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

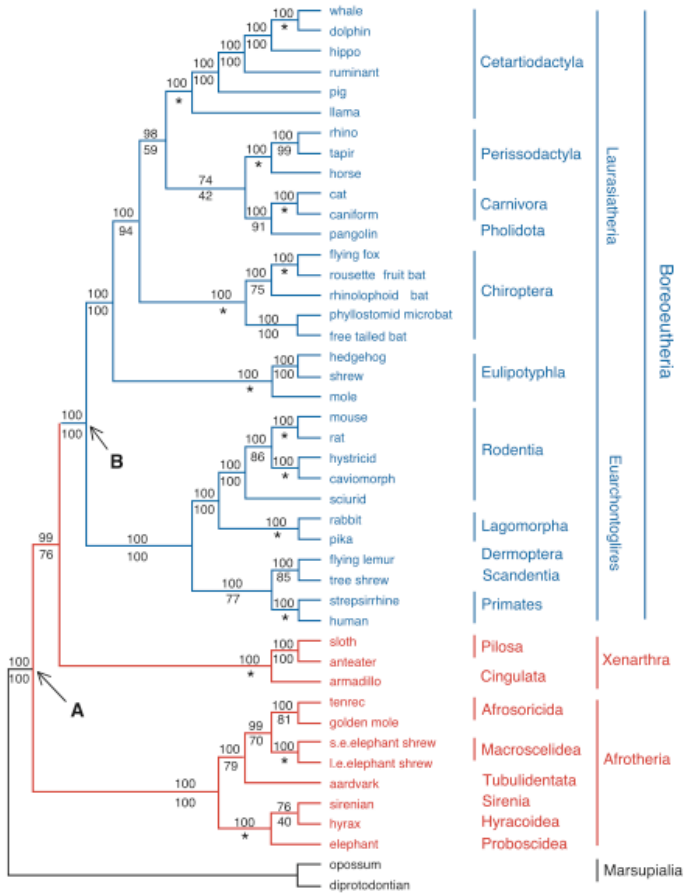
- Motivating Problem
- Building Trees Overview
- Software
- Sequence & Tree Formats

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$\leftrightarrow$G and C$\leftrightarrow$T) and tranversion (A$\leftrightarrow$C and G$\leftrightarrow$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$\leftrightarrow$G and C$\leftrightarrow$T) and tranversion (A$\leftrightarrow$C and G$\leftrightarrow$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

APPENDIX 2. Model parameters for the genes studied by Murphy et al. (2000):

		Base frequencies				Relative substitution rates						Proportion invariant	
Gene	Preferred model	A	C	G	T	AC	AG	AT	CG	CT	GT	sites	Alpha
Preferred model and estimated base frequencies for each gene						Model substitution and rate heterogeneity parameters for each gene							
ADORA3	K2P	0.25	0.25	0.25	0.25	1	3	1	1	3	1	0	
ADRB2	HKY+I+G	0.2	0.33	0.25	0.22	1	5.75	1	1	5.75	1	0.46	1.05
APP	GTR+I+G	0.25	0.24	0.18	0.33	1.6	3.66	0.47	0.72	2.65	1	0	0.78
ATP7A	GTR+I+G	0.33	0.21	0.19	0.19	1.11	5.33	0.68	0.92	4.43	1	0.2	1.56
BDNF	HKY+I+G	0.21	0.33	0.28	0.17	1	4.73	1	1	4.73	1	0.42	0.61
BMI1	GTR+I+G	0.29	0.15	0.16	0.4	2.35	7.08	0.64	1.77	5.71	1	0.14	0.82
CNR1	GTR+I+G	0.18	0.32	0.25	0.24	3.43	14	1.3	2.13	14.6	1	0.53	0.7
CREM	GTR+I+G	0.21	0.24	0.28	0.27	1.68	3.44	0.55	0.8	2.97	1	0.18	1.6
EDG1	HKY+I+G	0.17	0.36	0.27	0.2	1	4.93	1	1	4.93	1	0.44	0.72
PLCB4	GTR+I+G	0.3	0.27	0.19	0.24	0.94	2.77	0.59	0.56	2.33	1	0.04	2.88
PNOC	GTR+I+G	0.23	0.33	0.31	0.12	0.9	2.73	0.86	0.38	4.14	1	0.15	1.09
RAG1	GTR+I+G	0.21	0.3	0.29	0.19	2.04	5.59	1.01	0.67	9.09	1	0.49	1.07
RAG2	HKY+I+G	0.28	0.24	0.22	0.27	1	6	1	1	6	1	0.35	1.63
TYR	GTR+I+G	0.24	0.26	0.25	0.25	2.18	7.86	1.3	0.93	8.76	1	0.32	1.27
ZFX	HKY+I+G	0.35	0.23	0.18	0.23	1	7.94	1	1	7.94	1	0.49	1.24
VWF	HKY+I+G	0.2	0.34	0.28	0.18	1	4.41	1	1	4.41	1	0.15	0.92
BRCA1	GTR+I+G	0.33	0.22	0.23	0.22	1.15	4.38	0.75	1.17	4.75	1	0.04	3.4
IRBP	GTR+I+G	0.21	0.3	0.3	0.18	1.5	4.91	1.34	0.83	5.8	1	0.18	1.04
A2AB	GTR+I+G	0.17	0.34	0.3	0.18	1.02	3.59	0.93	0.62	3.71	1	0.3	1.29
mtRNA	GTR+I+G	0.34	0.2	0.21	0.25	5.86	14	3.85	0.58	29.3	1	0.41	0.53

(From Hillis *et al.* '05.)

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: Bayesan 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            TGCCTCTTCCGTTCAGTAATGAGGATGGACTACATGGTCTATTTCAGCTT
Diprotodontian     TGCCGCTTCCGCTCAGTTATGAGGATGGACTACATGGTCTATTTCAGCTT
Sloth              TGCAAATTCAGTTCCGTCATGAGAATGGACTACATGGTCTACTTCAGTTT
Armadillo          TGCAAATTCACTTCCGTCATGAGGATGGACTACATGGTGTACTTCAGTTT
Anteater           TGCAAATTCAGTTCCGTTGTGAGGATGGACTACATGGTCTACTTCAGTTT
Hedgehog           TGCCAATTCCGTTCTGTTGTGAGAATGGACTACATGGTGTTCTTCAGCTT
Mole               TGCAAGTTCCGCACAGTCGTGAGGATGGACTACATGGTCTACTTCAGCTT
Shrew              TGCCAGTTCCGCTCTGTGGTGAGGATGGACTACATGGTCTACTTCAGCTT
Tenrecid           TGCAAATTCCGTTCTACTATGAGAATGGACTACATGGTCTACTTCAGCTT
GoldenMole         TGCCAATTTCGTTCCGTAATGAGGATGGACTATATGGTCTACTTCAGCTT
...
```

Sequence Formats

Phylip:

```
 44 17028
Opossum      TGCCTCTTCC GTTCAGTAAT GAGGATGGAC TACATGGTCT ATTTCAGCTT
Diprotodon   TGCCGCTTCC GCTCAGTTAT GAGGATGGAC TACATGGTCT ATTTCAGCTT
Sloth        TGCAAATTCA GTTCCGTCAT GAGAATGGAC TACATGGTCT ACTTCAGTTT
Armadillo    TGCAAATTCA CTTCCGTCAT GAGGATGGAC TACATGGTGT ACTTCAGTTT
Anteater     TGCAAATTCA GTTCCGTTGT GAGGATGGAC TACATGGTCT ACTTCAGTTT
Hedgehog     TGCCAATTCC GTTCTGTTGT GAGAATGGAC TACATGGTGT TCTTCAGCTT
Mole         TGCAAGTTCC GCACAGTCGT GAGGATGGAC TACATGGTCT ACTTCAGCTT
Shrew        TGCCAGTTCC GCTCTGTGGT GAGGATGGAC TACATGGTCT ACTTCAGCTT
Tenrecid     TGCAAATTCC GTTCTACTAT GAGAATGGAC TACATGGTCT ACTTCAGCTT
GoldenMole   TGCCAATTTC GTTCCGTAAT GAGGATGGAC TATATGGTCT ACTTCAGCTT
...
```

Sequence Formats

FASTA:

```
>Opossum, 17028 bases, FC7ADFCB checksum.
TGCCTCTTCCGTTCAGTAATGAGGATGGACTACATGGTCTATTTCAGCTT
TTTCACATGGATCCTCATCCCTTTGGTCATCATGTGTGCCATCTATGTTG
ACATTTTCTATGTCATCCGGAACAAGCTCAGACAGAACTTCTCTGGCTCA
AAAGAGACAGGTGCATTCTATGGGAAGGAGTTCAAGACAGCCAAATCCCT
CTTTCTCATCCTCTTCTTGTTTGCCATATCCTGGCTGCCTTTATCCATCA
TCAACTGTATTTCTTATTTCTTCCCTAAGGCTGAGATA---CCTTCAGTT
TTGCTTGGGTTGGA?ATCCTGCTATCCCAT??????????????????????
??????????????????????????????????????????????????
??????????????????????????????????????????????????
??????????????????????????????????????????????????
??????????????????????????????????????????????????
??????????????????????????????????????????????????
??????????????????????????CCCGGGTGGTCATTTTGATGGTGTG
...
```

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`