Evolutionary Trees

Russ B. Altman

BMI 214
CS 274

Outline

1. Why build evolutionary trees?
2. Distance-based vs. character-based methods
3. Distance-based:
   - Ultrametric Trees
   - Additive Trees
5. Character-based:
   - Perfect phylogeny
6. Conclusions

Why build evolutionary tree?

Understand the lineage of different species.

Have an organizing principle for sorting species into a taxonomy.

Understand how various functions evolved.

Understand forces and constraints on evolution.

To do multiple alignment.

Types of trees

Distance based = tree is constructed using distances between species/objects (number of mutations, time, other distance metric).

Character based = tree is constructed based on the acquisition or loss of traits. Not connected explicitly to any measure of distance.

Roots

Rooted trees show a single common ancestor for every object within the tree. Requires more information.

Unrooted trees show objects as leafs, and internal nodes are some common ancestors, but there is insufficient information to tell whether or not a given internal node is a common ancestor of any 2 leaves.
Ultrametric Matrix and Tree

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

If $M$ is $n \times n$ symmetric matrix, then ultrametric tree $T$ for $M$ has properties:

1. $T$ contains $n$ leaves each corresponding to a row of $M$
2. Each internal node of $T$ is labeled by one entry from $M$ and has 2 children
3. The numbers along a path from root to leaf strictly decrease
4. $M(i,j)$ is label of least common ancestor of $i$ and $j$

Definition of Ultrametric matrix

$M$ is an ultrametric distance matrix, iff for every three indices $i$, $j$ and $k$ there is a tie for the maximum of $M(i,j)$, $M(i,k)$ and $M(j,k)$. That is, the maximum is not unique.

Evolutionary trees as ultrametric trees

$M(i,j)$ can be considered the time since $i$ and $j$ diverged in evolution:

“A and B had last common ancestor 8 million years ago”

Thus, $T$ becomes an evolutionary tree.

Molecular Clock Theory

(Zuckerkandl & Pauling, early 1960’s)

For any given protein, accepted mutations in the amino acid sequence for the protein occur at a constant rate.

Accepted = mutations that allow protein to function without death.

Theory implies that the number of accepted mutations occurring in a time interval is proportional to the length of the interval.

Molecular clock theory II

1. Rate of accepted mutations may be different for different proteins (depending on their tolerance for mutations).
2. Different parts of a protein may evolve at different rates.
Molecular clock theory III

The clock simplifies task of collecting ultrametric data:

If A and B differ by $k$ accepted mutations, then roughly $k/2$ mutations have occurred since divergence.

The $k/2$ mutations are proportional to the times since divergence (by clock theory), and so full matrix of data can be collected.

How can we collect this data?

1. Laboratory methods: mix single strands of DNA of related genes from different species, and measure how tightly they associate. The more complementary the DNA, the tighter the association.

   Used to study bird evolution, found nearly perfect ultrametric data. Thus, “DNA clock ticks at same rate for all bird species...”

2. Sequence analysis methods:

   Estimate number of mutations based on sequence comparisons, using a version of edit distance.

   Need to correct for multiple mutations, neutral mutations of DNA.

Ultrametric Matrices

Most data is not really ultrametric.

If data are ultrametric, then strong evidence for a molecular clock within that data, and that evolutionary tree is likely to be correct.

Un-ultrametric data sometimes can be fitted to “closest” ultrametric matrix nearby.

Additive Trees

Ultrametric are the best, but if data doesn’t fit, can go to weaker test: additive matrix.

Given $M$ matrix, $T$ is an additive tree if

* $T$ has at least $n$ nodes, with $n$ rows of $M$
* for each $(i,j)$ path from $i$ to $j$ has weight $M(i,j)$

Additive trees and evolution

Ultrametric trees are additive, but additive trees are not necessarily ultrametric.

(NOTE: If an additive tree has a node $v$ such that all leaves have the same distance to $v$, then it is ultrametric.)

Additive trees don’t have a root, but just indicate branching patterns, with no implied direction.
When data is imperfect...

Other distance methods:

1. Plain old cluster analysis--group species by similarity, form groups of related species. E.g. hierarchical cluster analysis (See Mount, UPGMA, p 261).

2. Least squares fitting of an error function that is minimal when observed and estimated distances on tree match. (See Mount, NEIGHBOR JOINING, p 260)

Character-based trees

Very often, we don’t have distance metrics between the species of interest.

But we do have a set of observable features.

Can use these features as the basis for building a tree.

Character-based trees

Rows = objects
Columns = traits

1 = object has trait (character)
0 = object does not have trait

Object has all traits on path from the root.

Character-based trees interpreted as evolutionary trees

1. Root of the tree represents an ancestral object that has none of the present $m$ characters $=[0 \ 0 \ ... \ 0]$.

2. Each of the characters changes from zero to one exactly once and never changes back.

Thus, each character labels one edge. Gives an evolutionary history by mutation event, but not time.

Where do characters (traits) come from?

Traditionally, morphological features like:

- has backbone
- has feathers
- loss of a certain bone in eye socket
- has a certain amino acid at position $i$
- walks on knuckles
- has certain nonfunctional DNA word in sequence
- whether a certain gap is present in a multiple alignment
- whether or not protein X regulates protein Y
Perfect phylogeny problem

Given the n x m binary matrix M, determine whether there is a tree for M, build it.

Theorem: Matrix M has a phylogenetic tree iff every pair of columns i, j the traits of i and j are either disjoint or one contains the other.

1 2 3 4 5
A 1 1 0 0 0
B 0 0 1 0 0
C 1 1 0 0 1
D 0 0 1 1 0
E 0 1 0 0 0

1 2 3 4 5
A 1 1 0 0 0
B 0 0 1 0 0
C 1 1 0 1 0
D 0 0 1 0 1
E 1 0 0 0 0

\[ O(nm) \] algorithm for solving perfect phylogeny problem

1. Consider each column of M as a binary number. Sort numbers in decreasing order, largest in column 1.
2. New matrix = \( M' \) and each trait is column #.
3. For each row, construct the string of traits it has.
4. Build a keyword tree \( T \) for the n strings from step 3.
5. Test whether \( T \) is a perfect phylogeny for \( M \).

Keyword tree

Rooted directed tree, edge labeled with character, branches with different labels, paths spell out keywords.

Can be used to solve dictionary problem—determine if a string is in the dictionary.

Keyword tree for \( M' \)

Keywords:
- A = 12
- B = 3
- C = 124
- D = 35
- E = 1

Using original character labels

Interpretation:
- D branched from B when D acquired new trait 5 (old 4).
- D and B had already acquired trait new 3.
Generalized Perfect Phylogeny
A character or trait is allowed to take on more than two states. E.g. Color of fur.

Very Hard Problem = NP-complete.

Maximum Parsimony Tree
For d characters, find tree with exactly d mutations. No more than is needed.

Very Hard Problem. Related to Steiner Trees.

Maximum Likelihood Methods
Have an underlying probabilistic model of the basic operators in evolution: mutation, insertion, deletion, translocation, etc...

Also have a set of sequences.

Can assess the probability of the data, given the model. Can compute the most likely model given the data, using Bayes’ Rule.

\[ P(M|D) = \frac{P(D|M) P(M)}{P(D)} \]

Multiple Alignment and Trees
Progressive alignment methods do multiple alignment and evolutionary tree construction at the same time.

Sequence alignment provides scores which can be interpreted as inversely related to distances in evolution.

Distances can be used to build trees.

Trees can be used to give multiple alignments via common parents.

Ribosomal Sequences: RDP

http://www.amnh.org/Exhibition/Expedition/Fossils/index.html

The Tree of Life
A distributed internet project containing information about phylogeny and biodiversity

David R. Maddison
University of Antwerp