20.181 Lecture 2

Contents

Homework 1 Review

- HW1 #2 can be done by hand, you are not required to code for this.

Going over UPGMA

D(i,j) = 1/(Size(i)*Size(j)) * [SUM(p in i, q in j) d(p,q)]
summing over all possible pairs in your subtree.

- One of the more common ways to go from a set of distances to a full tree.
- But its not always the right thing to use.
  - UPGMA assumes distance = time. This is called the *ultrametric* assumption.
  - What if the "true" tree has some branches much longer than others? (due to faster evolution in those species)
- How to traverse a tree (by hand) to find the distance between two nodes on it:
  - Walk up the tree from node1 to the common ancestor of those two nodes, walk down the tree to node2, adding the branch lengths as you go.
  - subtree = clade

- Is there a tree-building method that would get this tree correct?

There are many methods for tree-building,

But the simplest way to do it correctly in this case is the Neighbor Joining algorithm.

(accounts for long branches by subtracting from every distance the average distance to all other leaves... normalizing everything out. We won't get into this- we're going straight into the most sophisticated methods. But you can learn more about it here if you like.)

What happened to the dentist?

The wikipedia article about this was posted in Lecture1, but Eric encourages you to ignore that. Better coverage here: http://www.annals.org/cgi/content/full/124/2/255

What is phylogenetics good for?

- We have some gene X with unkonwn function, but known sequence. So we find a number of its homologs, and we see that some of them carry out function A, and some carry out function B.

So we *could* just say "gene X carries out function A or B". But can we do better than that?

- Let's cluster the genes based on their blast scores(?). (Clustering for now you can think of as drawing a link between two nodes if they have a distance less than a certain threshold). We find that all the A's cluster together, and all our B's cluster together, and geneX connects the two clusters by linking twice to clusterA and once to clusterB.
- Problem with clustering: definition of membership is always ad-hoc. We *could* say that geneX has function B based on the fact that
  - but if you look at phylogenetic tree instead, the answer is more clear. Look for simplest scenario.

Parsimony Principle: the simplest evolutionary history is likely to be the correct one.

Newick Notation

Newick notation: a way of representing trees in text format.
(You can use this if you want to email the answers to HW1.)

- A tree consists of: root, internal nodes, leaves
  - "**( )**" means a clade (descended from a common ancestor
  - "**,**" means a branchpoint
- This format can be extended to include branch lengths, using colons.
- Examples of writing trees in this notation:
  - http://evolution.genetics.washington.edu/phylip/newick_doc.html
  - http://evolution.genetics.washington.edu/phylip/newicktree.html