Species Tree Inference using SVDquartets

Laura Kubatko

Joint work with Dr. Julia Chifman
Wake Forest University School of Medicine

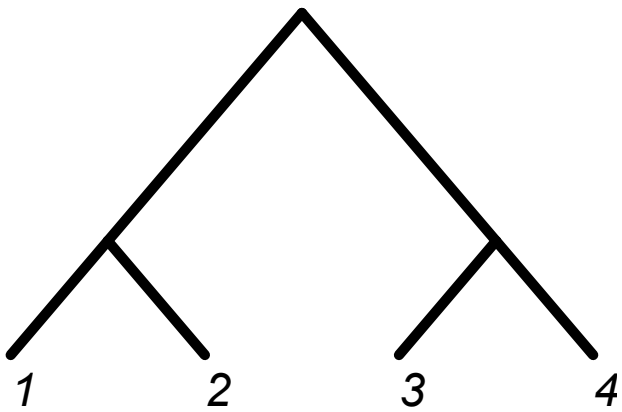
January 31, 2015

- In this tutorial, we'll discuss several different data types:
  - Multi-locus data – aligned DNA sequence data for many genes
  - SNP data – large number of SNPs sampled throughout the genome
  - Single-locus data – aligned DNA sequence data for a single gene
- In the first two cases, we'll assume that incongruence between gene trees and the species trees arises solely from the coalescent process
- In the third case, we assume that the locus under consideration is a single non-recombining unit

Goal: Estimate the underlying phylogenetic tree (species tree or gene tree)

Definition: splits

- **Definition:** A split of a set of taxa $\mathcal{L}$ is a bipartition of $\mathcal{L}$ into two non-overlapping subsets A and B, denoted $A|B$. A split $A|B$ is valid for tree T if the subtrees containing the taxa in A and in B do not intersect.



- Valid: $12|34$
- Not valid: $13|24$
  $14|23$

Definition: flattenings

$$p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l)$$

$$Flat_{12|34}(P) = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots \\ [AA] & p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots \\ [\cdots] & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

Theorem (Chifman and Kubatko 2014): Under the coalescent model and the GTR+I+Γ model and its sub models, we have the following:

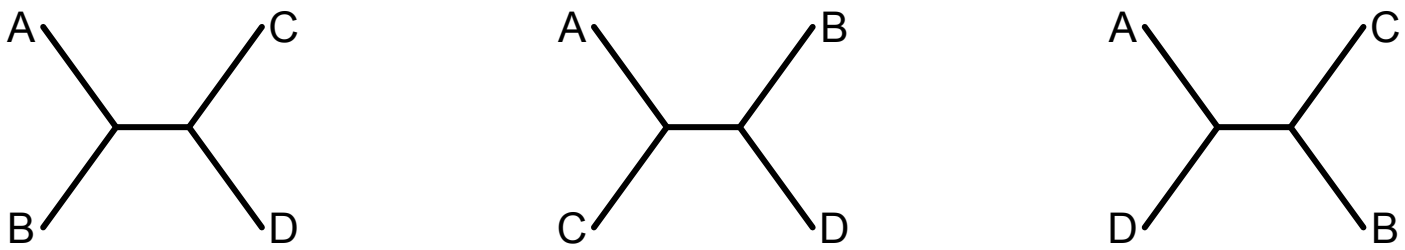
- If $A|B$ is a valid split for a tree T, then $rank(Flat_{A|B}(P)) \leq 10$.
- If $C|D$ is not a valid split for a tree T, then $rank(Flat_{C|D}(P)) > 10$.
- The species tree is completely determined by knowledge of valid splits on all quartets.

Extensions of the Main Result

- Arbitrary number of states, κ, under the coalescent model:
  - If $A|B$ is a valid split for a tree T, then $rank(Flat_{A|B}(P)) \leq \binom{\kappa+1}{2}$.
  - If $C|D$ is not a valid split for a tree T, then $rank(Flat_{C|D}(P)) > \binom{\kappa+1}{2}$.
  - The species tree is completely determined by knowledge of valid splits on all quartets.

- Single underlying gene tree (no coalescent assumption):
  - If $A|B$ is a valid split for a tree T, then $rank(Flat_{A|B}(P)) \leq 4$.
  - If $C|D$ is not a valid split for a tree T, then $rank(Flat_{C|D}(P)) = 16$.
  - The species tree is completely determined by knowledge of valid splits on all quartets.

Main idea: use the observed site pattern distribution to provide information about which of the three possible splits for a set of four taxa is the true split.



The program SVDquartets computes a score for each split in a given quartet of taxa and chooses the split with the best (lowest) score.

Algorithm

1. Generate all quartets (small problems) or sample quartets (large problems)
2. Estimate the correct quartet relationship for each sampled quartet
3. Use a quartet assembly method to build the tree
   - PAUP* uses the method of Reaz-Bayzid-Rahman (2014), called QFM, to build the tree.

- Variability in the estimated tree is assessed using nonparametric bootstrapping

- Multiple lineages are handled as follows:
  1. Sample four species
  2. Select one lineage at random from each species
  3. Estimate the quartet relationships among the four sampled lineages
  4. Restore the species labels (but lineage quartets are saved, too)

- Advantages:
  - Fast! How fast?
    - Rattlesnakes: ¡ 1 hour ($\sim$ 8500bp, 52 tips)
    - Soybeans: ¡ 1 day (6 million SNPs, 62 tips)
  - Scales well:
    - Number of quartets needed increases as number of species increases (but can be done in parallel)
    - Linear in number of sites (but this is just counting)
  - Potential for application to other data types
  - Natural way to handle missing data
- Disadvantages:
  - Only the (unrooted) topology is estimated – no parameters

- Described in the paper

  Chifman, J. and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model, *Bioinformatics*

- Implemented in PAUP* – thanks, Dave!

- Now on to the tutorial!