SEMESTER PROJECT PROPOSAL

BY

Claudia Kensela and Heli Patel

Data Structures and Algorithms

Oglethorpe University

APRIL 2015

Introduction

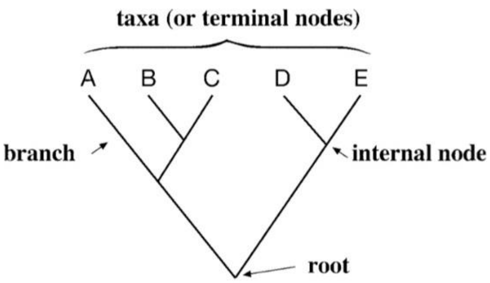
This proposal is intended to propose a sensible approach in reasonable computing time, to sub-classifying organisms by grouping them hierarchically into evolutionary *clades* or branches.

As this analysis is founded on solid evolutionary principles, it is important to begin by defining fundamental terminology. Firstly, *evolution* can be defined as the development, through natural selections and modifications, of a biological form. The driving force behind evolution is natural selection and the "survival of the fittest" mechanism that is allowed for by the genetic diversity produced from spontaneously occurring mutations (Xiong, p. 127).

Phylogenetics is the study of the evolutionary history of living organisms using tree-like diagrams called *phylogenetic trees* to represent the evolutionary relationships between the ancestry of a set of genes, species, or other taxa (Xiong, p. 127) The evolutionary relationships between organisms are inferred, and hypotheses about the evolutionary relatedness of species can be developed based on the similarity of biological sequence data.

Although phylogenetic trees constructed by computational methods are unlikely to perfectly resemble the actual historical species trees, computational phylogenetics is however, far more reliable than traditional phylogenetics. This is because traditional phylogenetics relies on often ambiguous morphological data obtained through measuring and quantifying the phenotypic properties such as average body size, lengths of particular bones or other physical features. On the other hand, the more recent field of molecular phylogenetics uses nucleotide sequences encoding genes or amino acid sequences encoding proteins, as the basis for classification (Contributors).

Furthermore, a few major assumptions must be made about our model. The first is that all molecular sequences used in phylogenetic construction are *homologous* - that is, they all subsequently diverged over time from a common origin. Phylogenetic divergence is assumed to be *bifurcating*, meaning that at any given point, a parent branch can split into only two daughter branches. (Xiong, p. 129) A typical bifurcating phylogenetic tree is shown below.



Such a tree is also known as a *rooted tree*, as all the taxa (sequences under study) have a common ancestor or *root node,* that has a unique evolutionary path that leads to all other nodes. Strictly speaking, the root of a historical species tree cannot be ascertained, as the common ancestor is already extinct, however in practice, we are able to explicitly define the root of a tree. This can be done by assigning a specific imputed sequence, known as the *most recent common ancestor (MRCA)* to the root node (Contributors).

Given the bifurcating and rooted nature of the phylogenetic trees, binary search trees are therefore the most appropriate data structures for the implementation of this molecular phylogenetic tree construction.

Process

The project will be divided into the following four main stages:

1. *Choosing a model of evolution*

We wish to study bacterial family as the organisms are fairly closely related and are ordinarily difficult to classify using traditional phylogenetics, which is heavily reliant on fossil morphology.

2. *Choosing molecular markers*

The next step in the phylogenetic construction is to decide whether to utilize discretely defined DNA or RNA nucleotide sequences encoding genes, or to utilize amino acid sequences encoding proteins, as the basis for classification, each having merits and limitations.

As we are studying closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, will be employed. In particular, we will use the noncoding regions of mitochondrial DNA. Additionally, with DNA-based phylogeny, the list of nucleic acid characters is simply {A, C, G, T} - much more manageable than the 20 amino acid variants.

3. *Performing multiple sequence alignment*

The third and "probably most critical step in phylogenetic analysis" is to construct sequence alignment to establish positional correspondence in evolution. This process results in aligned positions that are assumed to be genealogically related. (Xiong, p. 137)

Moreover, a successful alignment performs the following:

1. Ensures the matching of key nucleotide bases (at certain specific loci).
2. Measures the divergence between two sequences by counting the number of substitutions in the alignment.
3. Defines the observed distance between the two sequences using the proportion of substitutions.

Some foreseen difficulties include alignment errors and after deciding whether to use the full alignment or to extract parts of it, the design of a system to potentially remove unrelated or highly divergent sequences.

4. *Determining a tree building method*

This will most likely be done using the *Maximum Parsimony* algorithm that chooses to build the tree with the fewest evolutionary changes.

References

Contributors, W. (2015, January 23). *Computational phylogenetics.* Retrieved from Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Computational_phylogenetics&oldid=643818184

Contributors, W. (2015, January 23). *Computational phylogenetics.* Retrieved from Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Computational_phylogenetics&oldid=643818184

Xiong, J. (2006). In J. Xiong, *Essential Bioinformatics.* New York: Cambridge University Press.