Course Staff

Instructor:

• Mohammed El-Kebir (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

Piazza: (please sign up)
• https://piazza.com/illinois/spring2020/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 26 to Mar 11)
• Molecular biology and cancer biology
• Fundamental algorithms in computational biology
• Algorithms in computational genomics

Paper presentations (Mar 16 to Apr 13)
• Student presentation of research/survey paper

Course projects (Apr 15 to May 6)
• 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 write a short peer review
  • Summary
  • Major and minor comments
  • Outlook/future directions
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.

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

Cellular molecules:
1. DNA
2. RNA
3. Protein
DNA

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

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

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

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

5’…ACGTGACTGAGGACCGTG…

3’…TGCAGCTGACTCCTGGCAC…

Pair of strings from 4 character alphabet

5’…ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
TGATTTTTAAAAAATATT…

3’
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

...DTIGDWNSPSFFGIQLVSSVHT TLWYRENAFPVLGGFSWLSWFNW HNMGYYPVYHI GYPMIRCGLH VPMQFAFQSIARSFALVHWNAPM VLKINPHERQDPVFWPCLYYSVD IRSMHIGYPMIRCQYA...

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-Letters</th>
<th>1-Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>
Protein

• String of amino acids: 20 letter alphabet
• Folds into 3D structures to perform various functions in cells
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 $\rightarrow$ RNA $\rightarrow$ 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 20th Century

Hypothesis Generation and Validation

21st Century and Beyond

Algorithms

Hypothesis Generation and Validation

High throughput technologies
A Deluge of Data

What happened here?

Cost per Genome

Log Scale

$100M

$10M

$1M

$100K

$10K

$1K


National Human Genome Research Institute

genome.gov/sequencingcosts

Moore's Law
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 \textit{reads}
Reads: 30-1000 nucleotides

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

Interdisciplinary

Biology
Computer Science
Mathematics
Statistics

= FUN!

Why choose just 1?

<table>
<thead>
<tr>
<th>Best Jobs</th>
<th>Worst Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Actuary</td>
<td>200. Newspaper reporter</td>
</tr>
<tr>
<td>2. Audiologist</td>
<td>199. Lumberjack</td>
</tr>
<tr>
<td>3. Mathematician</td>
<td>198. Enlisted Military Personnel</td>
</tr>
<tr>
<td>4. Statistician</td>
<td>197. Cook</td>
</tr>
<tr>
<td>5. Biomedical Engineer</td>
<td>196. Broadcaster</td>
</tr>
<tr>
<td>6. Data Scientist</td>
<td>195. Photojournalalist</td>
</tr>
<tr>
<td>7. Dental Hygienist</td>
<td>194. Corrections Officer</td>
</tr>
<tr>
<td>8. Software Engineer</td>
<td>193. Taxi Driver</td>
</tr>
<tr>
<td>9. Occupational Therapist</td>
<td>192. Firefighter</td>
</tr>
</tbody>
</table>

Donald Knuth
Professor emeritus of Computer Science at Stanford University
Turing Award winner
“father of the analysis of algorithms.”

“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:

...ACTCGACTGAGGATTTGAGCAGATGA...

≈3.2 x 10^9 bp

Mouse Genome:

...ACTCAACTGAGATTCGAGCTTCAATGA...

≈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

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

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

Question: Can we recover how a tumor has evolved overtime?

https://scientificbsides.wordpress.com/2014/06/09/infering-tumour-evolution-2-comparison-to-classical-phylogenetics/
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)
Pet Peeve: Problem != Algorithm

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

Translating a biological problem into a computational biology

1. **Formulating a combinatorial problem**
2. **Analyzing complexity & combinatorial structure**
3. **Designing an algorithm**
4. **Interpreting solutions and validating the algorithm**

Central Question: Biological question
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]
Hallmarks of Cancer

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.
Inter-tumor heterogeneity: Every tumor is different!
### 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

- Single Nucleotide Variant (SNV)
- Small Insertion / Deletion (indel)
- Copy-Number Aberration (CNA)
- Structural Variant (SV)
- Whole-Genome Duplication (WGD)
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 \(\text{is a group of cells with the same mutations}\) {\(\text{, , }\)}

Clonal expansion
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]

Intra-Tumor Heterogeneity
Clonal Evolution Theory of Cancer
[Nowell, 1976]

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

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

These downstream analyses critically rely on accurate tumor phylogeny inference
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

**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>

![Smile](image1.png) ![Frown](image2.png) ![Pointed](image3.png)
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.
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

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

These downstream analyses **critically rely** on accurate tumor phylogeny inference.
Additional Challenge in Cancer Phylogenetics

- Tumor
- Normal

- 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

- human reference genome (3*10^9 bp)
- aligned read (100 bp)
- single nucleotide variant (SNV)
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)

Matrix $B$: $B = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}$

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)

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

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)

---

**Frequency Matrix F**

<table>
<thead>
<tr>
<th>m samples</th>
<th>n mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.80  0.62  0.53  0.03  0.01  0.02</td>
</tr>
<tr>
<td>P2</td>
<td>0.45  0.34  0.04  0.43  0.02  0.02</td>
</tr>
<tr>
<td>P3</td>
<td>0.66  0.04  0.89  0.02  0.15  0.30</td>
</tr>
<tr>
<td>P4</td>
<td>0.04  0.72  0.54  0.01  0.01  0.00</td>
</tr>
<tr>
<td>P5</td>
<td>0.02  0.67  0.03  0.06  0.03  0.11</td>
</tr>
<tr>
<td>M1</td>
<td>0.74  0.43  0.01  0.02  0.14  0.00</td>
</tr>
<tr>
<td>M2</td>
<td>0.01  0.54  0.15  0.25  0.32  0.38</td>
</tr>
</tbody>
</table>

---

**Binary Matrix B**

<table>
<thead>
<tr>
<th>m samples</th>
<th>n mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1  1  1  0  0  0</td>
</tr>
<tr>
<td>P2</td>
<td>1  1  0  1  0  0</td>
</tr>
<tr>
<td>P3</td>
<td>1  0  1  0  1  1</td>
</tr>
<tr>
<td>P4</td>
<td>0  1  1  0  0  0</td>
</tr>
<tr>
<td>P5</td>
<td>0  1  0  1  0  1</td>
</tr>
<tr>
<td>M1</td>
<td>1  1  0  0  1  0</td>
</tr>
<tr>
<td>M2</td>
<td>0  1  1  1  1  1</td>
</tr>
</tbody>
</table>

---

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)