**PRANK Notes**

* **Background**

1. Traditional MSA methods disregard the phylogenetic implications of gap patterns that they create and infer systematically biased alignments with excess deletions and substitutions, too few insertions, and implausible insertion-deletion-event histories.
2. When we study the functional and structural sequence changes leading to phenotypic differences between species.
3. The center to all analysis is accurate sequence alignment, which means the correct identification of homologous nucleotides or amino acids and the positioning of gaps indicating inserted and deleted sequences. However, this is still a highly error-prone step in comparative sequence analysis. Different MSA methods give different conclusions in both phylogenetic analyses and functional studies (online material), Alternative alignments of the same data can support entirely different mechanisms driving evolutionary and functional changes in sequences.
4. For ex: HIV and SIV gp120 familiar pattern of insertion and deletions squeezed compactly between conserved blocks of structurally important residues. Part of variable v2, High amino acid substitution rate and has shortened over time at a mutation hotspot, where overlapping sites have been independently deleted in different evolutionary branches 10/23 deleted.
5. PRANK

* **Parameters**
* **Hypothesis/The method**

1. Prevent the systematic errors by recognizing insertions and deletions as distinct evolutionary events.

* **Phylogeny Reconstruction**

Basic concepts used to describe trees and general features of tree reconstruction methods

1. **Distance and Parsimony Methods (3)**
2. **Likelihood and Bayesian Methods (4,5)**

* A phylogeny of a tree is a representation of the genealogical relationships among the species, among genes, among populations or even among individuals. In a tree we have root, node, and leaf. The tips/**leaves** or external nodes represent the present-day species or a column of the MSA sequence. The **nodes**/ internal nodes represent the ancestors/extinct ancestors for which no sequence data are available. The number of branches connected to the node is the *degree* of the node. When the degree is 2 then it is called *binary/bifurcating tree or fully resolved tree.* When the degree is more than 2 then it is called a multifurcating trees. A *clade* is a group of species. The **root** is the ancestor of the overall tree. The trees can be represented with or without root. When we represent a tree with root this is called a *rooted tree* and a tree in which the root is unknown is called an *unrooted tree.* The **branch length** represents the evolutionary rate/clock and if it is constant over time, then an assumption known as *molecular clock,* distance matrix and maximum likelihood methods can identify the ancestral paths. When we use this clock assumption it is called *molecular-clock rooting.* If we are not using any of these assumptions, then we use another strategy called *outgroup-rooting.*
* “Root the tree using ancient gene duplications that occurred prior to the divergence of all existing life forms (Gogarten et al. 1989; Iwabe et al. 1989)”
* The branching pattern of a tree is called the *topology* of the tree. A tree represented without any information about the branch lengths are called as *cladograms.* While a tree showing both tree topology and branch lengths are called *phylograms.* The trees are often represented with parenthesis notation in computer. The tree file is in Newick file format.
* Sometimes we are interested in how different two trees are. May be the difference among trees estimated from different genes, or the differences between the true tree and the estimated tree in a computer simulation conducted to evaluate a tree-reconstruction method.
* A commonly used measure of topological distance between the two trees is the 1. **Partition distance.** Limitations are: Does not recognize certain similarities between the trees, Ignores the branch length (solution: methods which incorporates the branch lengths in defining a distance between the two trees), partition distance may be misleading if we have multifurcating tree. If two trees are identical then the partition distance is 0, whereas the if they are extremely different then the Dmax will be 2(n-3), where n is the number of species in the tree. The partition distance to measure performance, P= 1 – D/ Dmax, where D is the distance between the true tree and the simulated tree/estimated tree.
* 2. **Consensus trees.** These trees summarize common features among a collection of trees.
* **Gene trees and Species trees.** There are several factors which may cause the gene tree to differ from the species tree. First, the estimation error may cause the estimated gene tree to be different from the species trees even if the true gene tree agrees with the species tree. Second, In the early stages of life that is near to the root of the tree there appears to have Lateral Gene Transfer (LGT). As a result, different proteins or genes may have different gene trees, in conflict with the species tree. Third, gene duplications, especially if followed by gene losses then this will result in an entirely different species tree. Fourth, when species are closely related the ancestral polymorphism or lineage sorting can cause gene trees to be different from the species tree.

**Classification of tree-reconstruction methods**

1. **Distance based**

In this the distances are calculated from **pairwise comparison of sequences**, forming a distance matrix, which is then used for further analysis. To construct tree from distance matrix clustering methods such as; UPGMA (unweighted pair-group method using arithmetic averages), Neighbor Joining (NJ).

1. **Character based and attempt to fit the characters observed in all species at every site to a tree.**

Maximum parsimony, maximum likelihood (ML), minimum evolution and Bayesian methods are all based on character based.

**Maximum parsimony**- Minimum number of changes, minimized over ancestral states.

**Maximum likelihood**- Log likelihood score, optimized over branch lengths and model parameters.

**Minimum evolution**- Tree length (sum of branch lengths, often estimated by least squares)

**Bayesian**- Posterior probability, calculated by integrating over branch lengths and substitution parameters.

1. **Algorithmic (Cluster methods) or Optimality based methods (Search methods)**

Algorithmic include UPGMA and NJ, which uses cluster algorithms to arrive at a single tree from the data (MSA, sequences) as the best estimate of the true tree.

Optimally based methods use an optimality criterion (objective criterion) to measure the tree’s fit to data, and the tree with the optimal score is the estimate of the true tree. In Maximum parsimony, the tree score is the minimum number of characters changes required for the tree, and the maximum parsimony tree or most parsimonious tree is the tree with the smallest tree score. The ML method uses the log likelihood value of the tree to measure the fit of the tree to the data, and the maximum likelihood tree is the tree with the highest log likelihood value. In Bayesian method, the Posterior probability of a tree is the probability that the tree is true given the data. The tree with the maximum posterior probability is the estimate of the true tree, known as the MAP tree. The optimality-based methods have to solve two problems: calculation of the criterion for a given tree and search in the space of all trees to identify the tree with the best score. To search heuristic methods are implemented to do the tree search.

1. **Model based**

Distance based methods use nucleotide or amino acid substitution models to calculate pairwise distances. Likelihood and Bayesian methods use substitution models (For ex. WAG in our case) to calculate the likelihood function. Parsimony does not make explicit assumptions about the evolutionary process.

**Stochastic Processes**

Collection of Random Variables indexed by Time. Time variable can be discrete and continuous. An alternative definition: Probability distribution over a space of paths. What are the dependencies in the Sequence of values? What is the long-term behavior on the sequence (Law of large numbers, Central limit theorem), What are the boundary events? Over a period of time kind of problems. Discrete Stochastic Processes **Random Walk.** Yi: iid RV.