Darwin’s regret: what maths tells us about the evolution of life

Talk outline
- Why trees?
- Why maths?
- What are some problems?
- Probability models

Pre-Darwin

Carl Linnaeus 1701-1778

First Notebook on Transmutation of Species, 1837.
Letter from Darwin to Lyell, 1860.
Evolution process mechanism

David Hillis lab ~3000 species rRNA sequences
mechanism – the theory

Sir Ronald Aylmer Fisher FRS (1890 –1962)
Sewall Green Wright (1889 – 1988)
J.B.S (‘Jack’) Haldane [1892-1964]

George Robert Price (1922 –1975)
William Donald "Bill" Hamilton FRS (1936 -2000)

Process: phylogenetic trees

1967 Walter Fitch and Emil Margoliash constructed phylogenetic trees from cytochrome c sequences from vertebrates that agreed well with the vertebrate fossil record.

David Sankoff (mathematician)

The growth of genomics/phylogenetics

Publications with “molecular” and “phylogenetic” in abstract
Further applications of phylogenetics
linguistics, stematology, psychology, whisky

Stammbaum for Indo-Europena. From Die Darwinische
Theorie und die Sprachwissenschaft (Schleicher’s 1863)

2. Why maths?

– Eugene Paul Wigner

“The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which.”
– Gian-Carlo Rota, (1986, in Discrete thoughts)

Why maths?

Analysing existing methods
Developing better methods
Help answer questions:
- Why do some methods lead to different estimated trees?
- How can we have confidence in a given tree? (or any tree?)
- Is evolution really ‘tree-like’?

First things about trees

Rooted (binary) tree
Unrooted (binary tree)
Counting trees

Species
- Kangaroo
- Chimpnzee
- Human
- Gorilla
- Hippopotamus
- Whale
- Lion
- Tiger

# trees = 1 x 3 x 5
# trees = 1 x 3 x 5 x 7 x 9 x 11 = 10,395 trees

# trees for 60 species is more than the number of atoms in the known universe

How do biologists build trees?

Old days (pre 1970s)
- Fossils
- Morphology, behavior, physiology

1970s onwards
- Amino acid/DNA sequences
- Gene order on chromosomes
- Presence/absence of genes/markers

Basic Idea: Compare species for difference in ‘characteristics’ they share.

Signal in data (and why it can be misleading...)

“Homoplasy” = reversals and/or convergent evolution

Life without homoplasy

10,395 trees

The ‘four is enough’ theorem

Every tree – on any number of species – is the only perfect phylogeny for some choice of four characters

A ‘perfect phylogeny’ … and the only one
Example of low-homoplasy data (SINEs)

Part 3: What are some problems?

- The ‘Felsenstein Zone’
- Lineage sorting
- Deep divergences – site saturation

The ‘Felsenstein Zone’

Species 1:....ACAACGT....
Species 2:....ACGACTC....

Does it happen?

J. Huelsenbeck 1998: Is the Felsenstein Zone a fly trap?
A different type of ‘zone’

Problem 2: lineage sorting: ‘gene trees vs. species trees’

The good (math) news....

(HC)G ~78%
(GC)H ~11%
(HG)C ~11%

(Ebersberger et al. MBE 2007)

The plot thickens...

Theorem 2006

Noah Rosenberg and James Degnan

Whenever you have five or more species, the most likely gene tree can be different from the species tree

Why bad?

But...

“Theorem” ~2008

There are simple ways that will find the species tree from gene trees that don’t get fooled!

Problem 3: Deep divergences

(the Metazoan phylogeny)

From Huson and Bryant, Applications of phylogenetic networks in evolutionary Coelomata vs Ecdysozoa studies, MBE. 2006
Deep divergences

"Theorem": The amount of data needed to resolve $\varepsilon$ is at least $\varepsilon^{\text{depth}}$.

Does it matter?

Great, we almost ‘knocked off’ Everest!

just epsilon feet from summit. Very good mister Hillary sir.

Part 4: Probability models

- The ‘shape’ of trees
- Predicting the loss of ‘evolutionary heritage’
- Where is the root of a tree?

An early example...

<table>
<thead>
<tr>
<th>Number of species in genus</th>
<th>Number of genera</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>1</td>
<td>131</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9 to 11</td>
<td>13</td>
</tr>
<tr>
<td>12 to 14</td>
<td>3</td>
</tr>
<tr>
<td>15 to 20</td>
<td>7</td>
</tr>
<tr>
<td>21 to 34</td>
<td>14</td>
</tr>
<tr>
<td>35 upwards</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>293</td>
</tr>
</tbody>
</table>

$Pr(N = n) \approx \frac{1}{n^{1+r}}$

Where do evolutionary trees come from?

Evolutionary tree     Reconstructed tree

Another viewpoint

Question 1: How long are the branches in a tree?

$L$?

Speciation rate = 1/million years
so the expected value of $L$ equals 1 million years

The bus 'paradox'

You turn up at a bus stop, with no idea when the next bus will arrive.

🌟 If buses arrive regularly every 20 mins what is your expected waiting time?

🌟 If buses arrive randomly every 20 mins what is your expected waiting time?
Length of a new branch

Expected value of $L$ is 1 million years

Tree puzzle (I):

A tree evolves with each lineage randomly generating a new lineage on average once every 1 million years (no extinction).

Look at the tree when it has 100 tips

What is the expected length of a randomly selected extant branch?

**Answer 1:** 1 million years? $\times$

**Answer 2:** 500,000 years? $\checkmark$

What about the interior lineages?

Tree puzzle (II):

Now suppose extinction occurs at the same rate as speciation. Suppose the 'reconstructed tree' we see today has 100 tips.

Does this tree look just like a tree grown with a slower speciation rate and zero extinction?

No

(A) Extinction=0

(B) Extinction = speciation

Real reconstructed trees generally look much more Model A than B trees

(Morlon et al. 2010; McPeek 2008)
Why does this matter?

Question:
If a random 10% of species from some clade were to disappear in the next 100 years due to current high rates of extinction, how much ‘evolutionary diversity’ would be lost?

Prediction is very difficult, especially about the future. Niels Bohr, Danish physicist (1885-1962)

Diversity

Predict the proportion of diversity that remains if each species (leaf) survives with probability $p$ (independently of others).

A much cited paper

“...80 percent of the underlying tree can survive even when approximately 95 percent of species are lost.”
Nee and May, Science, 1997

\[
\mu(p) = \frac{E \text{ of future diversity}}{E \text{ of present diversity}} = \frac{-p \log(p)}{1 - p}
\]
...80 percent of the underlying tree can survive even when approximately 95 percent of species are lost.

\[ \mu(p) = \frac{-p \log(p)}{1 - p} \]

...84 percent of the underlying tree is lost when approximately 95 percent of species are lost.

Question 2: Can we explain tree balance?

Evolutionary tree | Reconstructed tree
---|---
Grow a Yule tree till it has 101 leaves. Which is more likely?

Yule model

Balance:

Yule model

| 1/3 | 2/3 |

All trees equal model

| 3/15 = 1/5 | 12/15 = 4/5 |
Real trees


Question 4: Where’s the tree root?

Does the Yule model provide any clue?

Back to our trees

John Haigh 1970

Which was the original version?

Select the ‘most likely’ one. What’s the probability it is the original?

1 – log(2) ~ 0.307

Nine nodes: probability is >99.6%

~Independent of n

A different (but related) question

Theorem (McKenzie+S)

Pr(e_M = e_o) = 4log(4/3) - 1
~ 0.15

P(longer of longer < shorter)
Why maths? (again)…

I attempted mathematics, and even went during the summer of 1828 with a private tutor (a very dull man) to Barmouth, but I got on very slowly…. This impatience was very foolish, and in later years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics.

Some questions for the future

- How did life start?
- How did photosynthesis evolve?
- How much were genes transferred in early life?