Some mathematical issues in reconstructing evolutionary history from DNA

by Stephen J. Willson
Department of Mathematics
Iowa State University
Ames, Iowa 50011
The problem

Given DNA strings for extant species: (for example some out of 12,234 symbols)

snail          ATTACAGCTTTTCTTTTCCAAGACT
fruitfly       ATGATAACAAATTTATTTTCTGTA
mosquito       ATGATAACAAAACCTTGTTTTTCTGTT
nematode 1     ATTAACTAAT................
nematode 2     ATTAATCAA................
sea urchin 1   ATGACAATGACTATAACTGTAAGA
sea urchin 2   ATAACAGCAAGAATTTTGGGTCAG
lamprey        ATGACACTAGATATCTTTGACCAA
carp           ATGATAGTAAGCTTTTTTCGACCAA
tROUT          ATGACACTAAGCTTCTTCGACCAA
frog           ATGAACCTAAGCTTCTTCGACCAA
chicken        ATGAACCTAAGCTTCTTCGACCAA
opossum        ATGAACGAAAATCTATTTGCCCCA
blue whale     ATGAACGAAAATTTATTTGCCCCT
fin–back whale ATGAACGAAAATTTATTTGCCCCT
cow            ATGAACGAAAATTTATTTACCTCT
mouse          ATGAACGAAAATCTATTTGCCCTCA
rat            ATGAACGAAAATCTATTTGCCCTCA
lancelet       ATGATAGTAAGCTTTTTTCTGTCAG
Goal: reconstruct a "phylogenetic tree":

A phylogenetic tree shows which species are more closely related.

**Why are people interested in phylogenetic trees?**

1. Understand the history of evolution on earth
2. Identify disease sources

Example. In a courtroom case, experts are asked which individual was the source of A's infection.
3. Politicians and wild-life managers have only a limited budget for saving species.
They can choose to save
Option 1: C, D, E, F
Option 2: A, C, F, G
Option 3: B, F, J, N

Which option preserves the most biodiversity?
Option 1: C, D, E, F
Option 2:  A, C, F, G
Option 3: B, F, J, N
4. Geneticists know the role of a certain gene in all the species below except H. Which species is most likely to have the same role for the gene?
Mathematical trees
A mathematical graph consists of some vertices (dots) and some edges connecting pairs of vertices. It is a tree if it is connected and there are no cycles. Usually relationships among biological species are given by trees.

1. Vertices correspond to species.
2. Edges connect species with their closest descendent or ancestor species.
3. Vertices with only one edge, called leaves, correspond to known extant species and are labelled that way.
4. Other vertices correspond to ancestral species.
5. We focus on the branching pattern: there are no vertices with exactly two edges.
If there are $n$ leaves (n extant species), how many different trees are possible? (We look at trees in which each vertex has either 1 or 3 edges.)

Here they all are if $n = 4$: 

![Diagram of trees with vertices labeled 1, 2, 3, 4]
If $n = 5$ there are 15 trees:
Theorem. If there are \( n \) leaves (extant species) there are
\( (3) (5) (7) \ldots (2n-5) \)
different labelled trees.

If \( n = 16 \), then this is
\[
3 \ 5 \ 7 \ 9 \ 11 \ 13 \ 15 \ 17 \ 19 \ 21 \ 23 \ 25 \ 27 = 213,458,046,676,875 = 2.13 \times 10^{14}.
\]

If \( n = 19 \) then this is \( 6.32 \times 10^{18} \)

If \( n = 50 \), then this becomes \( 2.84 \times 10^{74} \).

Compare: A hard drive with 256 GB can store approximately \( 2.6 \times 10^{11} \) characters. If each tree could be stored in just one character, just listing all the trees for \( n = 50 \) would require \( \underline{__________} \) hard drives.
Answer:

$$2.84 \times 10^{74} / 2.6 \times 10^{11}$$

$$= \ 1063 \text{ hard drives}$$
Assumption: Evolution occurs by the accumulation of rare events. A substitution occurs along an edge each time a different nucleotide replaces a previous nucleotide. Substitutions should be infrequent.

Example. How many substitutions total took place in this tree?
Answer: 6
Which is the tree that best fits the data?

A **natural criterion**: The best tree is the one with the fewest substitutions.
A very degenerate example. Suppose that the DNA strings are
1 A
2 A
3 G
4 G
Which tree is best? (The best tree has the fewest substitutions.)
Answer: The first tree is best since it assumes the fewest substitutions.
Example. Suppose that the DNA strings are
1 ACACT
2 GTACC
3 ACTCT
4 GTTCC

For ((12)(34)) the best choice gives ___ substitutions.
Answer: 7 substitutions
For ((13)(24)) the best choice gives ___ substitutions.
Answer: 5 substitutions
For (14)(23) the best choice gives ___ substitutions.
Answer: 8 substitutions
Answer to the question: The best tree is $((13)(24))$ with 5 substitutions since this requires the fewest substitutions.
The length of a tree is the number of substitutions for the best choice of ancestors.

**Maximum Parsimony Criterion:** For each possible tree look at each possible assignment of DNA to the ancestors. Select as your answer the tree for which the fewest substitutions occur (the smallest length). This best tree is called the maximum parsimony tree.
Difficulties with the maximum parsimony approach:

(1) There are many trees to consider, too many for a study of each one.

Suppose there are 50 species. There are $2.84 \times 10^{74}$ possible trees. So you need to compute the length of each one of those trees.

If $2 \times 10^9$ trees could be tested in a second (2 GHz), then $n = 50$ trees all could be tested in $1.4 \times 10^{65}$ seconds. But 1 year is about $3 \times 10^7$ seconds. So it would take ____________________ years.

The earth is only $5 \times 10^9$ years old.
There may be several different maximum parsimony trees.

Example.

```
1   AA
2   GA
3   GG
4   AG
```

For ((12)(34)) the best choice gives 3 substitutions, so the length is 3.
For ((13)(24)) the best choice gives 4 substitutions.
For ((14)(23)) the best choice gives 3 substitutions.

Such occurrences of ties, rather than rare, are common. Working with real data, there might be several thousand different maximum parsimony trees.
(3) A maximum parsimony tree might have only a small number of substitutions fewer than other trees.

Maybe there is a unique maximum parsimony tree with length 2631. But there are 10 different trees with length 2632. Is the maximum parsimony tree so much better?
(4) Some substitutions might be regarded as more important than others.

For example A and G are chemically similar. So a substitution of T for A is
chemically a more important change than the substitution of G for A. In fact,
substitution of G for A generally occurs more frequently than substitution of T for A.
Our count ignores this kind of distinction.
Another approach:

**Find a mathematical model for how substitutions occur.**

Probability of substitution

<table>
<thead>
<tr>
<th></th>
<th>end</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>start A</td>
<td>.90</td>
<td>.05</td>
<td>.03</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>.04</td>
<td>.92</td>
<td>.02</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>.02</td>
<td>.03</td>
<td>.88</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>.01</td>
<td>.01</td>
<td>.04</td>
<td>.94</td>
<td></td>
</tr>
</tbody>
</table>
The Maximum Likelihood Criterion: Assume a probabilistic model for how substitutions occur. For each possible tree, find the "likelihood" that it occurs under this model. Choose the tree with the maximum likelihood. This is the "Maximum Likelihood Tree."
Assuming a particular model of molecular mutation (HKY 1985) and using the data from the example, we compute
\[ \ln \text{Likelihood}_1 = -31688.40 \]
\[ \ln \text{Likelihood}_2 = -31680.58 \]
\[ \ln \text{Likelihood}_3 = -31058.55 \]

Which tree has the highest likelihood?
ln Likelihood1 = -31688.40
ln Likelihood2 = -31680.58
ln Likelihood3 = -31058.55

means
Likelihood1 = e^{-31688.40}
Likelihood2 = e^{-31680.58}
Likelihood3 = e^{-31058.55}
The likelihood of H3 is much higher than that of H1 and H2. We accept H3 as true on the basis of maximum likelihood. In this case, the result agrees with our intuition. Note that the probability of obtaining these particular mitochondrial DNA for these species is still very low: $e^{-31058.55}$. 
Advantages of maximum likelihood criteria (over maximum parsimony):
   1. theoretical justification
   2. use of more of the data (maximum parsimony throws out a lot of information)

Difficulties with maximum likelihood criteria:
   1. There are still a huge number of trees to consider.
   2. There are also many possible models to use.
   3. There are many probability parameters to choose among.
**Distance methods**

1. Between each pair of taxa i and j find a distance $d(i,j)$ that corresponds to the amount of evolutionary change between them.

2. Find a tree $T$ with lengths on each edge so that the sum of lengths of the edges along paths between the leaves i and j is $d(i,j)$.

3. If this is not possible, do as well as possible.

**Example.** If the true tree is

```
   1
  / \ (3)
 /   (4)
/     / (2)
1---(4)---3
   /     /
  /   (3) /
 2     4
```

Then

- $d(1,2) = 4$
- $d(2,3) = 7$
- $d(1,3) = 9$
- $d(2,4) = 8$
- $d(1,4) = 10$
- $d(3,4) = 5$
Conversely, suppose we know the distances. One of the most common methods based on distances is **Neighbor-Joining**

**Example.** If

\[
\begin{align*}
    d(1,2) &= 4 & d(2,3) &= 7 \\
    d(1,3) &= 9 & d(2,4) &= 8 \\
    d(1,4) &= 10 & d(3,4) &= 5
\end{align*}
\]

We expect that 1 and 2 are "neighbors" since they are close.

\[
\begin{align*}
    d(1,x) + d(3,x) &= d(1,3) = 9 \\
    d(1,x) + d(2,x) &= d(1,2) = 4 \\
    d(2,x) + d(3,x) &= d(2,3) = 7
\end{align*}
\]

Solving, we get

\[
\begin{align*}
    d(1,x) &= 3 \\
    d(2,x) &= 1 \\
    d(3,x) &= 6
\end{align*}
\]
Similarly

\[ d(1, x) + d(4, x) = d(1, 4) = 10 \]
\[ d(1, x) + d(2, x) = d(1, 2) = 4 \]
\[ d(2, x) + d(4, x) = d(2, 4) = 8 \]

Solving, we get
\[ d(1, x) = 3 \]
\[ d(2, x) = 1 \]
\[ d(4, x) = 7 \]
Now
\[ d(3,x) = 6 \]
\[ d(4,x) = 7 \]
\[ d(3,4) = 5 \]

\[
d(x,y) + d(3,y) = d(3,x) = 6
\]
\[
d(x,y) + d(4,y) = d(4,x) = 7
\]
\[
d(3,y) + d(4,y) = d(3,4) = 5
\]

Hence
\[
d(x,y) = 4
\]
\[
d(3,y) = 2
\]
\[
d(4,y) = 3
\]

Result:
Advantages of distance methods
• The methods are fast
• The methods construct the trees rather than test each tree individually

Disadvantages of distance methods
• The initial estimated distances may be highly inaccurate.
• The same site may change more than once, and these extra changes may be "invisible."
• There are lots of ways to estimate the initial distances.
Consistency criteria:

Consider the following strings:
1. AAACCCCCC
2. CCCCCCCCCG
3. CCAAAACCC
4. AAAAACC
5. AACAAAACC

(a) There is a unique maximum parsimony tree 
((1 2) (5 (3 4))) with length 14.
(b) If we remove species 1 we expect to get the tree 
\(((2\ 5)\ (3\ 4))\). This has length 13.
(c) Yet the unique maximum parsimony tree for these four species is 
\(((2 \ 3) \ (4 \ 5))\) of length 12.

Summary: In this example, the maximum parsimony tree for \{1,2,3,4,5\} is
inconsistent with the maximum parsimony tree for \{2,3,4,5\}.

Maximum parsimony trees are often inconsistent.
Maximum likelihood trees are often inconsistent.

Goal: Find a criterion that leads to consistent trees.
A quartet consists of 4 taxa.

Suppose we want to build a tree for the set $S$ of taxa. 
(1) For each quartet $Q$ find the tree $T_Q$ best supported for just the taxa $Q$. 
(2) Fit the trees $T_Q$ together into a single tree for all the taxa. 

Example. Suppose that $S = \{1,2,3,4,5\}$. Here are the trees for each quartet. Fit them together into a tree for all of $S$. 
Problem: Fit together
Answer:
Theorem. Suppose that for each quartet $Q$ from the collection $S$ of taxa we are given a tree $T_Q$ with leaves $Q$. Then there is at most one tree $T$ consistent with all $T_Q$.

Call the resulting tree the **minimally inconsistent tree**.
Advantages:

(1) If Q has only four leaves, there are only 3 possible trees to distinguish between.

(2) The number of different quartets grows with the number of species, but not nearly as fast as the total number of trees. 
(The number grows like \( \frac{1}{24} n^4 \) rather than \( n^n \). 
If \( n = 50 \) we need to consider 230,300 different quartets, not \( 2.84 \times 10^{74} \) different trees.)

Disadvantages:

(1) Sometimes knowledge about a fifth species helps you to find the best tree for the first four species. So maybe finding the tree for a quartet is a tricky calculation.

(2) When there are lots of inconsistencies among the quartet trees, fitting the trees together to find the minimally inconsistent tree can be hard. There may be no single tree that fits all the quartet trees.
Summary

• Phylogenetic trees are used to describe the historical relationships among species as life evolved.
• DNA provides data to build the trees.
• When there are only a moderate number of extant species, the number of different possible trees is astronomically large. Accurately distinguishing them may take impossible amounts of time.
• There are different mathematical procedures to select the tree:
  (1) Maximum parsimony
  (2) Maximum likelihood
  (3) Minimal inconsistency
  (4) Distance methods

The first two methods in principle require too much time for exact solution. The third method makes some progress, but has difficulties in reconciliation. Distance methods are often very fast but plagued by inaccurate initial estimates for distances.
• Hence we must either:
  (1) find algorithms to approximate the solutions above; or
  (2) find new criteria with faster algorithms.