Comparative genomics

July 31, 2001, Taipei

Hsin-Fu Liu, PhD

Department of Medical Research, Mackay Memorial Hospital, Taipei, Taiwan

E-mail: hsinfu@ms1.mmh.org.tw
Tel: 02-28094661 ext. 2073
Fax: 02-28094679
Phylogenetic trees

A phylogenetic tree is a graphical representation of the evolutionary relations among organisms or operational taxonomic units (OTUs), also called taxa. The taxa can be species, populations, individuals, or genes. The tree is composed of nodes connected by branches.

A monophyletic group, called a clade, consists of individual taxa descending from a common ancestor.

True tree:
There is only one true tree.

Inferred tree:
A tree that is obtained by using a certain set of data and a certain method of tree reconstruction.

Tree topology (the branching pattern of a tree)

```
    external node
   /     \
internal node
```

Bifurcating tree
(fully resolved tree)

```
    \ninternal node
```

Multifurcating tree
(star-like tree)

Multifurcating trees:
• The true sequence of events.
• The exact order cannot be determined unambiguously with the current data set.
Some key points for performing a phylogenetic analysis

Select an informative region to analyze
  coding or non-coding region
  length of the fragment

Make an optimal sequence alignment
  the bias in selecting the taxa
  sequence homology
  insertions or deletions

Use different methods to construct the trees
  distance matrix methods
  discrete character methods
  substitution models

Statistical test for phylogenetic trees
  bootstrapping
Sequence alignment

Consider four sequences

PHYLOGENY
PHOLOGENY
PHLOGEYY
PHOLONY

Aligning the sequences

PHYLOGENY       PH-LOGE-NY
PHOLOGENY       PH-OLOGE-NY
PH-LOGEYY       PH--LOGEY-Y
PHOLO--NY       PH-OLO---NY

- The first step of any phylogenetic method implies the alignment of homologous sequences.

- Alignment requires the user to make assumptions regarding relative costs of substitutions versus insertions and deletions.

- In general, search for maximum similarity and minimize the number of insertions and deletions.

- Take a good look at the final alignments, as such alignments can be frequently improved by visual inspection.

- Exclude regions that can not aligned unambiguously.
Rooted and unrooted trees

Unrooted tree

Only specifies the relationships among the taxa.

Rooted trees

The direction of the evolutionary path is known and the root indicates the common ancestor of all the taxa. A root can be imposed to the tree by including an outgroup, a taxon clearly branching off earlier than the strains or taxa under study. In the absence of an outgroup, we may put the root at the midpoint of the longest pathway between 2 taxa by assuming that the rate of evolution was uniform over all the branches.

Number of possible rooted and unrooted trees

<table>
<thead>
<tr>
<th>number of taxa</th>
<th>number of unrooted trees</th>
<th>number of rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>954</td>
</tr>
<tr>
<td>7</td>
<td>954</td>
<td>10395</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
<td>135135</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
<td>2027025</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
<td>34459425</td>
</tr>
</tbody>
</table>
Methods for phylogenetic analysis

Distance matrix methods:

Unweighted pair group method with arithmetic average (UPGMA)
- Assumes a constant evolutionary rate
- Tree with equal branch lengths

Neighbor-joining method (NJ)
- Does not assume a constant evolutionary rate
- Tree with differential branch lengths

Discrete character methods:

Parsimony method (pars)
- Usually finds more than one tree and a consensus tree is made
- Only gives topology but no branch length

Maximum likelihood method (ML)
- Provides an estimate of the probability of a given branch
- Very time-consuming
Distance matrix methods

1 TCAAGTCAGGTTTCGA
2 TCCAGTTAGACTCGA
3 TTCAATCAGGCCCCGA

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.333</td>
<td>0.333</td>
<td></td>
</tr>
</tbody>
</table>

Convert dissimilarity to evolutionary distance by correcting for multiple events per site according to a certain model of evolution, e.g. Jukes and Cantor.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.44</td>
<td>0.44</td>
<td></td>
</tr>
</tbody>
</table>

A first idea about the relationship among different taxa can be viewed by their evolutionary distance. The simplest way to estimate the evolutionary distance is to count the total number of nucleotide differences between them and divide by the total number of available sites. An evolutionary distance obtained in such a way is called uncorrected distance. A more realistic estimation can be achieved by applying a particular model of evolution that is making assumptions on how nucleotides change during the evolutionary process.

Advantages:
• Very fast (at least some e.g. NJ method)

Disadvantages:
• Sequence information is reduced to one number
• Provide only one tree topology (e.g. NJ method)
• Dependent on the model of evolution used
Ancestral Sequence | Present Sequences
--- | ---
| 1 | 2
A | A A |
C | C — A |
T | T T |
G | G G |
A | G — A |
A | A A |
C | G — A |
G | G G |
T | A A |
A | A A |
A | T T |
C | C C |
G | G G |
C | C C |

Although only 3 nucleotide differences can be observed, 13 mutations have occurred during evolution.

During Evolution

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Substitution Type</th>
</tr>
</thead>
</table>
A | A | Single substitution |
C | C — A | Sequential substitutions |
T | T | |
G | G | |
A —> C —> T —> G | A | Coincidental substitutions |
A | A | |
C —> G | C —> A | |
G | G | Parallel substitutions |
T —> A | T —> A | |
A | A | |
A —> C —> T | A —> T | Convergent substitutions |
C | C | |
G | G | |
C | C —> T —> C | Back substitutions |
Correcting for unobserved mutations

How to correct the difference?

Apply a substitution model that tries to estimate the correct number of substitutions.
Nucleotide substitution models

Jukes and Cantor’s

Kimura’s 2 parameter

Substitution models

Jukes-Cantor (JC)  one rate of substitution
  equal base frequencies

Kimura 2-Parameter (K2P)  two types of substitution
  equal base frequencies

Hasegawa (HKY85)  two types of substitution
  unequal base frequencies

General Time Reversible (GTR)  six types of substitution
  unequal base frequencies
UPGMA and Neighbor-joining

UPGMA tree

NUJ tree
Maximum parsimony method

The principle of the parsimony method is to infer the nucleotide or the amino acid sequences of the ancestral species and choose a tree that requires the minimum number of mutational changes. Usually more than one tree with the same minimum number of changes are found and a consensus tree is made.

In theory all the possible tree topologies for n taxa should be evaluated in order to obtain the maximum parsimony tree. However, only a small number of all possible trees can be calculated when n is large. An approximate algorithms need to be used to find the parsimonious trees.

**Advantages:**
- Do not reduce all sequence information (e.g. distances)
- Evaluate different tree topologies
- Sequences of ancestral states can be estimated from a particular tree topology

**Disadvantages:**
- Can be slow for large data sets
- No correction for multiple mutations
- Sensitive to unequal rates of evolution
- Only give topology but no branch length
Only informative sites are used by parsimony method.

A site is phylogenetically informative when there are at least two different kinds of characters, each represented at least two times.
Finding the optimal tree

Exat algorithms

Exhaustive search:
Guaranteed to find the minimum tree because all tree topologies are evaluated. Not possible for more than \( \leq 10 \) sequences.

Branch and bound:
Guaranteed to find the minimum tree without evaluating all tree topologies. A larger number of taxa can be evaluated but still limited (depend on the data set).

Heuristic search
When a data set is too large to permit the use of exact methods, one must resort to heuristic approaches. Two basic strategies are used. An initial tree or set of trees is obtained by stepwise addition, then the tree is subjected to rearrangements (branch swapping) that attempt to find a shorter tree. Heuristic approaches are in favor of reducing computation time, therefore do not guaranteed to find the minimum tree.
Maximum parsimony

Investigate all possible tree topologies → Reconstruct ancestral sequences → Choose topology with smallest number of steps

1. Topology 1: 5 substitutions
   - ACTGA
   - ATTGA
   - GTGGA

2. Topology 2: 7 substitutions
   - ACTGA
   - ATTGA
   - GTGGA
   - ATTGA
   - GTGAC

3. Topology 3: 7 substitutions
   - ACTGA
   - ATTGA
   - GTGAC
   - GTGGC
   - ATTGA
   - GTGGC
Branch and bound
Stepwise addition

1

A
B
C
D

10 steps
8 steps
7 steps

2

A
B
C
D
E

20 steps
19 steps
22 steps

Add next sequence
Branch swapping

Evaluate number of steps again...
Consensus trees

Tree 1

Tree 2

Tree 3

Strict consensus tree

Majority-rule consensus tree
Maximum likelihood method

Unlike parsimony methods, ML does not assume that evolution is parsimonious but it rather tries each possible nucleotide at each node and calculates the likelihoods of all possible trees that could have produced the observed sequences using a given model of nucleotide substitution. The inferred phylogenies are those with the highest likelihood.

**Advantages:**
- Statistically well founded (branch length)
- Evaluate different topologies
- Use all sequence information

**Disadvantages:**
- Very slow (computationally intensive)
Statistical test for phylogenetic trees

The tree topology obtained by any of the tree-making methods should be viewed as only an estimate of the phylogenetic relation among the taxa. Therefore, it is essential to know how much confidence can be associated with the branch length or the appearance of a set of taxa as a monophyletic group in a given tree.

Bootstrap analysis

Bootstrap analysis is the most often used method for statistical evaluation of phylogenies. It is used to test the effects of sampling error on tree inference and the stability of tree nodes by calculating how often a particular cluster in a tree appears when nucleotide sites are re-sampled with replacements many times.

In general:

Bootstrap value > 95% :
Branch is considered to be robust.

Bootstrap value < 75% :
Not confident enough to fully support a topology.

Note:
If the original data set is biased for some reasons, a clade may be regarded as statistically significant even if it is a wrong one. Conversely, a clade may be a correct one even if its bootstrap value is less than 75%. This is because the original bias cannot be corrected by the resampling process.
The bootstrap technique

Original data set

1  AGGCTCCTAA...
2  AGGTTCGTAA...
3  AGCCCCGAGA...
4  ATTTCCGAGC...

bootsratp sample 1

1  GTACACCTAC...
2  GTACATTTAG...
3  GCACACCCAG...
4  TCACCTTCAG...

bootsratp sample 2

1  ATCACCCAAA...
2  ATTAGCGAAA...
3  ACCAGCGAAG...
4  CCTAGCGAAG...

bootsratp sample 3

1  CGAAATCGTC...
2  GGAAATTGTC...
3  GGAGACCCCC...
4  GTCGACTTCC...

... sample n
(100<n<1000)
Likelihood-mapping analysis
(quartet puzzling method)

Likelihood-mapping analyses can be used as a complementary approach to solve the controversial phylogenies.

- The method is based on an analysis of the maximum likelihoods for the three fully resolved tree topologies that can be computed for four sequences.

- The three likelihoods are represented as points inside an equilateral triangle.

- The triangle is partitioned into different regions.

- The centre of the triangle represents a star-like evolution whereas the three corners represent well-resolved phylogeny and the three intermediate regions between the corners reflect the difficulty in distinguishing between two of the three trees.

- For more than four sequences, the different strains can be grouped into four different subsets and all possible quartets generated by drawing one sequence from each subset can be evaluated.

- The more points distribute in a certain region of a particular corner, the bigger the support for the tree topology joining the four subsets represented by that corner.

- If most points locate in the center of the triangle, the four subsets are independent and related by a star-like tree.
Quartet puzzling method
The three corners of the triangle represent the three possible unrooted tree topologies for four taxa. L1, L2, and L3 represent the three likelihoods of the three trees, respectively. Each length of the perpendicular from point P to the triangle side is equal to the likelihood of the tree represented by the opposite corner.

Quartet Puzzling Support Values
The quartet puzzling support values can be interpreted in a similar way as bootstrap values (though they should not confused with them). Branches showing a quartet puzzling reliability > 90% can be considered strongly supported. Branches with lower reliability (> 70%) can in principle be also trusted but in this case it is advisable to check how well the respective internal branch does in comparison to other branches in the tree (i.e. check relative reliability). If you are interested in a branch with a low confidence it is also important to check the alternative groupings that are not included in the quartet puzzling tree.
Analysis of the phylogenetic signal

The quartet puzzling approach can be used also to visualize the phylogenetic content of a particular data set of n aligned sequences. For n sequences \(\binom{n}{4}\) possible quartets, exist. When \(n \geq 11\), a random sample of 10000 quartets is sufficient to obtain a comprehensive picture of the kind of phylogenetic signal present. The more the dots in the center of the triangle, the more the phylogenetic noise, reflecting star-like evolution, in the data set.
Molecular clock hypothesis
(the rate of evolution is constant)

If the evolutionary rate is constant, then it can be used to determine dates of species divergences. For example, knowing the evolutionary rate \( r \) for a certain gene and the nucleotide differences in that gene between two lineage or, more in general, their evolutionary distance \( k \), the time \( T \) at which the two lineages split is:

\[ T = \frac{k}{r} \]

Conversely, evolutionary rates can be calculated from the above equation when the divergence time between two lineages is known.

Problems in calibrate molecular clock

rate-constancy assumption is often seriously violated

• within primates, there is a well-documented slowdown in the lineage leading to human.

• the same virus can have different evolutionary rates in different host species or even in different individuals of the same host.
Problems in tree construction

- Systematic errors
- Long branches
- Unequal rates of evolution (lineages and sites)
- Bias in sequence content
- Sequence information content (too variable, too conserved)
- Recombination (e.g. viruses)
- Radiation of divergences (e.g. Cambrian explosion)
- Gene tree species tree

Systematic errors

Systematic errors are the errors caused by not fulfilling the assumptions crucial to a particular method.

e.g.
- Maximum parsimony: multiple mutations per site
- Distance methods: distances are not additive

Systematic errors can be the cause of wrong tree topologies!
Applications of phylogenetic analysis

- Taxonomy
- Evolution
- Molecular epidemiology
- Forensic science
- Biotechnology
- etc.
Direct sequencing of PBMC from the victim, suspect (IVDU), and controls (unrelated Swedish IVDUs and Swedish homosexual men).

**RT gene:**
Victim strain and suspect strain belong to the same cluster. They have a more recent common ancestor compared to the controls. However, the bootstrap support is weak (58%).

**gag gene:**
Victim strain and suspect strain clustered together with a strong bootstrap support (96%).

**Conclusion:**
The suspect is guilty if other circumstantial evidence is available.
A dentist from Florida seems to have infected 7 of his patients. The dentist died and the patients claim for insurance money.

Samples:
dentist, 7 patients, controls (HIV positive of the same area).

env gene:
HIV isolates from 5 patients and the dentist strain clustered together with sufficient bootstrap support (80%).
Two patients have different virus strains.

Conclusion:
The dentist has infected 5 of his patients.
The insurance company made a deal with the patients.
Hepatitis G virus in Belgian hemodialysis patients & HGV in Belgian HIV-positive IVDUs
Neighbor-joining tree
Parsimony tree

type 1a

type 2a

type 2b

(type 3)

Z78  Z89
Z90  Z11
Z27  Z28
Z77  Z80
Z53  Z60
Z93  Z98
Z42  Z39

Z78  Z89
Z90  Z11
Z27  Z28
Z77  Z80
Z53  Z60
Z93  Z98
Z42  Z39

(type 1b)

(type 1c)

(type 1d)
Likelihood-mapping analysis
The three human T-lymphotropic virus type I subtypes arose from three geographically distinct simian reservoirs.
References


European Workshop on Virus Evolution and Molecular Epidemiology, 1995-1999, Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium.

Workshop on Molecular Evolution, 1998, Marine Biological Laboratory, Woods Hole, Massachusetts, USA.
