How to read and make phylogenetic trees, part 2
+ Use of molecular phyletogenics in zoology

Zuzana Starostová
How to make phylogenetic trees?

Workflow:

- obtain DNA sequence
- quality check
- sequence alignment
  - calculating genetic distances
  - phylogeny estimation – topology and branch length
  - NJ, PM, ML, BA
  - reliability test (bootstrap)
  - tree visualization
Phylogeny estimation

Two types of methods:
- Character based (maximum parsimony, maximum likelihood, Bayesian analysis)
- Distance based (Neighbour-joining, UPGMA)

Two different approaches:
- Algorithm – number of specific steps resulting in one best tree
  Methods: UPGMA, Neighbour-joining

- Optimality criterion – consider and compare all theoretically possible trees
  based on selected criteria: number of evolutionary steps, likelihood value
**distances**

input is a matrix of distances between species

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>---</td>
<td>0.1</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>II</td>
<td>---</td>
<td>---</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>III</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.6</td>
</tr>
<tr>
<td>IV</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
proportional (p) distance

number of substitutions between sequences

\[ p = \frac{n_d}{n} \]

\[ p = \frac{4}{17} = 0.23 \]
real number of substitutions in the sequence over time is usually higher than observed p distance

we can see just 3 differences (p), but in fact there was 12 substitutions

Flegr, Evoluční biologie, Academia
In phylogenetic analyses we use "correction" of observed distances to estimate number of hidden changes (multiple mutations etc.).

Correction based on different substitution type (Ts, Tv), different substitution rate, frequencies of nucleotides.
Jukes-Cantor model (distance)
all substitution types and base frequencies are presumed equal

JC distance

\[ d_{JC} = -\frac{3}{4} \ln(1 - \frac{4}{3} p) \]

Kimura 2-parameter model (K2P):
transitions are more likely than transversions, equal base frequencies

K2P distance
\[ P = \frac{n_{TS}}{n} \]
\[ Q = \frac{n_{TV}}{n} \]

\[ d_{K2P} = 0.5 \ln\left(\frac{1}{1 - 2P - Q}\right) + 0.25 \ln\left(\frac{1}{1 - 2Q}\right) \]
methods

**Neighbour-joining (NJ)** - the fully resolved tree is "decomposed" from a fully unresolved "star" tree by successively inserting branches between a pair of closest neighbors and the remaining terminals in the tree; result is one tree.

- **other methods:** UPGMA (Unweighted Pair Group Method using Arithmetic means), **Minimal evolution**
conclusion, pros and cons

distance methods rely on evolutionary models (distance corrections) to estimate the numbers of multiple/parallel… substitutions – the result is dependent on how well the accepted models match the actual evolutionary properties of the sequences

only one tree is derived

discards the primary character data

problem with interpretation of branch lengths

very fast, ideal for the first insight
**Maximum parsimony:**

optimality criterion - parsimony score = minimum number of events (steps) required by a tree to explain the variation in the data

search for topologies that minimize the total tree length assuming a minimum number of base changes

“Occam’s Razor” – “keep it simple“

---

Using Maximum Parsimony to Choose Between Two Possible Trees

<table>
<thead>
<tr>
<th>Sample:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation:</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

![Diagram showing two possible trees with one change required for one tree and two changes required for the other, indicating that 1 change required results in a better tree, and 2 changes required result in a poorer tree.](image-url)
not all characters are good for parsimony: the alignment is checked for informative positions = a site must have the same character states in at least two taxa and must favor one topology over another
Maximum parsimony:
optimality criterion - parsimony score = minimum number of events (steps) required by a tree to explain the variation in the data

search for topologies that minimize the total tree length assuming a minimum number of base changes
“Occam’s Razor” – “keep it simple“

\[(2n - 3)!\]

We already know that there are a lot of possible trees- in most cases we can not compare all of them

\[2^{n-2}(n-2)!\]

<table>
<thead>
<tr>
<th>no. of taxa</th>
<th>no. of unrooted trees</th>
<th>no. of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>10 395</td>
<td>135 135</td>
</tr>
<tr>
<td>10</td>
<td>2 027 025</td>
<td>34 459 425</td>
</tr>
<tr>
<td>22</td>
<td>(3 \times 10^{23})</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>(3 \times 10^{74})</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>(2 \times 10^{182})</td>
<td></td>
</tr>
</tbody>
</table>
Tree searching

Exhaustive Searching

Branch and Bound Searching

Heuristic Searching

Starting tree

Local branch swapping
Swofford et al. (1996)

Global branch swapping
Maximum parsimony

in most cases we can not compare all trees

⇒ e.g. heuristic search

- create random tree
- calculate parsimony score
- rearranging of the tree,
- calculate parsimony score
- further the method works with the better (shorter) tree
- repeated rearranging and calculating scores
- at the end shortest tree

Sometimes (quite often) we find more equal trees
Consensus tree:

when multiple phylogenies are supported - a consensus tree shows only those relationships common to all trees (based on our settings)

- **strict consensus** (only relationships common to all trees)

- **majority-rule consensus** (relationships common to 50 or 70% of trees are shown)
Parsimony: pros and cons

+ works directly with characters
+ straightforward, well understood principle
+ relatively fast
+ does not need a model of evolution (but not really model free – change is rare)

- performs weakly on distantly related data
- long branch attraction
- can produce many trees with the same parsimony score
long branch attraction (LBA)

FIGURE 5.22. Long-branch attraction is a methodological artifact that can cause phylogenetic trees to inaccurately portray evolutionary history. The phenomenon causes errors in phylogenetic reconstruction when two (or more) of the entities being studied lie on the end of long branches in their "real" tree but are not sister taxa. (A) In this hypothetical "real" tree of five species, species 2 and 3 (which are not sister taxa, as indicated) have undergone higher rates of evolution than the other three, and thus sit at the end of longer branches. Many phylogenetic reconstruction methods used to infer the evolution of species will cause the long branches to appear to be closely related and thus produce an incorrect tree (as shown in B). (C) In studies of the evolution of microsporidia (a relative of fungi, left tree), long-branch attraction (LBA) is believed to have erroneously identified them as deeply branching eukaryotes (right tree). (The evolution of microsporidia is discussed in more detail on p. 198.) (D) In trees of anciently duplicated genes, long-branch attraction might have pulled bacteria down to the paralogs used to root the tree, because the paralogs are at the end of a long branch (right tree). This would occur if bacteria evolved at a higher rate than archaean and eukaryotes (as suggested in the left tree).

Maximum likelihood - ML

- method compares possible phylogenetic trees on the basis of their ability to predict the observed data. The tree that has the highest probability of producing the observed sequences is preferred.

- maximum likelihood reconstructs ancestors at all nodes of each considered tree, but it also assigns branch lengths based on the probabilities of mutations. For each possible tree topology, the assumed substitution rates are varied to find the parameters that give the highest likelihood of producing the observed sequences.

Likelihood describes how well the model predicts the data
- it prefers higher likelihood above the lower one
ML results in only 1 tree with branch lengths
Maximum likelihood - ML

- **ML uses model of sequence evolution (substitution model)**
- several programs (Modeltest, jModeltest, MrAIC…)
  programs examine the goodness of fit of the model to the data

- models differ in:
  • base frequencies
  • probability of nucleotides changes (transition x transversion)
  • heterogeneousness in different parts of sequence or in different position

Model examples:
Jukes-Cantor (JC),
Kimura 2-parametres model (K2P),
General time-reversible model (GTR)
Models of sequence evolution

- models are nested, one is a special case of the other
F = base frequencies; S = substitution type; I = proportion of invariable sites; G = gamma rates
Best model selection
program jModeltest (Modeltest)

Example of model:
Lset base=(0.3171 0.2948 0.1271) nst=6  rmat=(0.1710 5.8391 1.0000 0.1710 14.3282)
rates=gamma shape=0.3310 ncat=4 pinvar=0.4550;

1. relative base composition (4th is 1-(fr1st+fr2nd+fr3rd))
2. No. of substitution types (1 = same probability for all bases, 6 = every substitution has different probability)
3. substitution rate matrix – rate of changes of each type of bases in alignment
4. probability of changes distribution in individual positions
   (equal = equal for all position, gamma = with different gamma distribution, invgamma)
5. shape of gamma distribution
6. gamma distribution category
7. ratio (proportion) of invariable sites
Maximum likelihood

**pros**
- a lot of possible models of sequence evolution, robust to deviations from the model

**cons**
- computationally demanding, slow (nowadays not so big problem)

ML method can decrease effect of LBA

Swofford et al., *Systematic Biology*, 2001
reliability tests

- nonparametric resampling methods - bootstrapping, jackknifing

new data sets are created from the original data set by sampling columns of characters at random

- each site can be sampled again with the same probability as any of the other sites
Box 3
Bootstrap Analysis (Felsenstein, 1985)

```
s100 ..1010220112..
... ... ... ...
s3  ..0120401200..
s2  ..1000222003..
s1  ..1310110012..
A   ..AGGCCUCCAAA..
B   ..AGGGGUCCAAA..
C   ..AGCCCGGAAA..
D   ..AUUUCCGAAAC..

Tree based on original sequence alignment

100  
   B  
  /  
A   C

75
   D

sample 1 (s1)
A   ..AGGGGUCAAA..
B   ..AGGGGUCAAA..
C   ..AGGGGCCCCAAA..
D   ..AUUUUCCACC..

Bootstrap tree 1

sample 2 (s2)
A   ..AUUUCCCCAAA..
B   ..AUUUCCCCAAA..
C   ..ACCCCGGAAA..
D   ..ACCCCGGCCC..

Bootstrap tree 2

sample 3 (s3)
A   ..GgggUUUUCAAA..
B   ..GGGUUUUGAAA..
C   ..GCCCCCGGAAA..
D   ..UUUCCCCGAAA..

Bootstrap tree 3

100  
   B  
  /  
A   C

75
   D

sample 100 (s100)
A   ..AGUUCCAAAAAT..
B   ..AGUUCCAAAAAT..
C   ..AUCCCCAAAAAAT..
D   ..AUCCCCAAACCC..

Bootstrap tree 100

Bootstrap values superimposed on original tree
(2)

sample n (100<n<2000)

Bootstrap consensus tree
(1)
Bootstrap values:
< 50% - no - just by chance
> 75% ok
95-100% great
Bayesian inference/analysis

Bayesian inference of phylogeny uses a likelihood function to create a quantity called the **posterior probability** of trees using a model of evolution (substitution model), based on some prior probabilities (priors), producing the most likely phylogenetic tree for the given data
- uses Markov chain Monte Carlo (MCMC) algorithms

Based on theorem of Thomas Bayes (18. century) – Bayesian theorem
- describes the probability of an event, based on prior knowledge of conditions that might be related to the event
We will illustrate Bayesian inference using a simple example involving dice. Consider a box with 100 dice, 90 of which are fair and 10 of which are biased. The probability of observing some number of pips after rolling a fair or biased die is given in the following table:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Fair</th>
<th>Biased</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/6</td>
<td>1/21</td>
</tr>
<tr>
<td>2</td>
<td>1/6</td>
<td>2/21</td>
</tr>
<tr>
<td>3</td>
<td>1/6</td>
<td>3/21</td>
</tr>
<tr>
<td>4</td>
<td>1/6</td>
<td>4/21</td>
</tr>
<tr>
<td>5</td>
<td>1/6</td>
<td>5/21</td>
</tr>
<tr>
<td>6</td>
<td>1/6</td>
<td>6/21</td>
</tr>
</tbody>
</table>

The probability of a high roll is larger for the biased dice than for the fair dice. Suppose that you draw a die at random from the box and roll it twice, observing a four on the first roll and a six on the second roll. What is the probability that the die is biased?

A Bayesian analysis combines ones prior beliefs about the probability of a hypothesis with the likelihood. The likelihood is the vehicle that carries the information about the hypothesis contained in the observations. In this case, the likelihood is simply the probability of observing a four and a six given that the die is biased or fair. Assuming independence of the tosses, the probability of observing a four and a six is

$$\Pr[\text{Fair}] = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

for a fair die and

$$\Pr[\text{Biased}] = \frac{4}{21} \times \frac{6}{21} = \frac{24}{441}$$

for a biased die. The probability of observing the data is 1.96 times greater under the hypothesis that the die is biased. In other words, the ratio of the likelihoods under the two hypotheses suggests that the die is biased.

Bayesian inferences are based upon the posterior probability of a hypothesis. The posterior probability that the die is biased can be obtained using Bayes’ (1) formula:

$$\Pr[\text{Biased} \mid \text{Data}] = \frac{\Pr[\text{Data} \mid \text{Biased}] \times \Pr[\text{Biased}]}{\Pr[\text{Data} \mid \text{Biased}] \times \Pr[\text{Biased}] + \Pr[\text{Data} \mid \text{Fair}] \times \Pr[\text{Fair}]}$$

where \(\Pr[\text{Biased}]\) and \(\Pr[\text{Fair}]\) are the prior probabilities that the die is biased or fair, respectively. As we set up the problem, a reasonable prior probability that the die is biased would be the proportion of the dice in the box that were biased. The posterior probability is then

$$\Pr[\text{Biased} \mid \text{Data}] = \frac{\frac{24}{441} \times \frac{1}{10}}{\frac{24}{441} \times \frac{1}{10} + \frac{1}{36} \times \frac{9}{10}} = 0.179$$

This means that our opinion that the die is biased changed from 0.1 to 0.179 after observing the four and six.
Bayesian inference of phylogeny uses a likelihood function to create a quantity called the posterior probability of trees using a model of evolution (substitution model), based on some prior probabilities (priors), producing the most likely phylogenetic tree for the given data.

\[
\Pr(H|D) = \frac{\text{Likelihood} \cdot \text{Prior}}{\Pr(D)} = \frac{\Pr(D|H) \cdot \Pr(H)}{\Pr(D)}
\]

\(\Pr(D)\) is not possible to calculate as this is the \(\sum_H \Pr(D|H) \cdot \Pr(H)\). Too many different hypothesis.

- the hypothesis \(H\) is a combination of topology of branches, branch length and parameter of the substitution model

- we may approximate the posterior distribution for \(H\) using Markov Chain Monte Carlo (MCMC) methods
Bayesian analysis step-by-step:
- 4 chains
- 3D space (area) with all possible trees
- find (built) first tree, compute likelihood (L)
- second tree, compute L
- if L is better, jump to the second tree, if not, stay with the first one

Bayesian analysis
higher likelihood
local maximum
global maximum
Two types of chains:
**Cold** – conservative one, can jump only upwards, if finds better L value
**Warm** – three chains – can jump also downwards + jump accidentally + call cold one if find better topology

![Diagram showing higher likelihood and local maximum](image)

- Global maximum
- Local maximum

![Diagram showing higher likelihood](image)
If there are enough generations (i.e. search steps) cold chain finds the highest global $L$. 

- If there are enough generations (i.e. search steps) cold chain finds the highest global $L$. 

Diagram:

- Global maximum
- Local maximum
- Higher likelihood

Diagram image shows a landscape with multiple peaks, indicating the concept of global and local maxima in a likelihood function.
MrBayes run

- Output of **MrBayes** is file with all trees found by cold chain during the procedure. Usually every 100th tree from millions generation is saved.
- Usually we have two runs.

Trees at the beginning of run are not OK – we have to cut them (burnin)
**Posterior probability**

BPP (PP) is parameter of Bayesian analysis – instead of bootstraps

- BPP: represent the probability that the corresponding clade is true conditional on the model, the priors, and the data
- **below 0.95 – 0.9 topology is considered unreliable**
<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parsimony methods</strong></td>
<td></td>
</tr>
<tr>
<td>Simplicity and intuitive appeal</td>
<td>Assumptions are implicit and poorly understood</td>
</tr>
<tr>
<td>The only framework appropriate for some data (such as SINES and LINES)</td>
<td>Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution</td>
</tr>
<tr>
<td></td>
<td>Branch lengths are substantially underestimated when substitution rates are high</td>
</tr>
<tr>
<td></td>
<td>Maximum parsimony may suffer from long-branch attraction</td>
</tr>
<tr>
<td><strong>Distance methods</strong></td>
<td></td>
</tr>
<tr>
<td>Fast computational speed</td>
<td>Most distance methods, such as neighbour joining, do not consider variances of distance estimates</td>
</tr>
<tr>
<td>Can be applied to any type of data as long as a genetic distance can be defined</td>
<td>Distance calculation is problematic when sequences are divergent and involve many alignment gaps</td>
</tr>
<tr>
<td>Models for distance calculation can be chosen to fit data</td>
<td>Negative branch lengths are not meaningful</td>
</tr>
<tr>
<td><strong>Likelihood methods</strong></td>
<td></td>
</tr>
<tr>
<td>Can use complex substitution models to approach biological reality</td>
<td>Maximum likelihood iteration involves heavy computation</td>
</tr>
<tr>
<td>Powerful framework for estimating parameters and testing hypotheses</td>
<td>The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret</td>
</tr>
<tr>
<td><strong>Bayesian methods</strong></td>
<td></td>
</tr>
<tr>
<td>Can use realistic substitution models, as in maximum likelihood</td>
<td>Markov chain Monte Carlo (MCMC) involves heavy computation</td>
</tr>
<tr>
<td>Prior probability allows the incorporation of information or expert knowledge</td>
<td>In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify</td>
</tr>
<tr>
<td>Posterior probabilities for trees and clades have easy interpretations</td>
<td>Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge</td>
</tr>
<tr>
<td></td>
<td>Posterior probabilities often appear too high</td>
</tr>
<tr>
<td></td>
<td>Model selection involves challenging computation\textsuperscript{138,139}</td>
</tr>
</tbody>
</table>
Tree visualization:

nuclear and mitochondrial DNA maximum likelihood

0.05 changes
Tree visualization:
- Newick format

(A, (B, (C, (D, E))))

Different programs for tree visualization: TreeView, FigTree, Dendroscope
Take -Home Message!

- there are more methods how to calculate tree
- a phylogenetic tree is a hypothesis
- we have to test the reliability
- obtaining a good alignment is one of the most crucial steps towards a good phylogenetic tree

Software:

MP: PAUP*, TNT, Phylip, MEGA, …
ML: PAUP*, PHYML, GARLI, RAXML, Phylip, MEGA,…
BA: MrBayes
NJ: PAUP*, Phylip, MEGA, …
Table 1 | Functionalities of a few commonly used phylogenetic programs

<table>
<thead>
<tr>
<th>Name</th>
<th>Brief description</th>
<th>Link</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian evolutionary analysis sampling trees (BEAST)</td>
<td>A Bayesian MCMC program for inferring rooted trees under the clock or relaxed-clock models. It can be used to analyse nucleotide and amino acid sequences, as well as morphological data. A suite of programs, such as Tracer and FigTree, are also provided to diagnose, summarize and visualize results</td>
<td><a href="http://beast.bio.ed.ac.uk">http://beast.bio.ed.ac.uk</a></td>
<td>135</td>
</tr>
<tr>
<td>Genetic algorithm for rapid likelihood inference (GARLI)</td>
<td>A program that uses genetic algorithms to search for maximum likelihood trees. It includes the GTR + Γ model and special cases and can analyse nucleotide, amino acid and codon sequences. A parallel version is also available</td>
<td><a href="http://code.google.com/p/garli">http://code.google.com/p/garli</a></td>
<td>55</td>
</tr>
<tr>
<td>Hypothesis testing using phylogenies (HYPHY)</td>
<td>A maximum likelihood program for fitting models of molecular evolution. It implements a high-level language that the user can use to specify models and to set up likelihood ratio tests</td>
<td><a href="http://www.hyphy.org">http://www.hyphy.org</a></td>
<td>136</td>
</tr>
<tr>
<td>Molecular evolutionary genetic analysis (MEGA)</td>
<td>A Windows-based program with a full graphical user interface that can be run under Mac OSX or Linux using Windows emulators. It includes distance, parsimony and likelihood methods of phylogeny reconstruction, although its strength lies in the distance methods. It incorporates the alignment program ClustalW and can retrieve data from GenBank</td>
<td><a href="http://www.megasoftware.net">http://www.megasoftware.net</a></td>
<td>37</td>
</tr>
<tr>
<td>MrBayes</td>
<td>A Bayesian MCMC program for phylogenetic inference. It includes all of the models of nucleotide, amino acid and codon substitution developed for likelihood analysis</td>
<td><a href="http://mrbayes.net">http://mrbayes.net</a></td>
<td>71</td>
</tr>
<tr>
<td>Phylogenetic analysis by maximum likelihood (PAML)</td>
<td>A collection of programs for estimating parameters and testing hypotheses using likelihood. It is mostly used for tests of positive selection, ancestral reconstruction and molecular clock dating. It is not appropriate for tree searches</td>
<td><a href="http://abacusgene.ucl.ac.uk/software">http://abacusgene.ucl.ac.uk/software</a></td>
<td>137</td>
</tr>
<tr>
<td>Phylogenetic analysis using parsimony* and other methods (PAUP* 4.0)</td>
<td>PAUP* 4.0 is still a beta version (at the time of writing). It implements parsimony, distance and likelihood methods of phylogeny reconstruction</td>
<td><a href="http://www.sinauer.com/detail.php?id=8060">http://www.sinauer.com/detail.php?id=8060</a></td>
<td></td>
</tr>
<tr>
<td>PhyML</td>
<td>A fast program for searching for the maximum likelihood trees using nucleotide or protein sequence data</td>
<td><a href="http://www.atgc-montpellier.fr/phyml/binaries.php">http://www.atgc-montpellier.fr/phyml/binaries.php</a></td>
<td>53</td>
</tr>
<tr>
<td>RAxML</td>
<td>A fast program for searching for the maximum likelihood trees under the GTR model using nucleotide or amino acid sequences. The parallel versions are particularly powerful</td>
<td><a href="http://scoh-its.org/exelixis/software.html">http://scoh-its.org/exelixis/software.html</a></td>
<td>54</td>
</tr>
<tr>
<td>Tree analysis using new technology (TNT)</td>
<td>A fast parsimony program intended for very large data sets</td>
<td><a href="http://www.zmuc.dk/public/phylogeny/TNT">http://www.zmuc.dk/public/phylogeny/TNT</a></td>
<td>42</td>
</tr>
</tbody>
</table>

Note: all programs can run on Windows, Mac OSX and Unix or Linux platforms. Except for PAUP*, which charges a nominal fee, all packages are free for download. See Felsenstein's comprehensive list of programs at http://evolution.genetics.washington.edu/phylip/software.html. GTR, general time reversible; MCMC, Markov chain Monte Carlo.
But what to do if tree does not look „good“?

- add some more genes/sequences
- add some taxa
More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy

Antonis Rokas and Sean B. Carroll

- majority of taxa should have the most complete dataset!
How to estimate appropriate number of taxa?

- Software for measuring of phylogenetic representativeness

Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies

Federico Plazzi*¹, Ronald R Ferrucci² and Marco Passamonti¹
But what to do if the tree does not look „good“?

- add some taxa
- add some more genes/sequences
- change alignment parameters – the most important
- different model of sequence evolution for each gene
- model of sequence evolution can be different for all position in coding genes (COI, EF, Wg…)
- RY coding – R for purines (A,G), Y for pyrimidines (C,T)
  → partitioning analysis
- do not overpartition (Partition Finder v1.1.0 – Lanfear et al. 2012)
But what to do if the tree does not look „good“?

- add some taxa
- add some more genes/sequences
- change alignment parameters – the most important (contamination, „strange“ taxa)
- different model of sequence evolution for each gene
- model of sequence evolution can be different for all position in coding genes (COI, EF, Wg…)
- RY coding – R for purines (A,G), Y for pyrimidines (C,T)
- do not overpartition

- knowledge of secondary structure
  (compare with Munro et al. 2011
   - use MAFFT align. strategy)

- do more than one method
  (usually BA and ML (MP))
The CIPRES Science Gateway V. 3.1

The CIPRES Science Gateway V. 3.1 is a public resource for inference of large phylogenetic trees. It is designed to provide all researchers with access to large computational resources of the NSF TeraGrid through a simple browser interface. The CIPRES Science Gateway provides new hybrid parallel versions of RAxML (7.2.7) and MrBayes (3.1.2), as well as parallel GARLI (1.0) code to insure the fastest possible run times for submitted jobs. Through a collaboration with Alexandros Stamatakis and Wayne Pfeiffer, we now offer the fastest hybrid versions of RAxML [pdf] and MrBayes [pdf] currently available.

Use the CIPRES Science Gateway

High Performance Parallel Codes for Large Tree Inference on TeraGrid:
RAxML (7.2.7): MrBayes (3.1.2): GARLI (1.0)

High Performance Parallel Codes for Sequence Alignment on TeraGrid:
MAFFT (6.822)

Serial Codes for Tree Inference:

PAUP* (Parsimony); Poy (Simultaneous Sequence Alignment and Tree Inference).

Serial Codes for Sequence Alignment:
ClustalW, ContraLign, FSA, MUSCLE, PROCONS, PROBALIGN.

Learn more about: Requirements; Limitations; Architecture; Known Issues

CIPRES Portal V.1.15 and CIPRES Portal V.2.2 have now ceased operations.

http://www.phylo.org/sub_sections/portal/

Miller et al. 2010
Why use of molecular phylogetics in zoology?

- Phylogeny of different groups of taxa
- Definition of species boundary – use in taxonomy, cryptic species detection, character mapping and comparison
- Biodiversity implications
- Biogeography
- Conservation biology
- Disease prediction (Ebola, honey-bee pathogens, resistance etc.)

...
Phylogeny of Eucaryota (and Procaryota)
Adl et al. 2005, 2012 – review (several genes – genomes)

Derelle et al. PNAS 2015;112:E693-E699
Phylogeny of mammals

- 447 orthologous genes

Song et al. 2012 – PNAS
Closest relatives to primates

The diagram illustrates the evolutionary relationships among various groups, including primates, with a focus on identifying the closest relatives to primates. The sister group of primates is highlighted, and the diagram references the work of Janečka et al., Science 2007.
Phylogenetic analyses based on DNA data clarified the evolutionary relationships between humans and other primates.

- Darwin was the first to speculate on evolutionary relationships between humans and other primates.

- In 1960 from fossils, a paleontologist concluded that chimps and gorillas are our closest relatives and that the split occurred 15 MYA.

- Different molecular data put this split as much more recent - around 5 MYA.
Humans, chimpanzees, and bonobos are more closely related to one another than either is to gorillas or any other primate.

**Comparison of genomes:** humans and chimpanzees shared a common ancestor ∼5-7 MYA. The difference between the two genomes is ∼4%—comprising ∼35 million single nucleotide differences and ∼90 Mb of insertions and deletions.
Integrative taxonomy

• only 14–75% of estimated planet’s biodiversity is described (Mora et al. 2011, Costello, May & Sork 2013)

• limitation of morphological x molecular taxonomy

• integrative taxonomy (at first molecules and then morphology)
Cryptic species diversity in *Hemiphyllodactylus* geckos

- Previously known only 8 species and some subspecies, same appearance, loss of good diagnostic characters.

*Zoological Journal of the Linnean Society, 2013, 169, 849–880. With 9 figures*

Integrative taxonomy uncovers high levels of cryptic species diversity in *Hemiphyllodactylus* Bleeker, 1860 (Squamata: Gekkonidae) and the description of a new species from Peninsular Malaysia

L. Lee Grismer\(^{1,2}\), Perry L. Wood Jr\(^{3}\), Shahrul Anuar\(^{4}\), Mohd Abdul Muin\(^{5}\), Evan S. H. Quah\(^{6}\), Jimmy A. McGuire\(^{7,8}\), Rafe M. Brown\(^{9,10}\), Ngo Van Tri\(^{11}\) and Pham Hong Thai\(^{12}\)
Divergency plot – usually 18-30% in ND gene

Table 6. Uncorrected p-distances for the major lineages of the genus *Hemiphyllo daemonius* Bleeker, 1860 computed in MEGA v5.1 (Tamura, 2011)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N/A</td>
<td>0.318</td>
<td>0.304</td>
<td>0.243</td>
<td>0.294</td>
<td>0.265</td>
<td>0.316</td>
<td>0.251</td>
<td>0.290</td>
<td>0.039</td>
<td>0.290</td>
<td>0.288</td>
<td>0.259</td>
<td>0.270</td>
<td>0.275</td>
<td>0.272</td>
<td>0.272</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.286</td>
<td>0.303</td>
<td>0.293</td>
<td>0.252</td>
<td>0.269</td>
<td>0.316</td>
<td>0.285</td>
<td>0.287</td>
<td>0.046</td>
<td>0.290</td>
<td>0.290</td>
<td>0.275</td>
<td>0.275</td>
<td>0.275</td>
<td>0.275</td>
<td>0.275</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.290</td>
<td>0.283</td>
<td>0.288</td>
<td>0.301</td>
<td>0.285</td>
<td>0.265</td>
<td>0.186</td>
<td>0.290</td>
<td>0.099</td>
<td>0.181</td>
<td>0.209</td>
<td>0.288</td>
<td>0.278</td>
<td>0.278</td>
<td>0.278</td>
<td>0.278</td>
<td>0.278</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.106</td>
<td>0.137</td>
<td>0.173</td>
<td>0.280</td>
<td>0.283</td>
<td>0.193</td>
<td>0.015</td>
<td>0.283</td>
<td>0.283</td>
<td>0.015</td>
<td>0.283</td>
<td>0.283</td>
<td>0.283</td>
<td>0.283</td>
<td>0.283</td>
<td>0.283</td>
<td>0.283</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.287</td>
<td>0.288</td>
<td>0.301</td>
<td>0.285</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.285</td>
<td>0.287</td>
<td>0.300</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.287</td>
<td>0.288</td>
<td>0.300</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.287</td>
<td>0.288</td>
<td>0.300</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.287</td>
<td>0.288</td>
<td>0.300</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.287</td>
<td>0.288</td>
<td>0.300</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.289</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.315</td>
<td>0.160</td>
<td>0.150</td>
<td>0.251</td>
<td>0.277</td>
<td>0.268</td>
<td>0.269</td>
<td>0.268</td>
<td>0.269</td>
<td>0.266</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.268</td>
<td>0.015</td>
<td></td>
</tr>
</tbody>
</table>

Distances set in bold are intraspecific distances, and distances below the diagonal are interspecific distances.
Biogeography

Out of Australia and back again: the world-wide historical biogeography of non-pollinating fig wasps (Hymenoptera: Sycophaginae)

Astrid Cruaud¹*, Roula Jabbour-Zahab¹, Gwenaëlle Genson¹, Arnaud

- Fossil dating (*Idarnes* from Dominican Amber – 30-15My; endemic taxa to Mauritius (8My) and Solomon Islands (11-12My)
- Several genes, very well resolved topology
- *Ficus* origin 100-60My, *Ficus* pollinator origin 70-15My
- Sycophaginae origin (48-35My)
Out of Australia and back again!