Computational Phylogenetics for Variant Manuscripts in Sanskrit

A Seminar Report

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

by

Diptesh Kanojia
(154054002)

under the guidance of

Prof. Pushpak Bhattacharyya
and
Prof. Malhar Kulkarni

IITB - Monash Research Academy
Indian Institute of Technology Bombay
Mumbai

November 2016

Abstract

Over the last 15 or more years, the use of computational techniques for estimating evolutionary histories of languages (i.e. Computational Phylogenetics), has seen a tremendous increase. It is used to evaluate the relationships between various taxa, using computational methods. We plan to use the Computational Phylogenetic methods detailed in the report to collate multiple versions of a Sanskrit text, and thus prepare a critical edition from the said versions of the text. This work surveys different methods and different types of linguistic data that have been used to estimate phylogenies. It explains the scientific foundations of phylogenetic estimation, and presents methodologies for evaluating a phylogeny estimation method. Our work also surveys the techniques used to choose the best available model for tree construction and phylogenetic analysis. We have also included the recent papers related to the research area. With this work, eventually, we aim is to find how these methods would help us further our research for preparing critical editions for Sanskrit Manuscripts.

Contents

List of Figures

Chapter 1

Introduction

Phylogenetics is defined as the task of creating a Phylogenetic Tree which represents a hypothesis about the evolutionary ancestry of a set of genes, species or any other taxa. It is the study of evolutionary history and relationships among various taxa. A Taxon represents a group of one or more manuscripts written in Sanskrit in our case, where we analyze how the manuscripts are related to each other.

These relationships are discovered through phylogenetic methods that compute observed heritable traits in a manuscript, such as spelling errors, variations in text, text deletion, morphology of the text etc. under a model of evolution of these traits. The result of these analyses is a phylogeny (also known as a phylogenetic tree) – a diagrammatic hypothesis about the history of the evolutionary relationships of a group of manuscripts (usually belonging to the same text).

The Computational purview of our research problem deals with developing new methods for alignment estimation, phylogeny estimation, and species identification of metagenomic data (short reads). Computational historical linguistics, which involves the development of methods for estimating evolutionary histories of languages and of models of language evolution, is another research problem based on phylogenetics. Phylogenetic methods are designed to recover the ”true” evolutionary tree as often as possible. They do not guarantee to do so with high probability under reasonable conditions. Some which offer this guarantee vary considerably in their requirements (Warnow *et al.*, 2001). To rigorously establish the validity of such a phylogenetic approach, a fundamental question that must be addressed is whether the models in use are *identifiable*. From the theoretical distribution predicted by the model, is it possible to uniquely determine all parameters ?

Parameters for simple models include the topology of the evolutionary tree, edge lengths on the tree, and rates of various types of substitution, though more complicated models have additional parameters as well.

If a model is non-identifiable, one cannot show that performing inference with it will be statistically consistent. Informally, even with large amounts of data produced by an evolutionary process that was accurately described by the model, we might make erroneous inferences if we use a non-identifiable model. Under other models, many methods will be able to recover the tree if given long enough sequences. The latter methods are said to be *statistically consistent* under the model of evolution. Under some models of evolution, no method can be guaranteed to recover the true tree with high probability, due to unidentifiability.

For many models (such as the Jukes-Cantor model (Munro, 2012), the simplest four-state model, as well as more complex models, such as the General Markov (GM) (Steel, 1994) model), even simple distance methods are easily established to be *statistically consistent*.

Using the currently available models, finding optimal phylogenetic trees using compatibility criteria is in its general case NP-Complete (Warnow, 1993). Also, finding a Maximum Compatible Tree is NP-Hard (Roch, 2006). As a consequence, this will mean that efficient algorithms to solve the problem probably can not exist. On the other hand, by restricting the kinds of input to the problem, we may be able to solve it efficiently.

The trees generated by phylogenetic methods can be either *rooted* or *unrooted*, depending on **the input data** and **the algorithm used to build the tree**. A rooted tree is a directed graph that explicitly identifies a most recent common ancestor (MRCA), usually an imputed sequence that is not represented in the input. An unrooted trees plot the distances and relationships between input sequences without making assumptions regarding their descent. It can always be produced from a rooted tree, but a root cannot usually be placed on an unrooted tree without additional data on divergence rates (Mount, 2004).

For identifying a root, it is usually required that that inclusion in the input data of at least one ”outgroup” is know to be distantly related to the sequences of interest. Genetic distance measures can be used to plot a tree with the input sequences as leaf nodes and their distances from the root proportional to their genetic distance from the hypothesized MRCA.

An example of a phylogenetic tree is shown below in figure 1.1 where a phylogenetic tree is drawn manually from Malyalam manuscripts for Kāśikāvrtti.

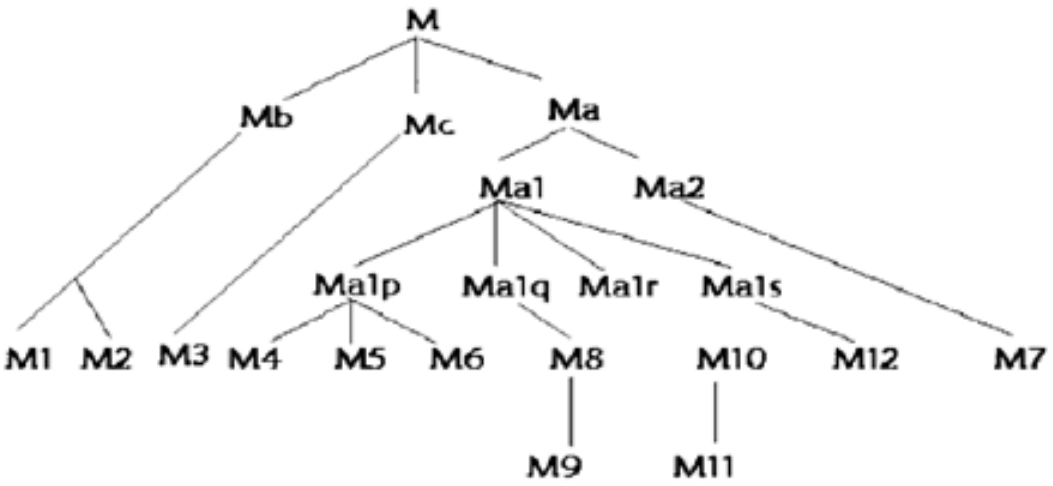

Figure 1.1: A Sample Stemma of Malayalam Manuscripts as shown by Kulkarni (2006)

Kulkarni (2006) show that M is the archetype source and Ma, Mb and Mc are its hyperarche child nodes. M is decided as a source based on the analysis made on the variant readings. In this process manuscripts have similar variants are grouped together and named as M1, M2, M3 . . . , M11.

The rest of the report is organized as follows: Chapter 2 discusses Computational Phylogenetics in detail by classifying phylogenetic trees, and the recent trends in molecular Phylogenetic Analysis. Chapter 3 entails brief description of the non-statistical Phylogenetic methods. The statistical methods for Computational Phylogenetics are detailed in Chapter 4. It also provides the basis of our model selection for phylogenetic tree construction, and the recent trends in the area. Chapter 5 enlists the limitations to the methodology, and some possible workarounds, and Chapter 6 concludes our survey, and provides a possible direction for future research.

Chapter 2

Phylogenetic Trees

The goal of Computational Phylogenetics is to assemble the phylogenetic tree based on the analysis of evolutionary data available through computational methods. Several types of methods viz. Distance-matrix based methods, Maximum Likelihood, Bayesian Inference already exist for drawing inference from evolutionary data and modeling the tree.

A phylogenetic tree shows the inferred evolutionary relationships among various taxa (biological species, languages, manuscripts etc.) i.e. their phylogeny, based on their similarities and differences in the characteristics. The taxa in the tree are implied to have descended from a common ancestor.

In a phylogenetic tree, every leaf represents a manuscript version. Nodes are labeled, either with manuscript names or the assigned values (also referred to as states) of their characters (the region it may have come from), and the edges represent the evolutionary connections.

It is important to note that there is usually a big difference between the leaf nodes, that represent real species, and the internal nodes, that in most cases represent the hypothetical evolutionary ancestors of the manuscripts in the data.

2.1 Tree Classification

Phylogenetic trees take several forms: They can be rooted or unrooted, binary or general, and may show, or not show, edge lengths. **A rooted tree** is a tree in which one of the nodes is stipulated to be the root, and thus the direction of ancestral relationships is determined. **An unrooted tree**, as could be imagined, has no pre-determined root and therefore induces no hierarchy. Rooting an unrooted tree involves inserting a new node,

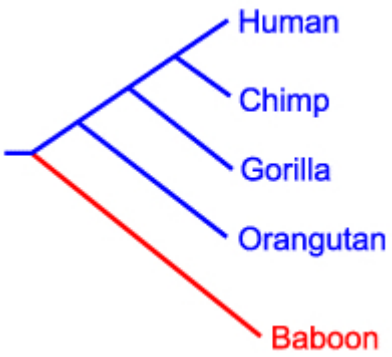

Figure 2.1: Rooted Tree

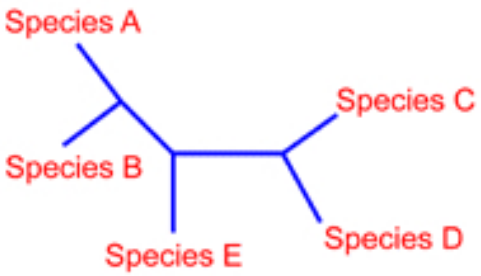

Figure 2.2: Unrooted Tree

which will function as the root node, between two existing nodes. Figures 2.1 and 2.2 show a rooted tree[1] and its unrooted counterpart[1], respectively.

A binary, or bifurcating, tree is of course a tree in which a node may have only 0 to 2 subnodes, that is, in an unrooted tree, up to three neighbors. It is sometimes useful to allow more than 2 subnodes (**multifurcation**), but the discussion in this lecture will be limited to binary trees.

A rooted bifurcating tree has exactly two descendants arising from each interior node (that is, it forms a binary tree), and an unrooted bifurcating tree takes the form of an unrooted binary tree, a free tree with exactly three neighbors at each internal node. A labeled tree has specific values assigned to its leaves, while an unlabeled tree, sometimes called a tree shape, defines a topology only. The number of possible trees for a given number of leaf nodes depends on the specific type of tree, but there are always more multifurcating than bifurcating trees, more labeled than unlabeled trees, and more rooted than unrooted trees. The last distinction arises because there are many places on an unrooted tree to put the root. For labeled bifurcating trees, there are:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!},\ for\ n >= 2 \tag{2.1}$$

total rooted trees and

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!},\ for\ n >= 3 \tag{2.2}$$

total unrooted trees, where n represents the number of leaf nodes (Felsenstein, 2004).

[1] `https://www.ncbi.nlm.nih.gov/Class/NAWBIS/Modules/Phylogenetics/phylo9.html`

Among labeled bifurcating trees, the number of unrooted trees with n leaves is equal to the number of rooted trees with $n - 1$ leaves.

The branches of a phylogenetic tree may be represented two different ways:

- **Scaled branches** : Branches will be different lengths based on the number of evolutionary changes or distance as shown in Figure 2.3.
- **Unscaled branches** : All branches in the tree are the same length as shown in Figure 2.4.

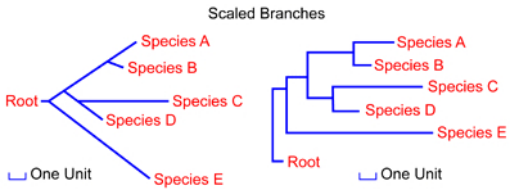

Figure 2.3: Scaled Tree

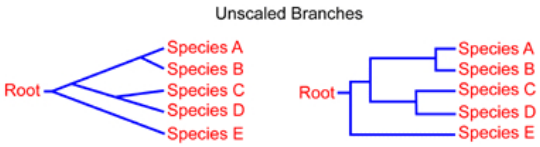

Figure 2.4: Unscaled Tree

A tree can show edge lengths, indicating the genetic distance between the connected nodes. We sometimes assume the existence of a molecular clock, a constant pace of the evolutionary processes. If this is the case, we could theoretically produce a phylogenetic distance-preserving tree which can be presented along a time-axis assigning to each node the time in which it "occurred" in the history of evolution. In such a "perfect" tree, the length of each edge would be the difference in time between the parent node and the child node.

2.2 Tree Construction

Phylogenetic trees, given a nontrivial number of input sequences, are constructed using computational phylogenetics methods. Distance-matrix methods which calculate genetic distance from multiple sequence alignments, are the simplest to implement, but do not invoke an evolutionary model.

Many sequence alignment methods such as ClustalW also create trees by using the simpler algorithms (i.e. those based on distance) of tree construction. Maximum parsimony is another simple method of estimating phylogenetic trees, but implies an implicit model of evolution (i.e. parsimony).

More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian Framework, and apply an explicit model of evolution to phylogenetic tree estimation (Felsenstein, 2004). Identifying the optimal tree using many of these techniques is NP-hard(Felsenstein, 2004), so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

Tree-building methods can be assessed on the basis of several criteria (Penny *et al.*, 1992):

- **Efficiency** : how long does it take to compute, how much memory does it need?
- **Power** : does it make good use of the data, or is information being wasted?
- **Consistency** : will it converge on the same answer repeatedly, if each time given different data for the same model problem?
- **Robustness** : does it cope well with violations of the assumptions of the model?
- **Falsifiability** : does it alert us when it is not good to use ? violated assumptions ?

Tree-building techniques have also gained the attention of mathematicians. Trees can also be built using T-theory (Dress *et al.*, 2001).

2.3 A Critical Edition

Maas (2009) create a phylogenetic tree for "Caraksamhita Vimanasthana" based on their findings in analysis of fifty-two manuscript versions of the text as shown in figure 2.5. They collate the results to find out that above manuscripts record more than 4000 variants, and more than 97% of all words and nominal stems in Trikamji's edition Trikamji (1941). They go on to detail that the manuscripts a large number of these variants can be ignored since they are insignificant scribal mistakes, but there are a significant number of changes in the text which has been inherited from the original source. Sanskrit Manuscripts like Caraksamhita were written many hundred years ago, and have been copied ever since for knowledge transfer. The copying was done by scribes who would travel to far away places, and bring manually wrote down the manuscripts over leaf / metal / wooden plates etc.

In copying a not too short passage of text a scribe can make mistakes and, at some instances, they may even deliberately change the wording of his exemplar i.e. source. In this way, they creates a new textual version which differs from the version of his exemplar in containing variant readings. This process of creating new versions with every new copy has probably kept changing the CS ever since the first copy of the final redaction by *Drdhabala* was prepared, perhaps about 1500 years ago. When a new version is copied, the scribe reproduces the variants which were created in the previous copy, and in addition, introduces new variants himself. The process of copying and recopying produces a hierarchical pattern of variants, so that some variant readings can be identified as being characteristics of whole lines of the transmission. Based on their identification, it is possible to create a genealogical tree, i.e., a "stemma," of all available and infer-able versions.

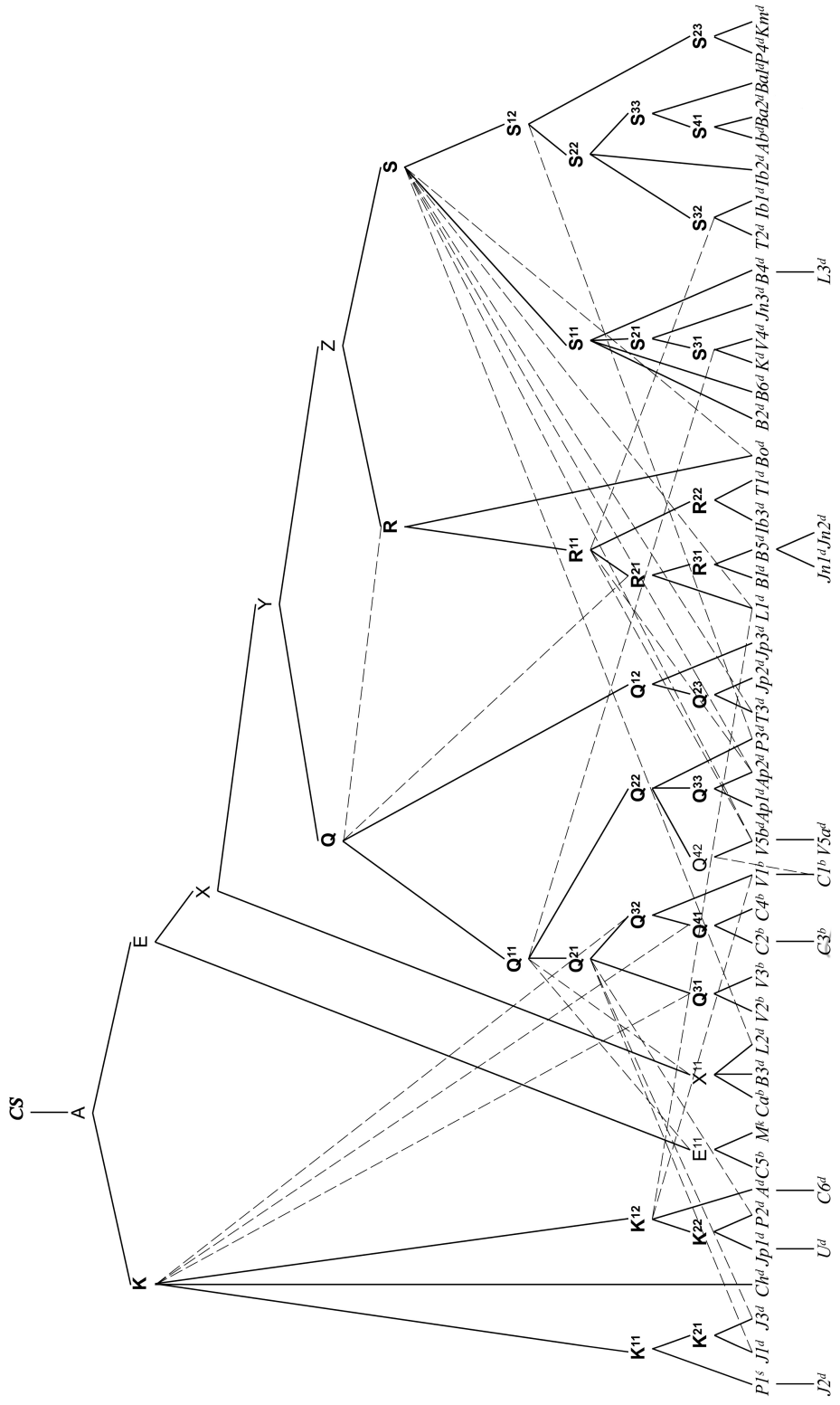

Figure 2.5: Phylogenetic Tree for a Critical Edition of *Caraksamhita*

Chapter 3

Phylogenetic Methods

The ”goal” of phylogenetic analysis is to recover ”bifurcating” trees, in which each taxon is linked to one other taxon through a node. Bifurcating trees are usually the most informative, because they tell which OTUs are most closely related. Polychotomous trees (those with multiple branches coming from one node) are usually less informative because they indicate that multiple OTUs are related to each other, but not how. (Some polytomies can be real, e.g. adaptive radiations). Various methods for performing it are listed in the section below.

3.1 Methods for Phylogenetic Analysis

Phylogenetic trees can be constructed using various computational phylogenetic methods. Phylogenetic Analysis has various methods(Geer R.C., 2002):

3.1.1 Distance Methods

A variety of distance algorithms are available to calculate pairwise distance, for example, Proportional (p) distances. Distance analysis compares two aligned sequences at a time, and builds a matrix of all possible sequence pairs. 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. These final estimates of the difference between all possible pairs of sequences are known as pairwise distances. Once the pairwise distances are calculated, they must be arranged into a tree. There are many ways to ”arrange” the species or genes according to their distances. One way to cluster or

optimize the distances is to join species or genes together according to their increasing differences, as embodied by their distances. Other ways use various coefficients to measure how well the branch lengths of the tree reflects the original pairwise distances. Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, and therefore they require an MSA (multiple sequence alignment) as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches (David, 2001). The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees (Felsenstein, 2004).

Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as the basis for progressive and iterative types of multiple sequence alignment. Various Distance-matrix based methods are:

Neighbor-joining

They apply general data clustering techniques to sequence analysis and uses genetic distance as a clustering metric. The simple version of neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a molecular clock) across lineages.

UPGMA and WPGMA

The Unweighted Pair Group Method with Arithmetic mean (UPGMA), and Weighted Pair Group Method with Arithmetic mean (WPGMA) methods produce rooted trees and require a constant-rate assumption i.e. they assume an ultrametric tree in which the distances from the root to every branch tip are equal.

Fitch-Margoliash method

This method uses a weighted least squares method for clustering based on genetic distance (Fitch *et al.*, 1967). Closely related sequences are given more weight in the tree construction process to correct for the increased inaccuracy in measuring distances between distantly related sequences. In practice, the distance correction is only necessary when the evolution rates differ among branches (Felsenstein, 2004) The distances used as input to the algorithm must be normalized to prevent large artifacts in computing relationships between closely related and distantly related groups.

The distances calculated by this method must be linear and the linearity criterion for distances requires that the expected values of the branch lengths for two individual branches must equal the expected value of the sum of the two branch distances. This property applies to biological sequences only when they have been corrected for the possibility of back mutations at individual sites. This correction is done through the use of a substitution matrix such as that derived from the Jukes-Cantor model of DNA evolution.

The least-squares criterion applied to these distances is more accurate but less efficient than the neighbor-joining methods. An additional improvement that corrects for correlations between distances that arise from many closely related sequences in the data set can also be applied at increased computational cost. Finding the optimal least-squares tree with any correction factor is NP-complete (Day, 1987), so heuristic search methods like those used in maximum-parsimony analysis are applied to the search through tree space.

Outgroups

Independent information about the relationship between sequences or groups can be used to help reduce the tree search space and assemble roots from unrooted trees. Standard usage of distance-matrix methods involves the inclusion of at least one outgroup sequence. It is known to be only distantly related to the sequences of interest in the query set (Mount, 2004). This usage can be seen as a type of experimental control. If the outgroup has been appropriately chosen, it will have a much greater genetic distance and thus a longer branch length than any other sequence, and it will appear near the root of a rooted tree. Choosing an appropriate outgroup requires the selection of a sequence that is moderately related to the sequences of interest; too close a relationship defeats the purpose of the outgroup and too distant adds noise to the analysis (Mount, 2004). Situations in which the

species from which the sequences were taken are distantly related, but the gene encoded by the sequences is highly conserved across lineages, should be avoided. Horizontal gene transfer, especially between otherwise divergent bacteria, can also confound outgroup usage.

3.1.2 Parsimony Method

Parsimony analysis is the second primary way to estimate phylogenetic trees from aligned sequences. Parsimony may be used to estimate ”species” or ”gene” phylogenies. In the parsimony approach (Farris, 1970), the goal is to identify that phylogeny that requires the fewest necessary changes to explain the differences among the observed sequences.

The above approach is based on the law of Maximum Parsimony, which is influenced by Occam's Razor (Gauch, 2003). It minimizes the total number of character-state changes is to be preferred. Under the maximum-parsimony criterion, the optimal tree will minimize the amount of homoplasy (i.e., convergent evolution, parallel evolution, and evolutionary reversals). In other words, under this criterion, the shortest possible tree that explains the data is considered best. The principle is akin to Occam's razor, which states that—all else being equal—the simplest hypothesis that explains the data should be selected.

The most naive way of identifying the most parsimonious tree is simple enumeration - considering each possible tree in succession and searching for the tree with the smallest score. However, this is only possible for a relatively small number of sequences or species because the problem of identifying the most parsimonious tree is known to be NP-hard (Felsenstein, 2004). Consequently, a number of heuristic search methods for optimization have been developed to locate a highly parsimonious tree, if not the best in the set. Most such methods involve a steepest descent-style minimization mechanism operating on a tree rearrangement criterion.

Branch and bound

This algorithm is a method used to increase the efficiency of searches for near-optimal solutions of NP-hard problems Hendy and Penny (1982). It is particularly well suited to phylogenetic tree construction because it inherently requires dividing a problem into a tree structure as it subdivides the problem space into smaller regions. It also requires as

input both a branching rule (in the case of phylogenetics, the addition of the next species or sequence to the tree) and a bound (a rule that excludes certain regions of the search space from consideration, thereby assuming that the optimal solution cannot occupy that region).

Identifying a good bound is a challenging aspect of the algorithm's application to phylogenetics. A simple way of defining the bound is a maximum number of assumed evolutionary changes allowed per tree. A set of criteria known as Zharkikh's rules (Ratner *et al.*, 1995) severely limit the search space by defining characteristics shared by all candidate "most parsimonious" trees.

The two most basic rules require the elimination of all but one redundant sequence (for cases where multiple observations have produced identical data) and the elimination of character sites at which two or more states do not occur in at least two species. Under ideal conditions these rules and their associated algorithm would completely define a tree.

Sankoff-Morel-Cedergren algorithm

The Sankoff-Morel-Cedergren algorithm was among the first published methods to simultaneously produce an MSA and a phylogenetic tree for nucleotide sequences (Sankoff *et al.*, 1973). The method uses a maximum parsimony calculation in conjunction with a scoring function that penalizes gaps and mismatches, thereby favoring the tree that introduces a minimal number of such events. An alternative view holds that the trees to be favored are those that maximize the amount of sequence similarity that can be interpreted as homology, a point of view that may lead to different optimal trees (De Laet, 2005). The imputed sequences at the interior nodes of the tree are scored and summed over all the nodes in each possible tree.

The lowest-scoring tree sum provides both an optimal tree and an optimal MSA given the scoring function. Because the method is highly computationally intensive, an approximate method in which initial guesses for the interior alignments are refined one node at a time. Both the full and the approximate version are in practice calculated by dynamic programming (Felsenstein, 2004).

MALIGN and POY

More recent phylogenetic tree/MSA methods use heuristics to isolate high-scoring, but not necessarily optimal, trees. The MALIGN method uses a maximum-parsimony technique to compute a multiple alignment by maximizing a cladogram score, and its companion POY uses an iterative method that couples the optimization of the phylogenetic tree with improvements in the corresponding MSA (Wheeler and Gladstein, 1994). However, the use of these methods in constructing evolutionary hypotheses has been criticized as biased due to the deliberate construction of trees reflecting minimal evolutionary events (Simmons, 2004). This, in turn, has been countered by the view that such methods should be seen as heuristic approaches to find the trees that maximize the amount of sequence similarity that can be interpreted as homology (De Laet, 2005, 2015).

3.1.3 Maximum Likelihood

Maximum likelihood is the third method used to build trees. Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree. The more probable the sequences given the tree, the more the tree is preferred. All possible trees are considered in this methodology. It is said to be computationally intense because the user can choose a model of evolution, the method can be useful for widely divergent groups or other difficult situations. 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.

It uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. It also requires a substitution model to assess the probability of particular mutations, roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites.

In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but it is believed to be computationally intractable to compute due to its NP-hardness (Chor and Tuller, 2005).

The ”pruning” algorithm is often used to reduce the search space by efficiently calculating the likelihood of subtrees (Felsenstein, 2004). The method calculates the likelihood for each site in a ”linear” manner, starting at a node whose only descendants are leaves (that is, the tips of the tree) and working backwards toward the ”bottom” node in nested sets. However, the trees produced by the method are only rooted if the substitution model is irreversible, which is not generally true of biological systems. The search for the maximum-likelihood tree also includes a branch length optimization component that is difficult to improve upon algorithmically; general global optimization tools such as the Newton-Raphson method are often used.

3.1.4 Bayesian Approaches

Based on maximum likelihood methods but incorporates prior probability. ”Prior Probability” - probability of hypthesis according to previous information. They Use complex sampling methods (i.e. Markov Chain Monte Carlo (MRMC) Methods), although the choice of move set varies. The selections used in Bayesian phylogenetics include circularly permuting leaf nodes of a proposed tree at each step Mau and Newton (1997) and swapping descendant subtrees of a random internal node between two related trees (Yang and Rannala, 1997) The use of Bayesian methods in phylogenetics has been controversial, largely due to incomplete specification of the choice of move set, acceptance criterion, and prior distribution in published work (Felsenstein, 2004) Bayesian methods are generally held to be superior to parsimony-based methods. They can be more prone to long-branch attraction than maximum likelihood techniques, although they are better able to accommodate missing data (Kolaczkowski and Thornton, 2009; Simmons, 2012).

Whereas likelihood methods find the tree that maximizes the probability of the data, a Bayesian approach recovers a tree that represents the most likely clades, by drawing on the posterior distribution. However, estimates of the posterior probability of clades (measuring their 'support') can be quite wide of the mark, especially in clades that aren't overwhelmingly likely. As such, other methods have been put forwards to estimate posterior probability (Larget, 2013). MCMC methods can be described in three steps: first using a stochastic mechanism a new state for the Markov chain is proposed. Secondly, the probability of this new state to be correct is calculated. Thirdly, a new random variable (0,1) is proposed. If this new value is less than the acceptance probability the new state is accepted and the state of the chain is updated. This process is run for either thousands

or millions of times. The amount of time a single tree is visited during the course of the chain is just a valid approximation of its posterior probability. Some of the most common algorithms used in MCMC methods include the Metropolis-Hastings algorithms, the Metropolis-Coupling MCMC and the LOCAL algorithm of Larget and Simon.

MrBayes Software

MrBayes is a free software that performs Bayesian inference of phylogeny. Originally written by John P. Huelsenbeck and Frederik Ronquist in 2001 (Huelsenbeck *et al.*, 2001). As Bayesian methods increased in popularity MrBayes became one of the software of choice for many molecular phylogeneticists. It is offered for Macintosh, Windows, and UNIX operating systems and it has a command-line interface. The program uses the standard MCMC algorithm as well as the Metropolis coupled MCMC variant. MrBayes reads aligned matrices of sequences (DNA or amino acids) in the standard NEXUS format (Maddison *et al.*, 1997).

MrBayes uses MCMC to approximate the posterior probabilities of trees (Metropolis *et al.*, 1953) The user can change assumptions of the substitution model, priors and the details of the analysis. It also allows the user to remove and add taxa and characters to the analysis. The program uses the most standard model of DNA substitution, the 4x4 also called JC69, which assumes that changes across nucleotides occurs with equal probability (Jukes and Cantor, 1969). It also implements a number of 20x20 models of amino acid substitution, and codon models of DNA substitution. It offers different methods for relaxing the assumption of equal substitutions rates across nucleotide sites (Yang, 1993) MrBayes is also able to infer ancestral states accommodating uncertainty to the phylogenetic tree and model parameters.

MrBayes 3.2 new version of MrBayes was released in 2012 (Ronquist *et al.*, 2012) The new version allows the users to run multiple analyses in parallel. It also provides faster likelihood calculations and allow these calculations to be delegated to graphics processing unites (GPUs). It also provides wider outputs options compatible with FigTree and other tree viewers.

3.1.5 Reliability Tests of the Tree Obtained

Not only can trees be estimated, but their reliability or robustness (i.e., accuracy) can be evaluated as well. Reliability refers to the probability that members of a clade will be

part of the true tree. Bootstrapping is the most common reliability test. In bootstrapping, resampling of the sites in the alignment is used to build new trees. 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. Multiple resamples (hundreds to thousands) are run for inferring using this method.

If there are m sequences, each with n nucleotides (or codons or amino acids), a phylogenetic tree can be reconstructed using some tree building method. From each sequence, n nucleotides are randomly chosen with replacements, giving rise to m rows of n columns each. These now constitute a new set of sequences. A tree is then reconstructed with these new sequences using the same tree building method as before. Next the topology of this tree is compared to that of the original tree. Each interior branch of the original tree that is different from the bootstrap tree the sequence it partitions is given a score of 0; all other interior branches are given the value 1. This procedure of resampling the sites and the subsequent tree reconstruction is repeated several hundred times, and the percentage of times each interior branch is given a value of 1 is noted. This is known as the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered ”correct” as detailed in Nei and Kumar (2000). This test is available for four different methods: Neighbor Joining, Minimum Evolution, Maximum Parsimony, UPGMA, and Maximum Likelihood.

Parsimony analysis is the second primary way to estimate phylogenetic trees from aligned sequences. Parsimony may be used to estimate ”species” or ”gene” phylogenies. 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.

Chapter 4

Model Selection and Recent Trends

Computational phylogenetics methods rely on a defined substitution model that encodes a hypothesis about the relative rates of mutation at various sites. At their simplest, substitution models aim to correct for differences in the rates of transitions. The use of substitution models is necessitated by the fact that the genetic distance between two sequences increases linearly only for a short time after the two sequences diverge from each other (alternatively, the distance is linear only shortly before coalescence).

The longer the amount of time after divergence, the more likely it becomes that two mutations occur at the same nucleotide site. Simple genetic distance calculations will thus undercount the number of mutation events that have occurred in evolutionary history. The extent of this undercount increases with increasing time since divergence, which can lead to the phenomenon of long branch attraction, or the misassignment of two distantly related but convergently evolving sequences as closely related (Sullivan and Joyce, 2005). The maximum parsimony method is particularly susceptible to this problem due to its explicit search for a tree representing a minimum number of distinct evolutionary events (Felsenstein, 2004).

4.1 Various Models

All substitution models assign a set of weights to each possible change of state represented in the sequence. The most common model types are implicitly reversible because they assign the same weight to, for example, a G>C nucleotide mutation as to a C>G mutation. The simplest possible model, the Jukes-Cantor model, assigns an equal probability to every possible change of state for a given nucleotide base. The rate of change between

any two distinct nucleotides will be one-third of the overall substitution rate (Felsenstein, 2004).

More advanced models distinguish between transitions and transversions. The most general possible time-reversible model, called the GTR model, has six mutation rate parameters. An even more generalized model known as the general 12-parameter model breaks time-reversibility, at the cost of much additional complexity in calculating genetic distances that are consistent among multiple lineages (Felsenstein, 2004). One possible variation on this theme adjusts the rates so that overall GC content - an important measure of DNA double helix stability - varies over time (Galtier and Gouy, 1998)

Models may also allow for the variation of rates with positions in the input sequence. The most obvious example of such variation follows from the arrangement of nucleotides in protein-coding genes into three-base codons. If the location of the open reading frame (ORF) is known, rates of mutation can be adjusted for position of a given site within a codon, since it is known that wobble base pairing can allow for higher mutation rates in the third nucleotide of a given codon without affecting the codon's meaning in the genetic code Sullivan and Joyce (2005). A less hypothesis-driven example that does not rely on ORF identification simply assigns to each site a rate randomly drawn from a predetermined distribution, often the gamma distribution or log-normal distribution (Felsenstein, 2004). Finally, a more conservative estimate of rate variations known as the covarion method allows autocorrelated variations in rates, so that the mutation rate of a given site is correlated across sites and lineages (Fitch and Markowitz, 1970).

4.2 Choosing the best model

The selection of an appropriate model is critical for the production of good phylogenetic analyses, both because underparameterized or overly restrictive models may produce aberrant behavior when their underlying assumptions are violated, and because overly complex or overparameterized models are computationally expensive and the parameters may be overfit (Sullivan and Joyce, 2005). The most common method of model selection is the likelihood ratio test (LRT), which produces a likelihood estimate that can be interpreted as a measure of "goodness of fit" between the model and the input data (Sullivan and Joyce, 2005). However, care must be taken in using these results, since a more complex model with more parameters will always have a higher likelihood than a simplified version of the same model, which can lead to the naive selection of models that are overly

complex (Felsenstein, 2004). For this reason model selection computer programs will choose the simplest model that is not significantly worse than more complex substitution models. A significant disadvantage of the LRT is the necessity of making a series of pairwise comparisons between models; it has been shown that the order in which the models are compared has a major effect on the one that is eventually selected.

An alternative model selection method is the Akaike information criterion (AIC), formally an estimate of the Kullback–Leibler divergence between the true model and the model being tested. It can be interpreted as a likelihood estimate with a correction factor to penalize overparameterized models (Sullivan and Joyce, 2005) The AIC is calculated on an individual model rather than a pair, so it is independent of the order in which models are assessed. A related alternative, the Bayesian information criterion (BIC), has a similar basic interpretation but penalizes complex models more heavily (Sullivan and Joyce, 2005)

A comprehensive step-by-step protocol on constructing phylogenetic tree, including DNA/Amino Acid contiguous sequence assembly, multiple sequence alignment, modeltest (testing best-fitting substitution models) and phylogeny reconstruction using Maximum Likelihood and Bayesian Inference, is available at Nature Protocol.

An non traditional way of evaluating the Pylogenetic Tree is to compare it with clustering result. One can use a Multidimensional Scaling technique, so called Interpolative Joining to do dimensionality reduction to visualize the clustering result for the sequences in 3D, and then map the phylogenetic tree onto the clustering result. A better tree usually has a higher correlation with the clustering result.

Chapter 5

Limitations and Workarounds

There are no methods to measure whether a particular phylogenetic hypothesis is accurate or not, Unless we have found the true relationships among the taxa being examined. The best result an empirical phylogeneticist can hope to attain is a tree with branches that are well supported by the available evidence Computational Phylogenetics (2016). Several potential pitfalls have been identified:

5.1 Homoplasy

Certain characters are more likely to evolve convergently than others; logically, such characters should be given less weight in the reconstruction of a tree (Goloboff *et al.*, 2008). Weights in the form of a model of evolution can be inferred from sets of molecular data, so that maximum likelihood or Bayesian methods can be used to analyze them. For molecular sequences, this problem is exacerbated when the taxa under study have diverged substantially. As time since the divergence of two taxa increase, so does the probability of multiple substitutions on the same site, or back mutations, all of which result in homoplasies. For morphological data, unfortunately, the only objective way to determine convergence is by the construction of a tree – a somewhat circular method. Even so, weighting homoplasious characters[how?] does indeed lead to better-supported trees (Goloboff *et al.*, 2008).

Further refinement can be brought by weighting changes in one direction higher than changes in another; for instance, the presence of thoracic wings almost guarantees placement among the pterygote insects because, although wings are often lost secondarily, there is no evidence that they have been gained more than once Goloboff (1997).

5.2 Horizontal gene transfer

In general, organisms can inherit genes in two ways: vertical gene transfer and horizontal gene transfer. Vertical gene transfer is the passage of genes from parent to offspring, and horizontal (also called lateral) gene transfer occurs when genes jump between unrelated organisms, a common phenomenon especially in prokaryotes; a good example of this is the acquired antibiotic resistance as a result of gene exchange between various bacteria leading to multi-drug-resistant bacterial species. There have also been well-documented cases of horizontal gene transfer between eukaryotes.

Horizontal gene transfer has complicated the determination of phylogenies of organisms, and inconsistencies in phylogeny have been reported among specific groups of organisms depending on the genes used to construct evolutionary trees. The only way to determine which genes have been acquired vertically and which horizontally is to parsimoniously assume that the largest set of genes that have been inherited together have been inherited vertically; this requires analyzing a large number of genes.

5.3 Species Split

The basic assumption underlying the mathematical model of cladistics is a situation where species split neatly in bifurcating fashion. While such an assumption may hold on a larger scale (bar horizontal gene transfer, see above), speciation is often much less orderly. Research since the cladistic method was introduced has shown that hybrid speciation, once thought rare, is in fact quite common, particularly in plants (Arnold, 1997; Wendel and Doyle, 1998). Also, paraphyletic speciation is common, making the assumption of a bifurcating pattern unsuitable, leading to phylogenetic networks rather than trees. Introgression can also move genes between otherwise distinct species and sometimes even genera, complicating phylogenetic analysis based on genes. This phenomenon can contribute to ”incomplete lineage sorting” and is thought to be a common phenomenon across a number of groups. In species level analysis this can be dealt with by larger sampling or better whole genome analysis. Often the problem is avoided by restricting the analysis to fewer, not closely related specimen.

5.4 Taxon sampling

Owing to the development of advanced sequencing techniques in molecular biology, it has become feasible to gather large amounts of data (DNA or amino acid sequences) to infer phylogenetic hypotheses. For example, it is not rare to find studies with character matrices based on whole mitochondrial genomes (16,000 nucleotides, in many animals). However, simulations have shown that it is more important to increase the number of taxa in the matrix than to increase the number of characters, because the more taxa there are, the more accurate and more robust is the resulting phylogenetic tree. This may be partly due to the breaking up of long branches.

5.5 Phylogenetic signal

Another important factor that affects the accuracy of tree reconstruction is whether the data analyzed actually contain a useful phylogenetic signal, a term that is used generally to denote whether a character evolves slowly enough to have the same state in closely related taxa as opposed to varying randomly. Tests for phylogenetic signal exist.

5.6 Continuous characters

Morphological characters that sample a continuum may contain phylogenetic signal, but are hard to code as discrete characters. Several methods have been used, one of which is gap coding, and there are variations on gap coding.

If more taxa are added to the analysis, the gaps between taxa may become so small that all information is lost. Generalized gap coding works around that problem by comparing individual pairs of taxa rather than considering one set that contains all of the taxa.

5.7 Missing data

In general, the more data that are available when constructing a tree, the more accurate and reliable the resulting tree will be. Missing data are no more detrimental than simply having fewer data, although the impact is greatest when most of the missing data are in a small number of taxa. Concentrating the missing data across a small number of characters produces a more robust tree.

Chapter 6

Conclusions and Future Work

We started with a formal definition of Computational Phylogenetics as a research area, and later discussed in detail the applications of the area, and the methodologies used to perform Phylogenetics. We provide in detail the construction of a phylogenetic tree, and its various classifications. We detailed various methodologies like Parsimony Method, Statistical methods like Maximum Likelihood, Bayesian Inference etc. and also brief about the algorithms and heuristics used in tree construction, and phylogenetic analysis. Later, we describe the methods used to select a model based on the available data, go some heuristics to check the accuracy of the tree constructed.

We have also surveyed for recent trends in phylogenetic analysis, and presented them in brief. Also, we studied the limitations of these methods, and provide some workaround for them. We conclude this work with all the possible methodologies used for Computational phylogenetics that can be used to infer the variances in Sanskrit manuscripts. We use some of the terms commonly used in molecular biology, in the methodologies section above as references to the kind of data we might encounter in building a critical edition.

In future, we aim to use the methodologies described above to collate the available Sanskrit manuscripts of the same text. We aim to analyze the collated work for variances, and would try to infer the time when they were created. Our goal, eventually, is to build the critical edition of the text for which we have data available with us.

References

Arnold, Michael L, 1997, *Natural hybridization and evolution* (Oxford University Press)

Chor, Benny, and Tamir Tuller, 2005, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics* **21**, i97–i106

Computational Phylogenetics, 2016, "Computational phylogenetics — Wikipedia, the free encyclopedia,"

David, W Mount, 2001, "Bioinformatics: sequence and genome analysis," *Bioinformatics* **28**

Day, William HE, 1987, "Computational complexity of inferring phylogenies from dissimilarity matrices," *Bulletin of mathematical biology* **49**, 461–467

De Laet, J, 2005, "Parsimony and the problem of inapplicables in sequence data," *Parsimony, phylogeny, and genomics*, 81–116

De Laet, Jan, 2015, "Parsimony analysis of unaligned sequence data: maximization of homology and minimization of homoplasy, not minimization of operationally defined total cost or minimization of equally weighted transformations," *Cladistics* **31**, 550–567

Dress, Andreas, Katharina T Huber, and Vincent Moulton, 2001, "Metric spaces in pure and applied mathematics," *Documenta Mathematica, Quadratic Forms LSU*, 121–139

Farris, James S, 1970, "Methods for computing wagner trees," *Systematic Biology* **19**, 83–92

Felsenstein, Joseph, 2004, *Inferring phylogenies*, Vol. 2 (Sinauer Associates Sunderland)

Fitch, Walter M, Emanuel Margoliash, *et al.*, 1967, "Construction of phylogenetic trees," *Science* **155**, 279–284

Fitch, Walter M, and Etan Markowitz, 1970, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochemical genetics* **4**, 579–593

Galtier, Nicolas, and Manolo Gouy, 1998, "Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of dna sequence evolution for phylogenetic analysis.." *Molecular biology and evolution* **15**, 871–879

Gauch, Hugh G, 2003, *Scientific method in practice* (Cambridge University Press)

Geer R.C., Alpi K. Bhagwat M. Chattopadhyay A. Gaedeke N. Lyon J. Minie M.E. Morris R.C. Ohles J.A. Osterbur D.L. Tennant M.R., Messersmith D.J, 2002, "Ncbi advanced workshop for bioinformatics information specialists [online],"

Goloboff, Pablo A, 1997, "Self-weighted optimization: Tree searches and character state reconstructions under implied transformation costs," *Cladistics* **13**, 225–245

Goloboff, Pablo A, James M Carpenter, J Salvador Arias, and Daniel Rafael Miranda Esquivel, 2008, "Weighting against homoplasy improves phylogenetic analysis of morphological data sets," *Cladistics* **24**, 758–773

Hendy, Michael D, and David Penny, 1982, "Branch and bound algorithms to determine minimal evolutionary trees," *Mathematical Biosciences* **59**, 277–290

Huelsenbeck, John P., Fredrik Ronquist, *et al.*, 2001, "Mrbayes: Bayesian inference of phylogenetic trees," *Bioinformatics* **17**, 754–755

Jukes, Thomas H, and Charles R Cantor, 1969, "Evolution of protein molecules," *Mammalian protein metabolism* **3**, 132

Kolaczkowski, Bryan, and Joseph W Thornton, 2009, "Long-branch attraction bias and inconsistency in bayesian phylogenetics," *PLoS One* **4**, e7891

Kulkarni, Malhar, 2006, "Malayalam manuscripts of malayalam manuscripts of the kāśikāvrtti: A study,"

Larget, Bret, 2013, "The estimation of tree posterior probabilities using conditional clade probability distributions," *Systematic biology* **62**, 501–511

Maas, Philip A., 2009, "On what to do with a stemma - towards a critical edition of the carakasamhitm vimanasthana 8," *Indian Journal of History of Science*, 163–185

Maddison, David R, David L Swofford, and Wayne P Maddison, 1997, "Nexus: an extensible file format for systematic information," *Systematic Biology* **46**, 590–621

Mau, Bob, and Michael A Newton, 1997, "Phylogenetic inference for binary data on dendograms using markov chain monte carlo," *Journal of Computational and Graphical Statistics* **6**, 122–131

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller, 1953, "Equation of state calculations by fast computing machines," *The journal of chemical physics* **21**, 1087–1092

Mount, David W, 2004, "Bioinformatics: sequence and genome analysis. 2004,"

Munro, Hamish Nisbet, 2012, *Mammalian protein metabolism*, Vol. 4 (Elsevier)

Nei, Masatoshi, and Sudhir Kumar, 2000, *Molecular evolution and phylogenetics* (Oxford university press)

Penny, David, Michael D Hendy, and Michael A Steel, 1992, "Progress with methods for constructing evolutionary trees," *Trends in Ecology & Evolution* **7**, 73–79

Ratner, Vadim A, Andrey A Zharkikh, Nikolay Kolchanov, Sergey N Rodin, Viktor V Solovyov, and Andrey S Antonov, 1995, *Molecular evolution*, Vol. 24 (Springer Science & Business Media)

Roch, Sebastien, 2006, "A short proof that phylogenetic tree reconstruction by maximum likelihood is hard," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **3**, 92

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck, 2012, "Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space," *Systematic biology* **61**, 539–542

Sankoff, David, CRISTIANE MOREL, and ROBERT J CEDERGREN, 1973, "Evolution of 5s rna and the non-randomness of base replacement," *Nature* **245**, 232–234

Simmons, Mark P, 2004, "Independence of alignment and tree search," *Molecular phylogenetics and evolution* **31**, 874–879

Simmons, Mark P, 2012, "Misleading results of likelihood-based phylogenetic analyses in the presence of missing data," *Cladistics* **28**, 208–222

Steel, Michael, 1994, "Recovering a tree from the leaf colourations it generates under a markov model," *Applied Mathematics Letters* **7**, 19–23

Sullivan, Jack, and Paul Joyce, 2005, "Model selection in phylogenetics," *Annual Review of Ecology, Evolution, and Systematics*, 445–466

Trikamji, Acharya Jadavaji, 1941, "Caraksamhita, chakrapani with commentary,"

Warnow, Tandy, Bernard ME Moret, and Katherine St John, 2001, "Absolute convergence: true trees from short sequences," in *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms* (Society for Industrial and Applied Mathematics) pp. 186–195

Warnow, Tandy J, 1993, "Constructing phylogenetic trees efficiently using compatibility criteria," *New Zealand Journal of Botany* **31**, 239–247

Wendel, Jonathan F, and Jeff J Doyle, 1998, "Phylogenetic incongruence: window into genome history and molecular evolution," in *Molecular systematics of plants II* (Springer) pp. 265–296

Wheeler, WC, and DS Gladstein, 1994, "Malign: a multiple sequence alignment program," *Journal of Heredity* **85**, 417–418

Yang, Ziheng, 1993, "Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites.." *Molecular biology and evolution* **10**, 1396–1401

Yang, Ziheng, and Bruce Rannala, 1997, "Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method.." *Molecular biology and evolution* **14**, 717–724