Hands-on Session I: Constructing Trees

Katherine St. John
Lehman College and the Graduate Center
City University of New York
stjohn@lehman.cuny.edu
Session Organization

• **Goal:** To be comfortable building trees from real data

• **Lecture:**
  – Standard Software Packages
  – Details on Web-based Software
  – Motivating Problem

• **Lab:**
  – Organized so you can use the DIMACS lab, or your own laptop
  – Welcome to work singly or in groups
Lecture Outline

• Motivating Problem
Lecture Outline

- Motivating Problem
- Building Trees Overview
Lecture Outline

- Motivating Problem
- Building Trees Overview
- Software
Lecture Outline

• Motivating Problem
• Building Trees Overview
• Software
• Sequence & Tree Formats
Lecture Outline

• Motivating Problem

• Building Trees Overview

• Software

• Sequence & Tree Formats

• Analyzing & Visualizing the Results
Motivating Problem: Which co-evolved?

Murphy et al.
“Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics,” Science ‘01
Motivating Problem: Which co-evolved?

- Murphy et al., Science ‘01, data set:
  44 taxa: (42 placentals + 2 marsupial for outgroups)
  22 genes: 19 nuclear + 3 mitochondrial
Motivating Problem: Which co-evolved?

- Murphy et al., Science ‘01, data set:
  44 taxa: (42 placentals + 2 marsupial for outgroups)
  22 genes: 19 nuclear + 3 mitochondrial

- Well-studied data set for underlying problem as well as methodology questions (over 300 citations).
Motivating Problem: Which co-evolved?

- Murphy et al., *Science* ‘01, data set:
  44 taxa: (42 placentals + 2 marsupial for outgroups)
  22 genes: 19 nuclear + 3 mitochondrial

- Well-studied data set for underlying problem as well as methodology questions (over 300 citations).

- For example: (Hillis et al., *Sys Bio*, 2005), is it better
  - to build trees on each gene sequence and take the consensus, or
  - concatenate the sequences and look at those trees?
Motivating Problem: Which co-evolved?

• For example: (Hillis et al., Sys Bio, 2005), is it better
  – to build trees on each gene sequence and take the consensus, or
  – concatenate the sequences and look at those trees?

• More tractable:
  – which of these genes co-evolved?
  – focus on several, or try all of them
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.
5. Analyze Results.
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.
5. Analyze Results.

We’ll focus on the last two today.
Models of Evolution

• Can make a significant difference when constructing trees.
Models of Evolution

- Can make a significant difference when constructing trees.
  - Jukes-Cantor (JC): simplest, all sites iid, equally likely, only parameter is the substitution rate
Models of Evolution

- Can make a significant difference when constructing trees.
  - Jukes-Cantor (JC): simplest, all sites iid, equally likely, only parameter is the substitution rate
  - Kimura-2-Parameter (K2P): distinguishes between the transition (A↔G and C↔T) and tranversion (A↔C and G↔T) rates all nucleotides occur at equal frequencies
Models of Evolution

- Can make a significant difference when constructing trees.
  - **Jukes-Cantor (JC):** simplest, all sites iid, equally likely, only parameter is the substitution rate
  - **Kimura-2-Parameter (K2P):** distinguishes between the transition (A↔G and C↔T) and tranversion (A↔C and G↔T) rates all nucleotides occur at equal frequencies
  - **Hasegawa-Kishono-Yano (HKY):** nucleotides occur at different frequencies
Models of Evolution

- Can make a significant difference when constructing trees.
  - Jukes-Cantor (JC): simplest, all sites iid, equally likely, only parameter is the substitution rate
  - Kimura-2-Parameter (K2P): distinguishes between the transition (A↔G and C↔T) and tranversion (A↔C and G↔T) rates all nucleotides occur at equal frequencies
  - Hasegawa-Kishono-Yano (HKY): nucleotides occur at different frequencies
  - General Time Reversible (GTR): assume symmetric substitution matrix (ie A changes to C at the same rate C changes to A).
## Models of Evolution

(From Hillis et al. ‘05.)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Preferred model</th>
<th>Base frequencies</th>
<th>Relative substitution rates</th>
<th>Proportion of invariant sites</th>
<th>Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preferred model and estimated base frequencies for each gene</td>
<td></td>
<td>Model substitution and rate heterogeneity parameters for each gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADORA3</td>
<td>K2P</td>
<td>0.25 0.25 0.25 0.25</td>
<td>1 3 1 1 3 1</td>
<td>0.46</td>
<td>1.05</td>
</tr>
<tr>
<td>ADRB2</td>
<td>HKY+I+G</td>
<td>0.2 0.33 0.25 0.22</td>
<td>1.6 3.66 0.47 0.72 2.65 1</td>
<td>0</td>
<td>0.78</td>
</tr>
<tr>
<td>APP</td>
<td>GTR+I+G</td>
<td>0.25 0.24 0.18 0.33</td>
<td>1.11 5.33 0.68 0.92 4.43 1</td>
<td>0.2</td>
<td>1.56</td>
</tr>
<tr>
<td>ATP7A</td>
<td>GTR+I+G</td>
<td>0.33 0.21 0.19 0.19</td>
<td>1 4.73 1 1 4.73 1</td>
<td>0.42</td>
<td>0.61</td>
</tr>
<tr>
<td>BDNF</td>
<td>HKY+I+G</td>
<td>0.21 0.33 0.28 0.17</td>
<td>2.35 7.08 0.64 1.77 5.71 1</td>
<td>0.14</td>
<td>0.82</td>
</tr>
<tr>
<td>BMI1</td>
<td>GTR+I+G</td>
<td>0.29 0.15 0.16 0.4</td>
<td>3.43 14 1.3 2.13 14.6 1</td>
<td>0.53</td>
<td>0.7</td>
</tr>
<tr>
<td>CNR1</td>
<td>GTR+I+G</td>
<td>0.18 0.32 0.25 0.24</td>
<td>1.68 3.44 0.55 0.8 2.97 1</td>
<td>0.18</td>
<td>1.6</td>
</tr>
<tr>
<td>CREM</td>
<td>GTR+I+G</td>
<td>0.21 0.24 0.28 0.27</td>
<td>1 4.93 1 1 4.93 1</td>
<td>0.44</td>
<td>0.72</td>
</tr>
<tr>
<td>EDG1</td>
<td>HKY+I+G</td>
<td>0.17 0.36 0.27 0.2</td>
<td>0.94 2.77 0.59 0.56 2.33 1</td>
<td>0.04</td>
<td>2.88</td>
</tr>
<tr>
<td>PLCB4</td>
<td>GTR+I+G</td>
<td>0.3 0.27 0.19 0.24</td>
<td>0.9 2.73 0.86 0.38 4.14 1</td>
<td>0.15</td>
<td>1.09</td>
</tr>
<tr>
<td>PNOC</td>
<td>GTR+I+G</td>
<td>0.23 0.33 0.31 0.12</td>
<td>2.04 5.59 1.01 0.67 9.09 1</td>
<td>0.49</td>
<td>1.07</td>
</tr>
<tr>
<td>RAG1</td>
<td>GTR+I+G</td>
<td>0.21 0.3 0.29 0.19</td>
<td>2.18 7.86 1.3 0.93 8.76 1</td>
<td>0.32</td>
<td>1.27</td>
</tr>
<tr>
<td>RAG2</td>
<td>HKY+I+G</td>
<td>0.28 0.24 0.22 0.27</td>
<td>1 6 1 1 6 1</td>
<td>0.35</td>
<td>1.63</td>
</tr>
<tr>
<td>TYR</td>
<td>GTR+I+G</td>
<td>0.24 0.26 0.25 0.25</td>
<td>1 7.94 1 1 7.94 1</td>
<td>0.49</td>
<td>1.24</td>
</tr>
<tr>
<td>ZFX</td>
<td>HKY+I+G</td>
<td>0.35 0.23 0.18 0.23</td>
<td>1 4.41 1 1 4.41 1</td>
<td>0.15</td>
<td>0.92</td>
</tr>
<tr>
<td>VWF</td>
<td>HKY+I+G</td>
<td>0.2 0.34 0.28 0.18</td>
<td>1.15 4.38 0.75 1.17 4.75 1</td>
<td>0.04</td>
<td>3.4</td>
</tr>
<tr>
<td>BRCA1</td>
<td>GTR+I+G</td>
<td>0.33 0.22 0.23 0.22</td>
<td>1.5 4.91 1.34 0.83 5.8 1</td>
<td>0.18</td>
<td>1.04</td>
</tr>
<tr>
<td>IRBP</td>
<td>GTR+I+G</td>
<td>0.21 0.3 0.3 0.18</td>
<td>1.5 3.59 0.93 0.62 3.71 1</td>
<td>0.3</td>
<td>1.29</td>
</tr>
<tr>
<td>A2AB</td>
<td>GTR+I+G</td>
<td>0.17 0.34 0.3 0.18</td>
<td>5.86 14 3.85 0.58 29.3 1</td>
<td>0.41</td>
<td>0.53</td>
</tr>
<tr>
<td>mtRNA</td>
<td>GTR+I+G</td>
<td>0.34 0.2 0.21 0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tree Building Software

Some Packages that perform multiple methods:

- **Phylogenetic Analysis Using Parsimony (PAUP 4.0):** Swofford ‘02
- **Phylogenetic Inference Package (Phylip 3.6):** Felsenstein ‘06
- **Molecular Evolutionary Genetic Analysis (MEGA 3.1):** Kumar, Tamura, & Nei ‘04
- **SplitsTree 4:** Huson & Bryant ‘06
Tree Building Software

Some specialized software:

- **MrBayes 3.1:** Bayesian inference of phylogeny, Huelsenbeck *et al.* ‘05
- **Bayesian Evolutionary Analysis Sampling Trees (BEAST):** Drummond & Rambaut ‘03
- **Quartet Puzzling:** Strimmer & Von Haeseler ‘96
Software with Web Interface

Web access available for:

- At the Pasteur Institute
  http://bioweb.pasteur.fr/intro-uk.html:
    Phylip, Quartet Puzzling, Weighbor, etc.
- SplitsTree (older version: 3.2) at:
  http://bibiserv.techfak.uni-bielefeld.de/splits/submission.html
Software for Today:

• Suggested that you use on-line software (quicker to get started, but will run slower)

• Or, you can download most programs to your laptops:
  – most freely available (notable exception: PAUP)
  – newer ones in Java and machine independent
  – most run on Unix (Linux & OS X), some run on Windows
Sequence Formats

- PAUP:
- Phylip:
- FASTA:
- Can use the program READSEQ to convert from one to another.
Sequence Formats

- PAUP:
- Phylip:
- FASTA:
- Can use the program READSEQ to convert from one to another. And EXTRACTSEQ (EMBOSS) to extract a region.
Sequence Formats

PAUP:

#NEXUS

Begin data;
Dimensions ntax=44 nchar=17028;
Format datatype=dna interleave gap=-;
Matrix
Opossum          TGCCTCTTCCGTTCAATGAGGATGGACTACATGGTCTATTTTCAGCTT
Diprotodontian  TGCCGCTTCCGCTCAATTATGAGGATGGACTACATGGTCTATTTTCAGCTT
Sloth           TGCAAATTTCAATCTCCGTCATGAGAATGGACTACATGGTCTACTTCAGTTT
Armadillo       TGCAAATTCACTTCCGTCATGAGGATGGACTACATGGTCTACTTCAGTTT
Anteater        TGCIAATTCAAGTTCCGTCATGAGGACTACATGGTCTACTTCAGTTT
Hedgehog        TGCCAAATTTCAATCTGTGTTGAGAATGGACTACATGGTCTACTTCAGTTT
Mole             TGCIAATTTGCACAGTGCATGAGGACTACATGGTCTACTTCAGTTT
Shrew           TGCIAATTTGCACAGTGCATGAGGACTACATGGTCTACTTCAGTTT
Tenrecid        TGCIAATTTGCATCTATAGAATGGACTACATGGTCTACTTCAGTTT
GoldenMole      TGCIAATTTGCATCTATAGAATGGACTACATGGTCTACTTCAGTTT
...
Sequence Formats

Phylip:

44 17028
Opossum   TGCCCTCTCC G TTCAGTAAT GAGGATGGAC TACATGGTCT ATTTTCAGCTT
Diprotodon TGCCGCTTCC GCTCAGTTAT GAGGATGGAC TACATGGTCT ATTTTCAGCTT
Sloth      TGCAAAATTCA G TTTCCGTCAT GAGAATGGAC TACATGGTCT ACTTCAGTTT
Armadillo  TGCAAAATTCA G TTTCCGTCAT GAGGATGGAC TACATGGTGT ACTTCAGTTT
Anteater   TGCAAAATTCA G TTTCCGTTGT GAGGATGGAC TACATGGTCT ACTTCAGTTT
Hedgehog   TGCCAAATTCC G TTTCTGGTGT GAGAATGGAC TACATGGTGT TCTTCAGCTT
Mole       TGCAAGTTCC GCACAGTCGT GAGGATGGAC TACATGGTCT ACTTCAGCTT
Shrew      TGCCAGTTCC GCTCTGTGGT GAGGATGGAC TACATGGTCT ACTTCAGCTT
Tenrecid   TGCAAAATTCC GTTCTACTAT GAGAATGGAC TACATGGTCT ACTTCAGCTT
GoldenMole TGCCAAATTTC GTTCCGTAAT GAGGATGGAC TATATGGTCT ACTTCAGCTT
...
Sequence Formats

FASTA:

>0possum, 17028 bases, FC7ADFCB checksum.
TGCCTCTTCCGTTCAGTAAATGAGGATGGACTACATGGTCTATTTTCAGCTT
TTTCACATGGATCCTCATCCCTTTGGTCACTATGTGTGCGCATCTATGTTG
ACATTTTCTATGTCACTCCGGAACAGCTCAGACAGAACTTCTCTGGCTCA
AAAGAGACAGGTGCAATTCTATGGGAAGGAGTTCAAGACAGCCAAATCCCT
CTTTCTCATCCTCTTTCTAATGGGCTATGCTTGCTGCCCTTTATCCATCA
TCAACTGTATTCTATTTCTCCCTAAGGCTGAGATA---CCTTCAGTT
TTGCTTGGGTGGGA?ATCCTGCTATCCCAT????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
????????????????????????????????????????????????
CCCCGGGTGTCATTTTGATGGTGTG
...
Visualizing Trees

Web access available for:

- Phylip: Felsenstein
- SplitsTree: Bryant & Huson
- Mesquite: Wayne & David Maddison
Getting Started

• Download the sequences to your machine.
Getting Started

- Download the sequences to your machine.
- Choose the subset you would like to analyze.
Getting Started

• Download the sequences to your machine.
• Choose the subset you would like to analyze
  (The PAUP file has the endpoints for each gene.)
Getting Started

- Download the sequences to your machine.
- Choose the subset you would like to analyze
  (The PAUP file has the endpoints for each gene.)
- Choose the methods you would like to apply
Getting Started

• Download the sequences to your machine.
• Choose the subset you would like to analyze (The PAUP file has the endpoints for each gene.)
• Choose the methods you would like to apply (Then convert sequences into the needed format.)
Getting Started

• Download the sequences to your machine.
• Choose the subset you would like to analyze (The PAUP file has the endpoints for each gene.)
• Choose the methods you would like to apply (Then convert sequences into the needed format.)
• Look at the resulting trees– do they support your hypothesis?
Helpful Websites

- Dataset for this tutorial:
  http://comet.lehman.cuny.edu/stjohn/dimacsTutorial

- The Pasteur Institute:
  http://bioweb.pasteur.fr/intro-uk.html

- SplitsTree: at:
  http://bibiserv.techfak.uni-bielefeld.de/splits/submission.html