Phylogenetics in the Age of Genomics: Prospects and Challenges
Antonis Rokas
Department of Biological Sciences, Vanderbilt University
http://as.vanderbilt.edu/rokaslab

“The Affinities of all the Beings [...] Represented by a Great Tree”

http://darwin-online.org.uk/
What is a Phylogenetic Tree?

A phylogenetic tree is the mathematical structure used to depict the evolutionary history of a group of organisms or genes.

Phylogenetic trees show historical relationships, not similarities.

Why are Phylogenetic Trees Useful?

Benefits to Science:
- The origin and history of life
- Evolution of molecules, phenotypes and developmental mechanisms

Benefits to Society:
- Human health (identification of disease agents & reservoirs)
- Agriculture (crops’ wild relatives)
- Biodiversity (conservation strategies)
Phylogenetics, Rokas (c) 2011

CSI: Phylogeny

- A gastroenterologist was accused of second-degree murder for injecting his former girlfriend with blood obtained from an HIV type 1 (HIV-1)-infected patient under his care.
- Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in the court.

Metzker et al. (2002) PNAS

The DNA Record

Glacial ice

Tundra and cold grassland

Woody vegetation

Horse (four digits lost; other bones fused)
Why Use the DNA Record?

- Genetic basis of characters known
- Larger amount of phylogenetic information
- Any groups of organisms can be compared
- Different rates of evolution
- Mathematical modelling
- Character delimitation is straightforward
- DNA is the ultimate level of information

The DNA Record is a Living Chronicle of Evolution

- Information in genomes is vital to reconstructing the processes and patterns of evolution
- Knowledge of evolution is a powerful guide to interpreting genomes
Fungi: a Model for Comparative Functional Genomics

**Medicine & Industry**
- Candidiasis (yeast infections)
- Aspergillosis
- Antibiotics (e.g. penicillin, lovastatin)
- Food (e.g. beer, wine, cheese)

**Model Organisms**
- Small genome size & high gene density
- Tractable systems for molecular and genomic studies

**Evolution & Ecology**
- Rich ecology and evolutionary history
- Vital in function of ecosystems (as decomposers)

---

Fungi: the Most Sequenced Eukaryotic Lineage

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
<th>Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protists</td>
<td>14%</td>
<td>34</td>
</tr>
<tr>
<td>Plants</td>
<td>9%</td>
<td>22</td>
</tr>
<tr>
<td>Metazoa</td>
<td>35%</td>
<td>87</td>
</tr>
<tr>
<td>Fungi</td>
<td>42%</td>
<td>103</td>
</tr>
</tbody>
</table>

Data from GOLD 3.0 (www.genomesonline.org) March 2011
“... there is, after all, one true tree of life [...]. It exists. It is in principle knowable. We don’t know it all yet. By 2050 we should – or if we do not, we shall have been defeated only at the terminal twigs, by the sheer number of species.”

Richard Dawkins

**The Problem of Incongruence**

Gene X

```
   a
  / 
 b   c
```

Gene Y

```
   a
  / 
 c   b
```

Species tree?

- **A: All organisms**
  - 39%
- **B: Mammals**
  - 35%
- **C: Insects**
  - 48%

*Incongruence is pervasive in the phylogenetics literature*
A Systematic Evaluation of Single Gene Phylogenies

Dataset: 106 genes on all 16 chromosomes totaling 127kb corresponding roughly to 1% of the genomic sequence, 2% of genes

Analyses: Maximum Likelihood (ML) & Maximum Parsimony (MP) on nt data sets and MP on amino acid data sets

Incongruence in Shallow Time

ML / MP

Rokas et al. (2003) Nature

Kellis et al. (2003) Nature

Plainly stated, taxonomists keep digging the same hole for themselves and falling down it; all that has changed, over the years, is the sophistication of the shovel

**Concatenation of 106 Genes Yields a Single Yeast Phylogeny**

- ML / MP on nt
- MP on aa

![Graph showing phylogenetic accuracy with increasing gene number](image)

**Phylogenetic Accuracy is Positively Correlated with Gene Number**

- Yeasts, 106 genes from 14 taxa
- Mammals, 20 genes from 42 taxa
- Angiosperms, 10 genes from 16 taxa

![Graphs showing phylogenetic accuracy with taxon number](image)

**References**

- Rokas et al. (2003) Nature
- Murphy et al. (2001) Science
- Zanis et al. (2002) PNAS
Assessing Robustness in Inference: Bootstrap

- Characters are sampled with replacement to create many bootstrap replicate data sets equal in size to the original one.
- Each bootstrap replicate data set is analysed (e.g., with parsimony, likelihood).
- Agreement among the resulting trees is summarized with a majority-rule consensus tree.
- The frequency of occurrence of clades, also known as bootstrap values, is a measure of support for those groups.

seq1: AAAATCGTGGGG
seq2: ATAATGGTGGCG
seq3: AAAATGGTGGCG
seq4: ATAATGGTGGCG

1234567...
Methods of Assessing Robustness are Subject to Systematic Error


An Example of Systematic Error in the Yeast Clade

Scannell et al. (2006) Nature
Gene Trees Can Differ from Species Trees


Inferring the Species Tree from Individual Gene Histories

Concordance Factor: The proportion of the genome for which a clade is true

S. cerevisiae
S. paradoxus
S. mikatae
S. kudriavzevii
S. bayanus
S. castellii
S. kluyveri
C. albicans

STEM: Maximum likelihood estimation of species trees
Kubatko et al. (2009) Bioinformatics

BEST: Bayesian estimation of species trees
Liu (2008) Bioinformatics

Lack of Resolution Due to Lack of Data or to Radiation?

**Lack of Resolution Among Most Metazoan Phyla**

Rokas et al. (2005) Science

50 genes, ML/MP

**How Can We Explain this Lack of Resolution?**

Cambrian Explosion Still Troubling to Evolutionists – 12/22/2005

Today’s New Reason To Believe – 01/24/2006
Hypothesis I: Mutational Saturation
the phylogenetic signal originally contained in these protein sequences has been erased by multiple substitutions

Hypothesis II: Evolutionary Radiation
the close spacing of cladogenetic events early in the history of metazoans limits their resolution
### Time of Origin of Metazoa & Fungi

1) **Fossils**

- **Metazoa:** Xiao et al. (1998) Nature
- **Fungi:** Yuan et al. (2005) Science

2) **Relative molecular clocks**

<table>
<thead>
<tr>
<th>Study</th>
<th>Metazoa</th>
<th>Fungi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Douzery et al. (2004) PNAS</td>
<td>849 Myr</td>
<td>727 Myr</td>
</tr>
<tr>
<td>Hedges et al. (2004) BMC Evol. Biol.</td>
<td>976 Myr</td>
<td>968 Myr</td>
</tr>
<tr>
<td>Heckman et al. (2001) Science</td>
<td>1177 Myr</td>
<td>1208 Myr</td>
</tr>
</tbody>
</table>

3) The way this set of 50 genes has evolved across the two Kingdoms is remarkably similar.

---

### A Remarkable Contrast in Phylogenetic Resolution

- **Human**
- **Mouse**
- **Zebrafish**
- **Tunicate**
- **Protostomida**
- **Anthropod**
- **Turbellaria platyhelminth**
- **Nematoda**
- **Mollusc**
- **Armillid**
- **Arthropoda oxidari**
- **Hydrozoan**
- **Hemicholinoid portlerian**
- **Calcarea portlerian**
- **Demoseponge**
- **Choanoflagellate**
- **Neurospora crassa**
- **Plasmodium**
- **Mycoplasma**
- **Candida**
- **Sacccharomyces cerevisiae**
- **Sacccharomyces bayanus**
- **Rabidobacter**
- **Ustilago maydis**
- **Pseudomonas**
- **Catenarius olivarius**
- **Phanerococcales chrysoasporium**
- **Cryptococcosis neoformans**

**Given:**
- lack of resolution in Metazoa (using lots of genes)
- contrast in resolution between Metazoa and Fungi (using identical genes)

*The early history of animals is best viewed as an evolutionary radiation*
Rapid Tempo of Cladogenesis Early in Metazoan History

<table>
<thead>
<tr>
<th>ProCambrian</th>
<th>Palaeozoic</th>
<th>Mesozoic</th>
<th>Cenozoic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendian</td>
<td>C</td>
<td>O</td>
<td>S</td>
</tr>
<tr>
<td>Cnidaria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Porifera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mollusca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brachiopoda</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctenophora</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pteropoda</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ophiuroidea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anthozoa</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annelida</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Echinodermata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chordata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemichordata</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tardigrada</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Incongruence in Deep Time

- Schierwater et al., 2009
- Philippe et al., 2009
- Dunn et al., 2008

Early-branching metazoans

Protostomes

Deuterostomes

Matus et al., 2006
Marklaz et al., 2006
Delsuc et al., 2006
Bourlat et al., 2006 (nuclear)
Bourlat et al., 2006 (mitochondrion)

What Makes Animals Hard to Resolve But Yeasts Easy?

Internode length: influences amount of phylogenetic signal (I)

Homoplasy: independent evolution of identical characters (•,•)
Homoplasy

the independent evolution of identical character states (e.g., amino acid residues) in different branches of a phylogenetic tree that are not directly inherited from a common ancestor

Quantifying Homoplasy Across the Tree of Life

Saccharomyces yeasts (106)

Aspergillus ascomycetes (200)

Fungal phyla (200)

Eukaryotic phyla (200)

Land plants (39)

Drosophila fruit-flies (200)

Cetartiodactyl mammals (12)

Metazoan phyla (239)

Paenungulata mammals (12)

Vertebrates (44)

Measuring Homoplasy Across the Tree of Life


<table>
<thead>
<tr>
<th>Pattern</th>
<th>Observed sites</th>
<th>Observed substitutions (observed sites x substitutions)</th>
<th>Expected sites</th>
<th>Expected substitutions (expected sites x substitutions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABBB</td>
<td>49</td>
<td>49 x 1 = 49</td>
<td>37</td>
<td>37 x 1 = 37</td>
</tr>
<tr>
<td>BABB</td>
<td>67</td>
<td>67 x 2 = 134</td>
<td>25.5</td>
<td>25.5 x 2 = 51</td>
</tr>
<tr>
<td>AABC</td>
<td>171</td>
<td>171 x 2 = 342</td>
<td>185.5</td>
<td>185.5 x 2e 371</td>
</tr>
<tr>
<td>BCAA</td>
<td>34</td>
<td>34 x 3 = 102</td>
<td>40</td>
<td>40 x 3 = 120</td>
</tr>
</tbody>
</table>

Observed Homoplasy = \( \frac{\text{ABAB} + \text{ABBA} \text{ observed substitutions}}{\text{Total number of observed substitutions}} \) = \( \frac{134}{1087} \) = 12.3%

Expected Homoplasy = \( \frac{\text{ABAB} + \text{ABBA} \text{ expected substitutions}}{\text{Total number of expected substitutions}} \) = \( \frac{51}{1035} \) = 5.0%
## An Overabundance of Homoplastic Substitutions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Obs(H)</th>
<th>Exp(H)</th>
<th>Obs(H)/Exp(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccharomyces yeasts</td>
<td>4.7%</td>
<td>2.0%</td>
<td>2.4</td>
</tr>
<tr>
<td>Aspergillus filamentous fungi</td>
<td>6.9%</td>
<td>2.5%</td>
<td>2.8</td>
</tr>
<tr>
<td>Fungal phyla</td>
<td>6.7%</td>
<td>3.5%</td>
<td>1.9</td>
</tr>
<tr>
<td>Eukaryote phyla</td>
<td>7.5%</td>
<td>3.5%</td>
<td>2.1</td>
</tr>
<tr>
<td>Plants</td>
<td>2.3%</td>
<td>0.9%</td>
<td>2.7</td>
</tr>
<tr>
<td><em>Drosophila</em> fruit-flies</td>
<td>2.6%</td>
<td>1.3%</td>
<td>2.0</td>
</tr>
<tr>
<td>Cetartiodactyls mammals</td>
<td>12.3%</td>
<td>4.9%</td>
<td>2.5</td>
</tr>
<tr>
<td>Paenungulata mammals</td>
<td>10.4%</td>
<td>5.0%</td>
<td>2.1</td>
</tr>
<tr>
<td>Metazoan phyla</td>
<td>7.4%</td>
<td>3.3%</td>
<td>2.3</td>
</tr>
<tr>
<td>Vertebrates</td>
<td>10.2%</td>
<td>3.2%</td>
<td>3.2</td>
</tr>
<tr>
<td><strong>Average:</strong></td>
<td>7.1%</td>
<td>2.3%</td>
<td>2.4</td>
</tr>
</tbody>
</table>


## Excess Homoplasy is Specific to Homoplastic Substitutions

<table>
<thead>
<tr>
<th>Clade</th>
<th>Parsimony-informative sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs(H)</td>
</tr>
<tr>
<td>Saccharomyces yeasts</td>
<td>3.3%</td>
</tr>
<tr>
<td>Aspergillus filamentous ascomycetes</td>
<td>10.3%</td>
</tr>
<tr>
<td>Fungal phyla</td>
<td>2.5%</td>
</tr>
<tr>
<td>Eukaryotic phyla</td>
<td>2.3%</td>
</tr>
<tr>
<td>Land plants</td>
<td>31.6%</td>
</tr>
<tr>
<td><em>Drosophila</em> fruit-flies</td>
<td>10.2%</td>
</tr>
<tr>
<td>Cetartiodactyls mammals</td>
<td>4.5%</td>
</tr>
<tr>
<td>Metazoan phyla</td>
<td>3.1%</td>
</tr>
<tr>
<td>*Paenungulata mammals</td>
<td>5.4%</td>
</tr>
<tr>
<td>*Vertebrates</td>
<td>2.9%</td>
</tr>
<tr>
<td><strong>Average:</strong></td>
<td>7.6%</td>
</tr>
</tbody>
</table>

*Excess Homoplasy is Specific to Homoplastic Substitutions*
Homoplasy Stems From Frequently Exchanged Amino Acids

Aspergillus filamentous ascomycetes

Cetartiodactyl mammals

190 possible interchanges among 20 amino acids
- 75 can be achieved via a single nucleotide substitution
- The other 115 require two or three substitutions

65% of observed interchanges is between the top 12 most frequently observed amino acid interchanges a single mutational step away

Conservative AA Substitutions Are Very Common in Alignments

GAC/T (D)
Aspartic Acid

GAA/G (E)
Glutamic Acid

What Are the Limits to Resolution?

For internodes in ~ 600 my animal phylogeny

<table>
<thead>
<tr>
<th>Time</th>
<th>Amount of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 my</td>
<td>700 variable sites</td>
</tr>
<tr>
<td>10 my</td>
<td>2,800 variable sites</td>
</tr>
<tr>
<td>1 my</td>
<td>28,000 variable sites</td>
</tr>
<tr>
<td>0.1 my</td>
<td>??????? variable sites</td>
</tr>
<tr>
<td>1 y</td>
<td>??????? variable sites</td>
</tr>
</tbody>
</table>
What If Mammals Had Diversified in the Cambrian?

Springer et al. (2003) PNAS

107 Myr tree
600 Myr tree
Phylogenetic Accuracy is Inversely Correlated with Elapsed Time

Failure to Resolve Internodes < 10 Million Years

Rokas et al. (2005) Science
**Why Are There So Few Bushes?**

Data in, fully-resolved phylogenetic tree out

**Buses in the Tree of Life**

**A**
The human / chimp / gorilla tree (5 - 8 million years ago)

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimpanzee</th>
<th>Orang-utan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>55% (64)</td>
<td>19% (18)</td>
<td>20% (20)</td>
</tr>
<tr>
<td>pi-characters</td>
<td>76% (564)</td>
<td>11% (130)</td>
<td>13% (143)</td>
</tr>
<tr>
<td>RSCs</td>
<td>80% (7)</td>
<td>12% (1)</td>
<td>0% (0)</td>
</tr>
</tbody>
</table>

**B**
The elephant / sirenian / hyrax bush (57 - 65 million years ago)

<table>
<thead>
<tr>
<th></th>
<th>Elephant</th>
<th>Sirenian</th>
<th>Hyrax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear genes (20)</td>
<td>5% (1)</td>
<td>30% (6)</td>
<td>10% (2)</td>
</tr>
<tr>
<td>Nuclear pi-characters (19)</td>
<td>21% (4)</td>
<td>63% (12)</td>
<td>26% (5)</td>
</tr>
<tr>
<td>Mitochondrial genes (12)</td>
<td>33% (4)</td>
<td>17% (2)</td>
<td>17% (2)</td>
</tr>
<tr>
<td>Mitochondrial pi-characters (127)</td>
<td>36% (46)</td>
<td>30% (30)</td>
<td>34% (43)</td>
</tr>
<tr>
<td>RSCs (1)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>100% (1)</td>
</tr>
</tbody>
</table>

**C**
The coelacanth / lungfish / tetrapod bush (370 - 390 million years ago)

<table>
<thead>
<tr>
<th></th>
<th>Tetrapod</th>
<th>Lungfish</th>
<th>Coelacanth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes (44)</td>
<td>36% (16)</td>
<td>27% (12)</td>
<td>34% (15)</td>
</tr>
<tr>
<td>pi-characters (234)</td>
<td>31% (22)</td>
<td>33% (96)</td>
<td>36% (106)</td>
</tr>
<tr>
<td>RSCs</td>
<td>0% (0)</td>
<td>100% (1)</td>
<td>0% (0)</td>
</tr>
</tbody>
</table>

**D**
The metazoan superbush (> 559 million years ago)

<table>
<thead>
<tr>
<th></th>
<th>Study I</th>
<th>Study II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes (507)</td>
<td>35% (175)</td>
<td>8% (39)</td>
</tr>
<tr>
<td>RSCs (84)</td>
<td>57% (48)</td>
<td>18% (15)</td>
</tr>
<tr>
<td>25% (21)</td>
<td>32% (320)</td>
<td>25% (254)</td>
</tr>
<tr>
<td>43% (434)</td>
<td>0% (0)</td>
<td>43% (434)</td>
</tr>
</tbody>
</table>

---

The Human / Chimp / Gorilla Tree

Inform. Sites

- Human 8561 / 11293
- Chimpanzee
- Gorilla
- Human 1302 / 11293
- Gorilla
- Chimpanzee
- Human
- Chimpanzee 1430 / 11293
- Gorilla

Patterson et al. (2006) Nature

Bushes in the Mammalian Tree

Inform. Sites

- Tree 1 (basal-Arhotharia) 22
- Boreotheria (Luwasa)
- Xenarthra (South America)
- Atolimia (Afroas)
- Xenarthra (South America)
- Atolimia (Afroas)
- Xenarthra (South America)
- Atolimia (Afroas)
- Boreotheria (Luwasa)

Nishihara et al. (2009) PNAS
The three major lineages first appeared within 20 – 30 million years ago, approximately 390 million years ago.


44 genes, ML/MP/NJ

Mind the Gap Between Real Data and Models

“One can use the most sophisticated audio equipment to listen, for an eternity, to a recording of white noise and still not glean a useful scrap of information”

Rodrigo et al. (1994) Chapter in: Sponge in Time and Space; Biology, Chemistry, Paleontology
Knowing that strikingly different groups form a clade and that the time spans between the branchings of these groups must have been very short, makes the knowledge of the branching order among groups potentially a secondary concern.

Acknowledgements

Wisconsin-Madison
Sean Carroll
Nicole King
Barry Williams
Cecile Ane
Bret Larget
David Baum

http://as.vanderbilt.edu/rokaslab