Algorithms in Bioinformatics: Lecture 15-16: Phylogeny Reconstruction

Lucia Moura

Fall 2010

Phylogeny Reconstruction: Introduction

These slides complement the lecture by Marcel Turcotte posted in the course web page.

Introduction, biological motivation and an overview of many approaches for phylogeny reconstruction should be read from Turcotte's lecture, first.

Here we add some extra material.

Overview of Methods

1. **Distance-Based Phylogeny Reconstruction Algorithms**
   A phylogeny tree is built based on the distance between the taxa (the more similar ones should be evolutionary more related).
   - UPGMA method (see Turcotte's slides, textbook 7.3.2)
   - additive tree reconstruction method (see textbook 7.3.3)
   - nearly-additivie tree reconstruction and neighbour-joining heuristic (see textbook 7.3.4)
2. **Character-Based Phylogeny Reconstruction Algorithms**
   Every taxon is described by a number of characters (number of fingers, presence/absence of protein, the nucleotide at a particular genome location, etc), each with a finite number of states.
   Goal: build a phylogeny tree that best explains the character matrix.
   Three types (see textbook) based on different optimization criteria:
   - parsimony (see Turcotte's slides, textbook 7.2.1)
   - compatibility (see textbook 7.2.2)
   - maximum likelihood (see Turcotte's slides, textbook 7.2.3)

Distance-Based Methods: valid distance matrices

A distance matrix M is said to be a **metric** if and only if :

- it is symmetric: $M_{ij} = M_{ji}$ and $M_{ii} = 0$, for all i, j; and
- it satisfies the triangle inequality: $M_{ij} + M_{jk} \geq M_{ik}$, for all i, j, k

From now on, we only consider metric distances.

A distance matrix is said to be **additive** if there exists a phylogeny tree T for S such that:

• every edge $\{u, v\}$ in T is associated with a positive weight d_{uv}; and

• for every $i, j \in S$, M_{ij} is equal to the sum of the edge weights along the path from i to j in T.

A distance matrix is said to be **ultrametric** if there exists a phylogeny tree T for S such that:

• additive properties are satisfied

• a root of the tree can be identified such that the distance to all leaves from the root is the same.

Recognizing additive and ultrametric distance matrices

Theorem (Buneman's 4-point condition)

M is **additive** *if and only if the 4-point condition is satisfied, that is, for any 4 taxa in S, we can label them as i, j, k, l such that*
$M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$

Theorem (3-point condition)

M is **ultrametric** *if and only if the 3-point condition is satisfied, that is, for any 3 taxa in S, we can label them i, j, k such that $M_{ik} = M_{jk} \geq M_{ij}$*

Three computational problems on a distance matrix:

1. **If M is ultrametric, reconstruct an ultrametric tree T for M in polytime.**
   Algorithm: UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
   principle: clusters similar taxa iteratively.
   Go back to Turcotte's lecture.
2. **If M is additive, reconstruct an additive tree T for M in polytime.**
   Will see next.
3. **if M is not exactly additive, find the nearest additive tree.**
   Will see next.

Additive Tree Reconstruction

Build the unique tree for 3 taxa i, j, k, creating node c with:

$$d_{ic} = \frac{M_{ij}+M_{ik}-M_{jk}}{2},\ d_{jc} = \frac{M_{ij}+M_{jk}-M_{jk}}{2},\ d_{kc} = \frac{M_{ik}+M_{jk}-M_{ij}}{2}$$

At each stage, given an additive tree T' for $k-1$ taxa, insert the k-th taxon into T' as follows:

• For every edge of T', check whether the k-th taxon splits this edge, using a similar condition as above.

Note: only one edge can be split by the kth taxon, so the tree reconstructed is unique.

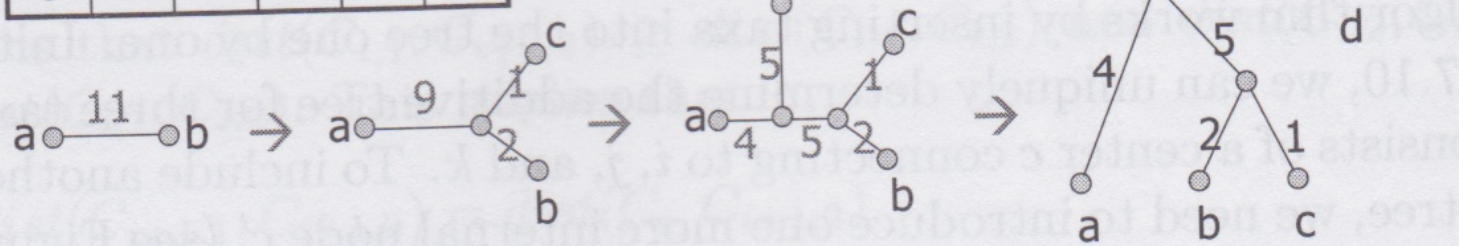
Time complexity:

At each iteration checks $O(k)$ edges, where k is the number of leaves; for each edge the check is constant.

So the running time is $O(1+2+\cdots+n) = O(n^2)$.

Additive Tree Reconstruction Example

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0



Nearly additive Tree Reconstruction

Suppose the distance matrix M is not additive.
Then, we want to find an additive matrix D that minimizes the sum of the squares of error $SSQ(D)$, where

$$SSQ(D) = \sum_{i=1}^{n} \sum_{j \neq i} (D_{ij} - M_{ij})^2.$$

The additive tree T corresponding to D is called the least square additive tree.

Computing T is NP-hard.

Heuristic method - **neighbour-joining (NJ) method** (1987):
similarly to UPGMA, it construct a larger cluster C by merging two nearest clusters A and B, with an extra constraint that the distance from A and B to other clusters should be as dissimilar as possible.

Neighbour joining algorithm

Neighbor-Joining algorithm

1: Let $Z = \{\{1\}, \{2\}, \cdots, \{n\}\}$ be the set of initial clusters;
2: For all $\{i\}, \{j\} \in Z$, set $dist(\{i\}, \{j\}) = M_{ij}$;
3: **for** $i = 2$ to n **do**
4: For every cluster $A \in Z$, set $u_A = \frac{1}{n-2} \sum_{D \in Z} dist(D, A)$;
5: Find two clusters $A, B \in Z$ which minimizes $dist(A, B) - u_A - u_B$;
6: Let C be a new cluster formed by connecting A and B to the same root r. Let r_A and r_B be the roots of A and B. The edge weights of (r, r_A) and (r, r_B) are $\frac{1}{2}dist(A, B) + \frac{1}{2}(u_A - u_B)$ and $\frac{1}{2}dist(A, B) + \frac{1}{2}(u_B - u_A)$, respectively;
7: Set $Z = Z \cup \{C\} - \{A, B\}$;
8: For any $D \in Z - \{C\}$, define $dist(D, C) = dist(C, D) = \frac{1}{2}(dist(A, D) + dist(B, D) - dist(A, B))$;
9: **end for**

Steps 2, 4, 5 take each $O(n^2)$, step 6 takes $O(1)$, step 8 takes $O(n)$.

Total running time $O(n^3)$.

Character-Based Methods

Every taxon is described by a number of characters (number of fingers, presence/absence of protein, the nucleotide at a particular genome location, etc), each with a finite number of states.

Goal: build a phylogeny tree that best explains the character matrix.

Three types based on different optimization criteria:

- parsimony
- compatibility
- maximum likelihood

Maximum Parsimony

Principle of parsimony:

> *"If there exist two possible answers to a problem or a question, then a simpler answer is more likely to be correct."*

When applied to phylogeny reconstruction, the aim is to build the phylogeny with the fewest number of point mutations.

The *parsimony length*, denoted by $L(T)$, is the minimum number of mutations required to explain tree T.

Parsimony problem:
Compute a phylogenetic tree T, for a set of taxa, that minimizes $L(T)$.

Computational problems

The small parsimony problem:
Given a tree topology T with each leaf labeled by a taxon, compute the parsimony length $L(T)$ and the corresponding labeling of internal nodes.
Solution by dynamic programming, see Turcotte's slides pages 83–98.

The large parsimony problem:
Given the character-state matrix M, compute the most parsimonious tree for M.
This problem is NP-hard.
Solutions:

- Branch and bound. (see Turcotte's slides: pages 99–112)
- Greedy local search. (e.g. nearest neighbour interchange; Turcotte's slides: pages 113–114)
- Approximation algorithms (see next pages)

A 2-approximation algorithm for the large parsimony problem

Best known approxim. ratio is 1.55 (Alon et al. 2008); we give one with ratio 2.

Algorithm Description:

INPUT: M, the character-state matrix.

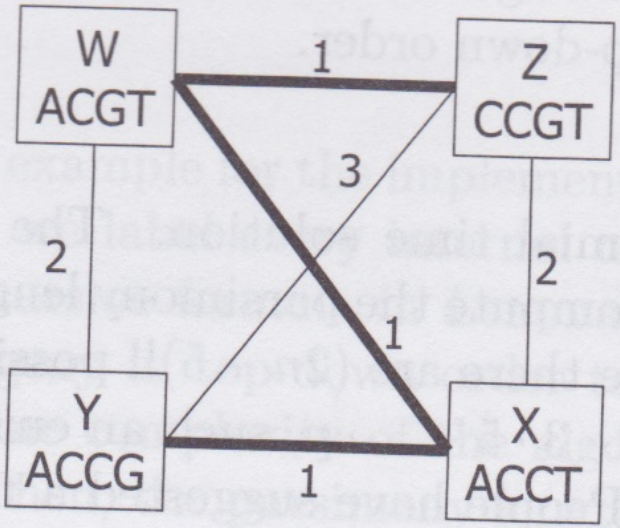
OUTPUT: a phylogenetic tree T with $L(T) \leq 2L(T^*)$, where T^* is a most parsimonious tree.

1. Let $G(S)$ be a graph whose vertex set is the set S of taxa, and the weight of every edge $\{a_i, a_j\}$ is the Hamming distance between the characters of a_i and a_j (number of changes).
2. Let T' be the minimum weight spanning tree of $G(S)$ (a tree that spans the vertex set and has minimum weight, among all such trees).
3. Covert T' to a phylogenetic tree T by replacing every internal node labeled by a taxa a with an unlabeled node and attaching a leaf labeled by a as its child. (Note that conversion non-binary to binary is easy)

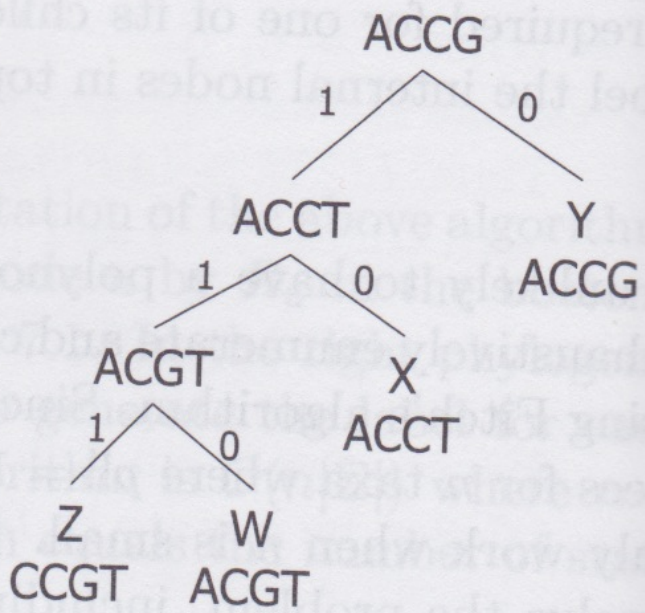
Example:

	1	2	3	4
W	A	C	G	T
X	A	C	C	T
Y	A	C	C	G
Z	C	C	G	T

S



G(S)



T

FIGURE 7.6: An example demonstrating the 2-approximation algorithm for the large parsimony problem. Given the set of 4 taxa S, we construct the graph $G(S)$ and compute the minimum weight spanning tree T' (bold edges in $G(S)$). From T', we construct a phylogenetic tree leaf-labeled by the 4 taxa.

Proving the 2-approximation ratio

Lemma

Let T' be a minimum spanning tree of $G(S)$. Then, the parsimony length of $L(T') \leq 2L(T^*)$, where T^* is the most parsimonious tree.

Proof: Create $\vec{T}$, a directed graph obtained from T^* by replacing each edge with one edge in each direction. Since the in-degree and out-degree of each node is the same, $\vec{T}$ has a Euler tour (a tour of the $\vec{T}$ that visits every edge once). So $w(C) = 2w(T*)$, since each edge of T^* appear twice in $\vec{T}$.

Let P be the path containing all nodes of $G(P)$ ordered by their first occurence in C. Since Hamming distance satisfy the triangle inequality $w(P) \leq w(C)$.

Since P is a spanning tree and T' is the minimum spanning tree, we get $w(T') \leq w(P)$.

Therefore, $L(T') = w(T') \leq w(P) \leq w(C) \leq 2w(T^*) = 2L(T^*)$.

Compatibility

Assumes that mutation is rare and most characters mutate at most once.

Computational problem:

Find a phylogenetic tree that maximizes the number of characters which have at most one mutation in the tree.

A character c is said to be compatible to a leaf-labeled tree T if there exists an assignment of states to the internal nodes of T such that at most one edge has a state change.

- **Perfect phylogeny problem:** determine if M admits a perfect phylogeny; which is equivalent to asking whether there exists a tree T such that all characters are compatible to T.
- **Large compatibility problem:** if M does not admit a perfect phylogeny, find the maximum set of mutually compatible characters and recover the corresponding tree.
  NP-hard; equivalent to the clique problem.

(See textbook Section 7.2.2 for details)

Maximum Likelihood

Assumes the observed character state matrix M is generated from some model of evolution.

A maximum likelihood method tries to estimate the phylogeny that best explains M, given the model.
For an informal discussion go to Turcotte's slides (pages 120–146).

Can tree reconstruction methods infer the correct tree?

- an *in vitro* experiment propagated bacteriophage T7 (a virus) in the presence of a mutagen, and the lineage was tracked - Hillis et al. (1992). A phylogeny of 9 taxa was constructed using: parsimony method, Fitch-Margoliash method, Cavalli-Sforza method, neighbor-joining method, UPGMA method. All five methods were able to reconstruct the true phylogeny.
- Leitner et al. (1996) collected HIV samples from people with known epidemiological relationships, and tested with various reconstruction methods. The following performed best: Fitch-Margoliash, neighbour joining, and maximum likelihood; maximum parsimony was in the middle; and UPGMA and KITSCH, which assume a constant molecular clock, performed the worst. All tended to overestimate small branches and underestimate large branches.

Conclusion: phylogeny reconstruction methods can correctly reconstruct the real evolution history, but we still cannot estimate well branch lengths.