COMP 5970/6970-004
Computational Biology: Genomics and Transcriptomics
Lecture notes 14: 3/1/2022

Haynes Heaton

Spring, 2022

Lecture Objectives

- Problem based learning
- Phylogenetic trees
- Distance metrics and distance matrices
- Phylogenetic tree inference

Problem based learning

Historically, the lecture method of teaching was developed when books were rare and expensive and most people could not read. With these restrictions, one person who could read would stand in front of everyone else and read from the available copy of the book.

Today thanks to universal public education in the first world, virtually everyone can read, and certainly everyone attending Auburn University can read. And thanks to technological advances of the printing press, later improvements, and more recently the internet, information in the form of the written word as well as educational online videos are cheap.

So far in this class I have given chalk talk lectures and have done live-programming of the algorithms both of which I think are much more engaging than the powerpoint style of lecture. But often, when it came time for the homework, many students did not understand the material. In particular, I lectured about and live-programmed the expectation maximization algorithm in 3 different lectures in 3 different contexts. But when I gave it as a problem in homework, about half of the class needed help from me and the other half simply didn't attempt it. Needing help is not a bad thing. I congratulate those who sought out help and eventually got the right solution. And struggling with a problem is not bad—struggling builds understanding. And all of this is not necessarily your fault. And it is not necessarily my fault. Perhaps it was the fault of the manner in which you were consuming the material—passively.

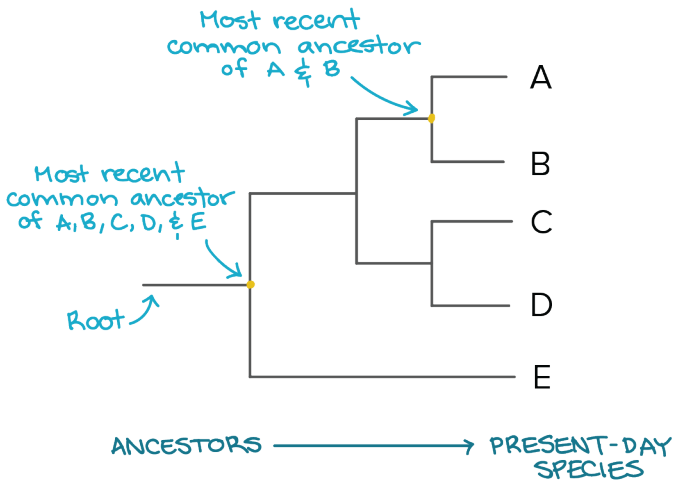
Studies show that retention of understanding from the lecture method is about 5%. Most students feel that they learn much more than 5% of the material, but lectures have a way of fooling you into thinking you understand something that you could not reproduce yourself.

So confronted with this reality, I sought help from a friend of mine who teaches math and I know really cares about teaching well. She sent me a manifesto on problem based learning and a discussion based class. The idea is that instead of the professor lecturing, providing motivation, working out example problems, the *student* will be the one developing the motivation, thinking about definitions, working out examples, and figuring out how to generalize the ideas to apply to new contexts. Some of these problems will be hard. You will get stuck, but that is a good thing! Getting unstuck with a problem is when you actually learn something. And how can you get unstuck if you never get stuck in the first place?

The following material will serve both as lecture note and homework 3 solutions (first 8 problems in this note, more solutions in the following notes).

1 Phylogenetic trees

Phylogenetic trees are diagrams describing the evolutionary or ancestral relationships between different organisms. These are represented as trees, often binary trees. They can either be rooted or unrooted. In a rooted binary tree, the root represents the most recent common ancestor of all of the leaf nodes and organisms represented by internal nodes. Often the observed organisms are current day and thus are all leaf nodes, but sometimes we have historical samples or ancient DNA which may represent some internal node of the phylogenetic tree.



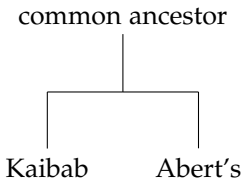
Question 1: You are looking at DNA sequences from different organisms and notice significant sequence similarity. List three potential reasons for these similarities.

Answer:

- Evolutionary relatedness
- Convergent evolution / functional relatedness
- Random chance (unlikely for long sequences)
- Mobile DNA element such as transposon, plasmid, or viral insert.

Question 2: Before the grand canyon formed, squirrels on either side of the Colorado river interbred freely. Over time, the physical barrier led squirrels on either side to not interbreed and eventually they became separate species with Kaibab squirrels on one side and Abert's squirrels on the other. Consider the common ancestor and these two species. Draw a diagram indicating this relationship.

Answer: In class we got a much more pictorial explanation with the grand canyon and a phylogentic tree over time overlayed on one another. All I was looking for is a simple tree.



Question 3: In graph theory, a tree is is an acyclic graph with exactly one path between any two nodes. Why are these good properties to model evolutionary relatedness?

Answer: The tree structure allows us to represent the branching nature of evolutionary divergence. While convergent evolution or in-breeding can result in two paths to the same node, the simple tree is a good model of evolution. If we allowed cycles in the directed tree, it would imply that transitively an organism is an ancestor of itself which due to the linear nature of time is nonsensical.

Question 4: Trees can be unrooted or rooted. Which of these most accurately models evolutionary relationships? Why?

Answer: A rooted tree most accurately models evolution because it is believed that all living organisms on earth are related to one another.

Question 5: Sequence alignment scores are a measurement of similarity—they are higher for more similar sequences. Sometimes we prefer to deal with distances—higher is more different. A distance must be non negative, symmetric ($D_{i,j} = D_{j,i}$), and satisfy the triangle inequality (for all i, j, and k, $D_{i,j} + D_{j,k} \geq D_{i,k}$). Recall our simple alignment score in which matches were +1, mismatches were -1, and indels were also -1. Can you find other values for these which makes this score a distance?

Answer: The simplest such metric would be matches = +0, mismatches = +1, indels = +1.

Question 6: A probabilistic scoring function we used was the log probability of the alignment under some model that assumed the sequences were related (it assigned higher probability to more similar sequences). Can you transform log(p(x,y)) to a distance metric?

Answer: -log(p(x,y))

Question 7: We define the distance between two nodes in a tree as the sum of the distance weighted edges in the path between those nodes. Given a tree T with nodes N, assume you can loop through the nodes, and loop through the edges from each node, and access the distance $D_{i,j}$ for any nodes i and j with an edge between them. You may also keep an auxiliary data structure which keeps track of visited nodes. Write an algorithm for finding the distance between any two nodes (use psuedocode or python or whatever as long as it is fully specific).

Answer: This could be done recursively, but in general due to the possibility of stack overflows with large data, I write the non-recursive version. In a general graph with cycles and edge weights, we would need *Dijkstra's* algorithm, but here breadth or depth first search will be fine.

$q \leftarrow$ empty queue
q.enqueue((i, 0))
visited $\leftarrow$ empty set
while !q.isempty() **do**

```
    (node, distance) ← q.dequeue()
    visited.add(node)
    for (neighbor, edge_distance) in node.neighbors() do
        if neighbor is j then
            return(distance + edge_distance)
        end if
        if !neighbor in visited then
            q.enqueue((neighbor, distance + edge_distance))
        end if
    end for
end while
```

Question 8: Given the following distance matrix, construct by hand an unrooted tree where all of the pairwise distances are conserved. You may add additional intermediate nodes which would represent common ancestors.

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0

Answer: