A study of tree construction methods in Phylogeny

Sukesh Pai

Dec 05, 2001

1 Introduction

Phylogeny is the problem of recreating the evolutionary history of a set of biomolecules from the information about the similarity of these molecules. The evolution mostly proceeds as a hierarchy or a tree. Given a phylogeny tree for a set of biomolecules, we can say that the biomolecule at the root of the tree evolved through mutation or speciation into each of the slightly different biomolecules at the leaves of the tree while going through the form of biomolecules at the intermediate nodes in the tree.
Multiple sequence alignment and other forms of data give information about the similarity or resemblence of a set of biomolecules. Such data is used to reconstruct the phylogeny of the biomolecules. There have been various methods, metrics and models used to study this problem and recreate the phylogeny. There are some very good methods that have been around for quite a while now, but recently there is a lot of interest in creating new methods that recreate the phylogeny from minimal information. In this report, we discuss some of the recent developments in phylogenetic methods. Until recently, the methods that reconstruct phylogeny assumed that the data had very restrictive properties like ultrametry or additivity. Moreover, most of the methods have serious problems scaling for a large number of biomolecules or taxa. Also, since accuracy in recreating the phylogeny is of utmost importance, the quality of the data had to be extremely good. Of late, new methods have come out which could use available data and generate more accurate phylogenies. These methods can work even with short sequence lengths while relaxing stringent constaints on the property of the data. Here is a summary of the methods that are studied in the report and their relative performances.

1.1 Summary

The first set of methods are used for the phylogeny problem in DNA sequences or protein molecules. They assume that a similarity score for each

pair of biomolecules is given in the form of a matrix under a general markov model. Further, the methods can work well as long as the matrices supplied are at least approximately additive (discussed further).

Method	Complexity	fast converging?	perf. on large taxa	perf. with rt. of evolution
Dyadic Closure	$O(n^5 logn)$	yes	unknown	unknown
Witness Antiwitness	$O(n^4 logn)$	yes	unknown	unknown
Neighbor Joining	$O(n^3 logn)$	no (unknown)	poor	poor
Harmonic Greedy Triplets	$O(n^5 + ln)$	yes	good	good
Disk-Covering NJ+MP	$O(n^6)$	yes	good	good

Here n is the number of biomolecules or sequences (taxa) and l is the sequence length.

The second set of methods are meant for studying the phylogeny problem for gene order rearrangement. This assumes that the evolution in a genome happens when a substring of genes are rearranged within the genome to create a new one. Under this model of genome evolution, we compare the following methods.

Method	Metric	Speed	perf. with # of genes	perf. with # of taxa
IEBP	Distance	use Neighbor Joining	accuracy reduces	accuracy reduces
Exact-IEBP	Distance	use Neighbor Joining	accuracy reduces	accuracy reduces
EDE	Distance	use Neighbor Joining	accuracy reduces	accuracy reduces
MPBE	Parsimony	comparable	constant accuracy	constant accuracy
MPME	Parsimony	slow	constant accuracy	constant accuracy

1.2 Analysis and Potential Improvements

Most of the good phylogenetic methods are heuristics for solving the NP-Hard phylogeny problem. A lot of effort is put into proving fast convergence of phylogenetic methods. The efforts have been fairly successful and now we have very good fast methods that can produce phylogeny trees with good accuracy. However, simple heuristics like Neighbor-Joining are still very

competitive with these very sophisticate methods. For a small problem, with limited number of taxa, Neighbor-Joining method is as good as any other top of the line method. The new methods have opened up avenues for studying a big set of taxa having reasonable length sequences.
The methods studied are based tightly on the kind of evolutionary model assumed. As we see, the model used for protein sequence evolution cannot be appropriate to study chromosome evolution which is predominantly through gene rearrangement. We have a different model to study those evolutions. Going ahead, there may be various other models that come up to suite the case at hand. There could be models to study evolution that adhere to some grammar in the sequences such that evolution produces only sequences compatible to certain grammatical rules. Under such models, the problem may turn out not to be NP-Hard, in which case elegant solutions would be available.
More complex models can be evolved out of existing ones by combining them. DNA molecules undergo some rate of mutations of the nucleic acids while some amount of inversion or transposition of the genes. Thus a combination of Jukes-Cantor and Nadeu-Taylor may be of interest.
There is probably some inherent limit on the size of the set of taxa that can be studied in one group. Certain molecules show increased mutation in a set of taxa during some period. The rate of evolution of the same set of molecule is not the same through out the history. So also, a group of species that were formed showed evolution with respect to some set of molecules and not with others. This means that we can use a set of molecules to study only a set of species. We may need to switch the set of molecules to study a different set of species. Now comes the problem of combining these trees such that we can create a bigger tree with all these species. For this, we can use the techniques used in calculating consensus trees.
The perspective in some of the phylogeny methods waver from the requirement of accuracy to the speed in solving the problem. More strict way of calculating accuracy need to be studied and experiments performed on existing methods on how they fare against the new metric.
We look at the tabulated methods in some detail while understanding the assumptions and conditions necessary for the use of these methods. We outline some distance based method and take a look at why the sequence length is an important criteria in designing a method. We study some quartet based methods and how they perform followed by breakpoint phylogeny methods and their performance.

2 Distance Based Methods

2.1 Convergence: Fast Convergence and Absolute Convergence

Let D be an $n \times n$ distance matrix. A method M is said to converge if it generates the true phylogeny tree for sufficiently long sequences at the leaves. The method is said to be combinatorially consistent if M(D) = D whenever D is additive. The method is continuous at d for L_∞ metric $(\max|d_{ij} - d'_{ij}|)$ if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $L_\infty(d, d') < \delta$ implies $L_\infty(M(d), M(d')) < \epsilon$ [1].
A method that is both combinatorially consistent and continuous on a neighborhood around every additive matrices is said to be **reasonable**. We want to show some bounds and conditions on reasonable methods for convergence. Bounds are defined for a method based on the length of the sequences used at the leaves. To see why the length of the sequences are important, consider a tree with n leaves. There could be $(2n-5)!!$ tree topologies for a binary tree with n leaves. If a method has to generate the true tree topology from a given set of sequences, the number of possible tree topologies that those sequences could generate has to be atmost one. The total number of n leafed trees that can be generated by sequences of length k is 2^{nk} assuming the sequences are binary. Thus we need at least the condition $(2n-5)!! \leq 2^{nk}$ for a method to be able to recreate the true tree.
For a given distance matrix d, computed under some distance model, we call it *statistically consistent* if each of the distance estimates d_{ij} converges to the true value λ_{ij} under the CF model (see Appendix), as the sequence length increases, with probability 1. This is true for General Markov model of which CF model is a part.
Theorem 1: Let (T, M) be a model tree in the General Markov model. Set $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda_e$. Assume that f, g are fixed with $0 < f \leq \lambda_e \leq g$ for all edges $e \in T$. Let $\epsilon > 0$, $\delta > 0$ be given. Then, there is a constant C such that, if the sequence length exceeds

$$Clogn e^{O(g.diam(T))} \tag{1}$$

then with probability at least $1-\delta$, we have $L_\infty(d, \lambda) = max_{ij}|d_{ij} - \lambda_{ij}| < \epsilon$, where d is the statistically consistent distance matrix, λ is the matrix of true distance, n is the number of leaves and $diam(T)$ is the topological diameter of T.
If the term in equation 1 is polynomial or polylogarithmic in n, then the method is said to be fast converging.
Definition 2: A phylogenetic reconstruction method Φ is absolute fast-converging (afc) for the General Markov model if, for all positive f, g, ϵ, there is a polynomial p such that, for all (T, M) in the General Markov

model, on set S of n sequences of length at least $p(n)$ generated on T, we have Probability$[\Phi(S) = T] > 1 - \epsilon$.

Definition 3: A phylogenetic reconstruction method Φ is relative fast-converging (rfc) for the General Markov model if, for all positive f, g, ϵ, there is a polynomial p such that, for all (T, M) in the model, on set S of n sequences of length at least $p(n)$ generated on T, we have Probability $[\Phi(S, A(f, g)) = T] > 1 - \epsilon$, where $A(f, g)$ denotes an oracle that provides information about f and g.

Most of the fast converging phylogeny methods are relative fast converging. That is their convergence for a given sequence length depends upon the values of the parameters f and g. Some of the absolute fast converging methods known are the Short Quartet methods and the Disk-Covering method.

A general technique for building the true phylogeny tree is to generate a set of good trees based on some criteria and evaluating each generated tree for the most optimal among them. Usually, the criteria used for generating good trees is to use some artificial parameter (specific to each method) that constraints the problem to give a more deterministic solution. This will be evident when we discuss some of these methods.

3 Neighbor-Joining Method

This method is among the first phylogeny methods to build consistent trees from additive matrices. The method has been around since 1987 and still is quite reliable in terms of accuracy and running time. [2]. Given an $n \times n$ distance matrix d, the method proceeds by identifying a pair of leaves as neighbors and joins them to a parent node, continuing the process until only two nodes are left. At every step the distance matrix is revised to "normalize" the distances from a node based on the average distance of the node from other nodes. At any given step, the two neighbors (i and j) selected are the pair of nodes which are closer than any other pair based on the current distance values. When these two nodes are joined, a new node k is created such that k is the parent node of i and j and the distance of k is the average of the distance $(d_{im} + d_{jm} - d_{ij})/2$ for all other nodes m in the set.

It has been proved that for additive matrices, the nodes with minimal distances are indeed the neighbors. The reliability of Neighbor-Joining method to cases where the distance matrix is not strictly additive is also proven in literature. Thus Neighbor-Joining method is one of the most robust, simple and efficient methods known in phylogeny. However, this method is not proven to be fast converging and needs sequences of exponential length to show convergence (this is the current known upper bound).

4 Short Quartet Methods

Short Quartet methods analyze simple small subtrees with four leaves to construct the true phylogeny tree [1]. See Appendix for definitions on Quartets and other tree properties.

4.1 Dyadic Closure Method

This method tries to build the phylogeny tree by computing the closure set for a given set of quartets by infering the rest from the given set [3]. Given a dissimilarity distance matrix d of size $n \times n$, a binary search over $q \in \{d_{ij}\}$ is done such that:
Four Point Method (FPM) is used to compute the set of quartet splits from d for quartets with d-width bounded by q. Let this set be A_q. Then, the method builds the closure set $cl(A_q)$ called *dyadic closure* of A_q based on the following two rules:
Rule 1: Given two quartet trees $ij|kl$ and $jk|lm$, infer the quartet trees $ij|km$ and $ik|lm$
Rule 2: Given two quartet trees $ij|lk$ and $ij|km$, infer the quartet tree $ij|lm$.
So, $cl(A_q)$ is the minimal set of quartet trees which contains A_q and is closed under the above two rules. It has been shown that the closure can be computed in $O(n^5)$.
A search through the values of q is made such that the $cl(A_q)$ contains exactly one tree on every quartet. The unique tree that comes out of this closure is returned as the final answer; otherwise the method returns failure. The Dyadic Closure Method is shown to have a running time of $O(n^5 logn)$ and is fast-converging for polylogarithmic length sequences.

4.2 Witness-Antiwitness Method

Witness-Antiwitness method is an improvement over Dyadic closure method as it gives an answer in more cases and also runs faster [4].

4.2.1 Witness-Antiwitness Tree Construction Algorithm, WATC

Given a set Q of quartet splits WATC outlines how to construct a tree T consistent with Q.
Stage I

- Start with every leaf of T defining an edi-subtree (edge-deletion-induced subtree).
- While there are at least four edi-subtrees, do:

- Form a graph G on vertex set given by the edi-subtree, and with edge set defined by siblinghood.
  - Case 1: *there are more than four edi-subtrees*: if the graph has at least one edge, select one and make the roots of the edi-subtrees, which are the vertices on the edge, children of a common root and replace them by the new subtree. If no component edge exists, return *Fail*
  - Case 2: *there are exactly four edi-subtrees*: Let the four subtrees be x, y, z, w. If the edge set of the graph G is $\{(x, y), (z, w)\}$, then construct the tree T formed by making the edi-subtrees x and y siblings, the edi-subtrees z and w siblings, and adding an edge between the roots of the two new edi-subtrees. If not, return *Fail*.

Stage II

- Verify that T is consistent with Q. If so, return T, else return *Fail*.

4.2.2 Witness-Antiwitness Method, WAM

WATC gives us the tree T if one exists for a given set Q of quartet splits. WATC is guaranteed to reconstruct the binary tree T if the set Q is T-forcing on T. Hence, some search strategies are developed so that we can always find a T-forcing quartet split on which to invoke the WATC algorithm.
Let Q_w be the quartet set bounded by a d-width of w. A sequential search strategy is to invoke WATC on each Q_w, $w \in \{d_{ij}\}$. This has a search space of $O(n^2)$. A better search strategy is an $O(logk)$ **sparse-high search**, where k is the sequence length.
The running time for WAM based on sequential search is $O(n^2k + n^6logn)$ and on sparse-high search is $O(n^2k + n^4lognlogk)$.

5 Harmonic Greedy Triplets Method

Harmonic Greedy Triplets Method [5] is a fast converging method with polynomial running time that always produces a tree. The algorithm proceeds constructively by building the tree at each step selecting a leaf not yet in the tree. When a leaf is selected, a new internal node is added such that it forms the center of the best triplet formed by the new leaf and two other leaves already in the tree. A triplet is a set of three node x, y and z such that their center p is at a distance $d_{xp} = (d_{xy} + d_{xz} - d_{yz})/2$. The triplet selected at each time is by the greedy approach of selecting a triplet with the new leaf node that has the maximal average closeness value. The closeness value for a set of three leaves is the harmonic mean of the pair-wise closeness values.

While selecting the triplet with the maximal closeness value, it is ensured that the closeness value is above a certain bound decided by the algorithm. The running time for this algorithm is shown to be $O(n^5 + ln^2)$, where n is the number of sequences and l is the length of the sequences used.

6 Disk Covering Method

Disk Covering Method or DCM is a meta-method that can be used in conjunction with other phlogenetics methods. It is a very general phylogenetic method booster that improves the accuracy of the phylogenetic methods used with it. Unlike the Short Quartet methods, DCM always reconstructs a tree with better performance [6].
Here, let us discuss DCM-Buneman method, that is Buneman method (see Appendix) used with DCM.

6.1 DCM Method

The method has two phases. In the first phase, it takes a phylogenetic method and a dissimilarity matrix and produces a collection of trees. In the second phase, the most appropriate tree is selected from this set.

6.1.1 Phase I

Given a distance matrix D, a set S of sequences and a phylogenetic base method Φ, the Phase I produces a set of trees from maximal cliques on a threshold graph as below:

- For each $w \in D_{ij}$
  - Let $E_w = \{(i,j) | D_{ij} \leq w\}$. Construct the threshold graph, $TG(D,w) = (S, E_w)$.
  - If $TG(D,w)$ is not connected, then let T_w be the star tree. Put T_w in the set of trees and continue for next in loop.
  - Triangulate $TG(D,w)$, minimizing $max\{D_{ij} | (i,j) added to (S, E_w)\}$, thus producing the triangulated graph $TG^*(D,w)$. (the heuristic step)
  - Compute the maximal cliques $C_1, C_2, \ldots, C_l$ of $TG^*(D,w)$ where $l \leq n$. For each i, $1 \leq i \leq l$, let $t_i = \Phi(C_i)$.
  - Merge the subtrees t_i using the strict consensus merger to produce the tree T_w. Add T_w to the set and continue. (Strict consensus merger method contracts a minimum set of edges in each tree in order to make them identical on the subtrees they induce. The

strict consensus of the induced subtrees is the maximally resolved tree that is a common contraction of the subtrees).

- return the set of trees $\{T_w | w \in \{D_{ij}\}\}$

6.1.2 Phase II

In Phase II, we return the most resolved tree T_w (the one with the most internal edges), and if a tie exists, it is broken by choosing the one associated with a higher w.
It has been shown that DCM-Buneman is fast converging. Later on, an improvement to this approach was suggested. This used Short Quartet Support method to replace phase II of DCM. This method DCM^* has theoretical guarantee of absolute fast convergence if the phylogenetic method used was at least relative fast converging [1].

6.2 Short Quartet Support Method

Short Quartet Support method decides the best tree T among a set of trees generated from a given dissimilarity matrix D over a set of sequences S.

- for each set of four sequences from S, compute the neighbor-joining quartets q; let Q be the set of all such quartets
- Return T_i from the set of trees such that the support for T_i with respect to Q, $s(T_i, Q)$ is maximum (see Appendix for the definition of support). Break the tie, if one exists, by selecting the smalllest i.

7 Performance Comparison

The methods studied above can be compared for various parameters like complexity, behavior of a method under short, long or very long sequences, behavior under varying number of sequences in the analysis etc. The metric biologists are most concerned with is the accuracy of each method. That is how closely does the tree generated by each method resemble the true tree. There is no true tree to compare for real sequences found in nature. But the accuracy of a method is judged based on some simulated evolution of a model tree and then using methods to regenerate the tree and comparing how close the generated tree is to the model tree. The quatification of accuracy is done by the **Robinson Foulds score (RF)** (see Appendix).

7.1 Performance of the Quartet based Methods

Dyadic closure, witness-antiwitness and disk closure methods were studied by researchers [7] [8] [9] and they found that disk closure method (DCM^*) in

conjunction with Neighbor-Joining (NJ) method performed the best among them. This is denoted as DCM^*-NJ. Base phylogeny methods were compared with the DCM-boosted version and it was found that DCM-boosted methods always outperfomed the base methods. Later studies comparing DCM^*-NJ, Harmonic Greedy Triplets method (HGT) and NJ were conducted. It was observed that for a given set of taxa, as the sequence length increased, the accuracy of DCM^*-NJ and HGT improved in relation to NJ. At smaller sequence lengths (< 4000), NJ outperformed the other two. It was also observed that DCM^*-NJ performed better than HGT, but overall, the gain in performance of these methods was not significant compared to NJ even for longer sequences.
Following these results, more experiments were done by changing the DCM method slightly. In the Phase II of DCM method, the criteria used to pick the final tree in case there is a tie is changed to pick a tree with the maximum score $s(T)$, where s is the number of thresholds w such that all quartets of diameter w agree with T. This method is named DCM-NJ+TS (TS for threshold support). Other variants were DCM-NJ+ML and DCM-NJ+MP with the tie in Phase II being broken by the maximum likelihood and maximum parsimony trees respectively.
It was observed that DCM-NJ+TS consistently performs at least as well as DCM^*-NJ. DCM-NJ+MP and DCM-NJ+ML performed equally well in all the test and they outperformed DCM-NJ+TS. Comparing DCM-NJ+ML/MP, the best of DCM methods with HGT and NJ shows that DCM-NJ-ML/MP outperformed NJ; and NJ outperformed HGT in experiments with varying sequence lengths (0 - 8000) for a fixed taxa size of around 100.
Similar experiments conducted with varying the taxa size (50 - 1600) for a fixed sequence length of 1000 shows that for low branch lengths, DCM-NJ-MP/ML outperforms NJ and HGT. The performance of NJ consistently degrades as the taxa size increases and at taxa size of 1600, HGT has better performance than NJ. Also, for higher branch lengths, the performance of NJ degrades faster while HGT and DCM-NJ+MP/ML continue to perform at almost constant accuracy.

8 Breakpoint Phylogeny

Hitherto we have been discussing the construction of phylogeny trees by looking at DNA or protein sequences of a set of organisms. The uderstanding being that the protein underwent some amount of mutation as speciation happened. These mutations were assumed to be point mutations where one nucleotide or amino acid changes into a different one with some probability. There are other kinds of changes that take place that are of interest to biologists. One such is the rare genomic changes of large-scale mutational events

in genomes that cause rearrangements including gene duplication or change in gene order. Researchers have studied such genomic changes in chormosomes (linear or circular ordering of genes) [10]. The kind of rearrangement of interest are inversions, transpositions and inverted transposition (see Appendix for definition).

8.1 Generalized Nadeau-Taylor Model

This model assumes that inversion, transposition and inverted transpositions are the three types of events that cause genome rearrangements so that all genomes retain equal gene content. Further, the number of each of the three types of events obey a Poisson distribution on each edge of the phylogeny tree. The relative probabilities of each type of event are fixed across the tree and the events of a given type are equiprobable. Concisely, this model can be represented as a triplet $(T, \{\lambda_e\}, (\gamma_I, \gamma_T, \gamma_{IT}))$, where the triplet $(\gamma_I, \gamma_T, \gamma_{IT})$ defines the relative probabilities of inversions, transpositions and inverted transpositions [11] [12].

8.2 Breakpoints

Consider two genomes $A = a_1, a_2, \ldots, a_n$ and $B = b_1, b_2, \ldots, b_n$ on the same set of genes $\{g_1, g_2, \ldots, g_n\}$. We say a_i and a_{i+1} are adjacent in A. If two genes g and h are adjacent in A but not in B, they determine a breakpoint in A. The breakpoint distance is the number of breakpoints in A relative to B and is equal to the number of breakpoints in B relative to A [13].
The problem in Breakpoint Phylogeny is, given a set of n genomes, reconstruct the phylogeny tree that minimizes the breakpoint distance at every edge of the tree. This has been shown as an NP-hard problem for the simplest case of three linear genomes by reducing this median problem to traveling salesman problem with m cities where m is the number of genes in the given genomes. An approach to reconstruct the phylogeny tree was conceived based on this idea. The outline of the method is as follows:

- Generate all tree topologies for a tree of n leaves
- each tree is given an internal node label through a heuristic iterative procedure that repeatedly finds the median of three neighboring labels.
- the median for three labels is computed by solving the corresponding traveling salesman problem.
- select the tree with the minimum breakpoint distance.

This constitutes the **BPAnalysis** algorithm [14].

There have been attempts to convert the breakpoint phylogeny into a distance based phylogeny problem through the use of suitable metrics that convert the number of breakpoints, inversions or transpositions into true evolutionary distances on the edge of a tree. Once we compute distances between every pair of genomes in this way, we can use any of the distance based method to reconstruct the phylogeny tree. There is extensive literature about various such metrics and demonstrates the relevance of the metric by using Neighbor-Joining method to reconstruct the tree.
Some of them include, IEBP, Exact-IEBP and EDE. IEBP or Inverting the Expected BreakPoint distance method approximates the expected breakpoint distance obtained after k random events in the generalized Nadeau-Taylor Tree. Using this, given two genomes, we can estimate the true evolutionary distance (t.e.d.) between them by selecting the number of events most likely to have created the observed breakpoint distance. The Exact-IEBP improves the accuracy by providing an exact calculation of the expected breakpoint distance [11]. EDE, Empirically Derived Estimator method estimates the t.e.d. by inverting the expected inversion distance [15].
There are other methods that are based on parsimony. MPBE or Maximum Parsimony on Binary Encodings [16] translates every genome into a binary sequence, where each site from the binary sequence corresponds to a pair of genes. The site takes a value 1 if the corresponding genes are adjacent in the genome or it take the value 0. Thus for n genes, we have $C(n, 2)$ sites in the binary sequence. From these sequences, the hamming distances for each pair of genomes is computed which is used to reconstruct the tree using maximum parsimony. The other parsimony based method is MPME or Maximum Parsimony on Multistate Encodings [15]. In this method, an n genes genome is translated to a sequence of length $2n$ such that for every i, $1 \leq i \leq n$, site i takes the value of the gene immediately following gene i and site $n+i$ takes the value of the gene immediately following gene $-i$. So the circular gene $(g_1, -g_4, -g_3, -g_2)$ becomes the sequence (-4, 3, 4, -1, 2, 1, -2, -3). This new sequence is then used to construct the parsimony tree.

8.3 Performance

All the above distance based and parsimony methods were used in a experiment to test for their accuracy [15]. Neighbor Joining was used as the tree construction method on all of the distance methods. It was found that EDE was the best among the distance methods, coming close to the parsimony methods. The parsimony method fared better than the distance methods in general, with MPME being the most accurate though quite slow.

9 Appendix

9.1 Different Distance Models

9.1.1 Jukes-Cantor

Jukes-Cantor model defines sequence evolution over a finite alphabet

$$A = \{a_1, \ldots, a_m\} \tag{2}$$

An *Evolutionary Tree* T for A is a binary tree with n leaves and an *edge mutation probability* p_e for each tree edge e. Also $p \in [0, 1 - 1/m]$ such that there exists f and g such that for every tree edge e

$$0 < f \leq p_e \leq g < 1 - 1/m \tag{3}$$

Let a sequence of length l at the root of the tree mutate at every edge of the tree to give n sequences at the leaves. Each such sequence can be represented as $s_1, \ldots, s_l \in A^l$. At every edge e each s_i remains unchanged with a probability of $1 - p_e$ and mutates with a probability of $p_e/(m-1)$ for each different symbol [5].
The edge transitions can be efficiently represented as an $m \times m$ symmetric matrix with the diagonal values being $1 - p_e$ and non diagonal values being $p_e/(m-1)$.
Jukes-Cantor model can be applied to nucleotide substitution (alphabet is {A, C, G, T}) or for amino acid substitutions among others. Note that the Jukes-Cantor matrix is symmetric and assumes that when mutation occurs, it happens with equal probability of substitution to a different alphabet. However, this may not be practical in the case of nucleotide substitution where there could be a higher probabilty of substitution among purines or among pyrmidines than a purine substituting a pyrmidine or the other way around. This is taken care by the **Kimura** model [2].

9.1.2 Simplified Jukes-Cantor: Binary characters

Most of the phylogeny methods discussed here use a simplified version of the Jukes-Cantor model with only two character. So, we have a 2×2 substitution matrix at each tree edge and p_e takes the value in the range $[0, 0.5]$.
This is sometimes called the Cavender -Felsenstein or Cavender-Farris Model [17]. An equivalent definition of the Cavender-Farris (CF) tree is $(T, \{\lambda_e : e \in E(T)\})$, where λ_e is the expected number of changes of a random site on edge e where the random variable for the number of changes on each edge is Poisson. It has been shown that $\lambda_e = 1/2ln(1 - 2p_e)$. CF trees are used in proving fast convergence of sequences. This is also sometimes referred as Neyman model in literature [3]. All these models belong to the General Markov model.

9.2 Distance Matrix Properties

The pairwise distances for the n sequences is given as an $n \times n$ matrix D. Note that this matrix has all the diagonal elements zero and is symmetric. Such a matrix is called a **dissimilarity matrix** [17].

9.2.1 Additive Matrix

Consider a tree T (like in CF model) with weights w_e on each edge e. We can define the distance between any two leaves as the distance $d_{ij} = \sum_{e \in P_{ij}} w_e$, sum of the weight on the edges along the path from i to j (where P_{ij} denotes the path from i to j). The $n \times n$ matrix constructed out of these d_{ij} distances will be a dissimilarity matrix. Such a matrix for which there exists a tree with positive edge weights and the distances satisfying the equation above is said to be an **Additive Matrix**.
It has been shown that given an additive $n \times n$ matrix, the unique tree consistent with the pair wise distances can be reconstructed in $O(n^2)$ time.

9.2.2 Ultrametric Matrix

Ultrametric matrix is an additive matrix with a rooted tree such that the distance from the root to all the leaves is equal. The **molecular clock hypothesis** held that the phylogenies formed ultrametric trees as the rate of mutation with time is a constant and hence the amount of mutation along any path to a leaf should be the same.

9.3 Centroid Matrix

Centroid matrix is an additive matrix which can be realized by edge-weighting a star topology (the root of the tree having an edge to each leaf).

9.4 Relationship between Additive, Ultrametric and Centroid Matrices

Let D be an additive matrix and X be a centroid matrix. Then D + X is an ultrametric matrix. This gives us a general strategy to obtain a nearby additive matrix for a given distance matrix d. We compute a suitable centroid matrix X and create a new distance matrix $d^{'} = d + \mathrm{X}$. Using one of the methods, we compute an ultrametic matix U close to $d^{'}$. From this we compute the required additive matrix D = U - X and then use this additive matrix to reconstruct the phylogeny tree.

9.5 Definitions for Quartet trees

A **Quartet** is a set of four leaves i, j, k, l from a given tree T, such that they can induce a subtree in T.

D-width of a quartet is the maximum of the pairwise distances D_{ij} for the four leaves, given an additive matrix D for the tree.
A **Short Quartet** around an edge e of a tree T is a quartet with leaves in each of the four subtrees around that edge of minimum D-width. The set of all such short quartets for all the edges in the tree is denoted by $Q(T)$.
Let T be a fixed tree leaf-labelled by the set S of n sequences, Q a fixed set of quartets on S, and D the distance matrix on S. The **support** of T with respect to Q denoted by $s(T, Q)$, is

$$max\{l|(q \in QandD - width(q) \leq l) \Longrightarrow q \in Q(T)\} \tag{4}$$

For a fixed set of quartets Q, given distance matrix D, Q_w is defined as $Q_w = \{q \in Q : \text{D-width}(q) \leq w\}$

9.6 Four Point Method

Four Point Method (FPM) Given a 4×4 distance matrix d, return the set of splits $ij|kl$ which satisfies $d_{ij} + d_{kl} \leq min\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}$ [3]. FPM can return one, two or three splits for each quartet based on whether the minimum is unique, two of them are same or all three are same respectively. If we use FPM on a truely additive distance matrix D, we will get unique split for each quartet and from the set of splits return by FPM, we can reconstruct the phylogeny tree uniquely in polynomial time. But, in most practical cases, the distance matrix is not truely additive and FPM may generate multiple splits for some quartets. Generating a tree from this set of quartet splits is hard since there could be conflicting splits that do not converge to the same tree.

9.7 Buneman Method

This method by Buneman, starts by infering topology of quartets using Four Point method (FPM) for a given dissimilarity matrix d. From this set of quartet splits Q, a *maximally resolved* tree is constructed satisfying the condition:
For all quartets $\{i, j, k, l\}$, if T restricted to i, j, k, l induces a binary tree (instead of a star), then the tree in Q on i, j, k, l is the same binary tree.

Let e be an edge in T. Deletion of the edge e from the tree T creates two rooted subtress T_1 and T_2. These are called edge-deletion-induced or edi subtrees. Consider that the leaves are numbered from 1 to n for the tree T. For each edi-subtree t, we define the unique leaf which is the lowest numbered among those closests topologically to the root of t as *representative* of t and is denoted as $rep(t)$.
For each $(n-3)$ internal edge of the n-leaf binary tree T, we pick a *representative quartet* $\{i, j, k, l\}$ such that the deletion of the edge and its end

points leaves behind four rooted subtrees and each i, j, k, l is the representative for each of these subtrees. Note that by definition, every representative quartet is a short quartet. Each representative quartet can be given a split according to the split of the subtrees each representative came from. The set of these representative quartets is denoted as R_T. We have

$$R_T \subseteq Q_{short}(T) \subseteq Q(T) \tag{5}$$

where $Q_{short}(T)$ is the set of short quartet splits realizable for the tree T.
It has been shown that R_T is consistent with a unique tree T, that is, all the quartet splits in R_T is valid for tree T and that is the unique tree for which it is true.
Also, if $R_T \subseteq Q$, then Q can only be consistent with the tree T.
So, for any Q such that $R_T \subseteq Q \subseteq Q(T)$ would mean that if Q is consistent with some tree, the tree is unique. The closure set used by Dyadic closure method had this property. Note, however that Q could be inconsistent, in which case there would be no tree consistent with Q.

Definition: Let T be a binary tree and Q a set of quartet splits on the leaves of T.

- Q has the **witness property for** T if, whenever t_1 and t_2 are sibling edi-subtrees of T and $T - t_1 - t_2$ has at least two leaves, then there is a quartet split in Q, $uv|wx$ such that $u \in t_1$, $v \in t_2$ and $\{w, x\} \cap (t_1 \cup t_2) = \emptyset$. This quartet split is called the *witness* to the siblinghood of t_1 and t_2.

- Q has the **antiwitness property for** T if, whenever there is a witness in Q to the siblinghood of two edi-subtrees t_1 and t_2 which are not siblings in T, then there is a quartet split in Q $pq|rs$ such that $p \in t_1$, $r \in t_2$and $\{q, s\} \cap (t_1 \cup t_2) = \emptyset$. This quartet split is called the *antiwitness* to the siblinghood of t_1 and t_2.

Definition: A set Q of quartet splits is said to be T-forcing if there exists a binary tree T such that

1. $R_T \subseteq Q \subseteq Q(T)$.

2. Q has the antiwitness property for T.

9.8 Definitions of rearrangements in Breakpoint Phylogeny

Let the genomes being studied be composed of genes from a set $G = \{g_1, g_2, \ldots, g_n\}$ of genes. Each genome is an ordering (circular or linear) of some multi-subset of these genes. Also, each gene can take either orientation and be positive (g_i) or negative ($-g_i$). A circular genome can

be represented as a linear genome under the implicit assumption that the permutation closes back on itself. The canonical representation of a circular genome is the linear representation where gene 1, g_1, is at the first position with positive sign. Thus a circular genome (g_1, g_2, g_3) is same as $(-g_1, -g_3, -g_2)$, but the first representation is canonical.
Let X be a genome with signed ordering $g_1, g_2, \ldots, g_k$. An **inversion** between indices a and b, for $a \leq b$, produces a genome with linear ordering

$$g_1, g_2, \ldots, g_{a-1}, -g_b, -g_{b-1}, \ldots, -g_a, g_{b+1}, \ldots, g_k$$

A **transposition** on the linear or circular ordering acts on three indices a, b, c with $a \leq b$ and $c \notin [a, b]$, picking up the substring $g_a, g_{a+1}, \ldots, g_b$ and inserting it immediately after g_c. Thus the genome X is replaced by

$$g_1, g_2, \ldots, g_{a-1}, g_{b+1}, \ldots, g_c, g_a, \ldots, g_b, g_{c+1}, \ldots, g_k$$

An **inverted transposition** is a transposition followed by an inversion of the transposed substring [16].

9.9 Calculating Accuracy of phylogenetic methods

If T is a model tree and $T^{'}$ is an estimation of the model tree by some phylogenetic method, we have

- Let $e \in E(T)$ be an internal edge of T, and let π_e be the bipartition of the set of sequences S induced by deleting the edge e from T. Let $C(T) = \{\pi_e : e \in E(T)\}$ be the set of all such bipartitions for T and similarly let $C(T^{'}) = \{\pi_e : e \in E(T^{'})\}$. These sets are called character encodings of T and $T^{'}$ respectively.
- The **false positives** are those bipartitions in $FP = C(T^{'}) - C(T)$, and the **false negatives** are those bipartitions
- **false negative rate** is defined as $|FN|/|E_I(T)|$ and **false positive rate** is defined as $|FP|/|E_I(T)|$, where $E_I(T)$ is the set of internal edges in T.

Robinson Foulds score (RF) is defined as the average of false negative and false positive rate [17]. Most performance studies on phylogeny reconstruction methods have used this score as the metric for finding accuracy of various methods.

References

[1] T. Warnow, B. Moret, and K. St. John. Absolute convergence: True trees from short sequences. *Proc. of 12th Annual Symposium of Discrete Algorithms*, pages 186–195, 2001.

[2] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis. *Cambridge University Press*, 1998.

[3] P.L. Erdos, M. Steel, L. Szekeley, and T. Warnow. A few logs suffice to build (alost) all trees - 1. *DIMACS Technical Report*, 97-71, 1997.

[4] P.L. Erdos, M. Steel, L. Szekeley, and T. Warnow. A few logs suffice to build (alost) all trees - 2. *DIMACS Technical Report*, 97-72, 1997.

[5] M. CSUROS and M.-Y. KAO. Recovering evolutionary trees through harmonic greedy triplets. *Proc. of 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[6] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *J. Computational Biology*, 6:369–386, 1999.

[7] K. St. John, T. Warnow, B. Moret, and L. Vawter. Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. *Proc. of 12th Annual Symposium of Discrete Algorithms*, pages 196–205, 2001.

[8] L. Nakleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. Designing fast converging methods in phylogenetics: An experimental study. *Intelligent Systems for Molecular Biology*, 2001.

[9] L. Nakleh, B. Moret, U. Roshan, K. St. John, J. Sun, and T. Warnow. The accuracy of fast phylogenetic methods for large datasets. *Proc. of 7th Pacific Symposium on Biocomputing*, 2002.

[10] M. Blanchette, T. Kunisawa, and D. Sankoff. Parametric genome rearrangement. *Gene*, page 172.

[11] L.-S. Wang. Exact-iebp: a new technique for estimating evolutionary distances between whole genomes. *Proc. 1st workshop Algorithms Bioinformatics WABI'01*, 2194:176–190, 2001.

[12] B. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps towards accurate reconstruction of phylogenies from gene-order data. *J. Computer System Science*.

[13] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. *Genome Informatics*, pages 25–34, 1997.

[14] D. Sankoff and M. Blanchette. The median problem for breakpoint in comparative genomics. *Computing and Combinatorics, Proceeding of COCOON'97*, 1997.

[15] L.-S. Wang, R. Jansen, B. Moret, L. Raubeson, and T Warnow. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. *Proc. of 7th Pacific Symposium on Biocomputing*, 2002.

[16] M. Cosner, R. Jansen, B. Moret, L. Raubeson, S. Wang, T. Warnow, and S. Wyman. A new fast heuristics for computing the breakpoint phylogeny and a phylogenetic analysis of a group of highly rearranged chloroplast genomes. *Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB00)*, pages 104–115, 2000.

[17] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation. *Intelligent Systems on Molecular Biology*, pages 196–205, 1999.