Phylogenies and Evolutionary Trees
Lecture notes for CSE 527, 19 November 2003
Notes taken by Dhileep Sivam

Complex question:
Given data about a set of species, infer phylogeny.

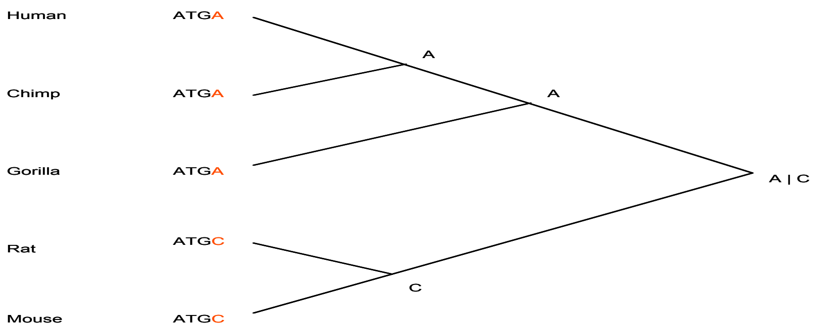
Simpler question:
Given data and a tree, evaluate how well the data fits the tree.

What computational biology aims to do:

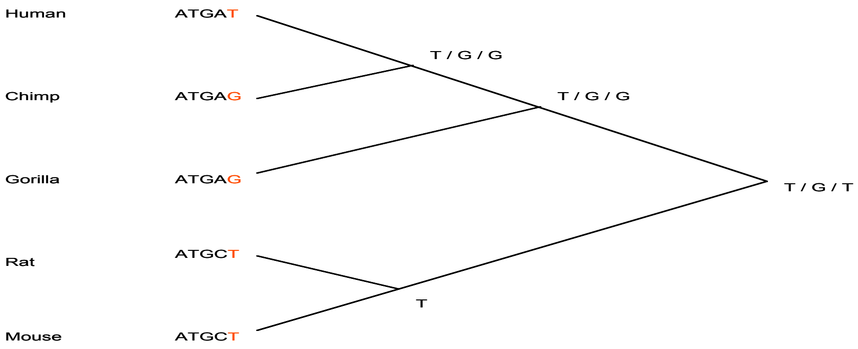
- make a set of trees
- score the trees
- Pick the best.

Parsimony:

- Similar to Occam's razor.
- States that best tree structure is such that the fewest possible changes need to be made in order to explain the tree.
- Most common method, but maybe not the best.
- There still might be more than one tree that is tied for least number of changes. While the trees may be tied, we can always determine the minimum number of changes that need to be made.



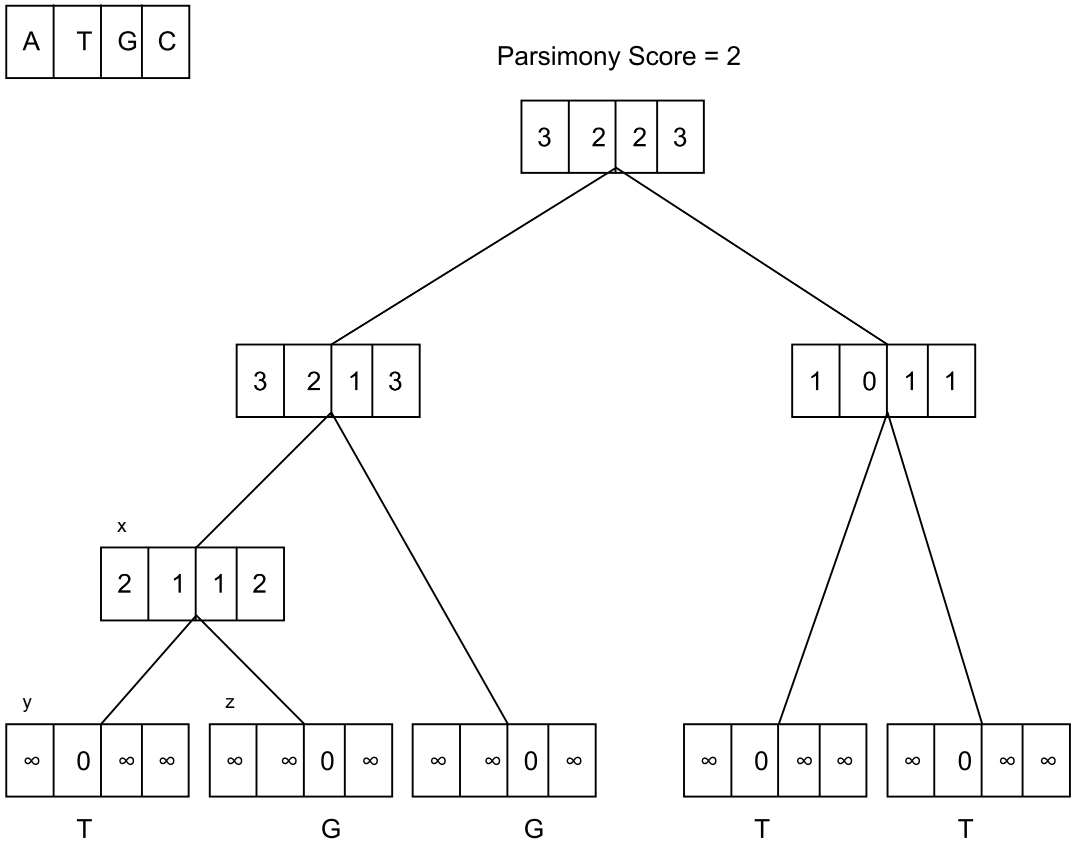
The above tree is organized such that the fewest number of change events are needed.



There is more than one way to explain the above tree, each with the same number of changes. The minimum number of changes needed to explain a tree can be calculated, however there can be multiple ways of explaining a tree given that minimum number.

How Do We Determine The Minimum Number Of Events To Construct A Given Tree?

Sankoff and Rousseau 1975: Dynamic Programming Algorithm.
Linear time: alphabet * tree size



Sankoff Rousseau Algorithm:

$$P_u(s) = \sum_{v = \text{all children of } u} \min(\text{cost}(s, t) + P_v(t))$$

Pu(s) = parsimony score for a given node u, assuming it is labeled s.
“min” is over all letters t in {A, T, G, C}
cost(s, t) is the cost of changing string ‘s’ to string ‘t’. Pv(t) is the parsimony score of string ‘t’ at the child node v.

Example (node x, y, z)

Nodes ‘y’ and ‘z’ are leaves. By definition the score at the leaves is 0 for the correct base pair and infinitely large for the others. Now we move to node ‘x’. At position ‘A’ we would would need to make 1 change (A to T) to minimize the score between ‘x’ and ‘y’ and 1 change to minimize the score between ‘x’ and ‘z’ for a total of 2 changes. At positions ‘T’ and ‘G’ we would need to make 1 change to minimize the score between the two children. At position ‘C’ we would need to make 2 changes by the same logic we apply at position ‘A’.

Discovery of Regulatory Elements by a Footprinting Algorithm:

- MEME, Gibbs sampler, etc do not exploit phylogenetic info.
- Closely related species are likely to have arbitrarily similar sequence.
- Widely diverged species are likely to have divergent sequence.
- Functionally relevant sequence is likely to be conserved across species, as mutations would be harmful.
- Conserved sequence across diverged species is likely to correspond to a motif.
- The notion of looking across species is particularly relevant when searching for gene specific regulatory regions. In such a case, there is likely to be only one/few protein binding site(s) in the genome, therefore methods such as Gibbs sampling will not work.

Mathieu Blanchette & Martin Tompa: Phylogenetic Footprinting Algorithm
"Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting". Genome Research 2002.

Example:

- Assume we have 10 primates and 2 rodents.
- We could run MEME?
- We could also align the sequences (e.g. ClustalW), but most likely the noise associated with such an alignment would wash out the short motif.

Goal of Phylogenetic Footprinting Algorithm

- Find protein binding sites.
- Short sequences (5 – 25bp).
- Up to 1000 bp upstream of coding sequence.

Given:

- phylogenetic tree *T*,
- set of orthologous sequences at leaves of *T*,
- length *k* of motif
- threshold *d*

Find:
Find each set *S* of *k*-mers*,* one *k*-mer from each leaf, such that the "parsimony" score of *S* in *T* is at most *d*.

This problem is NP-Hard.

Exact Algorithm

- Employs a generalization of the Sankoff –Rousseau algorithm.
- For each k-mer, calculate the minimum parsimony score across a phylogenetic tree.
- Total time O($n\ k\ (4^{2k} + l\ )$
    - n = number of species in the tree
    - k = motif length
    - l = average sequence length
- Better algorithm can restrict time to O($n\ k\ (4^{k} + l\ )$

Missing Motifs

- A Motif may not show up in every species.
- Search the tree for motifs with low parsimony that span a given width of the tree

Available at http://bio.cs.washington.edu/software.html

SEE "Phylogenetic Footprinting" Under the Lecture Slides section of the class websites for a detailed description of the Phylogenetic Footprinting Algorithm.