The Animal Phylogeny
and the Fundamental Importance of Taxon Sampling

Hervé Philippe, Henner Brinkmann, Denis Baurain

February 20\textsuperscript{th} 2006
The New Animal Phylogeny

71 slow-evolving genes x 37 animal taxa (20,705 sites) ; ML tree

The New Animal Phylogeny

71 slow-evolving genes x 37 animal taxa (20,705 sites); ML tree

Lack of resolution among most metazoan phyla

threshold = 70%

50 genes x 17 animal taxa (12,060 sites); ML/MP tree

Rokas et al. (2005) Science 310:1933-1938
Lack of resolution among most metazoan phyla

threshold = 70%

Rokas et al. (2005) Science 310:1933-1938

50 genes x 17 animal taxa (12,060 sites) ; ML/MP tree
Possible causes for the lack of resolution

- **long branch attraction** artifacts
  - removal of nematods and platyhelminths
- **presence of rogue taxa**
  - computation of leaf stability indices (bootstrap-based)
  - removal of poriferans and cnidarians
- **deviations in amino-acid composition**
  - removal of 6 scattered taxa violating the homogeneity assumption
  - computation of a LogDet NJ tree

No subset of taxa seems to decrease the resolution.
Effect of the removal of 3 long-branched taxa

ML tree; BS from ML/MP trees

<table>
<thead>
<tr>
<th>nd</th>
<th>clade</th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Chordata</td>
<td>96/60</td>
<td>94/66</td>
</tr>
<tr>
<td>6</td>
<td>Bilateria</td>
<td>74/32</td>
<td>100/72</td>
</tr>
<tr>
<td>7</td>
<td>mollusks + annelids</td>
<td>93/50</td>
<td>97/78</td>
</tr>
<tr>
<td>8</td>
<td>Protostomia</td>
<td>62/27</td>
<td>100/72</td>
</tr>
<tr>
<td>9</td>
<td>priapulids + arthropods</td>
<td>79/07</td>
<td>69/35</td>
</tr>
<tr>
<td>13</td>
<td>Cnidaria</td>
<td>57/68</td>
<td>55/63</td>
</tr>
</tbody>
</table>

supplementary data of Rokas et al. (2005) Science 310:1933-1938
Investigating the apparent systemic lack of resolution

- **amount of missing data**
  - design target: 20%
  - removal of priapulid (68%) and mollusk (54%)
- **total amount of data** (12,060 sites)
  - 56% of sites are variable
  - 31% are parsimony-informative
- **distribution** of informative sites among topologies
  - **likelihood mapping** (no resolution for diploblasts)
  - **resampling** up to 100,000 sites (slightly better support)

Nothing seems to explain the lack of resolution.
The rationale for comparing Metazoa and Fungi

- aim is to distinguish between 2 hypotheses
  - mutational saturation has erased phylogenetic signal
  - lack of resolution as a signature of closely spaced cladogenetic events in early animal evolution

- Why Fungi?
  - sistergroup relationship (Opisthokonta)
  - approximately same date of origin (according to fossils and molecular clocks)
  - same set of genes available (49 out of 50)
  - roughly same tempo (ML-estimated distances) and same mode of evolution (AA substitution matrices)
Fungi are better resolved than Metazoa

50 genes x (17 + 15) taxa (12,060 sites) ; ML/MP tree

Rokas et al. (2005) Science 310:1933-1938
The case for a radiation compressed in time

- molecular data
  - Fungi have a higher stemminess \((F = 0.201)\) than Metazoa \((F = 0.121)\)
  - Fungi have more parsimony-informative sites \((5,015 \text{ vs } 3,701)\) and less singleton sites \((1,518 \text{ vs } 3,080)\) than Metazoa

- paleontological data
  - Poriferans, cnidarians and bilaterian fossils appear within a 50 MY time-frame

Evidence for the so-called Cambrian explosion?
Simulation analysis of the mammalian radiation

The lack of resolution of the animal tree is interpreted as the signature of a radiation compressed in time.

covarion model; 16,000 sites; MP (NJ)

Rokas et al. (2005) Science 310:1933-1938
Debunking Rokas et al. (2005)...

• Rokas et al. (2005) are likely wrong
  – very few taxa
  – extensive use of MP
  – ignorance of previous studies
  – dismissing of alternative explanatory hypotheses
  – misleading simulations
  – biased conclusions

• We will demonstrate that...
  – animal evolution can be resolved (*no explosion*)
  – **taxon sampling** is of fundamental importance
A tree with more genes and a denser taxon sampling

133 genes x 47 animal taxa (31,092 sites)  
ML tree; bullets = 100% bootstrap support

Philippe et al. (unpublished)
A tree with more genes and a denser taxon sampling

Philippe et al. (unpublished)

ML tree; bullets = 100% bootstrap support

133 genes x 47 animal taxa (31,092 sites)
More genes or more/better taxa?

- Philippe et al. (unpublished) vs Rokas et al. (2005)
  - 133 (31,092 sites) vs 50 (12,060 sites)
  - among 10 random drawings of 50 genes, only one unresolved tree was retrieved
- Philippe et al. (2005)
  - 71 slow-evolving genes (20,705 sites) are able to resolve most internodes in the animal tree

A **large dataset is needed but not enough** to accurately resolve difficult phylogenetic relationships.
Let’s show we also need **adequate taxon sampling**.
Taxon sampling and non-phylogenetic signal

Philippe et al. (unpublished)
Taxon sampling and non-phylogenetic signal

2 conflicting signals in tree B
- phylogenetic signal (Ecdysozoa)
- non-phylogenetic signal (LBA)

hence,
- global loss of resolution
- artefactual topology

18 more nematodes added to tree B are unable to retrieve tree A.

One slow-evolving taxon is better than a bunch of fast-evolving taxa.

Philippe et al. (unpublished)
The Cambrian Explosion revisited