Algorithms for Computational Biology

Lecture Notes by Sindhu Raghavan

02-28-2008

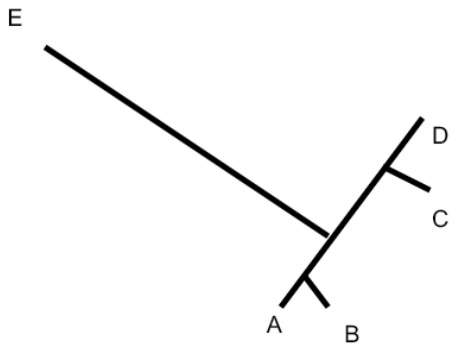
1 Statistical Consistency

A method m is said to be statistically consistent under model M if,
$\forall$ M model trees (T, Θ) and
$\forall \epsilon > 0$ there exists $K > 0$ such that
if sequences S of length $k > K$ are generated on (T, Θ), then $Pr[m(S) = T] > 1 - \epsilon$

Put in words, we say that a method is consistent under a model if the method can estimate the true tree with probability 1 as the length of the sequences increases.

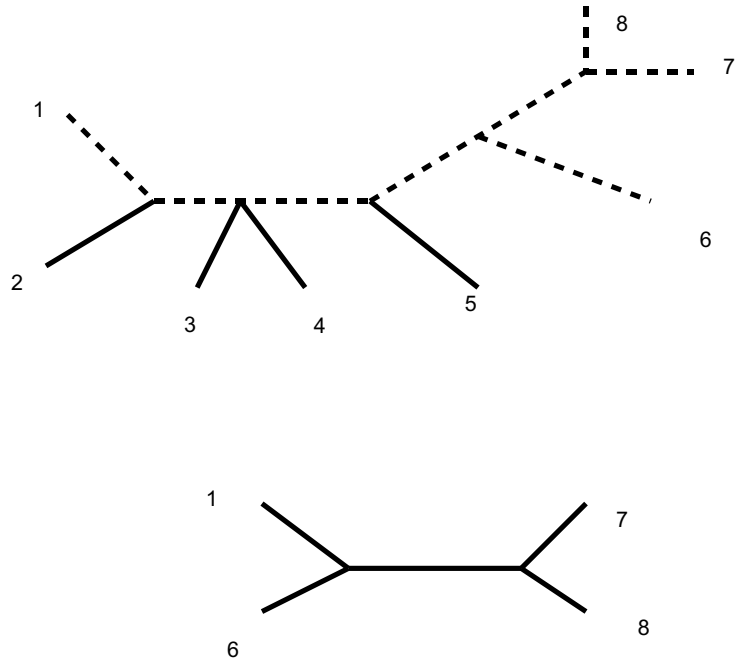
2 Rogue Taxon

E is called a rogue taxon in the fig below.



3 Agreement subtrees

A tree t is an agreement subtree of T_1, T_2, T_3, ... T_k if each T_i contains t as a homeomorphic sub tree.



In the above figure, agreement subtree 1,6,7,8 is obtained by taking all the edges between nodes 1,6,7,8 and supressing all nodes of degree 2.

Computational complexity of Maximum Agreement SubTree (MAST):

- If any of the trees T_i is a binary tree, then the MAST is also binary, in which case the MAST can be obtained in polynomial time. Thus, by making at least one of the input trees as binary, the computational complexity of the problem of finding MAST of the input trees can be reduced to polynomial time.
- General case - For any 3 trees, allowing any degree for a node, the problem is NP-hard [1].
- For an input of two trees, a dynamic programming solution in polynomial time is available to compute MAST of the input trees [2].

4 Felsenstein's proof for statistical inconsistency of Maximum Parsimony

In order to prove that MP is inconsistent, it is enough if we show a single model tree for which MP is inconsistent. We can show that MP fails to infer the correct tree when there are rogue taxa in

the data set. Consider the following model tree, called the Felsenstein zone quartet tree

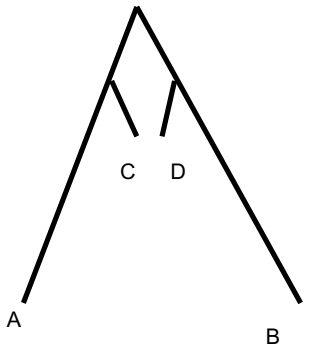


Fig 1

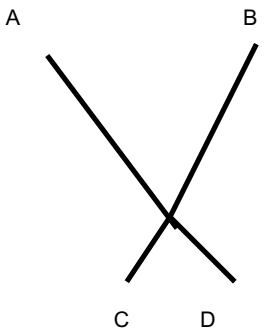


Fig 2

In the Fig 1, we see a tree in which A and C have diverged from a common ancestor and B and D have diverged from a common ancestor. However, we see that A and B have evolved rapidly when compared to C and D. This type of a tree is called Felsenstein zone quartet tree. A parsimony analysis on this data will invariably recover the wrong tree in which A and B are placed together and C and D are placed together. This is called long branch attraction. As a consequence, when we have a tree of this type (Fig 1), the more data we collect (i.e. the more characters we study), the more we tend towards the wrong tree as shown in Fig 2. Hence, we can prove that MP is statistically inconsistent under a given model of evolution.

Error rate for MP

As $k->$ infinity, FN error rate of MP $->1$ on a Felsenstein zone quartet tree, where
k - sequence length
As $k->$ infinity, FN rate of MP $->10$ percent on a caterpillar tree, where
k - sequence length

5 Taxon Sampling

Taxon sampling is the process of including more taxa in phylogenetic tree estimation. In the prosence of rogue taxa in our data sets, taxon sampling helps prevent the problem of long branch attraction when we use MP to analyze our data sets.

6 Phylogenetic networks

6.1 Implicit network

An implicit network is a graphical representation of bi-partitions that are not compatible.

6.2 Explicit network

An explicit network is a natural representation of HGT, hybridization or recombination.

6.3 Galled network

A galled network is a network that has no two cycles intersecting or overlapping.

6.3.1 Problem

Given t_1, t_2, t_3, ... t_k, find a galled network N, such that N contains a refinement of each t_i, i = 1,2,....k

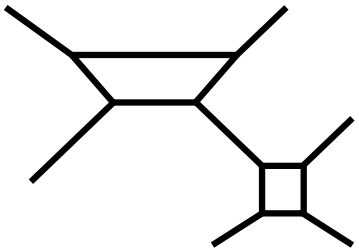


Fig a

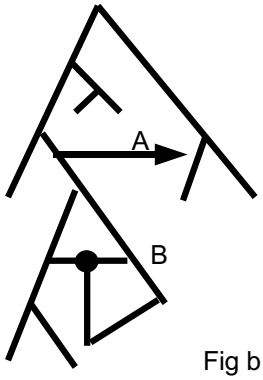


Fig b

A – Horizontal gene transfer

B – Hybridization event

Fig a – Implicit network

Fig b – Explicit network (It is also a galled network as the cycles do not overlap.

References

[1] Amihood Amir and Dmitry Keselman, *Maximum Agreement Subtree in a Set of Evolutionary Trees: Metrics and Efficient Algorithms.* SIAM Journal on Computing, Volume 26, Issue 6, pp. 1656-1669 1997.

[2] M. Steel, T. Warnow, *Kaikoura tree theorems: computing the maximum agreement subtree.* Information Processing Letters, 48, pp. 77–82, 1993.