>
<td>Pointed</td>
</tr>
</tbody>
</table>
Character-Based Phylogeny Reconstruction: Criterion

Question: Which tree is better?
Character-Based Phylogeny Reconstruction: Criterion

Parsimony: minimize number of changes on edges of tree
Why Parsimony?

• Ockham’s razor: “simplest” explanation for data
• Assumes that observed character differences resulted from the fewest possible mutations
• Seeks tree with the lowest parsimony score, i.e. the sum of all (costs of) mutations in the tree.

“All things being equal, the simplest solution tends to be the best one.”

William of Ockham
A Small and a Large Problem

**Small Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$ and tree $T$ with $m$ leaves, find assignment of character states to each internal vertex of $T$ with minimum parsimony score.

**Large Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree $T$ with $m$ leaves labeled according to $A$ and an assignment of character states to each internal vertex of $T$ with minimum parsimony score.
A Small and a Large Problem

**Small Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$ and tree $T$ with $m$ leaves, find assignment of character states to each internal vertex of $T$ with minimum parsimony score.

**Large Maximum Parsimony Phylogeny Problem:**
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree $T$ with $m$ leaves labeled according to $A$ and an assignment of character states to each internal vertex of $T$ with minimum parsimony score.

**Question:** Are both problems easy (i.e. in P)?
Large Maximum Parsimony Phylogeny

• This problem is NP-hard

• Heuristics using local search (tree moves)

1. Start with an arbitrary tree $T$.
2. Check “neighbors” of $T$.
3. Move to a neighbor if it provides the best improvement in parsimony/likelihood score.

Caveats:
Could be stuck in local optimum, and not achieve global optimum
Example: Nearest-Neighbor Interchange (NNI)

Rearrange four subtrees defined by one internal edge

Figure: Jones and Pevzner
Phylogenies are Key to Understanding Cancer

These downstream analyses critically rely on accurate tumor phylogeny inference.

Key challenge in phylogenetics: Accurate phylogeny inference from data at present time.
Additional Challenge in Cancer Phylogenetics

- Human reference genome (3*10^9 bp)
- Aligned read (100 bp)
Additional Challenge in Cancer Phylogenetics

- Human reference genome (3*10^9 bp)
- Aligned read (100 bp)
- Single nucleotide variant (SNV)
Additional Challenge in Cancer Phylogenetics:
Phylogeny inference from **mixed bulk samples** at present time.
Tumor Phylogeny Inference

Metastatic Colorectal Cancer (Patient CRC2) [Kim et al., Clin Cancer Res 21(19), 2015]:
- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)

```
\[ \begin{bmatrix}
P1 & 1 & 1 & 1 & 0 & 0 & 0 \\
P2 & 1 & 1 & 0 & 1 & 0 & 0 \\
P3 & 1 & 0 & 1 & 0 & 1 & 1 \\
P4 & 0 & 1 & 1 & 0 & 0 & 0 \\
P5 & 0 & 1 & 0 & 1 & 0 & 1 \\
M1 & 1 & 1 & 0 & 0 & 1 & 0 \\
M2 & 0 & 1 & 1 & 1 & 1 & 1
\end{bmatrix} \]
```

Binary Matrix \( B \)

Maximum Parsimony
Tumor Phylogeny Inference

**Metastatic Colorectal Cancer (Patient CRC2)**

[Kim et al., *Clin Cancer Res* 21(19), 2015]:
- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (homoplasy)

\[
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 & 1 \\
\end{pmatrix}
\]

\(m\) samples

\(n\) mutations

**Binary Matrix \(B\)**

**Maximum Parsimony**
Heuristic for Tumor Phylogeny Inference

Metastatic Colorectal Cancer (Patient CRC2)
[Kim et al., Clin Cancer Res 21(19), 2015]:
• 5 primary samples (P1-P5)
• 2 metastases (M1-M2)
• 412 single-nucleotide variants (SNVs)
• 41 mutate more than once (homoplasy)

Resulting sample tree is not representative of the division/mutation history or the migration history
Summary

• DNA, RNA and proteins are sequences
  • Central dogma of molecular biology: DNA -> RNA -> protein

• Problem != algorithm

• Key challenge in computational biology is translating a biological problem into a computational problem

• Cancer is a genetic disease caused by somatic mutations

• Inter-tumor heterogeneity and intra-tumor heterogeneity:
  • Not only is every tumor different, but so is every tumor cell...

• Reading:
  • “Biology for Computer Scientists” by Lawrence Hunter
    (http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)