General workflow of building phylogenetic tree: construct multiple sequences alignment, determine substitution model based on data, build tree, evaluate tree

Import Data (sequences or gene trees) 🡪 Create Species Tree using different methodologies, i.e. each of 4 main methods 🡪 Robustness tests 🡪 Visualization

1. Import Data
   1. Sequences
      1. Phybase has instructions for this, requires sequence files to be on your machine
         1. Must be a sequence file that includes DNA sequences, gene partitions, taxa names
         2. Can build MT, STAR, or STEAC species tree from gene tree vector or from tree file
   2. Gene Trees
      1. From Online
         1. TreeBASE package 🡪 actually use
         2. Rdryad 🡪 just mention
            1. From online data repository Dryad
      2. From Desktop Files
         1. Phylobase package 🡪 actually use
         2. Phyloch 🡪 mention
            1. Can load trees from BEAST, MrBayes, and other phylogenetic programs
2. Infer Species Trees
   1. Phangorn package 🡪 not sure
      1. Estimates phylogenetic trees and networks using ML, MP, distance etc
   2. Phybase 🡪 most likely
      1. Package to simulate and analyze gene trees and species trees
3. Robustness Test
4. Visualize
   1. Ggttree 🡪 package to use

Talk with Amy about Tree Building and Annotating

1. Slide 1: why do we want a tree? And can talk about that
2. Unpacking
   1. Look through last week’s ppt to get into ideas of theories
   2. What, why, goals, different ways to do it
   3. Simplify down to
   4. Data needed to build a tree
   5. Going from one gene to a gene tree
   6. Congruence tree
   7. Population trees
   8. Difference between species and gene trees
   9. Ultrametric trees
   10. Ecology: phylomatic 🡪backbone tree, send it taxa, match to genus, prune after
   11. Phylogenerator
   12. Models
       1. Is G to C same as G to C etc.
   13. Carol, tiffany, jeoue will be working with trees
   14. Barcode differences 🡪 why we use the different ones and how it relates
       1. DNA barcoding as a way to identify species
       2. Uses very short genetic sequence from standard part of the genome, similar to supermarket scanner and a barcode
          1. Standard barcode for almost all animal groups is a 648 bp region in mitochondrial cytochrome c oxidase 1 gene (“CO1”). COI is proving highly effective in identifying birds, butterflies, fish, flies and many other animal groups. COI is not an effective barcode region in plants because it evolves too slowly, but two gene regions in the chloroplast, matK and rbcL, have been approved as the barcode regions for plants.
       3. well-resolved molecular phylogenies derived from these DNA barcode sequences have the potential to improve investigations of the mechanisms underlying community assembly and functional trait evolution
   15. Differentce between identification and building trees
   16. Differing rates of evolution
   17. Tree as an endpoint 🡪 can get final results, or line it up to timing of biogeographical events, timing events in history to trees
       1. Nature, ecology and evolution looking at herbivory of leafs between beetles of south and
   18. Diversification analysis
   19. Microevolution 🡪 viral evolution
3. Look at 2 to 4 examples
   1. Solid presentation with examples and sample datasets
   2. And try to bring in our own data to try things on

Slides to Include

* **Use some of slides from bioinformatics lecture on ‘families’**
* What is phylogenetic systematics?
  + Phylogenetic systematics reconstructs evolutionary history and studies patterns of relationships among organisms
  + Because we cannot see this history we use clues left behind to try to reconstruct evolutionary history
* Why do we care?
  + All life on Earth is united by evolutionary history
* Goals of the process
* What do we mean by tree?
  + An evolutionary tree = phylogeny
  + Generally speaking phylogeny = evolutionary tree = phylogenetic tree = cladogram
  + Representation of evolutionary relationships among set of organisms or groups of organisms (taxa)
  + Ultimately trees depict clades 🡪 group of organisms that includes an ancestor and all descendants of that ancestor, can think of a clade as a branch on tree
* How to read an evolutionary tree
  + Tips of tree represent groups of descendent taxa/species
  + Nodes represent common ancestors of descendents
  + Two descendents that split from same node are sister groups
  + Many have an outgroup – a taxon outside group
    - Stems from base of tree and is more distantly related than any of the taxa of interest
    - Gives you a sense of of where on the bigger tree of life the main group falls
* How are evolutionary trees used
  + How to classify organisms based on evolutionary trees?
* Different Ways to reconstruct an evolutionary tree
* Phylogenetic Trees File Formats
  + Newick 🡪 .tre
  + Nexus 🡪 .nex
* Species Trees from Gene trees
  + Concept of phylogeny is as a cloud of gene histories
    - Discordance possible from horizontal transfer including hybridization, lineage sorting (deep coalescence), and gene duplication/extinction
  + all of the gene trees are part of the species tree, which can be visualized like a fuzzy statistical distribution
  + Gene trees inferred from sequence data, species trees inferred from gene trees
* Applications of gene trees to species
  + Species trees in area trees (biogeography)
  + Associate trees in host trees (coevolution)
  + If multiple parasites occupy a host or species occupying an area without competition, expected to behave more like duplicated genes and treated similarly

What to Actually Include

1. “Tree” Facts: Species vs. Gene Trees NCBI
   1. "Species" trees recover the genealogy of taxa, individuals of a population, etc.
      1. Internal nodes represent speciation or other taxonomic events.
      2. Species trees should contain sequences from only orthologous genes.
   2. "Gene" trees represent the evolutionary history of the genes included in the study.
      1. Gene trees can provide evidence for gene duplication events, as well as speciation events.
      2. Sequences from different homologs can be included in a gene tree; the subsequent analyses should cluster orthologs, thus demonstrating the evolutionary history of the orthologs.
   3. How to read an evolutionary tree
      * Tips of tree represent groups of descendent taxa/species
      * Nodes represent common ancestors of descendents
      * Two descendents that split from same node are sister groups
      * Many have an outgroup – a taxon outside group
        + Stems from base of tree and is more distantly related than any of the taxa of interest
        + Gives you a sense of of where on the bigger tree of life the main group falls
   4. Molecular Evolution Tips
      1. Rule Of Functional Constraint: As the biological (functional) importance of a gene or protein increases, the selective constraints on it increase, thereby lowering its evolutionary rate. Thus genes, proteins, and subsections thereof of greatest functional importance will evolve slowest and exhibit least polymorphism.
      2. Molecular Clock: Rate of sequence change is roughly proportional to time since last common ancestor of two species or genes.
   5. Methods of Analysis
      1. Distance Methods
         1. One of four primary approaches to analyze aligned sequences
         2. A variety of distance algorithms are available to calculate pairwise distance, for example:
         3. Proportional (p) distances
         4. Distance analysis compares two aligned sequences at a time, and builds a matrix of all possible sequence pairs.
         5. During each comparison, the number of changes (base substitutions and insertion/deletion events) are counted and presented as a proportion of the overall sequence length.
         6. These final estimates of the difference between all possible pairs of sequences are known as pairwise distances.
         7. Distance matrices used as non-parametric distance method
            1. These phylogenetic analysis methods rely on measure of genetic distance between sequences being classified, so need a MSA as input
            2. Distance defined often as fraction of mismatches at aligned positions, gaps are ignored or counted as mismatches
            3. Neighbor joining: applies general data clustering techniques to sequence analysis using genetic distance as clustering metric 🡪 produces unrooted trees but does not assume constant rate of evolution
      2. Parsimony Methods
         1. One of four primary methods
         2. Parsimony may be used to estimate "species" or "gene" phylogenies.
         3. In the parsimony approach, the goal is to identify that phylogeny that requires the fewest necessary changes to explain the differences among the observed sequences.
      3. Maximum Likelihood
         1. One of four
         2. Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree.
         3. The more probable the sequences given the tree, the more the tree is preferred.
         4. All possible trees are considered; computationally intense.
         5. Because the user can choose a model of evolution, the method can be useful for widely divergent groups or other difficult situations.
         6. The resultant value is the "probability" (technically, "likelihood") of the observed sequences, assuming a specific model of evolution given this tree. This "probability" is presented as a log likelihood, thus the less negative (or larger) the number, the greater the probability.
      4. Bayesian Approaches
         1. One of four
         2. Based on maximum likelihood methods but incorporates prior probability.
         3. "Prior Probability" - probability of hypthesis according to previous information.
         4. Use complex sampling methods (i.e.Markov Chain Monte Carlo Methods).
         5. MrBayes: [http://mrbayes.csit.fsu.edu/index.php/](http://mrbayes.csit.fsu.edu/index.php)
      5. Reliability Tests
         1. Not only can trees be estimated, but their reliability or robustness (i.e., accuracy) can be evaluated as well.
         2. Reliability refers to the probability that members of a clade will be part of the true tree.
         3. Bootstrapping is the most common reliability test.
            1. In bootstrapping, resampling of the sites in the alignment is used to build new trees.
            2. These extra samples are created with "replacement" - it is possible that some positions will be repeated in the subsample, while some positions will be left out.
            3. Multiple resamples (hundreds to thousands) are run.
            4. The closer the score is to 100, the more significant the grouping.
            5. Bootstrapping can be used with distance, parsimony and likelihood methods.
   6. Data Set Organization 🡪 2 ways

|  |
| --- |
| * + 1. "Sequential" or "interleaved" formats are ways that aligned sequence data may be prepared for phylogenetic analysis. |

|  |
| --- |
| * + 1. https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/images/phylo30.gif |

|  |
| --- |
| * + 1. The data set above is presented in "sequential" form.     2. Note that if a sequence for a particular species is quite long, it will begin again on the next line.     3. Long sequences may therefore be difficult to visually compare using the "sequential" format. Thus, one might use the interleaved format, as shown below.     4. The number 5 indicates the number of sequences; the number 42 represents the number of bases. |

|  |
| --- |
| * + 1. https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/images/phylo31.gif |

|  |
| --- |
| * + 1. Note that even when the sequence is long, it is easier to visually compare individual sequences in this interleaved format. |

1. Other Information
   1. Types of Trees
      1. They differ based on analysis
      2. Ultrametric vs. parametric trees
         1. Ultrametric has branch lengths proportional to time
            1. In distance methods i.e. gene based, it has root to branch tip distances as equal i.e. a constant rate based assumption
         2. Parametric has branch lengths
   2. Molecular clock and rate of evolution
2. Building Trees 🡪
   1. Data collection
      1. Code on importing sequences into R
      2. Ape pulls in data from GenBank
   2. Multiple Sequence alignment
      1. Programs to generate alignments try to maximize the number of matches among all sequences being aligned
      2. Msa package
   3. Determine substitution model based on data
      1. Rphast package
   4. Build Trees
      1. Have to pick a package:
         1. Ape
         2. Phybase
   5. Evaluate Tree
      1. Bootstrapping
         1. Rphylip package
         2. Ape package
   6. Examine traits for tree
      1. Treebase package
         1. Combine phylogeny data with traits, geography, or ecological community of taxa through Dryad and its repository
   7. Visualize Trees
      1. ggtree
3. Packages in R 🡪 pick 3 and then can show the cran site for different packages, can include at the end of worksheets
   1. Treebase
      1. Package to download trees from TreeBase, online repository of phylogenies and phylogenetic data
         1. Treebase is online repository of phylogenetic data from published pee-reviewed literature
         2. Trees of species, populations, or genes
      2. Can query web database from R in variety of ways and access all available phylogenies (won’t do that since it takes a full day, but still! So cool!)
      3. Data discovery
         1. Search for existing data that meet certain desired characteristics
         2. Take advantage of metadata 🡪 summary info describing data entries in repository
         3. Metadata can reveal trends in data deposition, useful in identifying patterns or biases in research or emerging potential types of data, but mostly interested in describing phylogenies
         4. Phylogenetic metadata:
            1. Number of taxa in tree, quality score if there, kind of tree (gene, species, or barcode), if phylogeny represents a consensus tree from distribution or single estimate
      4. Can combine phylogeny data with traits, geography, or ecological community of taxa through Dryad and its repository
   2. Phylobase
   3. Phybase
      1. Package to simulate and analyze gene trees and species trees
      2. Read, write, manipulate, simulate, estimate, and summarize phylogenetic trees (species and gene)
      3. Input/output can read and write in Nexus and Phylip
      4. Trees read as string and transformed to matrix to described relationship of nodes and branch lengths
   4. GGTree
      1. Visualizing associations in phylogenetic trees can help identify evolution patterns
      2. Next gen tree viewer that’s programmable and extensible
   5. Ape
      1. functions for reading and manipulating phylogenetic trees and DNA sequences, computing DNA distances, estimating trees with **distance-based methods**, and a range of methods for comparative analyses and analysis of diversification
   6. rphast
      1. provides R interface to PHAST software to conduct analyses in comparative and evolutionary genomics, such as estimating models of evolution from sequence data, score aligments for conservation or acceleration, and predicting elements based on conservation or custom phylogenetic trees
   7. msa
      1. The 'msa' package provides a unified R/Bioconductor interface to the multiple sequence alignment algorithms ClustalW, ClustalOmega, and Muscle.
      2. All three algorithms are integrated in the package, they do not depend on any external software tools and are available for all major platforms.
      3. The multiple sequence alignment algorithms are complemented by a function for pretty-printing multiple sequence alignments using the LaTeX package TeXshade.
   8. Rphylip
      1. Provides R interface for programs in PHYLIP phylogeny methods package
4. Can also have packages where I’m not running things, but just showing the final product and here’s a package for that, sample code if you’re interested in digging in