IMPACT OF MOLECULAR EVOLUTIONARY FOOTPRINTS ON PHYLOGENETIC ACCURACY – A SIMULATION STUDY

Dissertation

Submitted to

The College of Arts and Sciences of the UNIVERSITY OF DAYTON

In Partial Fulfillment of the Requirements for

The Degree

Doctor of Philosophy in Biology

by

Bhakti Dwivedi

UNIVERSITY OF DAYTON
August, 2009
ABSTRACT

IMPACT OF MOLECULAR EVOLUTIONARY FOOTPRINTS ON PHYLOGENETIC ACCURACY – A SIMULATION STUDY

Dwivedi Bhakti
University of Dayton

Advisor: Dr. Sudhindra R. Gadagkar

An accurately inferred phylogeny is important to the study of molecular evolution. Factors impacting the accuracy of a phylogenetic tree can be traced to several consecutive steps leading to the inference of the phylogeny. In this simulation-based study our focus is on the impact of the certain evolutionary features of the nucleotide sequences themselves in the alignment rather than any source of error during the process of sequence alignment or due to the choice of the method of phylogenetic inference. Nucleotide sequences can be characterized by summary statistics such as sequence length and base composition. When two or more such sequences need to be compared to each other (as in an alignment prior to phylogenetic analysis) additional evolutionary features come into play, such as the overall rate of nucleotide substitution, the ratio of two specific instantaneous, rates of substitution (rate at which transitions and transversions occur), and the shape parameter, of the gamma distribution (that quantifies the extent of
heterogeneity in substitution rate among sites in an alignment). We studied the implications of the following five sequence parameters, individually and in combination: sequence length, substitution rate, nucleotide base composition, the transition-transversion rate ratio and the rate heterogeneity among the sites. It is found that the transition-transversion rate ratio or kappa has a significant impact on phylogenetic accuracy, with a strong positive interaction with accuracy at high substitution rates, contrary to general belief. This work on known expected tree has implications for the researcher in field and would enable them to choose from among the multiple genes typically available today for an accurate phylogenetic inference. DNA sequences diverge from their ancestral sequences by means of evolutionary events (other than mentioned above) such as deletion (deletion of one more nucleotide from a sequence) or insertion (insertion of one more nucleotide to a sequence) events, commonly referred to as gaps in a sequence alignment. We have also investigated the relationship between the number of gaps and phylogenetic accuracy, when the gaps are introduced in an alignment to reflect indel (insertion/deletion) events during the evolution of DNA sequences. DNA sequence alignments were generated using computer simulation, while varying several sequence parameters and introducing both substitution and insertion/deletion events, along a 16-taxon model tree, and systematically varying the expected proportion of gapped sites. The resulting alignments were subjected to commonly used gap treatment methods and methods of phylogenetic inference. The results showed that in general, there is a strong almost deterministic relationship between the amount of gap in the data and the level of phylogenetic accuracy, when the amount of gap was high. Our results also suggest that,
as long as the gaps in the alignment are a consequence of indel events in the evolutionary history of the sequences, the accuracy of phylogenetic analysis is likely to improve if alignment gaps are categorized as arising from insertion events or deletion events and then treated separately in the analysis and if the phylogenetic signal provided by indels is harnessed, for example, by treating the gaps as binary characters in Bayesian or Maximum Parsimony analyses, or in an integrated manner along with substitution events.
ACKNOWLEDGEMENTS

I earnestly wish to express my reverential thanks to Dr. Sudhindra R. Gadagkar for providing me this opportunity and subsequent criticism, guidance, and skillful supervision. I really appreciate his patience in proofreading my abstracts, manuscripts, and thesis.

I would like to thank Dr. Panagiotis Tsonis for giving me the opportunity to work with him. His determination to publish the project report from my course work was critical in starting my career in the research world.

I am highly indebted to my committee members Dr. Jayne Robinson, Dr Mark Nielsen, Dr. John Rowe, and Dr. Dan Goldman for their benevolent supervision and valuable suggestions for all these years.

I am thankful to Karen Bahr, Lynda Routley for their unlimited help with all the paperwork.
I would like to thank my father Dr. Indresh H. Dwivedi, and mother Abha Dwivedi, for their unconditional and endless love and support in every endeavor of my life. I also would like to thank my brother Gyanesh H. Dwivedi, without his encouragement I do not think I would have been here, and my sister-in-law Swarnima Dwivedi for always being there for me.

I would remember the times that I spent at the University of Dayton, and would always be close to my heart. I hope the best for everyone. Thank you.
# TABLE OF CONTENTS

INTRODUCTION...........................................................................................................1

CHAPTERS

1. REVIEW OF LITERATURE

1.1 Phylogenetic inference.............................................................................................6

   1.1.1 Phylogenetic inference.....................................................................................9

   1.1.2 Data used for phylogenetic inference............................................................13

   1.1.3 Types of phylogenetic trees..........................................................................16

   1.1.4 Phylogenetic tree representation styles......................................................21

   1.1.5 Application of phylogenetic inference......................................................24

1.2 The process of phylogenetic inference............................................................30

   1.2.1 Data collection...............................................................................................33

   1.2.2 Sequence alignment......................................................................................36

       1.2.2.1 Pairwise sequence alignment.............................................................37

       1.2.2.2 Multiple sequence alignment.............................................................38

       1.2.2.3 Choosing an alignment method.........................................................50

       1.2.2.4 Alignment gap treatments....................................................................51

1.2.3 Model selection in phylogenetics.................................................................55

       1.2.3.1 Nucleotide substitutions.......................................................................55

       1.2.3.2 Model of nucleotide substitutions.......................................................57

       1.2.3.3 Choosing a model..................................................................................67

1.2.4 Phylogenetic tree reconstruction methods..................................................69
1.2.4.1 Distance-based methods.................................................................71
1.2.4.2 Maximum parsimony.................................................................80
1.2.4.3 Maximum likelihood.................................................................100
1.2.4.4 Bayesian analysis.................................................................104
1.2.5 Assessing phylogenetic accuracy.........................................................114
  1.2.5.1 Topological distances.......................................................115
  1.2.5.2 Assessing tree quality.......................................................117

2. RESEARCH ARTICLES

  2.1 Phylogenetic inference under varying proportions of indel-induced
      alignment gaps..................................................................................125

  2.2 The impact of sequence parameter values on phylogenetic accuracy........176

3. OTHER

  3.1 Whole genome comparison of H1N1 and H3N2 influenza A virus...........203

  3.2 Molecular Mimicry: structural camouflage of proteins and nucleic acids....
      .................................................................................................217

SUMMARY.................................................................................................244

LITERATURE CITED .............................................................................248
LIST OF FIGURES

CHAPTER 1.1

1. A completely resolved bifurcating phylogenetic tree.................................9
2. Example of Polytomy..................................................................................10
3. Orthologous and Paralogous genes..............................................................12
4. Comparison between Monophyletic, Paraphyletic, and Polyphyletic groups..13
5. Unrooted and Rooted tree.........................................................................17
6. Gene tree and Species tree........................................................................19
7. Phylogenetic tree representation styles.....................................................23

CHAPTER 1.2

1. The process of phylogenetic inference.........................................................32
2. An outline of phylogenetic tree reconstruction process using empirical data..34
3. Process of phylogenetic tree reconstruction using computer simulations......35
4. Dotplot matrix for aligning two sequences.................................................38
5. Steps involved in progressive sequence alignment....................................49
6. Transitional and transversional nucleotide substitutions..........................56
7. Nucleotide substitution saturation...............................................................58
8. Gamma distribution to model site rate heterogeneity...............................63
9. The first cluster created by the UPGMA algorithm.....................................73
CHAPTER 2.1

1. The model tree.................................................................133
2. Distribution of alignment gap percentages .......................................................... 141
3. Gap thresholds for different levels of phylogenetic accuracy ......................... 144
4. A. Pairwise comparison of inference methods under MD gap treatment ......... 151
   A. Pairwise comparison of inference methods under MD gap treatment ...... 155
   B. Pairwise comparison of inference methods when gaps are coded as distinct
      evolutionary events ...................................................................................... 160

5. Effect of the alignment gap percentage on phylogenetic accuracy, number of
   characters remaining in the alignment, and total alignment length after gap-
   introduction .......................................................................................................... 170

CHAPTER 2.2

1. The model trees .................................................................................................... 184
2. Joint effects of Kappa and substitution rates on the phylogenetic accuracy ....... 192
3. Combined effect of Kappa and Substitution rates on the phylogenetic accuracy,
   for the PhyML method .................................................................................. 194
4. Phylogenetic tree reconstruction based on the vertebrate mitochondrial COI
   sequences ........................................................................................................... 200

CHAPTER 3.1

1. Influenza A virus in detail .................................................................................. 208
2. Influenza A virus Eight RNA segments ............................................................ 209
3. GC content ........................................................................................................ 212
4. Disparity Index test ($I_D$) ............................................................................... 214
5. Non-synonymous to synonymous rate ratio ($\omega$) ......................................... 215
CHAPTER 3.2

1. Examples of camouflage in animals.................................................................220
2. Pathogen proteins that mimic DNA..............................................................225
3. Protein-DNA mimicry at transcriptional level...............................................228
4. RNA-tRNA mimicry....................................................................................233
5. Protein-tRNA mimicry................................................................................237
# LIST OF TABLES

## CHAPTER 1.1

1. The number of rooted and unrooted trees for a given number of taxa................18

## CHAPTER 1.2

1. Dynamic programming: the initialization step......................................................42
2. Dynamic programming: the matrix filling step.....................................................43
3. Dynamic programming: filled matrix....................................................................44
4. Dynamic programming: traceback step.................................................................45
5. Dynamic programming: traceback step contd.......................................................45
6. Dynamic programming- possible sequence alignment..........................................46
7. Summary of few currently available MSA program..............................................50
8. Commonly used alignment tasks and suggested MSA programs.........................51
9. Substitution rate matrix for Kimura-2-parameter model.......................................60
10. Substitution rate matrix for the HKY model.........................................................61
11. Substitution rate matrix for the GTR model..........................................................62
12. Distance matrix for the five sequences..................................................................73
13. Reduced distance matrix.....................................................................................74
14. Reduced distance matrix.....................................................................................74
15. Comparison of distance-based methods...............................................................79
16. Advantages and disadvantages of MP method.....................................................95
17. Advantages and disadvantages of likelihood method.........................................104
18. Advantages and disadvantages of Bayesian analysis..........................................110
CHAPTER 2.1

1. Sequence and indel parameter values used in the generation of gapped sequence alignments by computer simulation.................................................................135

CHAPTER 2.2

1. The parameter values used in the simulations.....................................................184
2. Marginal effects of low and high values of sequence parameters on phylogenetic accuracy..............................................................................................................189
INTRODUCTION

With the advances of sequencing technology, phylogenetic trees inferred using molecular (DNA or amino acid) sequences have become ubiquitous in the study of evolution (Hasegawa, Kishino and Yano 1987; Miyamoto, Slightom and Goodman 1987; Woese 1987; Giovannoni et al. 1988; Maeda et al. 1988; Cedergren et al. 1989; Lockhart et al. 1994; Arnason et al. 1996; Pakendorf and Stoneking 2005). An accurate phylogenetic inference is crucial for understanding a variety of biological processes (Hillis, Moritz and Mable 1996a; Graur and Li 1999; Nei and Kumar 2000). Applications of phylogenetic inference include understanding species biodiversity, divergence and speciation events, gene or genome evolution, epidemiology, prediction of protein and mRNA structure and function, development of antimicrobial drugs, vaccines and many others. Therefore, the accuracy of the phylogenetic reconstruction impacts our understanding of biology from various perspectives.

Molecular sequences diverge from their ancestral sequences by means of several evolutionary events. The two such main events are substitutions (replacement of one or more nucleotide by another nucleotide) and deletion (mutations in the form of deletion of one or more nucleotide from the sequence) or insertion events (addition of one or more nucleotide). While substitutions change the composition of a given sequence, indels (collective term for insertion/deletion) change the total sequence length. The
studies, have shown the effects of alignment error on the accuracy of tree reconstruction, however, none of them has accounted for the error generated as a function of the amount of gaps introduced in a true alignment of the sequences.

There are many examples in the literature of studies that have used molecular sequences (DNA and protein) with rather large gaps to infer phylogenies (Raymond et al. 2003; Lee and Wen 2004; Egan and Crandall 2008). It appears logical to expect an inverse relationship between the proportion of gapped sites in an alignment and the accuracy of the inferred phylogeny, particularly if the gaps are not treated as reflective of distinct evolutionary events, and thus, containing distinct phylogenetic signal. However, the relationship between the extent of “gappiness” in the data resulting from indel events in the evolutionary history of the sequences on the one hand, and phylogenetic accuracy on the other, has not been studied by introducing and systematically varying the number of gaps in the alignments in a biologically realistic manner, even as the literature on alignment gaps in the phylogenetic context has increased of late (Wiens 2003; Driskell and Christidis 2004; Philippe et al. 2004; Rivas 2005; Ogden and Rosenberg 2007; Hartmann and Vision 2008; Wiens and Moen 2008). Therefore, in this study, we investigated the effect of amount of alignment gaps on the accuracy of phylogenetic inference. Unlike other studies, we are not questioning the accuracy of an alignment, instead we assume that the sequence alignment obtained is perfect and truly represents the sequence homology. With this assumption, we determined the threshold above which the gap introduction results in inaccurate phylogenetic inferences this particular study (Chapter 2.1) investigates this main question using a computer simulation approach.
In the process of alignment, gaps are introduced in the sequences to account for the indels. Since gapped sites in a sequence alignment do not have any data (but may contain phylogenetic information), different methods have been devised for dealing with them during phylogenetic analysis, ranging from ignoring the gapped sites from the entire alignment to inferring, or differentially coding the state at each gapped site, using a number of different methods (see Swofford 2003; Ogden and Rosenberg 2007; Simmons, Muller and Norton 2007). Most of these treatment methods work reasonably well when the proportion of gapped sites in an alignment is small (Ogden and Rosenberg 2007; Simmons, Muller and Norton 2007). In this context, we also evaluated the performance of the total amount of gap in the alignment in the context of various gap treatment methods as well as different search algorithms—distance. We simulated non-coding DNA evolution, introducing nucleotide point substitutions (replacements) and insertion/deletion (indel) events along a balanced (symmetrical) 16-taxon model tree. The simulations were done while systematically varying the values of different sequence and indel parameters to introduce different amount of indels or gaps in the sequence alignment. All the simulation parameters were varied to include biologically realistic values (see chapter 2.1). Similarly, the ratio of insertion to deletion events was also varied based on published results (Saitou and Ueda 1994; Zhang and Gerstein 2003; Chen et al. 2007; Matthee et al. 2007). It was important to vary the ratio of insertions to deletions in order to determine if there was a differential effect on phylogenetic accuracy, since most of the commonly used gap treatment methods do not differentiate between gaps resulting from the two types of evolutionary events. Our results show that overall,
when the percentage of gapped sites (cells in the alignment matrix) in the alignment is low (≤20 percent), all the inference methods using any gap-treatment method perform well (90-100% accuracy). On the other hand, when the number of gapped sites increases in the alignment, differences exist among the inference methods and gap treatment approaches. Our results also show that gaps resulting from deletion events in the evolutionary history of the sequences appear to be harder to reconcile (when compared to those resulting from insertion events), leading to greater inaccuracies in phylogenetic inference. The results of this study are discussed extensively in chapter 2.1 of this dissertation.

In the second part of the study (Chapter 2.2), we instead of alignment gaps, focused on dissecting the features of DNA sequences to determine the optimal combinations of sequence parameters that are associated with accurately inferred phylogenies. DNA sequences can be characterized by summary statistics such as length and base composition. When two or more such sequences need to be compared to each other (as in an alignment prior to phylogenetic analysis) additional parameters come into play, such as the overall rate of nucleotide substitution, the ratio of two specific instantaneous rates of substitution: rate at which transitions (A↔G or C↔T) and transversions (all other changes) occur, and the shape parameter, of the gamma distribution (that quantifies the extent of heterogeneity in substitution rate among sites in an alignment). These sequence features are important parameters for the accurate reconstruction of a phylogeny. When an alignment of molecular sequences from different species is used to infer a phylogeny, what is actually being inferred is the
evolutionary history of the sequences in the alignment, with the expectation that it accurately reflects the evolutionary history of the organisms whose sequences are in the alignment (Nei and Kumar 2000; Felsenstein 2003). However, since different genes can produce different evolutionary histories (trees) (Nichols 2001; Gadagkar, Rosenberg and Kumar 2005), it is important to understand the individual and joint effects of the sequence parameters on the accuracy of phylogenetic reconstruction. Few studies in the literature have focused on understanding the effect of systematically varying each parameter, and the interaction among them, if any, on the accuracy of phylogenetic inference. Therefore, we conducted a simulation study wherein we systematically varied the values of the following sequence parameters: sequence length, overall rate of nucleotide substitution, base-composition, transition-transversion rate ratio, and the shape parameter, and studied their individual and joint effects on the accuracy of phylogenetic inference using different 16-taxon topologies and tree reconstruction methods. Overall, the results showed an increase in accuracy, for the sequence length, a decrease for the increase in substitution rate, no change for the base composition, a decrease for the shape parameter, and an increase for the increase in transition-transversion rate ratio, with the greatest difference in accuracy seen in the case of substitution rate, and the least in the case of base composition. The results of this study are explained in detail in Chapter 2.2 of this dissertation.
CHAPTER 1.1

PHYLOGENETIC INFERENCE
Charles Darwin (1809-1882)

Part of the only figure in the *Origin of Species* (Charles Darwin 1859). Darwin first used it to represent the divergence of variants within a species, showing successively more difference in a single lineage (a\(^1\)-a\(^{10}\)) and splitting into multiple lineages (m, s, i etc), some of which will become new species. Later, he expands the tree metaphor, explaining that, 

“limbs divided into great branches ... were themselves once, when the tree was small, budding twigs; and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups”
Ernst Haeckel (1834-1919)

Ernst Haeckel (1866) coined the word “phylogeny” and presented phylogenetic trees for most known groups of living organisms.
1.1.1. Phylogenetic Inference

*Phylogenetic inference* is the study of evolutionary relationship of different taxa or genes. A *phylogenetic tree* is a graphical representation of the evolutionary history (Figure 1).

![Figure 1. A completely resolved bifurcating phylogenetic tree.](image)

**Tree Representation**

A phylogenetic tree consists of two major components: nodes and branches (Figure 1). A branch is a line that connects any two nodes and is important in inferring the amount of evolution. Each branching event or *topology* shows the speciation or divergence event. The *branch lengths* correspond to the amount of evolution (or substitution differences for instance) between the sequences or nodes they connect. Thus, longer branches reflect distantly diverged sequences and short branches the recently diverged or closely related sequences. *Nodes* can be either external or internal. The external (terminal) nodes, the extant taxa, are often called *operational taxonomic*
units (OTUs). The internal nodes or hypothetical taxonomic units (HTUs) represent a common ancestor of two or more nodes.

An internal node is bifurcating if it has only two immediate descendant lineages but multifurcating if it has more than two immediate descendant lineages. In phylogenetic analysis it is assumed that speciation is a binary process that results in the formation of two species from a single ancestral species. However, this is not always the case. Most of the time, available data results in unresolved branching order, in this case node is multifurcating or leads to polytomy (Figure 2). Soft polytomy designate a lack of information about the order of divergence. Hard polytomy represents the hypothesis that multiple divergences occurred simultaneously.

![Polytomy Diagram](image)

**Figure 2. Example of Polytomy** (a) Partially unresolved tree (soft polytomy), (b) completely unresolved star tree (hard polytomy).

**Symbolic expressions**

A more convenient way of representing and distinguishing tree topologies is the newack format. For example, the topology of the bifurcating tree in Figure 1 can be
expressed as: $(((A, B), C), (D, E))$. A multifurcating tree as shown in Figure 2(a) and 2(b) can be expressed in the same manner: $((A, B), (C, D, E))$ and $(A, B, C, D, E)$, respectively.

**Homology**

"The natural system is based upon descent with modification.. the characters that naturalists consider as showing true affinity (i.e. homology) are those which have been inherited from a common parent, and, in so far as all true classification is genealogical; that community of descent is the common bond that naturalists have been seeking."


Homology is the similarity that is the result of inheritance from a common ancestor – the identification and analysis of homologies is central to phylogenetic inference. Homologs are most commonly defined as Orthologs, Paralogs, or Xenologs. Orthologues are the duplicated genes that are generated by gene duplication that occurred before the divergence of the two species or before speciation event (Figure 3). They are strictly vertically inherited genes, and their traces lead to their common ancestor. For example, gene a, b*, c and C*, B*, A* from the different species are orthologous genes. Paralogues are produced as a result of gene duplication (Figure 3) event within the lineage. For example, genes b* and B are paralogous genes. Xenologs are homologs resulting from the horizontal gene transfer of a gene between two species. The function of xenologs can be variable, depending on how significant the change in context was for the horizontally moving gene.
To infer a species phylogeny, one should use orthologous genes rather than paralogous genes, because only orthologous genes represent speciation events. In reality, however, it is difficult to distinguish between orthologous and paralogous genes. It is important to study the multigene families to know the evolutionary history of member genes and the process of gene duplication. For example, paralogs often are used as natural outgroup and were used to infer the root of the universal tree (Schwartz and Dayhoff 1978; Gogarten *et al.* 1989; Iwabe *et al.* 1989).

![Figure 3. Orthologous and Paralogous genes](image)

**Groupings**

A *clade* is a term used to describe a *Monophyletic* group, defined as a group consisting of a single common ancestor and all its descendants (Figure 4). The grouping of reptiles and birds (Aves) is generally believed to be monophyletic. The *Paraphyletic*
group is a group of taxa that includes the most recent common ancestor but does not contain all the descendants of a common ancestor (Figure 4). For example, Reptiles are a paraphyletic group. A group that does not contain the most recent common ancestor of its members is said to be Polyphyletic (Figure 4). The group of warm-blooded animals is polyphyletic, because it contains both mammals and birds, but the most recent common ancestor of mammals and birds was cold-blooded. Warm-bloodedness evolved separately in the ancestors of mammals and the ancestors of birds, so it is not a true phylogenetic grouping.

(Figure 4. Comparison between Monophyletic, Paraphyletic, and Polyphyletic groups.)

1.1.2. Data used for Phylogenetic Inference

To study the evolutionary relationships among different taxa or genes, different kinds of data can be used: Fossil, Morphological, Behavioral or Molecular. An ideal approach to reconstruct the evolutionary history of all organisms is by the use of fossil records. Fossils are important for correctly estimating relationships among living taxa in
that they hold clues to many ancestral character states. Also, the estimated dates of appearance of fossil in the fossil record can be important for testing and rooting phylogenetic trees. But since the fossil record is fragmentary and incomplete most investigators have used the classical way of estimating relationships between species by comparing their morphological features (such as presence or absence of fins, number of legs, lengths of legs, etc) or molecular data. The molecular data which is increasingly available, such as nucleotide or amino acid sequences has become an important tool to infer phylogenetic relationships. For extinct species, it is almost impossible to obtain molecular data. Using morphological or molecular characters from closely related fossils one can estimate their relationships. Good examples of this are viruses which do not leave any fossil records. The only way to know their past evolutionary relation is through the study of the evolution in the existing viruses using molecular sequences.

The greatest advantage of morphological studies is that they allow for much more thorough taxonomic sampling than is possible with molecular analysis. Sampling a large number of taxa for molecular studies can be difficult because of the cost of sequencing, the rarity (or extinction) of species, and the inaccessibility of the areas where certain taxa occur. Another advantage of morphology based phylogenetics is that each morphological character is probably encoded by a different gene or set of genes whereas in molecular datasets, many or all nucleotide characters may be drawn from the same gene. If the evolution of gene differs from that of a species, then tree reconstruction from molecular data may yield incorrect inference.
Molecular data have the following advantages over morphological data. (1) Molecular sequence data is abundant, (2) In cases, where morphology is very simple, for example bacteria, it is obvious to use a molecular data to infer bacterial phylogeny, (3) another advantage of molecular data is the wide range of evolutionary rate differences that exist across nucleotide or amino acid residues. This helps phylogenetic analysis of most distantly related and also of those with few or no morphological traits in common. For instance, the reported case of Florida dentist suspected of transmitting human immunodeficiency virus (HIV) to some of his patients (Ou et al. 1992). HIV phylogeny based on the sequence data showed the dental clade; the clade with a single evolutionary lineage that contained only viruses from the dentist and five patients he infected (Ou et al. 1992), (4) Also, genetic basis is usually known for molecular data, whereas for most morphological characters, is merely assumed, (5) molecular data can be used to infer ancestral states or sequences which may not be possible with morphological data unless provided by additional fossil records (Nei and Kumar 2000), for e.g., origin of whales (Graur and Higgins 1994; Gatesy 1997; Shimamura et al. 1997), evolution of color vision in mammals (Neitz, Neitz and Jacobs 1991), (6) molecular data assumes a mathematical model which can be used to estimate the rate of evolution of a given gene, can also be used to estimate the rate of divergence to detect the selection pressure in a given gene of an organism (for e.g., to understand the evolution of influenza virus in order to facilitate vaccine development and drug discovery), (7) Another advantage of the molecular approach is that character can be selected and defined in a relatively objective manner.
In morphological systematics, the characters must be discovered and delimited by the systematics, usually without any explicit criteria for characters selection or coding.

Whether to use morphological or molecular data for any particular evolutionary question has been a long-standing debate. The most common cause of incongruencies between molecular and morphological phylogenies is weaker support for either or both of the estimates, due to under sampling of characters and or taxa. Applying different reconstruction methods to the data may also lead to differences in tree estimates. Another source of conflict between findings of morphologists and molecular systematics is rooting. In morphology based studies, rooting of the phylogenetic trees is often based on a broad sampling of living and fossil taxa, focused on the closet relatives. In contrast, many molecular systematics use a single species to root their trees and this approach makes their estimates susceptible to long branches problems (Felsenstein 1985). Conflicts between phylogenies from molecular and morphological datasets may also be attributable to actual differences in phylogenetic histories, when phylogeny of the gene differs from the phylogeny of the species.

### 1.1.3. Types of Phylogenetic Trees

**Rooted and Unrooted Trees**

A rooted tree has a node as the root from which all the subsequent nodes descends. The direction of each path corresponds to evolutionary time, and the root is the
common ancestor of all the OTUs. Thus, a rooted tree specifies the evolutionary relationships among the taxa and does define direction of the evolutionary pathway. An *unrooted tree* lacks a root and only specifies the relationships among the OTUs but does not define evolutionary path. Furthermore, OTUs that may be next to each other on an unrooted tree need not be evolutionarily closely related. For example, Figure 5(a) is an unrooted tree. If we root the tree at mid–point (I), then the rooted tree would be that shown in Figure 5(b).

![Figure 5. (a) Unrooted and, (b) Rooted tree](image)

The number of possible bifurcating rooted and unrooted trees increases linearly with *n* OTUs (see Table 1). The number of possible unrooted trees for *n* OTUs (*n* ≥ 3) is given by,

\[ N_U = (2n-5)! / [2^{n-3}(n-3)!] \]

While the number of rooted trees for *n* OTUs (*n* ≥ 2) is

\[ N_R = (2n-3)! / [2^{n-2}(n-2)!] \]
Table 1. The number of rooted and unrooted trees for a given number of taxa (or OTUs).

<table>
<thead>
<tr>
<th>Possible Number of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of OTUs</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Unrooted trees</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Rooted trees</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

In an unrooted bifurcating tree of \( n \) taxa there are \( 2n-3 \) branches. Since there are \( n \) external branches connecting to \( n \) extant taxa, the number of internal branches is \( n-3 \).

The number of internal nodes is equal to \( n-2 \). In a rooted tree, the numbers of internal branches and internal nodes are \( n-2 \) and \( n-1 \), respectively, and the total number of branches is \( 2n-1 \).

Methods to root trees (Rooting)

Many methods of phylogenetic tree reconstruction yield unrooted trees. The root is then determined by other criteria such as \textit{outgroup} rooting, an external point of reference. An outgroup is a natural member of the group of interest (i.e. the ingroup). The outgroup can either be species (e.g., birds to root a mammalian tree) or previous gene duplicates (e.g., \( \alpha \)-globins to root \( \beta \)-globins). In the absence of an outgroup, the tree is rooted using additional taxonomical information such as the fossil record, which clearly indicates that it diverged earlier than the other OTUs under consideration. Another way
is to root the tree at the midway point between the two most distant taxa in the tree, as determined by the branch lengths.

**Species and Gene Trees**

A phylogenetic tree that represents the evolutionary history of a group of species or organisms is known as *species tree*. When a phylogenetic tree is inferred from one gene from each species, the inferred tree is a *gene tree*.

![Gene tree and Species tree diagram](image)

**Figure 6. Gene tree and Species tree**

In the case of unisexual organisms, the phylogeny of the gene is the phylogeny of the species. However, in other organisms the gene tree represents the evolutionary history of a gene but not necessarily that of the species in which it exists. It can differ from the species tree in several aspects: First, the divergence of two genes sampled from
two different species may predate the divergence of the two species (Figure 6), a process known as *lineage sorting*. Lineage sorting is a problem when divergence time of alleles is greater than the interval between the successive speciation events of the same lineage. In such cases, the evolutionary relationship between these alleles need not reflect the species phylogeny. However, if the alleles present in a lineage prior to the species divergence are monophyletic then allele phylogeny reflects species phylogeny. This however, will result in underestimation of the branch lengths. Secondly, the topology of a gene tree may differ from that of the species tree. Gene duplication events may also result in a species containing a number of distinct but related gene sequences. In addition, the incongruent phylogenetic trees for different genes may not an artifact of tree reconstruction method instead shows different evolutionary history of these genes derived by horizontal gene transfer (HGT). However, when multiple genes are sampled from individual species, reconstruction of a species tree becomes really impractical or difficult, particularly when each gene gives a different tree. However, a consensus tree of all genes could solve the problem, as it would reflect the evolutionary history of conserved gene core that have not been subjected to HGT and is stable in all of majority of the species, but the existence of such a stable core still remains questionable. Species trees are also best obtained from analysis that uses data from concatenated genes (ref), but because concatenation becomes inappropriate due to conflicts among individual gene trees, this approach may not produce a correct phylogenetic relationship. Phylogenetic congruence can also be obtained using recent computational methods such as supertree
approach (Emonds, Gittleman, and Steel, 2002, Elmonds 2004)) and other recent computation methods (e.g., Liu 2008; Kubatko, Carstens, and Knowles 2009).

**Expected and Realized tree**

Despite the large number of rooted and unrooted trees that can describe the relationships among the OTUs, only one of all possible trees can be represented as the true estimate of the actual evolutionary relationships. A tree that is obtained using the available dataset and using a method of tree reconstruction is called an inferred tree. An inferred tree may or may not be same as the true tree.

DNA sequences over time are subjected to a number of nucleotide or amino acid substitutions. Therefore, even if the expected number of substitutions increases linearly with time, a phylogenetic tree representing the observed number of changes could be different from what is expected. *Expected Tree* is the tree that can be constructed from the expected number of substitutions for each branch. *Realized Tree* is the tree that is actually being inferred from the actual number of substitutions. Unfortunately the expected and realized trees are almost never known for the real data and are often different from the tree reconstructed (or inferred tree) from the observed sequence data.

1.1.4. Phylogenetic Tree Representation Styles

Different types of phylogenetic tree representation can be used to depict different aspects of evolutionary history. The most common tree styles shown (Figure 7) are
Cladogram, Phylogram (or additive trees), Ultrametric trees, and Phylogenetic Networks. A tree-like network that represents an ancestor-descendent relationship is called a **cladogram**. A cladogram refers to the topology of a rooted phylogenetic tree (branch lengths are meaningless or do not represent time). A tree-like network representing number of character change through its branch lengths is called a **phylogram**. A phylogram includes phenetic relationships (topology and branch lengths). An **ultrametric** tree or chronogram is a phylogenetic tree that explicitly represents evolutionary time through its branch lengths. It is a tree with equal root-to-tip path lengths for all lineages, for example tree built with the assumption of a molecular clock. **Non-ultrametric** tree on the other hand does not assume a molecular clock, and thus the terminal nodes are not equally distant from the root. **Phylogenetic networks** are used to visualize evolutionary relationships between species or organisms. Traditionally, the evolution of species has been depicted using phylogenetic trees. This however, has been questioned by recent developments in molecular phylogenetics (Doolittle 1999; Diallo, Lapointe and Makarenkov 2006; Makarenkov, Kevorkov and Legendre 2006), when evolutionary events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are involved in the evolution of the lineage. Reticulate evolution shows the lack of independence between lineages. When a reticulation event occurs, two or more independent evolutionary lineages interact at some level of biological organization.
Figure 7. Phylogenetic tree representation styles. (a) Cladogram, (b) Phylogram, (c) Ultrametric tree, and (d) Phylogenetic Network.
1.1.5. Applications of Phylogenetic inference

Molecular phylogenetic analysis has a wide range of practical applications in the analysis of sequence data and is now an essential tool in areas ranging from population genetics to genomics to virology.

Systematics

In taxonomy, species phylogenies are now commonly inferred using reconstructed gene trees. This is particularly interesting when there are no morphological data, or when molecular and morphological data contradict. Different kind of data varies in their strength and weaknesses depending on the taxonomic group studied and the phylogenetic questions being addressed. Molecular data, particularly nucleotide sequences, cover huge data sets that are comparable across a wide taxonomic range. For example, two distantly related taxa share few morphological traits in comparison to numerous homologous genes. Inferring organism phylogeny may require combining data from different genes, or a separate analysis, or combined on conditions that there is no significant heterogeneity among the data set. Morphological data also provide information where extinct taxa are concerned. For extinct species, it is difficult to obtain molecular data and thus morphological characteristics such as fossils are usually the only way to estimate their relationships.

Phylogenies in Molecular epidemiology

Phylogenetic trees have become an important analytical tool to study the origin and spread of viral infections. In virology, phylogenetic trees have been used to define significant phylogenetic clusters of viruses. One of the famous examples in documenting
the source of particular infection events is - where trees of HIV sequences were used to show that a dentist in Florida had infected a group of his former patients with this virus (Ou et al. 1992). Similar studies have been performed on HIV, hantaviruses, blood-borne human pathogen, and highly evolving hepatitis C virus (HCV) (Hillis 1990; Hillis, Huelsenbeck and Cunningham 1994). Another very important example, which is also one of the major threat to public health nowadays is the infection caused by influenza virus. During the past years, Influenza A virus has caused several human epidemics, or pandemics, such as the Spanish influenza A (H1N1) pandemic of 1918, the Asian influenza (H2N2) pandemic of 1957, Hong Kong influenza (H3N2) pandemic of 1968, and the recent 2003 outbreak of highly pathogenic avian influenza (H5N1) in Asian countries. Influenza A virus occurs in a wide range of hosts including mammals and birds, and can be divided into 15 subtypes based on the surface glycoprotein HA and NA. Phylogenetic trees constructed on sequences from the nucleoprotein (NP) gene from all these species show that viral lineages are very species specific. Though influenza viruses do not cross species boundaries, but if they do, the results can be devastating. It has been shown that lineages with many mutations in one set of positively selected codons were usually ones which led to successful strains in subsequent seasons.

Inferences of Ancestral sequences

Inference of ancestral sequences: parsimony approach. Phylogenetic trees are important for the reconstruction of ancestral sequences when there is only one amino substitution at a site. When there are two or more amino acid substitutions at a site, it is not always possible to determine all ancestral amino acids. Maddison and Maddison
1992 developed a computer program for inferring ancestral amino acids using the principle of maximum parsimony. In addition, Maddison 1995 presented method that allows choosing the most likely ancestral sequence from several different possible ancestral sequences.

Inference of ancestral sequences: Bayesian approach. Yang 1997 developed a Bayesian approach in which the topology of the tree is assumed to be known, but the branch lengths are estimated by using the likelihood method. Zhang and Nei 1997 proposed a simplified version of Bayesian approach, estimating branch lengths by distance methods such as NJ. In another method (Koshi and Goldstein 1996), topology and branch lengths are estimated by the NJ method.

Inference of ancestral sequences: Likelihood approach. Chang et al. 2002 used maximum likelihood (ML) to reconstruct the sequence of visual pigments in the last common ancestor of birds and alligators; the protein was then synthesized in the laboratory (see Pupko et al. 2002 for a recent discussion of the methodology of ancestral-character-state reconstruction).

**Comparative studies**

Comparative analysis aims at establishing correlations between traits across taxa. It is important to discriminate dependencies between the investigated traits that are introduced merely by evolution from those representing true correlation. Felsenstein 1985 was the first to propose a statistical test for comparisons of continuous traits between organisms that used a phylogenetic hypothesis as a structural framework. He proposed using a series of independent contrast to search for correlations in traits among
terminal taxa and their ancestors. Further, Felsenstein 1985, 1988 proposed methods to account for incompletely resolved phylogenies and to estimate the branch lengths. Later, Grafen (1989, 1992) proposed phylogenetic regression, based on Felsenstein’s (1985) method but uses a likelihood approach to simultaneously estimate relationships between standardized independent contrasts and to transform branch lengths. Another assumption of Felsenstein method is that evolution proceeds through Brownian motion, so that expected variance of change in a trait is proportional to time. Consequently, many other methods have been developed for the analysis of correlated continuous as well as discrete characters based phylogenetic history.

**Detecting recombination and mutations**

New Bayesian methods (Suchard et al. 2002) can help determine which strains of human immunodeficiency virus-1 (HIV-1) arose from recombination. Identifying mutations is likely to be associated with disease. The lack of structural, biochemical and functional data from many genes implicated in disease means it is unclear which missense mutations are important. Fleming et al. 2001 used Bayesian phylogenetics to identify missense mutations in conserved regions and regions under positive selection in the breast cancer gene *BRCA1*. These data allowed them to prioritize these mutations for future functional and population studies.

**Detection of orthology and paralogy**

Phylogenetics is commonly used to sort out the history of gene duplications for gene families. This application is now included in even preliminary examinations of sequence data; for example, the initial analysis of the mouse genome (Consortium 2002)
included neighbour-joining trees to identify duplications in cytochrome P450 and other
gene families.

**Estimating divergence times**


**Detection of selection**

Amino-acid sites on the surface of influenza that are targeted by the immune system can be detected by an excess of non-synonymous substitutions (Bush *et al.* 1999; Anisimova, Bielawski and Yang 2002). This information might assist vaccine preparation.

**Determining the identity of new pathogens**

Phylogenetic analysis is now routinely performed after polymerase chain reaction (PCR) amplification of genomic fragments of previously unknown pathogens. Such analyses made possible the rapid identification of both Hantavirus (Hughes *et al.* 1993) and West Nile virus (Anderson *et al.* 1999; Lanciotti *et al.* 1999)

**Population History**

Phylogenetic trees also contain information about the demographic history of the population from which the sequences were sampled. Population size changes over time influence the shape and the branch lengths of trees. Coalescent theory (Donnelly and Tavare 1995) provides a probabilistic model for this process and thus allows the extraction of information concerning population history from an inferred tree (Pybus,
Rambaut and Harvey 2000; Strimmer and Pybus 2001). Trees also allow the
determination of the geographic origin of a population. For example, the common
ancestor of human mitochondrial DNA (mitochondrial Eve) has been shown to be of
African origin (Vigilant et al. 1991). In addition, phylogenetic trees provide information
about the time-frame of evolutionary events, e.g. about the age of the mitochondrial Eve
(Vigilant et al. 1991) or the time of the human-ape split (Hasegawa, Kishino and Yano
1985).
CHAPTER 1.2

PROCESS OF PHYLOGENETIC INFERENCE
Phylogenetic analysis of DNA or protein sequences has become an important tool for studying the evolutionary history of organisms from bacteria to humans (Hillis, Moritz and Mable 1996b; Li 1997; Nei and Kumar 2000; Felsenstein 2004a; Hall 2004b). Since the rate of divergence varies extensively with gene, one can study the evolutionary relationships of all groups of organisms. Darwin’s in his letter to Thomas Huxley (1857), wrote,

“The time will come I believe though I shall not live to see it, when we shall have fairly true genealogical (phylogenetic) trees of each great kingdom of nature”

Following are the steps involved in phylogenetic inference (see Figure 1):

1. Sequence collection
2. Sequence alignment
3. Model selection
4. Tree reconstruction
5. Testing the tree hypothesis
Figure 1. The process of phylogenetic inference.
1.2.1. Data collection

The first step in reconstructing a phylogenetic tree in real data analysis is obtaining the sequence dataset (Figure 2). This means retrieving the sequences from the public (or private) databases. The main repository for these data is the public nucleotide database, stored in the Genbank (http://www.ncbi.nlm.nih.gov/), EMBL (http://www.ebi.ac.uk/Databases/), and DDBJ (http://www.ddbj.nig.ac.jp/). Genome data are also available for example, from the National Centre for Biotechnology, NCBI (http://www.ncbi.nlm.nih.gov/Genomes/index.html), Institute for Genomic Research, TIGR (http://www.tigr.org/), the Joint Genome research Institute, JGI (http://www.jgi.doe.gov/), and Sanger (http://www.sanger.ac.uk/). When sequence data is collected, the first step is usually downloading other relevant (e.g., orthologs) sequences. Typically, a few outgroup sequences are included in a study to root the tree (that is, to indicate which nodes in the tree are the oldest), provide indication about the early ancestral sequences and improve the estimates of parameters in the model of evolution.
Simulated datasets

Simulated (or synthetic) dataset can be generated using Computer simulations. During simulation process, sequences are evolved on a given evolutionary model and tree (user-defined), and since one know these, simulated data can be used to test the methods and models.

Simulation starts with a random sequence and assumes a particular model of evolution and then generates nucleotide sequences according to the model, and to a phylogeny (see Figure 3). Here, we can manipulate factors such as evolutionary rate or overall evolutionary distances, while controlling for other factors that could influence the analysis. Thus, one can generate many replicate datasets under specified conditions. For example, we can look at the probability of obtaining the true topology under different tree reconstruction methods. In real data analysis, we almost never know the true phylogeny,
so it is difficult to study this problem empirically. However, if we use an appropriate mathematical model, we can simulate the evolutionary changes of DNA sequences following a given model tree. We can then reconstruct a tree by various methods using the artificially generated sequences and compare the topology of the tree (inferred tree) obtained with that of the model tree (used for the simulation). If this process is repeated many times, we can estimate the probability of obtaining the true topology and this probability can be used for comparing the efficiencies of different tree reconstructing methods.

![Diagram of tree reconstruction process](image)

**Figure 3. Process of phylogenetic tree reconstruction using computer simulations.**

The simulation and experimental phylogeny approaches are largely complementary, and both kinds of studies are necessary to evaluate the methods or models of phylogenetic analysis effectively. Simulations can also be used to explore virtually any conceivable phylogeny, and phylogenies can be replicated with speed.
Few simulation software: DAWG (Cartwright 2005), MYSSP (Rosenberg 2005b), PAML (Yang 1997; Yang 2007), EVOLVE (Hall 2008), etc.

1.2.2. Sequence Alignment

Phylogenetic analysis on nucleotide sequences requires that the sequences should be aligned so that the site-by-site differences among the sequences can be accurately measured. Alignment ensures that homologous parts of the sequences line up in order to measure nucleotide substitution differences. During this process gaps are usually introduced in the sequences, especially if they are of unequal length. This is done to account for deletion events (mutations in the form of deletion of one or more nucleotide or amino acid from the sequence) or insertion events (addition of one or more nucleotide or amino acid), since their divergence from a common ancestor.

Insertions and deletions are collectively referred to as indels (or so called gaps), because when a sequence involving either an insertion or a deletion is compared with the original sequence, a gap will appear in one of the two sequences. Gaps are differentiated into two types based on their length and origin. According to (Li 1997) length of gaps essentially exhibit a bimodal type of frequency distribution, with short gaps (upto 20-30 nucleotides) being mostly caused by errors during DNA replication and with long gaps occurring mainly because of unequal crossing over. However, the size of the gap is much less important than the fact if it is there at all, so alignment programs have separate penalties for inserting a gap (which is costly) and for making it bigger (relatively cheaper). Ideally, gap penalties differ for closely related versus distantly related
sequences, for different kinds of sequence, and for different regions of the same sequence, but this is mostly impractical. Therefore, all gap penalties are compromises, and an alignment can look very different depending on the penalties that are used.

1.2.2.1. Pairwise Sequence alignment

Pairwise alignments can only be used between two sequences at a time. The simplest way to align two sequences is by means of dot plot method or dynamic programming.

Dot-Matrix Method

In this method, the sequences are assigned, each to one axis of a graph, and a dot placed for each position where the two sequences have identical bases. The dotplot provides a way of quickly visualizing the similarities between all parts of two sequences simultaneously. Let us consider the following two nucleotide sequences:

1) ATGCGTCGTT
2) ATCCGCGAT

Sequence 1 has ten nucleotides and sequence 2 has nine nucleotides. Therefore, at least one gap must be introduced in the alignment of these sequences. A simple way of aligning these sequences is to do a two dimensional comparison, as shown in Figure 4. In this comparison (dot matrix), dots (or star) are given when the nucleotides in sequences 1 and 2 are identical. If the two sequences are identical, there will be diagonal line of dots.
If the sequences are identical except for a gap in one of the two sequences, the diagonal line will shift up or down in the middle of the line. Therefore, we can identify the gap.

![Dotplot matrix for aligning two sequences](image)

Figure 4. Dotplot matrix for aligning two sequences

However, it is obvious that the utility of dot plots is only for small sequences, even when automated (the plot becomes noisy). Dotplots can be generated using applet program ([www.isrec.isb-sib.ch/java/dotlet/Dotlet.html](http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html); [www.bip.bham.ac.uk/dotlet/Dotlet.html](http://www.bip.bham.ac.uk/dotlet/Dotlet.html)) that runs in a web browser. Dotplots are often useful to identify repeated domains or duplications in big proteins.

### 1.2.2.2. Multiple Sequence alignment

Multiple sequence alignment, the alignment of more than two sequences, is generally thought to lead to more accurate alignments than simple pairwise alignments (Duret and Abeddaim 2000). Most multiple sequence alignments are constructed by the method known as progressive alignment algorithm (Feng and Doolittle 1987; Feng and
Doolittle 1990), where similar sequences are aligned first and additional sequences are progressively added based on their divergence from the initial pair. Empirical and simulation studies have shown that multiple alignments perform better than pair wise alignments (Thompson, Plewniak and Poch 1999; Raghava et al. 2003; Rosenberg 2005a)

Let us again consider the two sequences,

1) AATCTATA  
2) AAGATA

Gaps can be introduced in various places, and in various combinations, as shown below.

<table>
<thead>
<tr>
<th>AATCTATA</th>
<th>AATCTATA</th>
<th>AATCTATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAG-AT-A</td>
<td>AA-G-ATA</td>
<td>AA--GATA</td>
</tr>
</tbody>
</table>

Clearly, if the sequences are long, it would become impossible for manual introduction of gaps and thus would need a computer to help us find the optimal gaps. Let us consider two nucleotide sequences of 100 nucleotides in length. If it takes $100^2$ seconds to exhaustively align these sequences, then it will take $100^4$ to align four sequences etc. Even computers needs more time than the universe has existed to align 20 sequences exhaustively. Clearly, the exhaustive method is not good. Therefore, in order to align sequences efficiently and in a reasonable amount of time, we perform alignment using heuristic methods such as dynamic programming and progressive alignment.
**Dynamic programming**

The idea in dynamic programming is to break up a large problem into smaller, sub-problems. This is done in a sequential manner such that at any given stage, solutions are known for the sub-problems. The main thing is that once a solution for a sub-problem is found, then it need not be computed again. This principle was applied by (Needleman and Wunsch 1970) for alignment of sequences. This algorithm is perhaps one of the first and most commonly used tools in bioinformatics. This alignment algorithm is for global alignment, that is, for aligning a pair of sequences in their entirety. Therefore, this algorithm did not consider gaps from the internal parts of the sequences any different from the gaps at the end of the sequence. Dynamic programming can also be applied local alignments via the Smith-Waterman algorithm (Smith, Waterman and Fitch 1981). Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

The following is an example of global sequence alignment using (Needleman and Wunsch 1970). Consider the following two sequences:

1) G A A T T C A G T T A

2) G G A T C G A
The two sequence lengths are, \( m = 11 \) and \( n = 7 \), respectively. Let us consider a simple scoring scheme where the score \((s)\) and gap penalty \((w)\) are determined as follows:

\[
\begin{align*}
  s_{ij} &= 1 \text{ if match} \\
  s_{ij} &= 0 \text{ if mismatch} \\
  w &= 0 \text{ if gap}
\end{align*}
\]

There are three steps in this algorithm:

1. Initialization (get the matrix ready)
2. Matrix filling (scoring)
3. Traceback (alignment)

1. **Initialization step**

The first step in the global alignment dynamic programming approach is to create a matrix with \( m + 1 \) columns and \( n + 1 \) rows where \( m \) and \( n \) correspond to the size of the sequences to be aligned. Since this example assumes there is no gap opening or gap extension penalty, the first row and first column of the matrix can be initially filled with 0 (Table 1).
Table 1. Dynamic programming: the Initialization step

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

2. **Matrix fill step**

Next, we start from the upper left hand corner and find the maximal score \((C_{i,j})\) for each position in the matrix, where \(i\) is the row number and \(j\) is the column number in the matrix.

\[
C_{i,j} = \text{Maximum} \begin{cases} C_{i+1,j-1} + s_{i,j} & \text{match/mismatch} \\ C_{i+1,j-1} + w & \text{gap in sequence 1} \\ C_{i+1,j} + w & \text{gap in sequence 2} \end{cases}
\]

Using this set of instructions, it gets the following scores for the first cell in the matrix Cell (1, 1):

\[
C_{1,1} = \text{maximum of} \begin{cases} C_{0,0} + s_{1,1} = 0 + 1 = 1 \\ C_{1,0} + w = 0 + 0 = 0 \end{cases}
\]
It has a G in both sequences and hence a match. Therefore, this cell (1,1) gets a score of 1. Since the gap penalty ($w$) is 0, the rest of row 1 and column 1 can be filled in with the value 1 (Table 2). Take the example of row 1. At column 2, the value is the max of 0 (for a mismatch), 0 (for a vertical gap) or 1 (horizontal gap). The rest of row 1 can be filled out similarly until we get to column 8. At this point, there is a G in both sequences (light blue in Table 2). Thus, the value for the cell at row 1 column 8 is the maximum of 1 (for a match), 0 (for a vertical gap) or 1 (horizontal gap). The value will again be 1. The rest of row 1 and column 1 can be filled with 1 using the above reasoning.

Table 2. Dynamic programming: the Matrix filling step

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After filling in all of the values the score matrix is as follows (Table 3):
Table 3. Dynamic programming: Filled matrix

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

3. **Traceback step**

After the matrix fill step, the maximum alignment score for the two test sequences is 6. The traceback step determines the actual alignment(s) that result in the maximum score. Note that with a simple scoring algorithm such as one that is used here, there are likely to be multiple maximal alignments. The traceback step begins in the $m,n$ position in the matrix (Table 4), i.e. the position that leads to the maximal score (in this case it is 6). Traceback takes the current cell and looks to the neighbor cells that could be direct predecessors. This means it looks to the neighbor to the left (gap in sequence #2), the diagonal neighbor (match/mismatch), and the neighbor above it (gap in sequence #1). The algorithm for traceback chooses as the next cell in the sequence one of the possible predecessors. In this case, the neighbors are marked in red. They are all also equal to 5.
Table 4. Dynamic programming: Traceback step

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Since the current cell has a value of 6 and the scores are 1 for a match and 0 for anything else, the only possible predecessor is the diagonal match/mismatch neighbor. If more than one possible predecessor exists, any can be chosen. This gives us a current alignment of

(Seq #1)   A
  |   
(Seq #2)   A

So now we look at the current cell and determine which cell is its direct predecessor. In this case, it is the cell with the red 5 (Table 5).

Table 5. Dynamic programming: Traceback step contd
The alignment as described in the above step adds a gap to sequence #2, so the current alignment is

(Seq #1)   T A
\[\text{|}\]
(Seq #2)   _ A

Once again, the direct predecessor produces a gap in sequence #2. After this step, the current alignment is

(Seq #1)   T T A
\[\text{|}\]
(Seq #2)   _ _ A

Continuing on with the traceback step, we eventually get to a position in column 0 row 0 which tells us that traceback is completed. One possible maximum alignment is (Table 6):

Table 6. Dynamic programming: Possible sequence alignment

One possible maximum alignment is:
There are more alternative solutions each resulting in a maximal global alignment score of 6. Since this is an exponential problem, most dynamic programming algorithms will only print out a single solution.

**Progressive alignment**

Most multiple sequence alignments are constructed by the method known as Progressive alignment algorithm (Feng and Doolittle 1987; Feng and Doolittle 1990). This is essentially a heuristic method and as such is not guaranteed to find the ‘optimal’ alignment. In this algorithm, pairs of sequences with small distances are first aligned, and the alignment of more distantly related sequences is done progressively for larger and larger groups (Figure 5b). In the next step, group of sequences are aligned with each other (Figure 5b). This is done by profile alignment algorithm, which is similar to the alignment algorithm for two sequences, except that the average distance is computed by considering all the nucleotides at every position in the two groups of sequences.

The order in which sequences are aligned in progressive alignment is determined by inferring a tree-like relationship (crude guide tree) among the sequences based on the matrix of pairwise distances scores between sequences (Figure 5a). The cardinal rule of progressive alignment is that once a gap is introduced, it is inserted in all the sequences of the same group, gap can only be added or enlarged, never moved or removed.
Most successful implementation of this algorithm is in CLUSTAL W (Thompson, Higgins and Gibson 1994), which is one of the most widely used alignment program, particularly for high-throughput genomic analysis and tends to be among the most accurate.

There are many additional multiple DNA sequence alignment algorithms and programs available, some of which use similar progressive alignment schemes as CLUSTAL W and T-COFFEE (Notredame, Higgins and Heringa 2000), and some of which use very different approaches, including statistical alignments based on maximum likelihood or Bayesian methods. T-Coffee calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence. It uses the output from Clustal as well as another local alignment program DIALIGN (Morgenstern et al. 1998), which finds multiple regions of local alignment between two sequences. The resulting alignment and phylogenetic tree are used as a guide to produce new and more accurate weighting factors.
Figure 5. Steps involved in progressive sequence alignment. (a) The first step is to reconstruct a guide tree (e.g., NJ). (b) Second step is to determine the order in which the sequences are added to the growing alignment.

Several other methods have been developed in recent years to achieve better performance in accuracy, speed or both for example MAFFT,(Katoh et al. 2002; Katoh et al. 2005), MUSCLE (Edgar 2004), PROBCONS (Do et al. 2005), DIALIGN (Morgenstern et al. 1998) etc. Table 7 provides a brief summary of these multiple sequence alignment program. For a detailed discussion on MSA, (Edgar and Batzoglou 2006).
1.2.2.3. Choosing an alignment method

There are three main considerations in choosing a MSA program: biological accuracy, execution time and memory usage (see Table 8). Table provides description of some of the recommended multiple sequence alignment methods under certain alignment tasks (Edgar and Batzoglou 2006). Multiple sequence alignment is an important tool in comparative genomics. It has been suggested that the quality of a tree may be dependent upon the methods of alignment than on the methods of phylogenetic reconstruction (Hall 2005a).
1.2.2.4. **Alignment Gap Treatments**

Multiple sequence alignment is a process of converting sequences of unequal length into sequences of equal length by introduction of gaps (which represent indel events) for a proper site-by-site comparison of regions among sequences that are homologous. The position of insertion and deletion events in molecular data sets can be useful phylogenetic information (see references in (Simmons and Ochoterena 2000; Simmons, Ochoterena and Carr 2001; Ogden and Rosenberg 2007; Simmons, Muller and Norton 2007), yet this information is rarely used, especially in large data sets with many indels. Some studies believe that indels may be unreliable as characters (Golenberg et al. 1993; Li 1997). However, it should be noted that because gaps are the product of the alignment procedure, and are not actually found in organisms or their sequences, sequences with gap characters do not have anything to compare with other sequences at the point where the gap occurs (Simmons, Ochoterena and Carr 2001). This becomes a
problem when reconstructing phylogeny. Thus, insertions and deletion events are treated in a variety of different ways during phylogenetic analysis (see (Swofford 2003a; Young and Healy 2003; Muller 2006; Ogden and Rosenberg 2007; Simmons, Muller and Norton 2007). Here, we discuss the most commonly used gap treatment methods.

Gapped sites can be treated as additional character state such as, fifth character state for bases in DNA (Eernisse and Kluge 1993) in maximum parsimony (Swofford 2003a). Therefore, adjacent gap characters are considered independently of their neighbors, although subsequent gap characters after the first may be weighted less heavily to reflect the possibility of longer indel regions. Essentially, each individual gap position is considered as if it were a separate indel event. This is not very realistic. Insertion or deletion events often consist of multiple bases. Since many gap characters do not arise independently of one another, counting each gap character as a separate event causes indel events to be considered multiple times in determining phylogenetic relationships. This over-weights the indels and can distort phylogenies. For these reasons, gaps should not be considered as a fifth character state for nucleotide characters.

Indels can also be treated as Binary characters or presence/absence characters, where the gapped sites in each column are coded as binary characters (1 if gap present, 0 if absent) available for the parsimony (Swofford 2003a) and Bayesian methods (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003).

Simple indel coding, SIC (Simmons and Ochoterena 2000). In this method, indels (or gaps) are coded as separate presence/absence characters in a data matrix, which is then considered along with the DNA base characters in phylogenetic analysis. Each
indel with different start and/or end positions is considered to be a separate character, which all of the sequences under consideration either have or lack. If one of the indels completely overlaps an indel contained within another sequence, the sequences containing the longer indel are coded as being inapplicable for the shorter indel. This is done because it is impossible to determine whether or not the shorter indel is present in the sequences containing the longer one. Simple indel coding method allows indels to be highly informative in determining a correct phylogeny (Simmons, Ochoterena and Carr 2001). This method attempts to better account for the fact that indels are evolutionarily related to one another, and that an indel region may be modified through additional insertion/deletion events to yield a different indel region in another sequence.

Complex indel coding, like simple indel coding, codes indels with different start and end positions as individual characters. However, overlapping indels may represent an evolutionary transition sequence (Simmons and Ochoterena 2000). Step matrices are constructed to accommodate this possibility. Complex indel coding utilizes more of the available information and never implies fewer steps than what is biologically realistic. However, this method generates some multi-state characters and step matrices and is thus more complicated to program. Also, the step matrices slow down phylogenetic programs, coding of the step matrices may now be easily automated using SeqState (Muller 2005). Later, (Muller 2006) proposed modified version of complex indel coding method collectively referred to as modified complex indel coding (MCIC; (Muller 2006)), which overcomes some of the limitation of previous complex indel coding method.
In contrast to each of the above approaches, which incorporate phylogenetic signal from indels into the data matrix, one may simply treat gapped positions as missing data and is inferred based on the optimization criteria, based on whether the inference method is distance-based, parsimony, likelihood, or Bayesian (Huelsenbeck and Ronquist 2001; Guindon and Gascuel 2003; Swoford 2003a). Treating gaps as unknown or missing data is the default option in PAUP.* The FAQ page for PAUP* at http://paup.csit.fsu.edu/paupfaq/paupfaq.html explains the working of this treatment under each inference method, and for PhyML and Bayesian, it is explained in Hulesenbeck and Ronquist 2001, Guindon and Gascuel 2003, Ronquist and Huelsenbeck 2003. Briefly, PAUP* deals with missing characters in the following manner: under the parsimony criterion, a missing character in a sequence is assigned the most parsimonious state given its placement in the tree. Under the likelihood criterion, a gapped site is assigned a state based on the likelihood which is computed by summing the likelihoods over all possible states – a strategy that is used by PhyML as well. For distance methods, PAUP* deals with the missing data by distributing the missing or ambiguous changes proportionally to each unambiguous change. Bayesian analysis was done using the program MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Altekar et al. 2004), which deals with gaps just as other Maximum Likelihood programs (Felsenstein 1989; Yang 1997; Swoford 2003a; Yang 2007).

Gapped positions can be excluded from the analysis altogether. Excluding gapped positions has the advantage of eliminating many nucleotide characters that contain inapplicable data (or “missing data”) from the data matrix, as these gaps introduce some
complications in distance estimation. Furthermore, sites with missing information can sometimes occur because of experimental difficulties, and they create the same problems as that for gaps. In MEGA (Kumar, Tamura and Nei 1993; Tamura et al. 2007), gapped sites are usually ignored during distance estimation in distance-based methods. There are two different ways to treat these sites. One way to deal with this problem is to delete all of these sites from data analysis. This option, which is called the Complete-Deletion option in MEGA, is generally desirable because different regions of DNA or amino acid sequences often evolve under different evolutionary forces. However, if the number of nucleotides involved in a gap is small and gaps are distributed more or less at random, one may compute a distance for each pair of sequences ignoring only those gaps that are involved in the comparison. This option is called the Pairwise-Deletion option.

Software examples: PAUP* (Swofford 2003a), PHYLIP-Erate (Felsenstein 1989; Rivas and Eddy 2008), GAPCODER (Young and Healy 2003), BARCOD (http://www.abi.snv.jussieu.fr/public/Barcod/), SEQSTATE (Muller 2005), MEGA (Tamura et al. 2007).

1.2.3. Model Selection in Phylogenetics

1.2.3.1. Nucleotide Substitutions

Molecular sequences diverge from their common ancestral sequences by means of several evolutionary processes. Substitution is the replacement of a nucleotide by another nucleotide, insertion or deletion is the insertion or deletion of one more
nucleotides. If insertions or deletions occur in a protein-coding gene, they may shift the reading frame of the nucleotide sequence. There insertions and deletions are called frameshift mutations. Nucleotide substitutions can be divided not types: *transition* and *transversions*. A transition is the substitution of a purine (adenine & guanine) for another purine or the substitution of a pyrimidine (thymine and cytosine) for another pyrimidine (Figure 6). A transversion is the substitution of pyrimidine for a purine or vice versa. Since there are four nucleotides, A, T, C, and G, in each sequence, there are 16 different types of nucleotides pairs. There are four pairs of identical nucleotides (AA, TT, CC, and GG), four transition-type pairs (AG, GA, TC, and CT), and eight transversion-type pairs (all the remaining pairs). If nucleotide substitution occurs at random among the four nucleotides, transversions are expected to be about two times higher than transitions. In practice, however, transitions usually occur more often than transversions (e.g., (Fitch 1967; Gojobori, Li and Graur 1982; Kocher and Wilson 1991).

![Figure 6. Transitional and transversional nucleotide substitutions.](image)

In the case of protein-coding genes, nucleotide substitutions that result in synonymous codons are called synonymous or silent substitutions, whereas those that result in nonsynonymous codons are called non-synonymous or amino acid replacement
substitutions. In addition there are mutations that result in stop codons, and they are called non-sense mutations. Most synonymous substitutions occur at the third nucleotide position of codons, but some occur at the first position. All nucleotide substitutions at the second position are either nonsynonymous or nonsense mutations. Often rate of synonymous nucleotide substitutions is lower than rate of non-synonymous substitution due to purifying selection. However, there are genes in which the rate of nonsynonymous substitutions is much higher than rate of synonymous substitutions. This is positive selection. In cases, when the rate of nonsynonymous substitutions is equal to the rate of synonymous substitutions, it is called neutral selection. Therefore, rate of synonymous substitution ($r_S$) & nonsynonymous substitution ($r_N$) is defined as the number of synonymous substitutions per synonymous site ($d_S$) and the number of non synonymous substitution per nonsynonymous site ($d_N$).

1.2.3.2. Model of Nucleotide Substitutions

A simple measure of the extent of sequence divergence is the proportion ($p$) of nucleotide sites at which the two sequences are different. $p$-distance is given by

$$p = \frac{n_d}{n}$$

where $n_d$ and $n$ are the number of different nucleotides between the two sequences and the total number of nucleotides examined, respectively.

$p$-distance gives a correct estimate when the sequences are closely related. However, when $p$ is large, it gives an underestimate of the number (uncorrected distances); because it does not take into account multiple substitutions, (backward and
Because of multiple substitutions at a site, the degree of divergence, $p$, between two compared sequences is smaller than the total number of substitutions that have occurred in two lineages since the divergence between the two sequences (Figure 7a). This problem is more serious for nucleotide sequences than for amino acid sequences because there are only four character states in nucleotide sequences. Therefore, to estimate the number of nucleotide substitutions, it is necessary to use a mathematical models of nucleotide substitutions (corrected distances). Model-based methods such as Maximum Likelihood ML and Bayesian analysis offer advantages when sequence divergence is large. Some of the important models of nucleotide substitution for nucleotide sequences analysis are discussed below. For a detailed review, see references from Swofford et al. 1996, Li 1997, Nei and Kumar 2000, Felsenstein 2004, Bos and Posada 2005.
Figure 7. **Nucleotide substitution saturation.** (a) Relationships of the p-distance with the proportion of nucleotide differences. (b) Evolutionary changes in nucleotide sequences.

**Jukes-Cantor (JC) model**

The Jukes and Cantor (1969) model assumes that nucleotide substitutions occur at equal frequencies (A=C=T=G=0.25). It also assumes that the rate of substitution is the same for every pair of nucleotides and thus they only have one parameter (given by α).

The JC distance \(d_{JC}\) between a pair of sequences is given by the following formula:

\[
\begin{align*}
    d_{JC} &= \left(-\frac{3}{4}\right) \ln \left[1 - \left(\frac{4}{3}\right)p\right] \\
    &= \left(-\frac{3}{4}\right) \ln \left[1 - \left(\frac{4}{3}\right)p\right]
\end{align*}
\]

This is one of the simplest models of DNA evolution and cannot be used when \(p\) is large (\(p \geq 0.75\)).

**Kimura two parameter (K2P or K80)**

As mentioned above, the rate of transitional nucleotide substitutions is often higher than that of transversional nucleotide substitutions in real data analysis. Therefore, Kimura (1980) introduced a model of nucleotide substitution that took transition and transversion differences into consideration. In this model, the rate of transitional substitutions per site (α) is assumed to be different from that of transversional substitutions (2β).
Table 9. Substitution rate matrix for Kimura-2-parameter model.

<table>
<thead>
<tr>
<th>Substitution From</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>β</td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>C</td>
<td>β</td>
<td>-</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>β</td>
<td>-</td>
<td>β</td>
</tr>
<tr>
<td>T</td>
<td>β</td>
<td>α</td>
<td>β</td>
<td>-</td>
</tr>
</tbody>
</table>

The Kimura 2-parameter distance ($d_{Kimura}$) for a pair of sequences is given by the following formula:

$$d_{Kimura} = -\left(\frac{1}{2}\right)\ln[1-2P-Q]-\left(\frac{1}{4}\right)\ln[1-2Q]$$

where, $P$ and $Q$ are the observed proportions of transitions and transversions, respectively, between the two sequences. This model assumes 2 parameters (alpha, $\alpha$ and beta, $\beta$) and thus called the Kimura 2-parameter model.

Hasegawa-Kishino-Yano (HKY85) model

The Hasegawa, Kishino and Yano (1985) model is a hybrid of Kimura-2-parameter model and equal input model and takes into account both the transition/transversion and GC content biases. The model allows the four nucleotides to be present in different frequencies. It assumes that transitions and transversions occur at different rates. One of the closely related model to HKY is the (Felsenstein 1984) or F84.
Table 10. Substitution rate matrix for the HKY model.

<table>
<thead>
<tr>
<th>Substitution From</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>$\beta_g$</td>
<td>$\alpha_g$</td>
<td>$\beta_g_T$</td>
</tr>
<tr>
<td>C</td>
<td>$\beta_g_A$</td>
<td></td>
<td>$\beta_g_T$</td>
<td>$\alpha_g_T$</td>
</tr>
<tr>
<td>G</td>
<td>$\alpha_g_A$</td>
<td>$\beta_g_C$</td>
<td></td>
<td>$\beta_g_T$</td>
</tr>
<tr>
<td>T</td>
<td>$\beta_g_A$</td>
<td>$\alpha_g_C$</td>
<td>$\beta_g_G$</td>
<td></td>
</tr>
</tbody>
</table>

$\beta_g, \beta_T, \alpha_g$, and $\alpha_T$ are the nucleotide frequencies.

**General Time Reversible (GTR) model**

Widely used models of nucleotide substitutions are usually time reversible; an $A \rightarrow T$ transversions is treated as equivalent to a $T \rightarrow A$ transversions. Thus six possible substitution types exist among the four nucleotides. Each of these transformations types may be treated as equivalent (Jukes and Cantor 1969), transitions may be treated separately from transversions (e.g., Hasegawa, Kishino and Yano 1985, Kimura 1980), all six may be treated as unique (Tavare' 1986), or any combination of the six types may be grouped. Furthermore, base frequencies may be assumed to be equal or allowed to vary.
Table 11. Substitution rate matrix for the GTR model.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution From</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>$a_{gC}$</td>
<td>$b_{gG}$</td>
<td>$c_{gT}$</td>
</tr>
<tr>
<td>C</td>
<td>$a_{gA}$</td>
<td>-</td>
<td>$d_{gG}$</td>
<td>$e_{gT}$</td>
</tr>
<tr>
<td>G</td>
<td>$b_{gA}$</td>
<td>$d_{gC}$</td>
<td>-</td>
<td>$f_{gT}$</td>
</tr>
<tr>
<td>T</td>
<td>$c_{gA}$</td>
<td>$e_{gC}$</td>
<td>$f_{gG}$</td>
<td>-</td>
</tr>
</tbody>
</table>

**Gamma distances**

The rate of nucleotide substitution is assumed to be constant for all nucleotide sites. In reality, this assumption rarely holds true, and the rate varies from site to site. For example, fast genes and slow genes, there are fast sites and slow sites within genes. Sites differ in how much they are free to vary. A site may be under strong selection and highly constrained; other sites, such as third codon positions, might be relatively unconstrained. Thus, to accommodate the among site rate variation (ASRV), model based on a gamma distribution has been proposed (Kocher and Wilson 1991; Tamura and Nei 1993; Wakeley 1993; Wakeley 1994). The gamma (or G) distribution is usually described with 2 numbers, $\alpha$ and $\beta$, that define the shape and mean of the distribution (Figure 8). The single parameter $\alpha$ is then inversely related to the extent of rate variation. The distribution with $\alpha \leq 1$ is L-shaped, meaning that most sites have very low substitution rates or are virtually invariable, while a few sites exist (substitution hotspots) with very high rates (Figure 8). The distribution with $\alpha > 1$ is bell shaped, meaning that...
most sites have intermediate rates while few sites have very low or very high rates (Figure 8). When \( \alpha \) approaches infinity (\( \infty \)), the model reduces to the case of a constant rate for all sites (Figure 8). By adjusting \( \alpha \), the gamma model can account for different levels of rate variation in real data (Yang 1996a).

![Gamma distribution to model site rate heterogeneity.](image)

That shape parameter \( \alpha \), is usually a free parameter in our models, and so it does not need to be provided, as it can be estimated from the data by maximum likelihood method (Yang 1994a) or other methods (e.g., (Kocher and Wilson 1991; Sullivan, Holsinger and Simon 1995; Yang and Kumar 1996). Yang (1996) has computed the \( \alpha \) value for various nuclear and mitochondrial genes. Among-site rate variation can be incorporated into models of nucleotide evolution (for example, we can specify a GTR+G model). Gamma distances are generally more realistic than no-gamma distances, but they
have larger variances than the latter. For estimating branch lengths of a tree, gamma
distances generally give better results.

The above-mentioned model parameters all work at the individual nucleotide
level and therefore treat each nucleotide as an independent unit. However, for protein
coding DNA sequence this is not the case. Whether or not a substitution changes an
amino acid depends on the other nucleotides in that codon when the substitution occurs,
thus individual nucleotide sites in protein coding sequence are not independent. To
accommodate this, nucleotide models that treat a codon triplet as an independent unit
have been formulated to more accurately model coding DNA (Goldman and Yang 1994;
Muse and Gaut 1994; Pedersen, Wiuf and Christiansen 1998). Variations of these
models provide parameters to account for transition bias, codon frequency, rate variation
among codon positions, and different rates for non-synonymous substitutions (Yang et al.
2000).

The performance of a model-based phylogenetic method may depend on the fit of
the model to the data (Huelsenbeck and Crandall 1997). When working with more
divergent sequences, the use of one model over another can alter the results of analysis,
and even lead to strong support for the wrong tree topology (Kelsey, Crandall and
Voevodin 1999), a fact that underscores the importance of using the best fit model for a
particular data set. Due to the wide diversity in size, variation and rates of evolution
among different data sets, there is no single best-fit model suited for use in any data set.
Use of inadequate, overly simplistic models selected without statistical validation often
leads to biased estimation of evolutionary genetic parameters (Huelsenbeck and Hillis
The model parameter with one of the strongest influences on genetic distance and phylogenetic estimation is among-site rate variation. Rate variation among sites is particularly problematic and misleading when substitution rates also vary among branches in the tree (e.g. non-clock-like evolution) (Kuhner and Felsenstein 1994). When both types of variation are present, use of the best fit model seems to be essential to obtain the correct tree topology (Yang 1996a). Except in cases with strong rate variation among both sites and lineages, tree topology estimation is relatively robust to violations of model assumptions (Yang 1994b; Yang, Goldman and Friday 1995). Unfortunately the same robustness does not extend to estimation of parameters such as substitution rates, branch lengths and genetic distance. Failing to include rate heterogeneity among sites results in underestimation of the number of substitutions at highly mutable sites (Yang 1996a).

Simplifying the assumptions of a model by failing to include a factor for transition bias can also adversely alter the outcome of analysis. A transition bias is found universally among DNA sequences (Wakeley 1994) and inclusion of this parameter is essential for accurate estimates of genetic distance for NJ analysis (Tamura 1992; Tajima and Takezaki 1994). Similarly, failure to incorporate transition bias will result in underestimation of branch lengths in ML phylogeny estimation (Yang, Goldman and Friday 1994). Aside from the inherent problems of branch length and genetic distance underestimation, these factors can alter the tree topology and lead to erroneous conclusions regarding the dates of lineage splitting (Tamura and Nei 1993). There is also
an correlation between transition bias and among site rate variation, so that the level of among-site rate variation is underestimated (overestimation of a) using models that exclude a transition bias (Yang, Goldman and Friday 1994). Intuitively, when transitions occur with higher frequency than transversions, many transitional substitutions are expected at the fast-changing sites. If the among-site rate variation is ignored, some of the transitions will be overlooked and the transition to transversion rate ratio will be underestimated (Kuhner and Felsenstein 1994; Yoder and Yang 2000). One of the major advantages of using models is the ability to more accurately estimate the actual number of substitutions that have occurred in a set of sequences. The alternative way of dealing with sites or sequences which are suspected of saturation of substitutions, is simply to eliminate them from consideration. While this does effectively eliminate the influence of homoplasy at those sites, any information that can be gleaned from those sites is also lost and the size of the sample is decreased, exposing the analysis to the increasing effects of sampling error or bias. The amount of data required for consistent phylogenetic analysis depends on the shape of the tree, numbers of taxa and levels of diversity. If the tree shape is not symmetric and branch lengths are very long, then analysis of data with less than 500 nucleotides will generally not be reliable, especially for more general models (Huelsenbeck and Hillis 1993; Sullivan and Swoford 2001). Consistency and reliability of phylogenetic inference is expected to increase by analyzing longer sequences and additional taxonomic sampling.
1.2.3.3. **Choosing a model**

Given that model choice is critical in phylogeny estimation and the vast array of potential models from which to choose, one is faced with the decision of how to select from among these. In general, use the simplest model that adequately explains the data. If a more complex model yields a greater improvement in tree score (or other measure of goodness of fit to data) than would be expected if applied to random data, then use the more complex model. Models differ in their free (i.e. adjustable) parameters. We want to find out a model that best fits the data. The more free parameters, the better the fit of the model to the data. However, the more free parameters, the higher the variance, and the less power to discriminate among competing hypothesis. Model complexity can often lead to computational intractability, so pragmatic concerns sometimes outweigh statistical ones (for example, NJ and parsimony are mainly justifiable by their speed).

The goodness of fit of a model to observed data can be examined by using the likelihood ratio (LR) test (Felsenstein 1988) or Akaike Information Criterion (AIC) (Akaike 1974), and Bayesian Information Criterion (BIC) (Schwarz 1978). When there are two models, models 1 and 2, and model 1 is a special case of model 2, model 1 is said to be nested in model 2. When the correct topology is known and model 1 is nested in model 2, one can compute the log likelihood ratio using the following equation:

\[ LR = 2(lnL_2 - lnL_1) \]
where $\ln L_1$ and $\ln L_2$ are the ML values for models 1 and 2, respectively. Therefore, we can test whether model 2 is significantly better than model 1 or not. In general, the likelihood ratio test cannot be used unless the two models compared are nested. However it is possible to compare the two nonnested models by using Akaike Information Criterion (AIC) (Akaike 1974) as long as the topology considered remains the same. AIC is defined as:

$$AIC = -2\ln L + 2p$$

where $\ln L$ is the log likelihood value for a given model and $p$ is the number of free parameters to be estimated. It has been proposed that the statistical predictability of a model is higher when AIC is low than when it is high. In actual data analysis, the AIC value is almost always lower when a sophisticated model (such as the HKY model) is used than when a simple model (K2P) is used. However, empirical studies with known phylogenies have shown that the AIC value has virtually no correlation with the probability of obtaining the correct topology (Russo, Takezaki and Nei 1996). Theoretical and simulation studies have also indicated that a sophisticated model does not necessarily give the correct topology with a higher probability than a simple model (Takahashi and Nei 2000).

Other model selection strategies such Bayesian Information Criterion (BIC) (Schwarz 1978) also attempt to find the appropriate level of complexity on the basis of the available data. ModelTest is a computer program written by (Huelsenbeck and Crandall 1997; Posada and Crandall 1998) that automates the process for the selection of
model of nucleotide substitution that best fits the data. The program chooses among 56 models, and implements three different model selection frameworks: hierarchical likelihood ratio tests (hLRTs), Akaike information criterion (AIC), and Bayesian information criterion (BIC). The program also implements the assessment of model uncertainty and tools for model averaging and calculation of parameter importance, using the AIC or the BIC. Programs such as Prottest (Abascal, Zardoya and Posada 2005), and Modelgenerator (Keane et al. 2006) are closely related to Modeltest. In addition, review articles by (Posada and Buckley 2004; Sullivan and Joyce 2005) provide an excellent overview of model choice in phylogenetics.

1.2.4. Phylogenetic Tree Reconstruction Methods

Phylogenetic analysis of DNA or amino acid sequences has become an important tool for studying the evolutionary history of organisms. Since the rate of divergence varies extensively with gene, one can study the evolutionary relationships of all groups of organisms. There are two processes involved in the phylogenetic inference: estimation of tree topology (branching pattern of a tree) and estimation of branch lengths for a given tree topology.

There are two main approaches for inferring phylogenies. The first one, called the phenetic approach, is the study of relationships among a group of organisms on the basis of the degree of similarity between them, be that similarity molecular, phenotypic or anatomical. A tree-like network representing phenetic relationships (topology and
branch lengths) is called a phenogram. It accepts all monophyletic, paraphyletic and polyphyletic groups. There is no restriction on the number or type of characters (data) that can be used, although all data must be first converted to a numerical value, without any character weighting. Each organism is then compared with every other for all characters measured, and the number of similarities (or differences) is calculated. The organisms are then clustered in such a way that the most similar are grouped close together and the more different ones are linked more distantly. The taxonomic clusters, called phenogram, that result from such an analysis do not necessarily reflect genetic similarity or evolutionary relatedness.

The second one, called the cladistic approach, considers possible pathways of evolution, inferring the features of the ancestor at each node and choosing an optimal tree according to some model of evolutionary change. The basic assumption behind cladistics is that members of a group share a common evolutionary history. Thus, they are more closely related to one another than they are to other groups of organisms. A tree-like network that represents an ancestor-descendent relationship is called a cladogram. A cladogram refers to the topology of a rooted phylogenetic tree (branch lengths are meaningless). The phenetic approach is based on similarity whereas the cladistic approach is based on genealogy.

Several methods of estimating phylogenetic trees are available. Some of the more commonly used methods include Neighbor Joining (NJ) (Saitou and Nei 1987), Maximum Parsimony (MP) (Fitch and Margoliash 1967; Fitch 1971) and Maximum
Likelihood (ML) (Felsenstein 1981), and Bayesian method (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). While several differences exist, one common feature that unites NJ, ML and Bayesian methods is the use of explicit statistical models of nucleotide evolution. For a comprehensive discussion of the methods for inferring phylogenies see Hillis, Moritz and Mable 1996, Nei 1996, Li 1997, Nei and Kumar 2000, Felsenstein 2004, and Hall 2004.

1.2.4.1. Distance-based Methods

In distance methods, evolutionary distances (number of nucleotide or amino acid substitutions) are computed for all pairs of taxa, and a phylogenetic tree is constructed by using an algorithm based on relationships among these distance values. There are several distance methods for reconstructing phylogenetic trees, but those are commonly based on the principles of least squares (Cavalli-Sforza and Edwards 1967) and minimum evolution (Fitch and Margoliash 1967).

There are four main distance matrix methods:

1) UPGMA
2) Least square (LS)
3) Minimum Evolution (ME)
4) Neighbor Joining (NJ)
1) UPGMA

The unweighted pair group method using arithmetic averages (UPGMA) is one of the simplest method for tree reconstruction, was originally developed in the early 1960s (Rohlf 1993) for evolutionary analysis of morphological characters and thus tree constructed by this method is sometimes called a phenogram. However, it can be used to infer molecular phylogeny when the rate of evolution of gene(s) is constant (Molecular clock assumption) and the evolutionary distance is large for all pairs of sequences.

*Molecular Clock Hypothesis* states that the rate of nucleotide substitutions, or amino acid substitutions (if proteins are being compared) is approximately constant over evolutionary time, that is, the degree of difference between two sequences can be used to assign a date of the time at which their ancestral sequence diverged. This was first noted by Zuckerkandl and Pauling (1962,1965). The rate of molecular change however, differs among groups of organisms, among genes, and even among different regions of the same gene. Furthermore, a molecular clock requires calibration with fossils to determine timing of origin of clades, and thus their accuracy is crucially dependent on the fossil record. However, the concept of the molecular clock has a long history of controversy, and it is still hotly debated by evolutionist (see (Nei and Kumar 2000)). If it applies for a certain group of organisms, it is still very useful for studying evolutionary relationships of organisms or for estimating time of divergence between different organisms.

In UPGMA, evolutionary distances are computed for all pairs of taxa or sequences, and the distance values are presented in the form of matrix. For example, suppose we want to construct a tree of five species A, B, C, D and E using UPGMA
method. The matrix shown below (Table 11) is the distance-matrix for these five species; these are the pair wise distances that can be obtained from any of the distance methods.

Table 12. Distance matrix for the five sequences.

<table>
<thead>
<tr>
<th>Species</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>12</td>
<td>11</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>

UPGMA work’s by clustering of taxa’s with the smallest distance. Here, $d_{AB}$ (distance between A and B) is smallest among all the distance values in the above matrix. Then taxa A and B are clustered together with the branch point 1(Figure 9).

![Figure 9. The first cluster created by the UPGMA algorithm.](image)

Here, we assume that the lengths of the branches leading from this branch point to taxa A and B are the same. Taxa A and B are then combined into a single composite taxon or cluster (A,B) and again the distance between this cluster (A,B) and another taxon is computed as follows:

\[
\text{dist}_{(A,B),C} = \frac{\text{dist}_{AC} + \text{dist}_{BC}}{2} = 5.5 \\
\text{dist}_{(A,B),D} = \frac{\text{dist}_{AD} + \text{dist}_{BD}}{2} = 9 \\
\text{dist}_{(A,B),E} = \frac{\text{dist}_{AE} + \text{dist}_{BE}}{2} = 11.5
\]

Thus, we have the following new matrix (Table 12).
Table 13. Reduced distance matrix.

<table>
<thead>
<tr>
<th>Species</th>
<th>A,B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>5.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>9</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>11.5</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>

Here the distance between \( d_{(AB,C)} \) is smallest and hence the taxa AB and C are combined into a new composite taxon or cluster (Figure 10).

![Figure 10. The second cluster created by the UPGMA algorithm.](image)

The distance between the newly created cluster and each of the remaining taxa is now computed by

\[
\text{dist}_{(AB,C),D} = \frac{\text{dist}_{AD} + \text{dist}_{BD} + \text{dist}_{CD}}{3} = 9.3
\]

\[
\text{dist}_{(AB,C),E} = \frac{\text{dist}_{AE} + \text{dist}_{BE} + \text{dist}_{CE}}{3} = 12
\]

We then have,

Table 14. Reduced distance matrix

<table>
<thead>
<tr>
<th>Species</th>
<th>A,B,C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>9.3</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

As the distance between D and E is the smallest in the above distance matrix, we will then combine D and E. Thus, the tree obtained from UPGMA method for the five species is:
A tree obtained by UPGMA is usually presented as rooted tree (Figure 11). However, one can also construct an unrooted UPGMA tree in order to test the reliability using the Bootstrap (Felsenstein 1985) and Interior branch test methods (Nei, Stephens and Saitou 1985). Both methods test the positiveness of each of the internal branches in a tree. Once it is proved the tree is regarded as reliable on statistical basis. UPGMA assumes a molecular clock, and thus is very sensitive to unequal evolutionary rates.

2) Least-Square (LS) methods

Least square method is used to obtain the topology in cases where rate of nucleotide substitution varies from one evolutionary lineage to another. The most commonly used LS methods are general LS (Cavalli-Sforza and Edwards 1967) and weighted LS method (Fitch and Margoliash 1967). Both methods compute the minimum sum of squared differences between pairwise distances and estimate pairwise distances for a given topology and the topology with the smallest value is chosen as the best tree.

However, LS methods often give negative estimates of branch lengths, and mainly for this reason probability of obtaining correct tree topology is often lower than
other distance methods (Saitou and Nei 1987; Saitou and Imanishi 1989; Kuhner and Felsenstein 1994).

3) Minimum-Evolution (ME) methods

This method follows an optimality criterion to choose among the different possible tree topologies for a given set of sequences. The topology that satisfies the criterion is considered the best tree. In this method, sum of all branch lengths (S) is computed by least square method (Kidd and Sgaramella-Zonta 1971; Rzhetsky and Nei 1992) or Fitch and Margoliash’s algorithm (Saitou and Imanishi 1989) for all possible topologies and the topologies that has the smallest S value is chosen as the best estimate of the true tree. One of the good statistical property of ME method is that if the distance estimates are unbiased estimates of the true distance then branch length for the true tree is smaller than sum of branch lengths for any other tree.

Although, ME method is statistically appealing, it requires large amount of computational time particularly when the number of sequences (m) is large (>10).

4) Neighbor-Joining (NJ) method

Saitou and Nei (1987) developed the NJ method that does not examine all possible topologies, but at each stage of taxon clustering uses a minimum evolution (ME) principle. NJ is based on the concept of neighbors, which are defined as two taxa that are connected by a single internal node in an unrooted tree. For example, in the unrooted tree
shown below (Figure 12), species A and B are neighbors; likewise C and D are neighbors.

![An unrooted phylogeny of four taxa](image)

**Figure 12. An unrooted phylogeny of four taxa**

NJ method starts with a star tree, which is produced under the assumption that there are no internal nodes and no clustering of taxa (see Figure 13). The sum of the branch lengths of the star tree is computed ($S_0$). For example, if we begin with the star tree of 6 species as shown below (Figure 13) and consider A, B as neighbor’s, then they are joined and a new sum of branch length is computed again ($S_1$). Similarly all possible combinations are computed. The pair of taxa with the smallest branch length is considered neighbors and this result in adding of one internal node to the tree. Likewise next pair of neighbors is determined until all the taxa are joined (Figure 13).
Figure 13. Illustration of the computational process in the neighbor-joining method.

Neighbor-joining method generates only one possible tree (an additive tree) with branch length estimates and is much faster as compared to other methods. Some other distance-based methods include the Bio-NJ (Gascuel 1997a), and Unweighted Neighbor-Joining (Gascuel 1997b) etc.

Distance Measures

Theoretically, if the total number of substitutions between any pair of sequences \((d)\) is known, all the above distance methods produce the correct phylogenetic tree (additive tree). In reality, this number is almost always unknown, and thus many different methods for estimating this number have been proposed. The simplest distance estimate is the proportion of observed nucleotide differences \((p\) distance\) between two sequences per site. However, when \(p\) is large (or when sequence divergence is high), it
gives an underestimate of the number, because it does not correct for multiple substitutions or saturation. To estimate the number of nucleotide substitutions, it is necessary to use mathematical model of nucleotide substitution such as the Jukes-Cantor model, Kimura distances, and others (e.g., HKY85) to produce the correct phylogeny. However, if $d$ is large and if the substitution rate varies extensively among sites and with evolutionary lineage, gamma distance based on the gamma distribution (Tamura and Nei 1993; Wakeley 1993; Wakeley 1994) expected to produce better trees. When a phylogenetic tree is reconstructed from the coding regions of a gene, the distinction between synonymous ($d_S$) and non-synonymous substitutions ($d_{NS}$) may be helpful in reconstructing a tree, because the rate of synonymous substitutions is usually much higher than that of nonsynonymous substitutions.

Table 15. Comparison of distance-based methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbor-Joining</td>
<td>Fast- suitable for analyzing large data sets ($m &gt; 500$), and the bootstrap can be done very easily.</td>
<td>Information is lost – given only the distances it is impossible to derive the original sequences.</td>
<td>PAUP</td>
</tr>
<tr>
<td></td>
<td>A large number of models are available with many parameters- improves estimation of distances.</td>
<td>Only through character based analyses can the history of sites be investigated e.g., most informative positions be inferred.</td>
<td>PHYLIP</td>
</tr>
<tr>
<td></td>
<td>Can use ML to test the fit of model to data</td>
<td>Generally outperformed by likelihood based methods.</td>
<td>MEGA</td>
</tr>
<tr>
<td>Minimum-Evolution</td>
<td>Uses model to correct for unseen changes.</td>
<td>Distance corrections can break down when distances are large.</td>
<td>PAUP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PHYLIP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MEGA</td>
</tr>
</tbody>
</table>
1.2.4.2. Maximum Parsimony

Parsimony is "a principle that states that the simplest explanation that explains the greatest number of observations is preferred to more complex explanations". The theoretical basis of maximum parsimony (a tree-building strategy) comes from the parsimony principle given by William of Ockham's. The data used in MP is in the form of "character data". A character could be binary (two state) e.g. presence or absence of a feature (e.g. wings), or it could be a multistate e.g. the nucleotide or amino acid residues at a particular site in the gene sequence. These differences in character states can be explained by evolutionary changes and thus are useful in inferring the evolutionary relationships of organisms.

In Maximum Parsimony (MP), the preferred phylogenetic tree is the tree that requires the least number of evolutionary changes. MP was originally developed for morphological data (Hennig 1950; Eck and Dayhoff 1966) first used MP for reconstructing trees from amino acid sequences. Later, (Fitch 1971; Hartigan 1973) developed an algorithm for nucleotide sequences. With reference to the molecular data, in MP, the nucleotides of ancestral taxa are inferred separately at each site for a given topology under the assumption that mutational changes occur in all directions among the four nucleotides. The smallest number of nucleotide substitutions that explains the evolutionary process for the topology is computed. This is done for all the possible topologies. The topology that requires the smallest number of changes or substitutions to explain the relationships among the taxa considered is chosen as the best tree. Often two
or more trees with the same minimum number of changes are found, so that no unique tree can be inferred. Such trees are said to be equally parsimonious.

Nucleotide sites that have the same nucleotide for all the taxa are known as the *invariable sites* and the sites at which one or more nucleotide varies are known as *variable sites*. In parsimony, we remove the invariable sites and use only variable sites. Even in variable sites not all sites are useful for finding the MP tree. *Singleton sites* (these are the sites with unique nucleotides) are not informative because the nucleotide variation at the site can always be explained by the same number of substitutions in all the topologies. So, basically we can not prefer one topology over the other. For a nucleotide site to be informative there should be at least two different types of nucleotides and each occurring at least two times. These sites are called *informative sites* (Fitch 1977).

Consider the following example of four sequences:

<table>
<thead>
<tr>
<th></th>
<th>123456789012345678901</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>CTTCGTTGGATCAGTTTGATA</td>
</tr>
<tr>
<td>B</td>
<td>CCTCCTTGGATCATTTTGATA</td>
</tr>
<tr>
<td>C</td>
<td>CTGCTTTGGATCAGTTTGAC</td>
</tr>
<tr>
<td>D</td>
<td>CCGCCTTGGATCAGTTTGAC</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Invariant</th>
<th>**</th>
<th>*</th>
<th>*****</th>
<th>*****</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant</td>
<td>**</td>
<td>*</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Informative</th>
<th>**</th>
<th>**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-inform.</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

There can be three possible unrooted trees for these four sequences (Figure 14). Site 1 is not informative because all sequences at this site have C, such that no change is required in any of the possible trees.
Figure 14. **Procedures of the maximum parsimony method.** The figure shows the number of nucleotide substitutions for each of the three possible tree topologies using informative (site 2 and 3), and non-informative sites (site 5).

For site 5, each of the three possible trees requires 2 changes and so it is not informative, thus parsimony uninformative site (Figure 14). At site 2 and site 3, two trees require 2 changes, whereas one tree requires only 1 change (Figure 14). These sites are parsimony-informative.
Therefore, for a nucleotide site to be informative for constructing an MP tree there must be at least two different kinds of nucleotides at the site, each of which must occur in at least two of the sequences (taxa). In the final step of MP analysis, we sum the number of changes over all the informative sites for each possible tree, this gives the tree length. The topology with the lowest tree length is the tree with the least number of changes – the parsimony tree.

However, sometimes even the singleton sites are informative for topological tree reconstruction in other tree building methods. Some studies include the singleton substitutions to estimate the tree length (Nei and Kumar 2000). It does not affect the identification of MP tree because the number of singleton substitutions is same for all the topologies. It is always good to find out whether the tree length of a given topology is based only on parsimony informative sites or on all variable sites. Even invariable sites are used for phylogenetic inference in distance and likelihood method. In case of homoplasy (explained below), however, MP trees would not be reliable even though they have many informative sites. For this reason, (Kluge and Farris 1969) proposed an estimate consistency index (c), retention index and the rescaled consistency index (Farris 1989) to measure the extent of homoplasy for a given dataset.

MP methods can be divided into weighted MP and unweighted MP methods. In unweighted methods, nucleotide or amino acid substitutions are assumed to occur with equal or nearly equal probability. However, in reality certain substitutions like transitions occur more often than transversions. Maximum parsimony that assigns different weights to different types of substitutions is called weighted parsimony (Farris 1969; Tateno,
Takezaki and Nei 1994). If transitions are completely ignored and only transversions are used the method is called *transversions parsimony*. Under weighted parsimony, slowly evolving sites or substitutions are informative only when distantly related sequences are studied. However, when closely related sequences are used, the fast evolving sites or substitutions are more informative.

**Character evolution**

Observed similarities between two taxa can be due to evolutionary relationships (e.g., shared ancestral characters, shared derived characters) or homplasy (independent evolution of the same character). Phylogenetic trees can be used to distinguish between the ancestral and derived character states. This could be used to infer the evolutionary events such as nucleotide substitutions that happened in the past.

**Plesiomorphic State:** In a given tree (Figure 15), if a sequence has the same base as the common ancestor of all the descendent sequences under consideration then it is the ancestral or primitive state.

![Figure 15. Tree showing shared ancestral state](image)
**Apomorph State:** If a sequence has a different base as compared to the ancestor of the sequences being studied, then it is a derived or apomorphic state. It can be of two types: *Autapomorphic* or unique derived character state (Figure 16a) and *Synapomorphic* or shared derived states (Figure 16b)

**Figure 16. Tree showing derived ancestral states.** (a) Autopomorphic (Unique derived shared character state), (b) Synapomorphic (shared derived character state).

Shared derived characters are very useful for phylogenetic tree reconstruction. In addition, we can infer a rooted tree without an outgroup species. In cladistic parsimony, only shared derived characters are used to reconstruct phylogeny. Most parsimony analysis allows the reversibility of the character states. Methods such as strict cladistic analysis (Hennig 1950) and (Camin and Sokal 1965) parsimony are based on the assumption of irreversibility.
**Homoplasy-misleading evidence of phylogeny**

Occurrence of the same base or amino acid in two gene sequences could be a result of homology or homoplasy. Homoplasy does not reflect the shared ancestry, thus a poor indicator of evolutionary relationships (Figure 17).

![Figure 17. Tree showing homoplasy.](image)

Homoplasy can result from the following:

- a) Convergent Evolution, an independent evolution of the same feature in two unrelated sequences derived from different ancestors.

![Figure 18. Homoplasy-Convergent evolution](image)
b) **Parallel Evolution** results from the independent evolution of same characters, though inherited from same common ancestor.

![Figure 19. Homoplasy-Parallel evolution](image)


c) **Secondary loss**, homoplasy that results from the loss of an ancestral feature such that it resembles the ancestral conditions.

![Figure 20. Homoplasy-Secondary loss](image)
Tree Search Strategies

When number of taxa is small \((m<10)\), it is possible to compute the tree lengths for all the possible topologies based on the procedure described above to find the MP tree. This type of search is the \textit{exhaustive search}. However, when \(m\) is large \((m>10)\) it becomes nearly impossible to examine all topologies. In those cases, we can compute the tree length only for potentially correct topologies only if we know the incorrect topologies. This type of search is called the \textit{specific-tree search}. In cases, when \(m>10\) and incorrect topologies are not known, following two approaches are used, \textit{branch-and-bound method} and \textit{Heuristic search}.

Branch-and-Bound Search

The branch and bound algorithm was first introduced by (Hendy and Penny 1982) in parsimony analysis since then several different versions have been developed (Swofford 2003a). The differences are very small in terms of algorithm and efficiency. In Kumar version (Kumar, Tamura and Nei 1993), the search for the MP tree starts with an initial core tree of three taxa which has only one unrooted tree. This three taxa tree is chosen such that the length of the tree is largest or largest among all possible three taxon trees. The remaining taxa are added to this tree one by one according to a certain order and then the tree length of the new tree is computed at each stage of the taxon addition. If the addition of the taxon to a particular branch results in a tree length greater than a predetermined upper bound of tree length (LU), this topology and all the subsequent topologies that can be generated by adding more taxa to this core are ignored.
Core Tree and Order of Taxon Addition

The initial core tree for the search is chosen such that the length (L) of the tree is largest among all possible three-taxon trees. This is to make L closer to the length (L_M) of the MP tree so that one can reach the MP tree faster. In order to determine the order of taxon addition, one of the remaining taxa is added on one of the three branches of the initial core tree and a tree length is computed by the MP procedure. This computation is repeated for the two remaining branches and minimum value of the three tree lengths is recorded. This process is then repeated for all the remaining taxa. The taxon that shows the maximum value of the minimum tree lengths is the first to be added to the initial core tree. This algorithm is known as max-mini algorithm. To find the next taxon, the same algorithm is applied for the remaining taxa using the tree for the first four taxa as the next core tree. Again, the taxon that shows the maximum of the minimum tree lengths is added to the initial core tree of three taxa. This process is repeated until the final taxa is added or until the addition order of all taxa is determined. Since the maximum of the minimum values is closer to L_M than many other value, this order to taxon addition is expected to speed up the search for the MP tree.

Searching the MP trees

Based on the initial core tree and order of taxon addition, we begin the next step of the algorithm. A predetermined upper bound of tree length (L_U) is computed for a temporary MP tree. This value represents the temporary minimum number of substitutions which is likely to be slightly larger than the real minimum number of
substitutions ($L_M$). This value is determined by heuristic search called the stepwise addition or the branch-and-bound-like algorithm.

Consider an example of four taxa, taxa A, B, and C form the initial core tree (A1), and taxa D is added to this tree (Figure 21). There are three ways of adding D to the core tree (trees B1, B2 and B3). We first compute the tree length ($L$) for tree B1. If this $L$ is greater than $L_U$ we ignore this topology. However, if $L$ is equal to $L_U$ for this tree, we save the tree as potential MP tree and move onto the tree B2. If in case $L$ is smaller than $L_U$, then this tree will become the next temporary MP tree and $L_U$ is now replaced by this
new L value. We then move on to C. We apply the same procedure to the tree C. Once we get the tree with the smallest $L_U$, that tree is chosen as the MP tree. There may be two or more equally parsimonious trees. This algorithm is faster as not all trees are examined. However, even this method becomes time consuming when $m \geq 20$.

**Heuristic Search**

Several algorithms of the heuristic search for MP trees exists (Maddison and Maddison 1992; Swofford 2003a). The principle behind this algorithm is to first construct a provisional MP tree using *stepwise addition algorithm* and then subject this provisional tree to different *branch swapping* methods to find the most parsimonious tree.

**Stepwise addition algorithms**

In this set of algorithm, the initial core tree that is the provisional MP tree is obtained following the same procedure as mentioned above in the branch-and-bound search. This provisional tree usually has a longer tree length than that ($L_M$) of the MP tree. Therefore, this tree is subjected to branch swapping procedures to find a tree that has a smaller L value than the provisional tree. (Swofford 2003a) describe various ways of producing the provisional MP tree considering the order of taxon addition. (a) *As-is option*: The simplest option, in which the initial core is produced by the first three taxa given in the data set, and the following taxon addition is done according to the taxon order in the dataset. Usually this method is not very effective for finding a tree with a small L. (b) *Random option*: In this option, pseudorandom numbers are used to
determine the order of the taxon addition and this is repeated many times to obtain a provisional tree. (c) **Closet option:** Initial core tree is produced by examining all triplets of taxa and choosing the one that shows the smallest L value and in subsequent steps, a taxon whose addition to the previous core tree shows the smallest increase in L is chosen.

**Branch Swapping**

Once a provisional tree is produced, the tree is subjected following methods of branch swapping: 1) Nearest Neighbor Interchanges, NNI, 2) Subtree Pruning Regrafting, SPR, and 3) Tree Bisection-Reconnection, TBR(Swofford 1993), see Figure 22.

1) **Nearest Neighbor Interchanges (NNI)**

In this method of branch swapping, all the alternative topologies or trees that are different from the provisional MP tree by a topological distance of 2 are examined (Figure 22a). This algorithm is related to the close neighbor interchange (CNI) algorithm in relation to ME method.

2) **Subtree Pruning Regrafting (SPR)**

In the SPR algorithm, a branch of a provisional tree is cut into two parts, a pruned subtree and the residual tree (Figure 22b). The cutting point of the pruned subtree is than grafted onto each branch of the residual tree to produce a new
topology. This is done for all the branches of the residual tree to produce more trees to be examined.

3) *Tree Bisection Reconnection (TBR)*

In the TBR search, a provisional tree is cut into two subtrees at a branch and then uses two subtrees are then reconstructed by joining two branches, one from each subtree to generate a different topology (Figure 22c). This is done for all possible pairs of branches of the two subtrees that can generate different topologies. In case of SPR, however, only the cutting point of subtree is regrafted to a branch of the residual tree, whereas in the TBR all possible combinations of branches from the two subtrees are considered for reconnection. Therefore, the number of topologies examined in the TBR is larger than SBR, when the number of taxa is greater than five. Since the TBR search examines larger number of trees than the NNI or SBR, many studies use this method. However, even this method examines only a limited number of trees when the number of taxa is large (Maddison 1991).
(a) Nearest neighbor interchange (NNI)

(b) Subtree pruning and regrafting (SPR)

(c) Tree bisection and reconnection (TBR)

Figure 22. Three different methods of branch swapping for finding MP trees.
MP method tends to give incorrect topologies when the number of sequences used \((m)\) is large and the number of nucleotides used \((n)\) is small. (Nei, Kumar and Takahashi 1998) suggested that a relatively crude method of finding MP tree(s) gives essentially the same conclusion about phylogenetic inference as the exhaustive search when the accuracy of the tree obtained is examined by the bootstrap test. In fact, computer simulation (Takahashi and Nei 2000) has shown that for randomly generated model trees of 48 sequences with \(n = 1000\) the NNI search of MP trees is as efficient as the TBR search in inferring the true tree. This indicates that MP trees are often incorrect and that there is no need to spend an enormous amount of time for obtaining MP trees.

Table 16. Advantages and disadvantages of MP method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsimony</td>
<td>MP is a simple method and does not depend on model of evolution.</td>
<td>Parsimony analysis is statistically inconsistent. In parsimony analysis, it is difficult to treat the phylogenetic inference in a statistical framework because there is no way to compute the means and variances of minimum number of substitutions.</td>
</tr>
<tr>
<td></td>
<td>Gives more reliable trees than other methods when the extent of sequence divergence is low (Miyamoto and Cracraft 1991), rate of nucleotide substitution is more or less constant (molecular clock), number of nucleotides (informative sites) examined is large, and Homoplasy is either rare or widely (randomly) distributed on the tree.</td>
<td>With more data the certainty that parsimony will give the wrong tree increases - so that parsimony is statistically inconsistent. Advocates of parsimony initially responded by claiming that Felsenstein’s (1978) result showed only that his model was unrealistic. It is now recognised that the long-branch attraction (in the Felsenstein Zone) is one of the most serious problems in phylogenetic inference.</td>
</tr>
<tr>
<td></td>
<td>Consequence of parsimony analysis is that it gives us an insight into ancestral sequences. MP can easily take care of insertions and deletions of nucleotides, which may provide phylogenetic information.</td>
<td>Parsimony analysis generates an incorrect topology, if homoplasy (backward and parallel substitutions) is common or concentrated in particular parts of the tree.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phylogenetic reconstruction using parsimony is excellent when divergences are small. If the divergences are very small, it might even be difficult to fit a model due to lack of variation in the data. However, model-based methods such as ML (maximum likelihood) and Bayesian analysis offer advantages when divergences are large.</td>
</tr>
</tbody>
</table>

PAUP

PHYLIP

MEGA
Long Branch attraction

This is a well known problem in molecular phylogenetics, where unrelated long branches (or short branches) tend to join together or attract each other in the reconstructed tree (Figure 23). This phenomenon is often called long-branch attraction (Felsenstein 1985; Hendy and Penny 1989) or short-branch attraction (Nei 1996).

Consider the evolution of one site on the tree shown below (Figure 23). At the root of the tree, the character state is an A. Over the short branches, to short1 and short2, it remains an A. However, on the longer branches it has had time to be hit by mutations. The ancestral state A will be preserved in short1 and short2, but the character states will differ in taxa long1 and long2. Four different patterns might arise in the terminal taxa. Those patterns will be where character states in long1 and long2 are –

- Both the same as short1 and short2
- One the same and one different
- Both different and different from each other
- Both different but the same as each other
All the patterns possible are either uninformative or misinformative. A parsimony analysis will tend to group the long branches together, and tend to do so more if more data is added.

**Phylogenetic Congruence**

MP methods often produce several equally parsimonious trees. Congruence between trees can be investigated using consensus and agreement subtrees method.

**Consensus trees**

MP methods often produce several equally parsimonious trees. In this case, consensus tree summarizes the agreement between two or more parsimonious trees. A consensus tree is not an optimal tree instead they are derivative trees. There are different types of consensus trees (Swofford 2003a), but the most commonly used are the strict consensus trees and majority-rule consensus trees.
A Strict Consensus tree includes only those groups that are common to all the trees being considered. Consider the following two equally parsimonious trees (Figure 24a-c) obtained by an MP method. Here, the relation between taxa A, B & C is not common to all the trees and thus the strict consensus of these trees contains a polytomy as shown in Figure 24d.

![Consensus Trees](image)

**Figure 24. Examples of consensus trees.** The figure is obtained from Nei and Kumar 2000

A Majority Consensus tree contains all groups occurring in at least majority of the trees. In the case of two trees, the majority rule consensus tree is identical to the strict consensus tree but for three or more trees the resulting consensus is different. The most commonly used is the 50% majority-rule consensus tree (or semi-strict consensus tree). In this tree, a branching pattern that occurs with a frequency of 50% or more is adopted. Therefore, the 50% majority-rule consensus tree for the same set of MP trees is given by Figure 24e. If we use 70% majority rule, the tree is identical to strict consensus tree
(Figure 24f). The 100% majority-rule consensus tree is always identical with the strict consensus tree (Figure 24d).

**Agreement Subtrees**

A tree containing the largest subset of sequences for which the relationships among sequences are invariant across all the phylogenies is included (Swofford 2003a). Agreement subtrees are more resistant to ‘rogue sequences’ (one or a few sequences that are difficult to place on the tree); the presence of such sequences can make a consensus tree relatively unresolved, even when there is considerable agreement on the relationships between the other sequences (Figure 25).

![Agreement Subtrees Diagram](image)

**Figure 25. Illustration of greatest agreement subtrees method**
1.2.4.3. **Maximum Likelihood**

Felsenstein (1981) developed maximum likelihood algorithms for amino acid and nucleotide sequence data. The likelihood value of a phylogenetic tree is the probability of observing the data under a given tree and a specified substitution model. The aim of maximum likelihood method is to find the tree (from all possible trees) with the highest likelihood value. In this method the bases (nucleotides or amino acids) of all sequences at each site are considered separately (as independent), and the likelihood of having these bases are computed for a given topology. This likelihood is added for all sites and the sum of the log-likelihood is maximized to estimate the branch length of the tree. This procedure is repeated for all possible topologies and the topology that shows the highest likelihood is chosen as the final tree. This method takes into account various parameters of the evolutionary process, such as the relative probabilities of transitions versus transversions, or the degree to which rate of evolution differs across sites. These parameters are estimated in the tree evolutionary process from the data set, thus need not to be known.

Let us consider the following example, of four taxa given in Figure 26 and assume that DNA sequences are \( n \) nucleotides long and are aligned with no insertions/deletions. In this case, there are three different topologies, A, B, and C, as shown in Figure 26. The nucleotides at the first position of these sequences are all A (Figure 26), but we have to consider all the four possible nucleotides at each nodes 5 and 6.
Figure 26. Unrooted phylogenetic tree for four taxa to explain the maximum likelihood method of phylogenetic reconstruction. $v_i = r_i t_i$, where $r_i$ is the rate of nucleotide substitution and $t_i$ is the evolutionary time for branch $i$. Let $P_{ij}$ be the probability that nucleotide $i$ becomes nucleotide $j$ at a given site. Here $i$ and $j$ refers to any of the four nucleotides A, C, G, and T. In the likelihood method, rate of substitution ($r$) is allowed to vary from branch to branch and is expressed in terms of number of substitution differences ($v = rt$). Therefore, $v_i = r_i t_i$ denotes the expected number of substitutions for the $i$-th branch. This figure is obtained from Nei and Kumar (2000).

Therefore, if we assume node 5 as the starting point, the likelihood of having the observed nucleotides is obtained by the following equation.

$$L_1 = g_A P_{AA}(v_1) P_{AA}(v_2) P_{AA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$$
$$+ g_T P_{TA}(v_1) P_{TA}(v_2) P_{TA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$$
$$+ g_C P_{CA}(v_1) P_{CA}(v_2) P_{CA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$$
$$+ g_G P_{GA}(v_1) P_{GA}(v_2) P_{GA}(v_5) P_{AA}(v_3) P_{AA}(v_4)$$
$$+ ...$$

The total number of terms in this equation is 16, because both nodes 5 and 6 can take four different nucleotides. The second nucleotide site also has A for all sequences (Figure 26), so $L_2$ is the same as $L_1$. The third nucleotide site has G for all the sequences. Therefore $L_3$ is written as
\[ L_3 = \text{g}_A \, P_{AG}(v_1) \, P_{AG}(v_2) \, P_{AG}(v_5) \, P_{AG}(v_3) \, P_{AG}(v_4) \\
+ \ldots \\
+ \text{g}_G \, P_{GG}(v_1) \, P_{GG}(v_2) \, P_{GG}(v_5) \, P_{GG}(v_3) \, P_{GG}(v_4) \]

This equation also has 16 terms. This computation is done for all \( n \) number of nucleotide sites and for each site likelihood function (\( \ln L \)) is computed. The next step of the ML method is to assume nucleotide frequencies (\( g_A, g_C, g_G, \) and \( g_T \)) and maximize \( \ln L \) by varying \( v_1, v_2, \ldots, \) and \( v_5 \). The \( v_i \)'s that maximize \( \ln L \) are the ML estimates of branch lengths of tree shown in Figure 26. To determine the ML tree, the same computation is done for the other possible tree topologies. The tree with the highest likelihood (\( \ln L \)) value is the ML tree for the dataset.

**Search strategies for ML trees**

The search strategies for an ML tree is very time consuming and are very similar to those (e.g., NNI and TBR) used for obtaining ME or MP trees (see above). However, the efficiencies of these algorithms in obtaining the correct topology are not necessarily the same for the ME, MP, and ML methods (Nei, Kumar and Takahashi 1998). Recent computer simulation (Takahashi and Nei 2000) suggested that the stepwise addition + NNI search is usually as efficient as the more extensive TBR search in finding the true tree. Like ME, and MP methods, ML methods tend to give incorrect topology when \( m \) is large and \( n \) small (Nei 1996).
**Model for evolution for likelihood**

Recent computer programs such as PAML (Yang 1997; Yang 2007) and PAUP (Swofford 2003a) include various substitution models such as the Jukes-Cantor (1969), Kimura (1980) in addition to other complex models such as Hasegawa, Kishino, and Yano (1985) model, general reversible model (1994). The rate of nucleotide substitutions varies extensively from site to site and approximately follows the gamma distribution. (Yang 1993; Yang 1994a) incorporated this feature into the ML method of phylogenetic inference, and the computational algorithm taking into account this feature are included in the current phylogenetic software packages (Kumar, Tamura and Nei 1993; Felsenstein 1995; Guindon and Gascuel 2003; Swofford 2003a; Yang 2007).

When choosing an evolutionary model, one should not assume a model; instead find a model that best fits the data. We want to choose the model that gives the highest likelihood, but penalized by the number of model free parameters.

However, for obtaining the true topology, sophisticated models do not necessarily give better results than simple models such as the Jukes-Cantor model, though the likelihood value for an ML tree is almost always higher for the former model than the latter. Simulation studies have shown that under specified conditions a simple model gives a higher probability of obtaining the true tree than a complicated model, even if a sequence is evolved following the latter model.
Table 17. Advantages and disadvantages of likelihood method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum likelihood</td>
<td>ML methods are known to be consistent and powerful basis of statistical inference.</td>
<td>Require long computation time to construct a tree. Even if the number of sequences is about ten, it requires an enormous amount of computational time. This problem however, have been solved to certain extent in recent likelihood packages such as PhyML (Guindon and Gascuel 2003)</td>
<td>PAUP</td>
</tr>
<tr>
<td></td>
<td>The likelihood fully captures what the data tell us about the phylogeny under a given model.</td>
<td>ML estimation may be subject to systematic errors. This happen if the model of evolution used to evaluate the likelihood of given a tree does not reflect the actual evolutionary processes. In this context, the program ModelTest (Posada and Crandall 1998) can help remove some of the tedium of choosing a model.</td>
<td>PHYLIP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PAUP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PAML</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PhyML</td>
</tr>
</tbody>
</table>

1.2.4.4. Bayesian analysis

Phylogenetic analysis using a Bayesian approach has become very widely used in the few years since its introduction (Rannala and Yang 1996; Yang and Rannala 1997; Mau, Newton and Larget 1999). This method is closely related to maximum likelihood. The optimal hypothesis is the one that maximizes the posterior probability. The posterior probability for a hypothesis is proportional to the likelihood multiplied by the prior probability of that hypothesis. In a Bayesian analysis, inferences of phylogeny are based upon the posterior probabilities of phylogenetic trees. The posterior probabilities of the $i$-th phylogenetic tree (conditional on an alignment of DNA sequences (X)) can be calculated using Bayes theorem:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}.$$
where,

- $P(A)$ is the prior probability or marginal probability of $A$. It is prior in the sense that it does not take into account any information about $B$.
- $P(A|B)$ is the conditional probability of $A$, given $B$. It is also called the posterior probability because it is derived from or depends upon the specified value of $B$.
- $P(B|A)$ is the conditional probability of $B$ given $A$.
- $P(B)$ is the prior or marginal probability of $B$, and acts as a normalizing constant.

The same model of nucleotide substitution used in maximum likelihood analyses (Swofford et al. 1996) can be used in a Bayesian analysis of phylogeny. Bayesian inference uses complex models and has many parameters, and is too big to calculate analytically. The likelihoods and the prior probabilities at various points in the distribution are relatively easy to calculate, but the marginal likelihood of these multidimensional distribution problems becomes complex and intractable. Fortunately there are ways to approximate the posterior distribution that do not require calculation of the marginal likelihood, and the most common way uses a Markov chain Monte Carlo (MCMC) approach (Hastings 1970; Green 1995). This approach depends only on making posterior probability ratios, and so while the likelihoods and priors need to be calculated, the marginal likelihoods in the ratio cancel out, and need not be calculated.
Markov chain Monte Carlo (MCMC)

The posterior probabilities are approximated by the Markov chain Monte Carlo (MCMC) method using MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). The basic MCMC algorithm works as follows (Figure 27): first, a new state for the chain is proposed using a stochastic mechanism. Second, the acceptance probability for this new state is calculated. The acceptance probability is equal to minimum of one or the likelihood ratio times, the prior ratio times, the proposal ratio, where the likelihood ratio is the ratio of the likelihoods of the new state to the old state, the prior ratio is the ratio of the prior probability of the new state to the old state, and the proposal ratio is the ratio of the probability of proposing the old state to the probability of proposing the new state. Third a uniform (either 0 or 1) random variable is drawn. If the random number is less than the acceptance probability, then the new state is accepted and the state of the chain is updated. Otherwise the chain remains in the old state. The process of proposing and accepting/rejecting new states is repeated many thousand or million of times so that it builds up a picture of the most probable trees and parameters.

MCMC is a computational machine that takes samples from this posterior distribution. The more samples you let it take, the better it’s approximation. It is able to handle complex models and lots of parameters, to be more biologically realistic. The result of an MCMC is a large number of samples of the parameters and tree topologies, and that leaves us with the easily surmountable problem of how to digest and summarize those samples to extract some meaning. A bigger and more difficult problem is to find
out whether it has run well, and whether it has run long enough. We can interpret the results in a very direct way. For example, the highest posterior probability tree is the one that gets sampled most often.

Figure 27. Diagram showing Markov chain Monte Carlo (MCMC) robot rules.

MCMC convergence

The chain only works properly after it has converged or equilibrated. The plot of the number of generation versus the log probability of the data (the log likelihood values) provides a good indication of whether the chain has reached a plateau or not (Figure 28). This is a widely used but unreliable way of assessing convergence. Other potential indicators of convergence are potential scale reduction factor (PSRF), and averaged standard deviation of split support (ASDOSS).
MCMC burn-in

The first samples are often discarded as “burn-in”. In the convergence diagnostic shown below (Figure 28), convergence had been reached after 2000 generations, it makes sense to discard 25 % of the samples obtained during the 10,000 generations. Since we sampled every 10th generation, there are 1,000 samples (1,001 total) and 25 % translates to 250 samples. If we are at stationarity, this plot should look like ‘white noise’, that is, there should be no tendency of increase or decrease over time.

![MCMC likelihood plot](image)

**Figure 28. An example of MCMC likelihood plot**

**MCMCMC**

MrBayes introduced the variant of MCMC called Metropolis-coupled Markov Cahin Monte Carlo (MCMCMC) (Geyer 1991). (MC)$^3$ runs $n$ chains, $n-1$ of which are heated. A heated chain has the steady-state distribution. After all $n$ chains have done one step, a swap is attempted between two randomly chosen chains. If the swap is accepted,
then the two chains switch states. Inferences are based only on states sampled by the cold chain. The heated chains can more easily explore space of phylogenetic trees; the effect of heating is to lower peaks and to fill in valleys. The cold chain can effectively leap across deep valleys in the landscape of trees when a successful swap is made between the cold chain (local optimum of trees) and a heated chain that is exploring another peak.

**Bayesian and ML inference-Differences**

Bayesian approach differs from the ML method in the following ways: One of the main ways that the Bayesian approach differs from ML is that the Bayesian approach deals with uncertainty in a more explicit way. For example, nuisance parameters such as branch lengths and model parameters will each have some uncertainty associated with them. The ML approach is to find the ML values for all of those nuisance parameters, but the Bayesian approach is to retain that uncertainty in the result. This means that the result of a Bayesian analysis is usually not a single point, but rather is a probability density or distribution. Being a distribution might mean that it is awkward to summarize the result to pick out the message that you want to get from it, but even so it can be considered an advantage over ML. While in ML a picture of the uncertainty involved is often done after the analysis, often laboriously, as in for example the bootstrap, in a Bayesian analysis the uncertainty is part of the initial result. Another difference is that Bayesian analysis explicitly relies on prior probabilities. The prior probability might be known with confidence, but for many problems we have only a vague idea of what the prior
probabilities are, and since a Bayesian analysis forces us to be explicit about them we may have to make something up. This can be considered both strength and a weakness, and is certainly controversial. Of course the implementation details differ between Bayesian and ML analysis. ML uses hill-climbing algorithms to get the result to any required precision, while Bayesian analyses generally require an approximating algorithm such as the MCMC. From a computational point of view, the Bayesian MCMC can handle more parameters than ML, which means that you can solve bigger problems with more realistic models. While ML expresses itself in terms of the probability of the data given the model, the Bayesian approach expresses the result as the probability of the model given the data. The probability of the model, of hypothesis, is more likely what the investigator wants, and this directness is one of the main attractions of Bayesian analysis.

Table 18. Advantages and disadvantages of Bayesian analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>It has an acceptable speed, and can handle big datasets with parameter rich-models (see Larget and Simon, 1999). Has a strong connection to the maximum likelihood method; might be a faster way to assess support for trees than maximum likelihood bootstrapping.</td>
<td>The prior distributions for parameters must be specified. In some circumstances, the prior probability can influence the results. It can be difficult to determine whether the Markov chain Monte Carlo (MCMC) has run for long enough (or have reached convergence).</td>
<td>MrBayes</td>
</tr>
<tr>
<td>BAMBE</td>
<td>Reliability of Phylogenetic methods</td>
<td></td>
<td>BAMBE</td>
</tr>
</tbody>
</table>

Reliability of Phylogenetic methods

Phylogenetic methods (e.g. distance, parsimony, likelihood) can be evaluated in terms of their general performance, particularly their (1) Computational speed, (2) consistency, (3) statistical tests of phylogenetic trees, (4) probability of obtaining the true
topology, and (5) reliability of branch length estimates. Studies of these properties can be analytical or by simulation or based on known phylogenies.

In terms of computational speed, UPGMA and NJ are superior to most other inference methods currently available. These methods can handle a very large number of sequences \(m>500\). The NJ, ME, and LS methods are all consistent estimators if unbiased estimates of nucleotide substitutions are used as distant measures. When the number of sequences examined \(m\) is small, \(m<5\), it is possible to evaluate probability of obtaining the correct topology for the distance and parsimony methods (Nei and Kumar 2000). Studies have shown that when the evolutionary rate is more or less constant for all four or five sequences, NJ gives slightly better accuracy than MP (Fitch and Margoliash 1967). However, analytical studies have shown that it is very difficult when \(m\) is large, and thus the conclusion based on these studies may not apply to a wide variety of situations. For this reason, comparison of accuracy of obtaining the correct topology among different methods is usually done by computer simulation. In computer simulation probability of obtaining a correct tree can be estimated for a variety of evolutionary conditions. Numerous simulation studies have been performed using different evolutionary models, inference methods, and computer programs, most of them have been reviewed by Nei 1991, 1996, Nei and Kumar 2000.

There have been many arguments that ML methods are best because they have desirable statistical properties, such as consistency (Nei and Kumar 2000). However, ML does not always have these properties, for example, if the model is wrong/inadequate (fortunately this is testable to some extent). Simulations studies have shown that ML
methods generally outperform distance and parsimony methods over a broad range of realistic conditions (Whelan, Lio and Goldman 2001). But, most simulations cover a narrow range of very (unrealistically) simple conditions such as fewer number of taxa (typically four), and few parameters (models such as JC, K2P) (Felsenstein 1981). Some other simulations with more realistic models of larger number of sequences (Saitou and Nei 1986) have shown that parsimony is often less efficient than NJ or the methods of minimum evolution and maximum likelihood, though it is a good method under certain circumstances.

All phylogenetic methods make assumptions, whether explicit or implicit, about the process of nucleotide substitution (Felsenstein 1988). Consequently, all the methods (except maximum parsimony) of phylogenetic inference depend on their underlying models. Simulation studies have shown that the shape of the underlying true tree has an enormous impact on the importance of model choice. In the ideal case (Figure 29), the underlying tree shape is such that all existing methods estimate phylogeny accurately; Maximum likelihood estimation is very robust to violations of model assumptions, and model choice is not critical (e.g., (Sullivan and Swofford 2001). However, model choice is critical in the Felsenstein zone (Figure 29) when branch lengths or rates are highly unequal, and that observation is widely accepted (Felsenstein 1981). Weighting can improve the performance of parsimony (i.e. reduce the size of the Felsenstein zone).
Figure 29. The effect of topology on robustness. At the center of the continuum, phylogenetics signal is strong and model choice is not critical (i.e., maximum likelihood is robust to violations of model assumptions). In the Felsenstein zone (left), model selection is critical, as is also the case for the inverse Felsenstein zone (right) (Sullivan and Joyce 2005).

Although, it is generally difficult to know the true topology in real data analysis, there are a few cases where phylogenetic trees are produced experimentally in the laboratory with T7 phages and used to test methods of phylogenetic analysis (Hillis et al. 1992). Hillis (1992) concluded that when realistic conditions are considered, the maximum parsimony method is generally superior to other methods such as the Neighbor-Joining (NJ) method and UPGMA in the recovery of a true phylogeny. However, there are cases where phylogenetic tree for a group of organisms is established based on the paleontological and morphological bases. For example, phylogenetic tree reconstructed based on the 13 protein coding mitochondrial genes from the 11 vertebrate species (Russo, Takezaki and Nei 1996) were compared with the true tree. In this study, few genes produced the correct topology, and some genes produced incorrect trees, regardless of the method and algorithm used. These results showed that some genes are
more suitable than others in phylogenetic inference and that all the tree building methods
tend to produce the same topology whether the topology is correct or not.

Both simulation and experimental studies indicate that many methods of
phylogenetic analysis are powerful enough to reconstruct evolutionary histories with a
high degree of accuracy, as long as the rate of nucleotide substitutions are appropriate for
analysis. This emphasizes the importance of methods that evaluate whether rates of
evolutionary change in target sequences are appropriate for phylogenetic analysis (Hillis
and Huelsenbeck 1992). Experimental phylogenies also indicate that many methods may
be fairly robust to violations of the underlying assumptions, such as non-independence
among nucleotide sites or deviations from simple models of evolution.

1.2.5. Assessing Phylogenetic Accuracy

When a phylogenetic tree is reconstructed, it is important to know the reliability
of tree obtained. There are two types of errors in a phylogenetic tree: topological and
branch length errors. The topological differences are differences in branching pattern
between an inferred tree and the true tree, and the latter are deviations of estimated
branch lengths from the true branch lengths. If the true topology is known as in the case
of computer simulations, it is possible to evaluate the extent of topological errors by the
topological distance method. However, if the true topology is not known (which is
almost always the case), the reliability of the topology obtained is usually tested by
examining the statistical confidence of each branching pattern of the topology (e.g.,
bootstrap tests). The reliability of branch length estimates can be tested by either
analytical (Nei, Stephens and Saitou 1985; Li 1989; Dopazo 1994) or the bootstrap (Felsenstein 1985).

1.2.5.1. **Topological Distances**

*Topological distances* are the topological differences between the reconstructed tree and the possible closely related alternative trees. The topological distance is the most widely used method to measure tree distance (Robinson and Foulds 1981; Penny and Hendy 1985). The distance that is computed is also known as symmetric distance (Robinson and Foulds 1981) or partition matrix (Penny and Hendy 1985). This distance measure does not use branch length information, only the tree topologies.

For unrooted bifurcating trees, this distance is twice the number of internal branches at which sequence partition results in difference between the two trees compared. In general, an unrooted bifurcating tree for \( m \) taxa has \( (m-3) \) possible internal branches and a topological distance \( (d_T) \) of \( 2(m-3) \). For example, consider the two unrooted trees for seven taxa with four internal branches in Figure 30.

![Figure 30. Two unrooted trees for seven taxa.](image)
The above trees can be divided into two groups based on the cut at any of the possible four internal branches. Here, cutting at some internal branches results in the same partition of taxa in both the trees. For example, cutting at branch 1 yields two same subgroups (A, B) & (C, D, E, F, G) in both trees. However, cut at branch 3 produces different partitions in the two trees. Therefore, the topological distance between these two trees is given by:

\[ d_T = 2 \times 1 = 2 \]

To compute the distances \( d_T \) in case of multifurcating trees, we use (Rzhetsky and Nei 1992) formula as follows:

\[ d_T = 2[\text{Min}(q_1, q_2) - p] + |q_1 - q_2| \]

where,

\[ q_1, q_2 = \text{possible number of interior branches in both trees compared.} \]
\[ p = \text{number of partitions that are same for both trees.} \]

In general, the trees that are different from the reconstructed tree with \( d_T = 2 \) and 4 are examined for the most likely trees.

Some other methods to compute tree distance are the Branch Score Distance (Kuhner and Felsenstein 1994) method, CompareTrees method (Hall 2004a). These method uses branch lengths, and can only be calculated when the trees have lengths on all branches. CompareTrees computes three scores for each comparison: a topology score, a branch length score, and a tree score. The topology score is the number of internal branches that are present in both the model tree and the inferred tree. The branch length score is calculated as follows; first the absolute difference in branch lengths for
each branch that is present in both the model tree and the inferred tree is computed. This number is then subtracted from 1 to produce the score for that branch. Later all branch scores are averaged to produce the branch length score for the tree. The product of the topology score and the branch length score is the tree score.

1.2.5.2. *Assessing Tree Quality*

Several methods have been proposed that attach numerical values to internal branches in trees that are intended to provide some measure of the strength of support for those branches and the corresponding groups. These methods include:

- character sampling methods (bootstrap and jackknife)
- comparisons with suboptimal trees- decay analyses
- comparison of phylogenetic hypothesis (Tests of two trees)

**Bootstrapping**

One of the most commonly used tests of the reliability of an inferred tree is bootstrap test (Felsenstein 1985). In this test, reliability of the inferred tree is evaluated by using (Efron 1982) bootstrap re-sampling technique (shown in Figure 31). Bootstrapping essentially tests whether the given dataset supports the given tree. This is done by taking $n$ random subsamples (or columns) of the dataset with replacement from the original set of sequences, building trees from each of these and calculating the frequency with which the branching pattern of the given tree are reproduced in each of these random subsamples. If a group is found in every subsample tree, then its bootstrap
support is 100%, if it is found in only two-thirds of the subsample trees, its bootstrap support is 67% (Figure 31). This number is known as the bootstrap confidence value or the bootstrap value.

A high bootstrap value (e.g., > 85%) is indicative of strong phylogenetic signal in the data. On the other hand, a low bootstrap value need not mean the relationship is false, only that it is poorly supported. Bootstrap analyses of known phylogenies show that it is generally dependable measure of phylogenetic accuracy, and that value of 70% or higher are likely to indicate reliable groupings (Hillis and Bull 1993). For more discussion of the interpretation of bootstrap proportions, see (Nei and Kumar 2000; Baldauf 2003; Soltis and Soltis 2003)
**Figure 31. Bootstrap analysis.** The original data set of three taxa (A-C) each with 9 nucleotide characters is bootstrapped across characters with replacement to produce bootstrap pseudo-replicates. Each pseudo-replicate contains each of the three original taxa, but some original characters are represented more than once and some not at all. Each bootstrap dataset is then analyzed phylogenetically. The trees from all three datasets are summarized in a consensus tree with 67% bootstrap support.

*Limitations of the Bootstrap*

The usefulness of bootstrap values for assessing even the relative confidence in clades is limited by the application of the bootstrapping procedure to topologies rather than single variables (i.e., nodes), the effects of varying numbers of characters on bootstrap values, statistical bias with increased taxon sampling and the underlying assumptions and properties of the data and phylogeny reconstruction algorithms (e.g., (Harshman 1994; Farris *et al.* 1996; Sanderson and Wojciechowski 2000)

*Parametric Bootstrapping*

An alternative to the standard, nonparametric bootstrapping for testing specific hypothesis of relationships is the parametric bootstrap (Efron 1985), in which a single dataset can be used to parameterize a model of sequence evolution. This model is then used to simulate new, independent data sets, each of which is analyzed in turn to generate a distribution against which a specific hypothesis can be tested (Bull *et al.* 1993). Likelihood ratio tests with parametric bootstrapping allow for tests for specific
hypothesis that cannot be addressed via nonparametric bootstrapping (Huelsenbeck and Crandall 1997).

**Jackknifing**

Jackknifing is very similar to bootstrapping and differs only in the character re-sampling strategy. Jackknifing re-sampling, in which either characters or taxa are re-sampled without replacement (Miller 1974; Efron 1982; Efron and Gong 1983). Some proportions of characters (e.g. 50%) are randomly selected and deleted. Replicate datasets are then analyzed and the results are summarized with a majority consensus tree. Jackknifing and bootstrapping tend to produce broadly similar results and have similar interpretations (Felsenstein 1985). PAUP* (Swofford 2003a) also has a jackknife option that can be used with parsimony, maximum likelihood or distance based phylogeny reconstruction.

**Decay analysis**

In parsimony analysis, a way to assess support for a group is to see if the group occurs in slightly less parsimonious trees also. The length difference between the shortest trees including the group and the shortest trees that exclude the group (the extra steps required to overturn a group) is the decay index or Bremer support (Bremer 1988; Bremer 1994; Morgan 1997; Giribet 2003). Decay indices (DI) for each clade can be determined by, (1) Saving increasingly less parsimonious trees and producing corresponding strict consensus trees until the consensus is completely unresolved, (2)
analyses using reverse topological constraints to determine shortest trees that lack each clade, (3) with the Autodecay or TreeRot programs (Sorenson 1999; Swofford 2003a) ((in conjunction with PAUP*). Decay analysis can be extended to any optimality criterion and to other relationships.

Generally, the higher the decay index the better the relative support for a group. Like bootstrap values decay indices may be misleading if the data is misleading. Unlike bootstrap proportions decay indices are not scaled (0-100) and it is less clear what an acceptable decay index is. The magnitude of decay indices and bootstrap values generally are correlated (i.e. they tend to agree). Only groups found in all most parsimonious trees have decay indices > zero.

Tests of two trees (or more)

These are the tests that allow us to determine if one tree is statistically significantly better than another. Tests are of the null hypothesis that the differences between two trees (A and B) are no greater than expected from sampling error. Some of the commonly used methods to evaluate alternative phylogenetic hypotheses are Winning sites (Prager and Wilson 1988), Templeton (Templeton 1983), Kishino-Hasegawa (Kishino and Hasegawa 1989), Shimodaira-Hasegawa (Shimodaira and Hasegawa 1999), t-test (Swofford et al. 1996) etc.
Winning sites

The simplest wining sites test sums the number of sites supporting tree A over tree B and vice versa (those having fewer steps on, and better fit to, one of the trees). Under the null hypothesis characters are equally likely to support tree A or tree B and a binomial distribution gives the probability of the observed difference in numbers of winning sites.

Templeton test

Templeton’s test is a non-parametric Wilcoxon signed ranks test of the differences in fits of characters to two trees. It is like the ‘winning sites’ test but also takes into account the magnitudes of differences in the support of characters for the two trees. Recent studies of the relationships of turtles using morphological data have produced very different results with turtles grouping either within the parareptiles (H1) or within the diapsids (H2) the result depending on the morphological data. Parsimony analysis of the most recent data favoured H2. However, analyses constrained by H2 produced trees that required only 3 extra steps (<1% tree length). The Templeton test was used to evaluate the trees and showed that the slightly longer H1 tree found in the constrained analyses was not significantly worse than the unconstrained H2 tree.

Kishino-Hasegawa test

The Kishino-Hasegawa (KH) test is similar in using differences in the support provided by individual sites for two trees to determine if the overall differences between
the trees are significantly greater than expected from random sampling error (Kishino and Hasegawa 1989). It is a parametric test that depends on assumptions that the characters are independent and identically distributed (the same assumptions underlying the statistical interpretation of bootstrapping). It can be used with parsimony and maximum likelihood - implemented in PHYLIP (Felsenstein 1995) and PAUP* (Swofford 2003a).

![Figure 32. Distribution of step/likelihood differences at each site](image)

Under the null hypothesis the mean of the differences in parsimony steps or likelihoods for each site is expected to be zero, and the distribution normal (Figure 32). From observed differences we calculate a standard deviation. If the difference between trees (tree lengths or likelihoods) is attributable to sampling error, then characters will randomly support tree A or B and the total difference will be close to zero. The observed difference is significantly greater than zero if it is greater than 1.95 standard deviations. This allows us to reject the null hypothesis and declare the sub-optimal tree significantly worse than the optimal tree (p < 0.05).
Problem with tests of two trees

To be statistically valid, the Kishino-Hasegawa test should be of trees that are selected \textit{a priori}. However, most applications have used trees selected \textit{a posteriori} on the basis of the phylogenetic analysis. When we test the 'best' tree against some other tree, the KH test will be biased towards rejection of the null hypothesis. Only if null hypothesis is not rejected will result be safe from some unknown level of bias. Comparatively, the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999) is a more statistically correct technique for testing trees selected \textit{a posteriori} and is implemented in PAUP*. However it requires selection of a set of plausible topologies. The Approximately Unbiased test (implemented in CONSEL (Shimodaira and Hasegawa 2001) may be the best option under specified conditions.
CHAPTER 2.1

PHYLOGENETIC INFERENCE UNDER VARYING
PROPORTIONS OF INDEL-INDUCED ALIGNMENT GAPS

This chapter is submitted to the *BMC Evolutionary Biology* journal for publication as research article.
Phylogenetic inference under varying proportions of indel-induced alignment gaps

Bhakti Dwivedi¹ and Sudhindra R Gadagkar¹*

¹Department of Biology, University of Dayton, 300 College Park, Dayton, OH 46469-2320, USA

*Corresponding author

Email addresses:

Bhakti Dwivedi: dwivedbz@notes.udayton.edu
Sudhindra R. Gadagkar: gadagkar@notes.udayton.edu
Abstract

Background

The effect of gaps in sequence alignments on phylogenetic accuracy has been the subject of many studies. In this study, we have investigated the relationship between the number of alignment gaps and phylogenetic accuracy, when the gaps are introduced by computer simulation to reflect indel (insertion/deletion) events during the evolution of DNA sequences, and thus carry phylogenetic signal. The resulting (true) alignments were subjected to commonly used gap treatment and phylogenetic inference methods.

Results

(1) In general, there was a strong – almost deterministic – relationship between the amount of gap in the data and the level of phylogenetic accuracy when the alignments were very “gappy”, (2) gaps resulting from deletions (as opposed to insertions) contributed more to the inaccuracy of phylogenetic inference, (3) the probabilistic methods (Bayesian, PhyML & ”MLε” a method implemented in DNAML in PHYLIP) performed better at most levels of gap percentage when compared to parsimony (MP) and distance (NJ) methods, with Bayesian analysis being clearly the best, (4) methods that treat gapped sites as missing data yielded less accurate trees when compared to those that
attribute phylogenetic signal to the gapped sites (by coding them as binary character data – presence/absence, or in the MLε method), and (5) in general, the accuracy of phylogenetic inference depends upon the amount of available data when the gaps result from mainly deletion events, and the amount of missing data when insertion events are equally likely to have caused the alignment gaps.

**Conclusions**

When gaps in an alignment are a consequence of indel events in the evolution of the sequences, the accuracy of phylogenetic analysis is likely to improve if: (1) alignment gaps are categorized as arising from insertion events or deletion events and then treated separately in the analysis, (2) the evolutionary signal provided by indels is harnessed in the phylogenetic analysis, and (3) methods that utilize the phylogenetic signal in indels are developed for distance methods too. When the true homology is known and the amount of gaps is 20 percent of the alignment length or less, the methods used in this study are likely to yield trees with 90-100 percent accuracy.
Background

DNA sequences are used routinely to infer phylogenies (Nei and Kumar 2000; Felsenstein 2004b; Hall 2004b). The sequences within lineages (branches of the phylogenetic tree) evolve independently over time by means of several evolutionary processes, including point replacements of nucleotides (base substitutions), and insertion and deletion (indel) events. While base substitutions change the nucleotide composition of a given sequence, indels are likely to change the total length of the sequence. If indel events have occurred during the course of evolution of the molecular sequences being studied, it becomes necessary to align the corresponding homologous regions among the sequences for a proper site-by-site comparison among them, before phylogenetic analysis. In the process of alignment, gaps are introduced in the sequences to account for the indels. Different methods have been devised for dealing with gapped sites during phylogenetic analysis, ranging from ignoring the gapped sites from the alignment to inferring or differentially coding the state at each gapped site, using a number of different methods (for a list of methods, see Swofford 2003, Ogden and Rosenberg 2007, Simmons, Muller and Norton 2007). Most of these treatment methods work reasonably well when the proportion of gapped sites in an alignment is small (Ogden and Rosenberg 2007; Simmons, Muller and Norton 2007).

There are many examples in the literature of studies that have used molecular sequences (DNA and protein) with rather large gaps to infer phylogenies (Raymond et al. 2003; Lee and Wen 2004; Egan and Crandall 2008). It appears logical to expect an inverse relationship between the proportion of gapped sites in an alignment and the
accuracy of the inferred phylogeny, particularly if the gaps are not treated as reflective of distinct evolutionary events, and thus, containing distinct phylogenetic signal. However, the relationship between the extent of “gappiness” in the data resulting from indel events in the evolutionary history of the sequences on the one hand, and phylogenetic accuracy on the other, has not been studied by introducing and systematically varying the number of gaps in the alignments in a biologically realistic manner, even as the literature on alignment gaps in the phylogenetic context has increased of late (Wiens 2003; Driskell and Christidis 2004; Philippe et al. 2004; Rivas 2005; Ogden and Rosenberg 2007; Hartmann and Vision 2008; Wiens and Moen 2008). For example, several studies investigating the relationship between the amount of alignment gap and phylogenetic accuracy have done so in the context of aligning sequence fragments such as ESTs (e.g., (Philippe et al. 2004; Hartmann and Vision 2008)), using computer simulation to first generate the alignments and then introduce gaps, such that the gaps do not contain any phylogenetic signal (e.g., (Wiens 2003; Wiens and Moen 2008)); are in the context of only empirical data (e.g., (Egan and Crandall 2008)); or where the emphasis was more on levels of divergence among the taxa (e.g., (Cantarel, Morrison and Pearson 2006)). Furthermore, the relative performance of the gap treatment methods that are common among inference methods has also not been compared in this context. For example, all inference methods allow gaps to be treated as missing data or “MD” (although the treatment of the missing data differs among the methods, with the state at the gapped sites inferred in parsimony and distance-based methods of phylogenetic analysis, based on criteria that are specific to each method, while in likelihood and Bayesian analyses, the
likelihoods are summed over all four possible assignments of a nucleotide to a given gapped site). It is not known how the data inferred under these criteria work in conjunction with each of the respective inference methods to influence the accuracy of phylogenetic inference, when the gaps reflect indel events in the alignment.

We obtained sequence alignments for this study by means of simulating non-coding DNA evolution, introducing nucleotide point substitutions (replacements) and insertion/deletion (indel) events along a balanced (symmetrical) 16-taxon model tree (Fig. 1). (Simulations were done along random and pectinate 16-taxa trees as well, but we report the results only from the balanced model tree shown in Fig. 1 for reasons explained below.) The simulations were done while systematically varying the values of different sequence and indel parameters. All the simulation parameters were varied to include biologically realistic values. For example, the rate of introduction of indels included the range observed in non-coding sequences (Parsch 2003; Zhang and Gerstein 2003; Yamane, Yano and Kawahara 2006; Matthee et al. 2007). Similarly, the ratio of insertion to deletion events was also varied based on published results (Saitou and Ueda 1994; Zhang and Gerstein 2003; Chen et al. 2007; Matthee et al. 2007). It was important to vary the ratio of insertions to deletions in order to determine if there was a differential effect on phylogenetic accuracy, since most of the commonly used gap treatment methods do not differentiate between gaps resulting from the two types of evolutionary events.

We assessed the accuracy of phylogenetic inference as the topological correctness of the inferred tree when compared to the model tree. Our results show that overall,
when the percentage of gapped sites (cells in the alignment matrix) in the alignment is low (≤ 20 percent), all the inference methods (using any gap-treatment method) perform well (with 90-100% accuracy). On the other hand, when the number of gapped sites increases in the alignment, the probabilistic methods (particularly Bayesian analysis) are clearly more accurate, although at the highest gap levels, NJ and MP are sometimes better. Our results also show that gaps resulting from deletion events in the evolutionary history of the sequences appear to be harder to reconcile (when compared to those resulting from insertion events), leading to greater inaccuracies in phylogenetic inference, evidently because of the loss of the phylogenetic signal present in the sites deleted. When compared to MD method of treating gaps, a much higher accuracy was seen in our study when the gaps were coded separately as in the BC (Binary Character state) treatment in conjunction with the Bayesian and MP methods, or as in the DNAMLɛ package (Rivas and Eddy 2008).

**Methods**

**Computer simulations**

True DNA sequence alignments were generated by simulating evolution along a 16-taxon model tree using the computer program, Dawg, version 1.2 (Cartwright 2005). The model tree topology used for the simulations was a balanced, non-ultrametric tree with random branch lengths (Fig. 1), borrowed from Ogden and Rosenberg 2006. The nucleotide substitution model used was HKY (Hasegawa, Kishino and Yano 1985), with
rate heterogeneity among sites. Simulations were done to mimic nucleotide sequences with different properties by systematically varying the sequence and indel parameters (in a fully factorial manner, see below). Simulations were also done with 16-taxon random-branching and pectinate trees, and these alignments were also subjected to the same analyses that the alignments from the balanced tree of Figure 1 were. The results of these analyses showed that while the alignments from the random-branching tree were essentially the same as that from balanced tree, those from the pectinate tree were not. These differences have been pointed out at appropriate parts in the text, while presenting the results only from the balanced tree of Figure 1.

Figure 1 – The model tree. The 16-taxon balanced tree, obtained from Ogden and Rosenberg (2006), was used as a model tree for the simulations of DNA evolution, the results of which we
have presented in this paper. Two more sets of simulations were also done – one with a random-branching tree and the other with a pectinate tree (both 16-taxon too), but the results from these analyses have not been shown, for reasons explained in the text. The scale given in the figure refers to the number of nucleotide substitutions per site. During simulation, the total number of substitutions to be made for a given branch, for a given parameter-combination, was obtained as the product of the branch length in the model tree, the rate multiplier and the sequence length. Values from the latter two parameters were obtained from Table 1.

The values of the sequence and indel parameters used in the simulations are given in Table 1. These values were varied based on several studies (e.g., (Yang 1996a; Yang 1998; Zhang and Gerstein 2003; Matthee et al. 2007)) to ensure the generality of the conclusions from this study. The sequence parameters varied were: initial sequence length ($l$), transition to transversion rate ratio ($\kappa$), and the rate of nucleotide substitution ($r$) as the number of substitutions per site. The shape parameter ($\alpha$) of the gamma distribution was set to 0.5 to specify the extent of rate heterogeneity among sites. The nucleotide base frequencies were kept constant (A% = T% = 30%; G% = C% = 20%) throughout the simulations, which were based on the literature (Zhang and Gerstein 2003; Yamane, Yano and Kawahara 2006). All other options in the program pertaining to the sequences were set to default during simulation. Note also that the results, when compared between $l = 500$ and $l = 2500$, produced similar patterns (except for an increase in accuracy), as also when compared between $\kappa = 2.0$ and $\kappa = 5.0$. Therefore, results have been presented in this paper for only $l = 500$ and $\kappa = 2.0$. 

134
Table 1: Sequence and indel parameter values used in the generation of gapped sequence alignments by computer simulation. The sequence length, \( l \), is measured as the number of nucleotides. The indel rate, \( \lambda \), refers to the number of indel events per nucleotide substitution, and is expressed as a proportion (for example, a \( \lambda \) value of 0.03 indicates that there were three indel events for every 100 substitutions), \( r \) is a multiplier, that, when multiplied by a given branch length in the model tree and the sequence length, yields the number of substitutions to be introduced in that branch during simulation.

<table>
<thead>
<tr>
<th>Sequence Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial sequence length (( l ))</td>
<td>500, 2500</td>
</tr>
<tr>
<td>Transition-transversion rate ratio (( c )):</td>
<td>2.5</td>
</tr>
<tr>
<td>Gamma distr. shape parameter (( a )):</td>
<td>0.5</td>
</tr>
<tr>
<td>Nucleotide subst. rate multiplier (( r )):</td>
<td>0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.0, 1.25, 1.5, 1.75, 2.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indel Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indel rate (( \lambda )):</td>
<td>0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.25</td>
</tr>
<tr>
<td>Insertion/Deletion rate ratio:</td>
<td>1:1, 1:3</td>
</tr>
</tbody>
</table>

The rate at which indels were introduced during simulation, \( \lambda \), was varied as a function of the substitution rate (Table 1). For example, a \( \lambda \) of 0.03 refers to an average of 3 indels per 100 substitutions. \( \lambda \) was also varied to include the range typically observed in empirical sequence data (Parsch 2003; Zhang and Gerstein 2003; Yamane, Yano and Kawahara 2006; Matthee et al. 2007). In addition, in order to mimic the very large number of gaps that can potentially be seen in introns and other non-coding sequences, a few higher indel rates were also added (with corresponding increased substitution rates). Although phylogenetic analysis is typically done without
differentiating between insertions and deletions, the insertion to deletion ratio was set to either 1:3 (Zhang and Gerstein 2003; Matthee et al. 2007) or 1:1 (Saitou and Ueda 1994; Schaeffer 2002; Chen et al. 2007), at a given indel rate, in order to accommodate differing opinions about the ratio of deletions and insertions, and to determine if the two have different impacts on phylogenetic accuracy. The size distribution of insertions/deletions was as per mammalian pseudogene data (Zhang and Gerstein 2003) and ranged from 1 to 60bp in length. This distribution of the indel length can be observed in other non-coding sequences, such as chloroplast inter-genic regions (Yamane, Yano and Kawahara 2006), and nuclear DNA sequences of primates (Saitou and Ueda 1994). Each set of sequence and indel parameters (44 sets and 20 sets, respectively) was replicated 100 times, thus producing 88,000 16-taxon non-coding sequence alignments.

**Phylogenetic analysis**

Phylogenetic analysis was done on the alignments obtained using Neighbor-Joining (NJ) and Maximum Parsimony (MP) methods as implemented in PAUP* version 4.0 b10 (Swofford 2003a). Maximum Likelihood analyses (PhyML) was done using the program, PhyML version 2.4.4 (Guindon and Gascuel 2003), because of its speed (Guindon and Gascuel 2003). Finally, Bayesian analysis was done using MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Altekar et al. 2004) with default settings.
Maximum Likelihood HKY pair-wise distances were used for the NJ analyses. In PhyML analysis, the initial tree was built using BIONJ (Guindon and Gascuel 2003). The parameters of the HKY substitution model (the four base frequencies and the transition/transversion rate ratio) along with the proportion of invariable sites and the gamma distribution shape parameter were estimated from the simulated data for both NJ and PhyML analysis. For the MP analysis, a heuristic search was done using the stepwise addition algorithm for the provisional tree and subsequent branch swapping using the Nearest-Neighbor Interchange (NNI) method. (NNI results for MP are known to be as good as those from the more thorough – and time-consuming – Tree Bisection Reconnection (TBR) searches (Takahashi and Nei 2000; Piontkivska 2004). In addition, our TBR and NNI results for a representative subset of the simulations yielded essentially the same results). All other settings were set to default. Similarly, results analyzed from the PhyML version 2.4.4 (Guindon and Gascuel 2003) using NNI were not different from the recent PhyML version 3.0 (Guindon and Gascuel 2003) with SPR (Subtree Pruning and Regrafting) tree search.

For the Bayesian analysis, the nucleotide substitution model used was HKY with invariant sites and rate heterogeneity of rates across sites. The number of generations was set to 50,000 with a sampling frequency of 50. In cases when convergence was not obtained (typically for high substitution and indel rates), the number of generations was increased to 100,000 with a sampling frequency of 100. Burn-in was set to 25 percent of the generations and the inferred tree was estimated as the consensus of all compatible
groups of the post burn-in trees. The inferred tree was then compared to the model tree and topological distances were measured using PAUP* version 4.0b10 (Swofford 2003a).

**Treatment of gaps**

Gapped sites in our (true) alignments were subjected to the following gap treatment methods during phylogenetic analysis: a) “MD” (Missing Data) – in this treatment the nucleotide state at each gapped site is treated as a missing character based on the optimization criteria, based on whether the inference method is distance-based, parsimony, likelihood, or Bayesian (Huelsenbeck and Ronquist 2001; Guindon and Gascuel 2003; Swofford 2003a). Treating gaps as unknown or missing data is the default option in PAUP* The FAQ page for PAUP* at http://paup.csit.fsu.edu/paupfaq/paupfaq.html explains the working of this treatment under each inference method, and for PhyML and Bayesian, it is explained in Huelsenbeck and Ronquist 2001, Guindon and Gascuel 2003, Ronquist and Huelsenbeck 2003. Briefly, PAUP* deals with missing characters in the following manner: under the parsimony criterion, a missing character in a sequence is assigned the most parsimonious state given its placement in the tree. Under the likelihood criterion, a gapped site is assigned a state based on the likelihood which is computed by summing the likelihoods over all possible states – a strategy that is used by PhyML as well. For distance methods, PAUP* deals with the missing data by distributing the missing or ambiguous changes proportionally to each unambiguous change. Bayesian analysis was done using the program MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003;
Altekar et al. 2004), which deals with gaps just as other Maximum Likelihood programs (Felsenstein 1989; Yang 1997; Swofford 2003a; Yang 2007). (b) "BC" (Binary Character state) – the gapped sites in each column are coded as binary characters (1 if gap present, 0 if absent), available for the MP and Bayesian methods (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Swofford 2003a). This treatment can be invoked in PAUP* for MP analysis with the commands “GapMode=Missing”, and “Symbol=01” under “Format” and “Options”, and providing a matrix of symbols reflecting gapped sites. In Bayesian analyses (using MrBayes), binary characters are included as a separate binary restriction data partition, using the command “coding=variable” under “lset”. (c) “ML$\varepsilon$”, a probabilistic model implemented in the DNAML package (Rivas and Eddy 2008) of PHYLIP (Felsenstein 1989) program, that incorporates insertion and deletion events in addition to substitution events in the evolutionary model.

Assessing phylogenetic accuracy

The accuracy of the inferred trees was measured as the percentage of internal branches reconstructed correctly in the inferred tree, obtained as $P_C = \left[1 - \frac{d_T}{(2m-6)}\right]100$, where $d_T$ is the topological distance between the inferred and model trees (Robinson and Foulds 1981; Penny and Hendy 1985) and $m$ is the number of sequences in the alignment (16). $P_C$ values were averaged over all the (100) replicates for each parameter.
combination, to give $P_c$. In MP analysis, when multiple equally parsimonious trees were recovered, all comparisons were made between the model tree and a randomly chosen single tree from among equally parsimonious inferred trees, since the strict consensus of these trees tended to produce a star tree, especially at high substitution rates (Wiens 2003).

**Results**

We first describe the manner in which the alignment gaps were quantified in this study, and the effect of different simulation parameters on the number of gaps in the alignment.

*Quantification of alignment gaps*

The amount of gap in an alignment was determined as a percentage, in the following manner. First, the number of gapped sites was determined for each sequence and then obtained as an average among all the sequences in the alignment. (Note that our definition of a gapped site is common to all methods: a single cell in the alignment matrix. Thus, a gap that covers three cells (the space of three bases) in a given sequence, even if contiguous, is counted as three gapped sites in that sequence.) This was expressed as a percentage of the altered length (as a result of indel introduction) of the alignment. This percentage was then averaged across the replicates, for a given treatment, as a simple arithmetic mean (since the change in length due to the introduction of the indels varied minimally among them). We refer to the gap percentage by the term
G/S (for Gap percentage per Sequence) throughout the paper. The gap percentages thus obtained (G/S) were used to compare the relative performances of the phylogenetic methods (PhyML, MP, NJ, and Bayesian analysis) under the different gap treatment methods.

Figure 2 shows the G/S distribution of the total number of gaps in an alignment, expressed as a percentage of the total length of the alignment. Panels A and B refer to simulations where the rate ratio of insertions to deletions was 1:1 and 1:3, respectively. In each panel, the distribution of gaps has been plotted separately for each substitution rate \(r\), as a function of the rate of indel introduction \(\lambda\) that in turn, was varied as a function of the substitution rate. Both panels of Figure 2 show that the average gap percentage increases non-linearly with increase in \(\lambda\) for all \(r\). The gap percentage was minimum (~2%) when \(\lambda = 0.03\), and \(r = 0.025\), and maximum (~90%) when \(\lambda \geq 0.19\) and \(r \geq 1.0\). No noticeable differences were seen in the percentage of gaps in the alignments.
Figure 2 - Distribution of alignment gap percentages. The percentage of alignment gaps, G/S, given as the number of gapped sites per sequence length averaged over all the sequences in the alignment and expressed as a percentage of the alignment length altered after indel introduction, is plotted as a function of the indel rate (λ) for each nucleotide substitution rate, r, whose values are shown by means of different markers; see legend below figure. G/S is a suitable quantification of gaps when they are treated as missing data, binary characters or in the MLε method. G/S values were averaged over all the other parameter values for sequence length, l = 500, gamma distribution shape parameter, α = 0.5, and transition-transversion rate ratio, κ = 2. The gap distributions are shown for insertion-deletion rate ratios of 1:1 (Panel A), and 1:3 (Panel B).

when different values of sequence length (l) and transition-transversion rate ratio (κ) were used in the simulations (not shown). However, as expected, the gap percentage varied considerably with the relative proportion of insertions and deletions, with more gaps seen when the ratio of insertions to deletions was 1:1 and fewer when the ratio was 1:3, especially at low to medium substitution rates. This difference in the number of gaps is because an insertion event in a single sequence adds a gap of the size of the insertion to all the other sequences during alignment, whereas a deletion in a sequence results in a gap in only that sequence and no other, especially if the insertion/deletion event is recent in the evolutionary history of the affected sequence. The distribution of gap percentages for the random and pectinate trees was largely similar to that shown for the balanced tree in Figure 2.
Finding the gap threshold

Using the above measures of phylogenetic accuracy, it is possible to determine thresholds of gap percentages for given levels of phylogenetic accuracy. These thresholds are shown in Figure 3, which is arranged such that there are two panels for each inference method, one for the insertion-deletion rate ratio of 1:1 and the other for the ratio 1:3. The horizontal and vertical axes in each panel reflect the rate of nucleotide substitution and rate of indel introduction, respectively. However, in order to relate to empirical phylogenetic analyses (where these rates are not routinely determined), the background in this figure has been color-coded based on the percentage of alignment gaps (which can be easily determined), and the contour lines of phylogenetic accuracy have been drawn against this background. Thus, one can trace the level of accuracy of phylogenetic reconstruction based on the percentage of gaps in the alignment rather than on the rates of substitution or indel introduction. Such a representation also makes it easier to determine gap thresholds for phylogenetic accuracy in empirical studies, to determine the expected level of accuracy given a certain percentage of gaps in an alignment.
Figure 3 – Gap thresholds for different levels of phylogenetic accuracy. The gap thresholds are shown for NJ (Panels A, B), PhyML (C, D), MLε (E, F), Bayesian (G, H), and MP (I, J) methods, for various levels of phylogenetic accuracy. The background in each chart is color-coded to reflect the gap percentage thresholds; see legend at bottom of figure. Each chart also shows values of $P_c$ plotted against different gap thresholds. In Panels E-F, and G-J, white dotted lines refer to the MLε method of Rivas and Eddy (2008), and the Binary Character state (BC) treatment, respectively. The dark red dotted lines refer to the Missing Data (MD) method in all panels. Each $P_c$ value reflects one of all possible combinations of values of substitution rate, $r$, and indel rate, $\lambda$, (see Table 1), sequence length, $l = 500$, transition-transversion rate ratio, $\kappa = 2$, and the gamma among-site rate variation shape parameter, $\alpha = 0.5$, averaged over 100 replicates, for a total of 110 data points in each graph. The left panels show the results for the insertion-deletion rate ratio of 1:1 and the right panels for the ratio 1:3.

In Figure 3, Panels A and B show the level of phylogenetic accuracy for MD in the NJ analysis. These results are remarkable for several reasons. First, the contour lines of accuracy typically follow specific gap percentage ranges, as indicated by the color of the background. In other words, there appears to be a somewhat deterministic relationship between the number of gaps in an alignment and the level of phylogenetic accuracy one can expect in an NJ analysis. This appears to be true in the case of PhyML also (Panels C and D). Furthermore, both methods can be seen to be doing better in the 1:1 than in the 1:3 graphs, showing that the relative proportions of insertions and deletions matter in determining the accuracy.
Panels E and F show the results for MLε analysis. Here, we see that the minimum accuracy is approximately 90% and 70% for the 1:1 and 1:3 cases, respectively. Clearly, the MLε analysis has higher accuracy when compared to the MD analysis in conjunction with any inference method. The integrated method incorporating both substitutions and indels, MLε appears to be equivalent in accuracy to the BC method in Bayesian analysis, in the case of the 1:1 ratio of insertions and deletions. However, the accuracy of MLε is lower for datasets with larger deletion biases (as in the 1:3 ratio) and is in keeping with the other indel-coding methods (Panels G-J).

The Bayesian and MP analyses are shown in the panels, G, H, and I, J, respectively, with the dark red and white dotted contour lines within each panel representing the accuracy when the gaps are treated as Missing Data (MD) and Binary Characters (BC), respectively. As in the case of the other methods, the relationship between accuracy and G/S is clearly strong here too. Furthermore, this apparent cause-and-effect relationship appears to hold, whether the treatment method is MD or BC, especially at larger G/S values (towards the red end of the background color). It must, however, be noted that the actual relationship between the percentage of gaps in an alignment and the level of phylogenetic accuracy that can be expected is vastly different between the two gap treatment methods, MD and BC. Thus, even when the G/S value exceeds 80 percent of the length of the alignment (orange color background), as much as 70 percent of the branches are reconstructed accurately by MP (Panel I), and 90% by the
Bayesian method (Panel G), when the gaps are treated as binary characters (BC), and the insertion-deletion ratio is 1:1. In contrast, only approximately 40 percent of the branches are accurately inferred in either analysis under the MD treatment (same panels). The only case where the relationship between the percentage of gaps and phylogenetic accuracy is not as straightforward is at very high accuracy levels; the contour lines for 95 percent accuracy cross color (gap percentage) boundaries or are confined to small portions of the gap percentage range of 0-20 percent. The reconstruction accuracy for the Bayesian and MP methods in alignments where the gaps are largely due to deletion events (1:3) is worse than when compared to alignments where there is equal contribution from insertions and deletions to the gaps (1:1). Panels G and H (Bayesian analysis) and I and J (MP) show that there is a 10-20 percent difference in accuracy for a given level of gaps in an alignment between the two insertion-deletion ratios, whether for MD or BC treatments. Thus, when the gaps exceed 90 percent of the alignment length, the reconstruction accuracy is seen to be around 90 percent (BC treatment) and approximately 40 percent (MD treatment) when the insertion-deletion ratio is 1:1, whereas it is less than 80 percent (but more than 60 percent; BC treatment) and 20 percent (MD treatment) when the ratio is 1:3 (Panels G and H).

Increasing the sequence length does not appear to change the pattern of these results very much, except that there is greater accuracy when the sequence length is 2500 nts. (not shown). This improvement in accuracy, however, is not uniform across the breadth of the gap percentage landscape, being higher at the low gap percentage levels. For example, when the sequence length is 500 nts., and the gaps are equal to or greater
than 80 percent, the accuracy is approximately 80 percent (Figure 3, Panel I; BC
treatment). The corresponding accuracy when the sequence length is 2500 nts., is
approximately 99 percent – a 9-10% difference between the two lengths. Sequence
length is known to be an important determinant of phylogenetic accuracy (Graybeal
1998; Rosenberg and Kumar 2001; Gadagkar, Rosenberg and Kumar 2005) and its
influence is not being investigated in this study.

Figure 3 also shows that there are some differences among the inference methods
with respect to the gap threshold. First, the level of accuracy at a given gap percentage is
higher in the Bayesian, PhyML and MP analyses when compared to the other analyses,
especially at higher gap percentages, when the comparison is made for the MD treatment
(which is common among the inference methods, except of course, ML, which is an
integrated gap-coding/phylogenetic analysis method). Thus, when the gaps amount to
more than 80 percent in the alignment (for insertion-deletion rate ratio of 1:1), the
Bayesian, MP, and PhyML-analyzed trees are inferred with approximately 60% , 40%,
and 40% accuracy, respectively, whereas the accuracy in the NJ analyses is less than 20
percent. The comparison here also reflects the differences among the criteria used in
treating the gaps as missing data in the three inference methods. From these results, it
appears that the MD treatment in NJ infers the states at the gapped sites less accurately
than does the corresponding treatment in MP and the method used in assigning the
likelihood in Bayesian and PhyML analyses. Furthermore, the tightness of the
relationship between the contour lines of phylogenetic accuracy and the gap percentages
on the one hand, and the lower accuracy at high gap percentages in the PhyML and NJ
analyses on the other, imply that while these two methods appear to be more capable of overcoming other sources of error in phylogenetic inference (such as homoplasy in the case of MP), they fall victim to poorer treatment of gaps as missing data.

**Phylogenetic accuracy of different inference methods under varying gap percentages**

We first compare the phylogenetic accuracy of all the inference methods, taken two at a time, for the MD gap treatment since this is available for all the methods. The results are shown in Figure 4A. In each panel in Figure 4A, the average phylogenetic accuracy, $\bar{P}_c$, for one method is plotted against that of another, so that if identical, the two $\bar{P}_c$ values will lie on the diagonal. Values above the diagonal refer to cases where the method plotted on the vertical axis has relatively higher $\bar{P}_c$ values and those below the diagonal to the cases where the method on the horizontal axis has the higher $\bar{P}_c$ values.

For each comparison between inference methods, the left panel shows the results for the insertion-deletion ratio 1:1, and the right panel for the ratio 1:3. In each chart, the circles for $\bar{P}_c$ values are also color-coded to reflect the gap percentage (G/S) against which the $\bar{P}_c$ values have been measured, with the color ranging from light blue (for the lowest G/S values) to red (for the highest G/S values). In general all the inference methods do rather poorly when G/S is extremely high and very well when G/S is very low. Furthermore, as seen in Fig 3, all methods yield more accurate trees for a given G/S
value when the gaps are caused by insertions and deletions in equal proportions (left panels), when compared to alignments where the gaps result from largely deletion events (right panels). This is evident by noting that the dots of a given color (G/S value) are higher in the charts in the left panel and lower in the right, for any given pair of inference methods being compared. However, there are distinct differences among the methods
Figure 4A – Pairwise comparison of inference methods under MD gap treatment. $P_c$ values are compared, in a pairwise fashion, for four inference methods: NJ, PhyML, MP, and Bayesian analysis. As elsewhere, the left and right columns refer to the 1:1 and 1:3 insertion-deletion rate
ratios. In each panel, the average phylogenetic accuracy, $\bar{P}_c$, for one inference method is plotted against that of another. The dots in each graph are color coded to reflect the gap percentage (G/S %) against which the $\bar{P}_c$ values have been measured, ranging from light blue (for the lowest G/S values) to red (for the highest G/S values); see legend below figure. Each $\bar{P}_c$ value reflects one of all possible combinations of values of substitution rate, $r$ and indel rate, $\lambda$, (see Table 1), sequence length, $l = 500$, transition-transversion rate ratio, $\kappa = 2$, and the gamma among-site rate variation shape parameter, $\alpha = 0.5$, averaged over 100 replicates, for a total of 110 data points in each graph. The paired $t$-test ($p < 0.05$) results are shown with a letter (within the dot) that signifies if a particular method is statistically better than the other in a given comparison (J – Neighbor-Joining, P – PhyML, M – Maximum Parsimony, and B – Bayesian analysis) for the parameter combination. The $t$-test results that were not statistically significant are presented with no symbol (letter) within each dot. The results of the Z test ($p < 0.001$) over 110 data points for each method-method comparison is shown with a letter followed by an asterisk (J* – Neighbor-Joining, P* – PhyML, M* – Maximum Parsimony, and B* – Bayesian analysis).

too, and they are brought out in these pairwise comparisons. For instance, it is clear that, irrespective of whether the insertion-deletion ratio is 1:1 or 1:3, in general, the Bayesian, MP and PhyML methods are somewhat comparable, while NJ does the poorest in the presence of gaps, especially when G/S is large. However, comparing the relative performance of the methods from such graphs becomes subjective. Therefore, we conducted the paired $t$-test (at 5% level of significance) for each of the 110 data points (parameter combinations or “genes”) in each of the charts. The results of the $t$-test are
given by means of a letter that signifies if a particular method is statistically better than the other in a given comparison (B – Bayesian analysis, P – PhyML, L – MLε, M – Maximum Parsimony, and J – Neighbor-Joining). We also determined which method was better, overall, in each of the panels, using the Z test, and this is shown by the corresponding letter with an asterisk in the upper triangle.

Thus, in the comparison between NJ and PhyML, we see that PhyML shows a significantly greater overall accuracy (Z test; \( p < 0.001 \)), and that this difference is almost always statistically significant for the individual comparisons. The superiority of PhyML over NJ in the presence of gaps is very clear when the insertion-deletion ratio is 1:1 (all 110 comparisons statistically significant; \( t \) test; \( p < 0.05 \)). When the ratio is 1:3, again, PhyML is better than NJ almost all the time; NJ is found to be significantly better only in two instances out of 110; two comparisons were not significant. The comparison between MP and NJ also yields similar results, with MP being clearly superior most of the time (for 78 “genes”, with 32 comparisons turning out non significant; NJ is never better than MP.) in the left panel. The result in favor of MP is more pronounced at higher G/S values, with the graph deviating away from the diagonal. In the comparison between MP and PhyML, PhyML is superior to the other in a majority of cases, irrespective of the insertion-deletion ratio. Interestingly, MP superiority is seen only at very high G/S values, while PhyML is better almost everywhere else. This is particularly evident in the right panel (insertion-deletion rate ratio of 1:3). In both panels, PhyML is significantly better, overall (Z test; \( p < 0.001 \)).
In Figure 4A (contd.), we show the results of the Bayesian method under the MD treatment compared to PhyML, MP, and NJ methods. The figure shows that, irrespective of the insertion-deletion rate ratio and G/S%, the Bayesian method is more accurate than MP, NJ or PhyML, overall ($p < 0.001$). When compared individually, it is seen to be better than NJ in all the genes in the left panel and all but one of the genes in the right panel. Next, the Bayesian method is seen to be better than PhyML, overall, but it is statistically better ($p < 0.05$) 48 times (out of 110 comparisons across the entire spectrum of G/S values), with PhyML outperforming it ($t$ test; $p < 0.05$) in 22 cases, with neither being better than the other in the remaining 40 genes, in the right panel. In the left panel (1:1) similar results were obtained, where PhyML and Bayesian were each statistically better ($t$ test; $p < 0.05$) than other roughly equal number of times (around 30 cases each), with the two methods occupying different “niches” (Bayesian doing better at high gap percentages and PhyML at the low to intermediate levels of gap percentage.). When Bayesian and MP methods are compared, the Bayesian analysis is statistically better ($t$ test; $p < 0.05$) in almost 90 cases, whereas MP is better almost never, irrespective of the insertion-deletion rate ratio.
Insertion:Deletion = 1:1  
Insertion:Deletion = 1:3

Bayesian analysis+MD vs. PhyML+MD
Bayesian analysis+MD vs. MP+MD
Bayesian analysis+MD vs. NJ+MD

<50  50-60  60-70  70-80  80-90  90-100 (G/S%)
Figure 4A contd – Pairwise comparison of inference methods under MD gap treatment.

$P_c$ values are compared, in a pairwise fashion, for four inference methods: NJ, PhyML, MP, and Bayesian analysis. As elsewhere, the left and right columns refer to the 1:1 and 1:3 insertion-deletion rate ratios. In each panel, the average phylogenetic accuracy, $P_c$, for one inference method is plotted against that of another. The dots in each graph are color coded to reflect the gap percentage (G/S %) against which the $P_c$ values have been measured, ranging from light blue (for the lowest G/S values) to red (for the highest G/S values); see legend below figure. Each $P_c$ value reflects one of all possible combinations of values of substitution rate, $r$ and indel rate, $\lambda$, (see Table 1), sequence length, $l = 500$, transition-transversion rate ratio, $\kappa = 2$, and the gamma among-site rate variation shape parameter, $\alpha = 0.5$, averaged over 100 replicates, for a total of 110 data points in each graph. The paired t-test ($p < 0.05$) results are shown with a letter (within the dot) that signifies if a particular method is statistically better than the other in a given comparison (J – Neighbor-Joining, P – PhyML, M – Maximum Parsimony, and B – Bayesian analysis) for the parameter combination. The t-test results that were not statistically significant are presented with no symbol (letter) within each dot. The results of the Z test ($p < 0.001$) over 110 data points for each method-method comparison is shown with a letter followed by an asterisk (J* – Neighbor-Joining, P* – PhyML, M* – Maximum Parsimony, and B* – Bayesian analysis).

It is important to note that the MD treatment is different among the inference methods. (In the case of the probabilistic methods – PhyML and Bayesian analysis, the likelihood is summed over all four nucleotides at the gapped sites (Huelsenbeck and Ronquist 2001; Guindon and Gascuel 2003; Ronquist and Huelsenbeck 2003), while for a
distance method like NJ, the nucleotide state at each gap is inferred by distributing the missing changes to unambiguous changes, and finally, for the MP method, a given state is assigned to each gapped site in a sequence if it is the most parsimonious, given the placement of the taxon in the tree (see the FAQ on the PAUP*website (http://paup.csit.fsu.edu/paupfaq/paupfaq.html). Hence, it appears that the level of accuracy seen in this study for the different inference methods is also attributable to the accuracy with which the state is inferred/likelihood is computed at the gaps by the corresponding MD methods.

Next, we compare the accuracy of the different inference methods, again, taken two at a time, when gaps are treated not merely as missing data but as information that is included in the phylogenetic analysis. Under the MP and Bayesian methods, gaps can be treated as binary characters (BC), with sites in a given column being scored as a 1 if gapped and 0 if not. Among the recent advances in the modeling of molecular sequence evolution is the integration of insertion and deletion events along with base substitution processes in a probabilistic framework for phylogenetic inference (Rivas and Eddy 2008). We have used this method for maximum likelihood analysis of our data, to compare this treatment (which we refer to as MLε) with the BC treatment in MP and Bayesian analyses. These comparisons (again, pairwise as in Fig. 4A) are shown in Figure 4B.

As in Figure 4A, here too the average phylogenetic accuracy, $P_c$, for one method is plotted against that of another in each panel (1:1 and 1:3), and the color-coding scheme is the same as well. Figure 4B shows that the differences among the inference methods
in accuracy in the presence of indel-induced gaps, is much more evident when gaps are included in the phylogenetic analysis and not treated as missing data (that is, when compared to the results in Figure 4A). Furthermore, the accuracy is generally much higher, even in the method with the lower accuracy (note the $\overline{P}_c$ value ranges in the axes, which have been optimized for maximum spread of the points within each graph for greater visibility). Finally, it can be seen, just as in Figure 4A, but in a more pronounced manner, that the accuracy of both methods is higher for a given G/S level when the gaps result from equal proportion of insertions and deletions (panels in left column), as opposed to when they are largely from deletion events (right column), as evidenced by a comparison of the heights of the dots of a given color between the two panels, particularly at the mid to higher G/S values (light green, yellow and orange colored dots).

The different indel-coding methods are compared for their performance, in conjunction with the corresponding inference methods, in Figure 4B. It is immediately obvious that accuracy is much higher (and in a tight range of values) in all the left panels (1:1 ratio) for the probabilistic methods (Bayesian and ML$\epsilon$) when compared to the right side panels (1:3 ratio), providing a compelling case for the association between phylogenetic accuracy and evolutionary origin of alignment gaps (insertion or deletion) – at least for the probabilistic methods. This fact is even more obvious in the middle left panel where the two probabilistic methods are compared. When MP and ML$\epsilon$ are compared (Fig. 4B, top panels), it is clear that ML$\epsilon$ has a much higher accuracy in the
medium to high range of G/S values when the insertion-deletion ratio is 1:1, with MP
doing better at low to medium G/S values. This difference between the two methods is
much more pronounced in the right panel (insertion-deletion ratio, 1:3), where ML\(E\) is
Figure 4B – Pairwise comparison of inference methods when gaps are coded as distinct evolutionary events. $\overline{P_c}$ values are compared, in a pairwise fashion, for the inference methods: MP, Bayesian analysis, and the MLε analysis, when the gaps were treated as binary characters or
by the DNAMLε method. As in Fig. 4A, the average phylogenetic accuracy, $\overline{P}_C$, for one method is plotted against that of another in each panel. For each pairwise comparison between the inference methods, the left panel shows the results for the insertion-deletion rate ratio is 1:1 and the right panel when it is 1:3. The dots in each graph are color coded to reflect the gap percentage (G/S) against which the $\overline{P}_C$ values have been measured, ranging from light blue (for the lowest G/S values) to red (for the highest G/S values). Each $\overline{P}_C$ value reflects one of all possible combinations of values of substitution rate, $r$ and indel rate, $\lambda$, (see Table 1), sequence length, $l = 500$, transition-transversion rate ratio, $\kappa = 2$, and the gamma among-site rate variation shape parameter, $\alpha = 0.5$, averaged over 100 replicates, for a total of 110 data points in each graph. The paired $t$-test ($p < 0.05$) results are shown with a letter (within the dot) that signifies if a particular method is statistically better than the other in a given comparison (B – Bayesian analysis, L – MLε, and M – Maximum Parsimony) for the parameter combination. The paired $t$-test results that were not significant are presented as dot with no symbol. The results of the Z test ($p < 0.001$) over 110 data points for each method-method comparison is shown with a letter followed by an asterisk (M* – Maximum Parsimony, B* – Bayesian analysis, and L* – DNAMLε).

better only at the highest G/S values and MP clearly the better of the two elsewhere. The middle panels show that Bayesian analysis produced more accurate trees when compared to MLε in 77 (right panel) and 96 (right panel) out of 110 comparisons. Just as in the MP, MLε comparison (top panels), MLε again outperforms Bayesian at the highest G/S values. The difference between the two panels is quite evident, with both methods
varying in accuracy in a very tight range in the left panel (when compared to the 1:1 ratio). Note that the distribution of $\hat{P}_c$ values is somewhat similar when the MP and Bayesian methods are each compared to ML$\varepsilon$ (top and middle panels for the 1:3 ratio), suggesting a similar pattern between MP and Bayesian, under BC treatment of gaps, although the Bayesian method appears to be doing better than MP against ML$\varepsilon$. These two methods are compared in the bottom panels. As mentioned above, it is immediately apparent that the accuracy for the Bayesian method in the left panel is much higher (minimum 90%) when compared to the right panel. These panels also show that whenever the difference in accuracy between MP and the Bayesian method is statistically significant ($t$ test; $p < 0.05$), the latter is always better (with 30 percent of the cases being non significant). Furthermore, the “genes” where the difference in accuracy is statistically significant are mostly spread across medium to high G/S values. In summary, the Bayesian method is superior to the MP and ML$\varepsilon$ methods under the gap coding approach, irrespective of the relative proportions of insertions and deletions in the alignment.

The results shown in Figures 4A and 4B have been obtained from our analyses of the alignments obtained from simulations done on the balanced (symmetric) model tree (Fig. 1). We also obtained sequence alignments from simulations done with 16-taxon random-branching and pectinate trees for a subset of parameter values that, however, spanned the range of parameter values used in this study (Table 1). All the analyses shown in Figures 4A and 4B were done on these alignments as well (not shown),
including the pairwise comparisons among the inference methods and the paired $t$-tests. The results in those analyses showed that while the inference methods compare among themselves for the random-branching tree just as they did for the balanced tree, there are some differences in the case of the pectinate tree. In the case of the MD analysis, while Bayesian was the better method overall, the performance of PhyML in the case of the pectinate tree was glaringly different. In the case of the balanced tree, PhyML showed greater accuracy than NJ in essentially all the cases, irrespective of the insertion-deletion ratio. However, for the pectinate tree, the roles are exactly reversed, with NJ better than PhyML in essentially all the cases – again, irrespective of the insertion-deletion ratio. Similarly, while PhyML and MP were each better than the other roughly equal number of times in the case of the balanced tree, MP accuracy was superior for the pectinate tree in essentially all the genes studied, irrespective of the insertion-deletion rate ratio.

When indel coding was used, again, the random tree results are quite similar to those from the balanced tree. Interestingly, the results from the pectinate tree were not glaringly different from those from the balanced tree, but rather, the two were largely similar, except that the overall accuracy was lower by about 20 percent.

**Discussion**

We undertook this study to investigate the relationship between the number of gapped sites in a sequence alignment and the accuracy of phylogenetic inference, and furthermore, to understand the impact of different gap treatment methods, phylogenetic
inference methods, the ratio of insertions to deletion events in the evolutionary history of
the sequences, and other sequence parameters such as sequence length and the transition-
transversion rate ratio, on this relationship. Using the computer program, Dawg v1.2
(Cartwright 2005) we simulated DNA evolution along a 16-taxon model tree (Fig.1),
incorporating both nucleotide substitution events and insertion and deletion (indel) events
(the latter as a function of the substitution rate.). The resulting DNA alignments were
then subjected to three gap treatment methods, namely, MD, BC, and ML\(\varepsilon\) (see
Methods), and the phylogenetic analysis was done using popular phylogenetic inference
methods – distance (NJ), parsimony (MP), likelihood (PhyML) and Bayesian analysis.

A remarkable result in this study is the strong, almost deterministic, dependence
of the accuracy of phylogenetic inference on the percentage of gapped sites in the
alignment, irrespective of the inference method, gap treatments, or insertion-deletion rate
ratio, when the percentage of gapped sites was high (Fig. 3). This made the assignment
of gap thresholds for specific levels of phylogenetic accuracy fairly straightforward,
without being necessarily concerned with other determinants of phylogenetic accuracy. It
was only at lower gap levels that the relationship was not as straightforward, and other
factors (e.g., substitution rate) began to play a part in directly influencing the accuracy of
the inferred trees (as evidenced by the contour lines of accuracy crossing gap percentage
thresholds in Fig. 3).

Earlier studies that have compared gap treatment methods have been confined to
comparing their relative performances within a given inference method, particularly MP
Therefore, this study was undertaken to provide users with a comparison of other commonly used inference methods as well. We find that the probabilistic methods are clearly superior to MP and NJ, irrespective of whether gaps are treated as missing data or binary characters. Treating gaps as binary characters implies the assignment of unambiguous phylogenetic signal to them in the evolutionary history of the sequences. Therefore, the number of gaps has little bearing on the distortion of the phylogenetic signal under the BC method. On the other hand, the MD method requires the inference of the missing state at each gapped site (or the summation of the likelihood for all four nucleotides at the gapped sites), a process that is bound to be strained with increasing number of gaps in the alignment. Therefore, it is easy to understand the relative superiority of the BC gap treatment method. It must be noted, of course, that this method can only contribute to phylogenetic accuracy as long as the alignment gaps are known without error (as in this study). Thus, the importance of the accuracy of sequence alignment cannot be underestimated.

The ML$\epsilon$ method performed well in our study, although the Bayesian method was better, especially when the insertion-deletion ratio was 1:3 (Fig. 4B). When compared to MP analysis (the other inference method that incorporated the BC), ML$\epsilon$ was much better when the number of gaps was high, irrespective of the insertion-deletion ratio. Such methods hold the potential for more accurate reconstruction of phylogenies in the presence of large alignment gaps (also see Rivas 2005, Loytynoja and Goldman 2008).
In addition to the MD treatment and the gap-coding treatments such as BC, other treatment methods exist, although not widely used anymore. One of these is pairwise deletion, a gap treatment method that is meaningful only when sequences are compared in a pairwise fashion, as in distance methods of inference, such as NJ. Moreover, it is an extremely rapid method that is suited to the speed of NJ. The other is complete deletion of entire columns of gapped sites from the alignment, which is a gap-treatment method that is applicable to any phylogenetic inference method. We did these analyses as well, because there is sometimes an uncertainty about which of these two methods is better (Nei and Kumar 2000; Chang et al. 2008). The complete deletion of gaps posed a problem in our study as the number of sites that needed to be removed from the alignment, especially at higher substitution rates, caused the remaining sequence length to become so small that often at least one of the four nucleotides failed to be represented in the alignment. Therefore, we used this method only when the substitution rate was very low \((r \leq 0.2)\), and when the alignment length (remaining after complete deletion) for each replicate of a given sequence combination was at least 100nts.

Since the complete deletion treatment could be used only for low substitution rates, the comparison between the two treatments is also made only across this range. Furthermore, since the pairwise deletion method can only be used in conjunction with the NJ method in this study, we compared the two methods only for NJ. Both methods are comparable at low to moderate gap percentages, but diverge thereafter in the accuracy of phylogenetic inference (not shown). It must also be noted that the gap percentage does not reach very high levels in the pairwise deletion as it does in the complete deletion
method. Thus, while for a given gap percentage, the two treatment methods may be comparable in terms of phylogenetic accuracy, the pairwise removal of gaps appears to be better since the gap percentage is much lower with this method.

A comprehensive list and analysis of gap treatment methods may be found in Ogden and Rosenberg (Ogden and Rosenberg 2007) and Simmons Muller and Norton (Simmons, Muller and Norton 2007). However, they did not compare among phylogenetic inference methods, even for those gap-treatment methods that were common to multiple inference methods. In this study, while we do compare among gap-treatment methods, our emphasis is also on comparing among inference methods, insertion-deletion ratios, and the effect of the amount of gap on phylogenetic accuracy under varying parameters.

In order to better understand the influence of the alignment gaps on phylogenetic accuracy, we performed the same simulations, but with only base substitutions and no indels. As there were no gaps in the alignments, the data were subjected to phylogenetic analysis without any processing by means of gap treatment methods. The results of this analysis showed that, as expected, Maximum Likelihood and Bayesian analysis produced the most accurate trees, particularly at the highest substitution rates (not shown).

Another notable finding in this study is the differential influence of insertions and deletions on phylogenetic accuracy. Most of the commonly used gap treatment methods do not distinguish between insertions and deletions. Our results show that phylogenetic accuracy was lower when the insertion-deletion ratio was 1:3. Even the probabilistic methods (PhyML, MLE and Bayesian), which produced the most accurate trees when
insertions and deletions were introduced in equal numbers, performed somewhat poorly when the ratio was 1:3 (Figs. 3, 4A, and 4B). It therefore, appears important to develop methods that first distinguish between insertion and deletion events in the evolutionary history of the sequences in an alignment, and then treat them separately to add distinct signals to the phylogenetic analysis.

In this study, the metric we have used to measure the accuracy against is the percentage of gaps in the alignment, and this in turn has been measured mainly as G/S. Some studies have found that it is not the amount of data missing but rather the amount of data remaining that matters in determining the accuracy of the phylogeny being inferred (Wiens 2003; Philippe et al. 2004; Wiens 2006). In order to compare our results with the results from these studies, we show the accuracy, $\bar{P}_c$, the remaining number of nucleotides after the gaps are removed from the alignment, and the total length of the alignment resulting from the introduction of indels during the evolution of the sequences, all plotted as a function of G/S (Figure 5). The layout of Figure 5 is the same as that of Figure 3, with the left and right columns referring to insertion-deletion ratios of 1:1 and 1:3, respectively, and the inference methods arranged one below the other, in the same order, namely, NJ, PhyML, MLEC, Bayesian analysis and MP.

One of the first things that stand out in Figure 5 is the general accuracy of the MD method when the G/S is low and poor accuracy when G/S is high, irrespective of the inference method. Interestingly, when the accuracy curve in each graph is compared to the curve of the remaining number of nucleotides, there seems to be little
relationship between the two in the left panels (1:1), again, irrespective of the inference method. Thus, even as the number of remaining nucleotides (red triangles) continues to be high for large G/S values, the $P_e$ value (for MD treatment; red circles) plummets down to close to zero. This is because, although the remaining number of nucleotides is high, this
Figure 5 - Effect of the alignment gap percentage on phylogenetic accuracy, number of characters remaining in the alignment, and total alignment length after gap-introduction
The average accuracy, $\overline{P}_c$, (dark red circles for the MD treatment, open diamonds for the ML$\varepsilon$ method, and open circles for the BC treatment), remaining number of characters in the alignment after the gaps are removed (red inverted triangles), and the total length of the alignment (including gaps; symbolized by green, upright triangles), are each shown as a separate function of the average gap percentage, for NJ (Panels A, B), PhyML (C, D), ML$\varepsilon$ (E, F), Bayesian (G, H), and MP (I, J) inference methods. The left panels show the results for the insertion-deletion rate ratio of 1:1 and the right panels for the ratio 1:3. Each data point in the graph was obtained as the corresponding value for one of all possible combinations of values of $r$, the substitution rate and $\lambda$, the indel rate (see Table 1), sequence length ($l = 500$), transition-transversion rate ratio ($\kappa = 2$), and the gamma among-site rate variation shape parameter ($\alpha = 0.5$), averaged over 100 replicates, for a total of 110 data points in each graph.

largely a consequence of insertion events having added nucleotides to the sequences. Thus, although there is data, there is little phylogenetic information in it, since homology across sequences at these levels becomes nebulous, at best, leading to low accuracy. On the other hand, the curve of remaining nucleotides itself drops with increase in the G/S value in the right column panels (1:3) – a reflection of the greater proportion of deletion events. Therefore, while the $\overline{P}_c$ values drop with increase in G/S in spite of an abundance of data in the left column panels, they do so in the right column panels evidently because of the loss of data as G/S increases (note the scale on the secondary Y axis). Thus, while the remaining amount of data may be an important determinant of accuracy (as in the right column in Fig. 5, and as mentioned in (Wiens 2003; Philippe et al. 2004; Wiens
2006)), this is true only when homology among the sequences in the alignment can be established in the remaining character data. If, however, the remaining character data is largely a result of insertion events, the relationship is unlikely to hold, as seen in the left panels.

On the other hand, if the gaps are coded separately (e.g., as BC), then the phylogenetic signal present in the gaps (if the alignment is accurate) increasingly becomes the only information for the inference method to rely on, as G/S increases. The loss of signal from the character data is reflected in the decreased phylogenetic accuracy at high G/S values (left column). The greater loss of phylogenetic accuracy at medium G/S values in the right column panels of Figure 5 can be attributed to fewer deletion events that are distinct and non-overlapping when compared to insertion events that are more likely to be distinct and non-overlapping, as the increase in the total length of the alignment with indel introduction will be much higher when the insertion-deletion rate ratio is 1:1.

In this study, we also found that the alignments from the random-branching tree yielded essentially the same results as those from the balanced tree, while those from the pectinate tree were different (not shown). The analyses from the pectinate tree data in general showed lower accuracy than the corresponding analyses from the balanced tree datasets. Furthermore, the relative performances of the different inference methods were not the same between the two model topologies. In particular, the relative performance of the PhyML method was worse when the topology contained pectinate branching.
This is a simulation-based study and is confined to certain specific simulation parameters and methods of gap treatment and phylogenetic inference used in this study. However, the choices of the parameter values have been made based on empirical studies in the literature. This included the size distribution of indels as well (Zhang and Gerstein 2003), which may not be a critical feature as far as the BC treatment is concerned, but may be important when the state is inferred at the gaps or coded. Therefore, we believe that the results obtained in this study are sufficiently general to be useful to the community of molecular phylogeneticists. However, we must add a note of caution that while it is likely that the general results of this study will hold, the particulars may be dependent on the specific choices of simulation and other parameter values. Finally, the relationships between phylogenetic accuracy and gap percentage in this study were derived based on two unlikely events in empirical studies – knowledge of the true tree and a perfect alignment. These certainly are sources of uncertainty and/or error in real data analysis, and must be accounted for, in empirical studies. However, the utility of simulation-based studies such as this is that they serve to provide an assessment and quantification of relationships in the absence of confounding factors.

**Conclusions**

The presence of gaps in molecular sequence alignments is common-place in the literature. Our simulation-based results show that, when the alignment gaps reflect indel events without error, and the number of gapped sites per sequence is ≤20 percent of the
sequence length, all the inference methods used (NJ, PhyML, MLε, Bayesian analysis and MP) perform well in accurately inferring the phylogeny. However, when the number of gaps is large (≥80 percent), the Bayesian method clearly outperforms the other inference methods when the gaps are treated as Missing Data (MD), although it must be noted that since each inference method uses a different criterion in treating gaps as missing data (see Methods section), the higher accuracy for the Bayesian and PhyML method can perhaps also be attributed to a more accurate integration of the state at each of the gapped sites. Within the MP and Bayesian methods, the inference of the phylogeny was significantly more accurate when each gapped site was treated as a Binary Character state (BC) than when the gaps were treated as MD. When the sequences in an alignment contain a large number of gaps, as in the case of highly diverged sequences, coding gaps as in likelihood analysis (MLε) may be more efficient than Bayesian or MP in combination with the BC method. Finally, our results also show that it is more difficult to accurately infer the phylogeny from an alignment where a greater proportion of gaps reflect deletion events rather than insertion events in the evolutionary history of the sequences in the alignment.

Acknowledgments

We thank Ohio Supercomputer Resources (OSC) for help with computational analyses. We would also like to thank Reed Cartwright for help with his simulation
program, and the anonymous reviewers for insightful suggestions. This research was supported by start-up funds from University of Dayton (to SRG) and summer research funding from the Graduate School, University of Dayton (to BD).
CHAPTER 2.2

THE IMPACT OF SEQUENCE PARAMETER VALUES ON PHYLOGENETIC ACCURACY

This chapter is published in *Biology and Medicine*, 1(3): 50-62, 2009
The Impact of Sequence Parameter Values on Phylogenetic Accuracy

Bhakti Dwivedi\textsuperscript{1} and Sudhindra R. Gadagkar\textsuperscript{1,2}\textsuperscript{*}

\textsuperscript{1}Department of Biology, University of Dayton, Dayton, OH 45469-2320
\textsuperscript{2}Present Address: Central State University, Wilberforce OH 45384-1004

\textsuperscript{*}Corresponding author:

E-mail: sgadagkar@centralstate.edu
Abstract

An accurately inferred phylogeny is important to the study of evolution. Factors affecting the accuracy of an inferred tree can be traced to several sequential steps leading to the inference of the phylogeny. We have examined here the impact of some features of nucleotide sequences in alignments on phylogenetic (topological) accuracy rather than any source of error during the process of sequence alignment or choice of the method of inference (as is usually done). Specifically, we have studied (using computer simulation) the implications of changing the values of the following five parameters, individually and in combination: sequence length (l), nucleotide substitution rate (r), nucleotide base composition (θ), the transition-transversion rate ratio (κ), and the substitution rate heterogeneity among the sites (α). An interesting, and unexpected, result was that κ has a strong positive relationship with phylogenetic accuracy, especially at high substitution rates. This simulation-based work has implications for empirical researchers in the field and should enable them to choose from among the multiple genes typically available today for a more accurate inference of the phylogeny being studied.

Keywords: molecular evolution, phylogenetic inference, nucleotide substitution rate, transition-transversion rate ratio, phylogenetic accuracy, substitution saturation
Introduction

Phylogenetic reconstruction from an alignment of molecular sequences is the last in a series of consecutive steps that include obtaining the required sequences (either from DNA extracted from tissue or from a databank), aligning them, and employing a method of inference to reconstruct the phylogeny. Obviously, the correctness of the inferred phylogeny depends upon accuracy at each of these steps, and therefore, most studies that evaluate the determinants of the accuracy of phylogenetic inference have focused on understanding the contribution of these steps, particularly the latter two, to the accuracy of the inferred tree (Hillis 1995; Nei 1996; Takahashi and Nei 2000; Raghava et al. 2003; Huelsenbeck and Rannala 2004; Hall 2005; Rosenberg 2005a; Rosenberg 2005b; Ogden and Rosenberg 2006)

In the present study, however, rather than assess the contribution to phylogenetic accuracy of the three steps mentioned above, we have instead, focused on dissecting the features of the DNA sequences to determine the optimal combinations of sequence parameters that are associated with accurately inferred phylogenies. With the advances in sequencing technology already yielding DNA sequences for hundreds of taxonomic groups, and with the promise of much more to come with the popularization of next-generation sequencing (Mardis 2008) and metagenomics (Singh et al. 2009), several studies have focused on improving phylogenetic inference methods to handle large datasets (e.g., (Guindon and Gascuel 2003; Tamura, Nei and Kumar 2004). However, with the new technologies yielding large sets of diverse sequences and projects such as the Tree of Life (Maddison, Schulz and Maddison 2007) utilizing them to compare across
extremely large times of divergence, it is also necessary to understand the effect of large differences in various parameters of these sequences on the accuracy of phylogenetic inference.

When an alignment of molecular sequences from different species is used to infer a phylogeny, what is actually being inferred is the evolutionary history of the sequences in the alignment, with the expectation that it accurately reflects the evolutionary history of the organisms whose sequences are in the alignment (Nei and Kumar 2000; Felsenstein 2003). However, since different genes can produce different evolutionary histories (trees) for a group of taxa (Nichols 2001; Gadagkar, Rosenberg and Kumar 2005), it is important to understand the individual and joint effects of the sequence parameters on the accuracy of phylogenetic reconstruction.

DNA sequences can be characterized by summary statistics such as length and base composition. When two or more such sequences need to be compared to each other (as in an alignment prior to phylogenetic analysis) additional parameters come into play, such as the overall rate of nucleotide substitution (replacement of one nucleotide by another nucleotide), the ratio of two specific instantaneous rates of substitution: rate at which transitions (A→G or C↔T) and transversions (all other changes) occur, and the rate variation among sites. These comprise some of the sequence parameters that are important for the accurate reconstruction of a phylogeny.

For example, the amount of evolution (substitution rate) among sequences is usually a deciding factor in reconstruction of a phylogeny, such that slow-evolving gene sequences (e.g., elongation factor-1α, small subunit ribosomal RNA) are used to infer the
relationships among distantly related taxa (Regier and Shultz 1997; Struck \textit{et al.} 2007), and fast-evolving sequences (such as animal mitochondrial genes, virus genomes, and the third codon position of protein coding genes) are used to infer the phylogeny among closely related organisms (Hillis \textit{et al.} 1992; Ou \textit{et al.} 1992; Yang 1996a, 1996b; Yoder, Vilgalys and Ruvolo 1996.) The reason fast-evolving genes cannot be used to infer phylogenies of distantly related taxa is of course, because higher evolutionary rates lead to multiple substitutions at the same site, and thus, a saturation of the phylogenetic signal, leading to incorrect tree reconstruction (Halanych and Robinson 1999; Xia 2000; Struck, Hessling and Purschke 2002; Struck \textit{et al.} 2007). Therefore, different approaches have been developed to detect saturation in order to exclude entire genes or parts thereof from the phylogenetic analysis (Xia \textit{et al.} 2003; Struck \textit{et al.} 2008). For example, the ratio of the numbers of transitions to transversions plotted against the genetic distances ($p$) for all pairwise sequence comparisons in the alignment is a common test to determine whether sequences have experienced substitution saturation (Halanych and Robinson 1999; Struck, Hessling and Purschke 2002; Xia \textit{et al.} 2003; Struck \textit{et al.} 2007). In general, since the frequency of transitional substitutions is known to be higher than transversional substitutions in the genome (Wakeley 1996), this test detects saturation when the plot shows no further increase in the observed number of transitions despite increasing genetic distances. Thus, saturation of transitions at high levels of sequence divergence indicates saturation in the data.

Most studies that have investigated the influence of sequence parameter values on phylogenetic accuracy have varied one sequence parameter at a time, and therefore, have
failed to record the influence of the interaction among the parameters, the inference methods, and any other factors considered, such as the topology of the model tree. Therefore, we simulated evolution along different model trees (Fig. 1) to generate non-coding DNA sequences, varying different parameters across wide ranges (that are, however, biologically realistic), and compared the performance of different inference methods in reconstructing the phylogeny. During the simulation process, we systematically varied the following sequence parameters: sequence length ($l$), overall rate of nucleotide substitution ($r$), nucleotide base-composition ($\theta$), transition-transversion rate ratio ($\kappa$), and the heterogeneity of substitution rates among sites ($\alpha$), in order to study their individual and joint effects on the accuracy of phylogenetic tree reconstruction, using the following inference methods: Neighbor-Joining (NJ), Maximum Parsimony (MP) and Likelihood-based methods (ML and PhyML). In addition, we simulated evolution along tree topologies of different size, shape, and relative branch lengths.

While most of our results agree with those in the literature, there is one notable exception, namely, those involving $\kappa$, which shows a positive relationship with phylogenetic accuracy, thus appear to contradict previous studies (Yang 1998; Rosenberg and Kumar 2003). Our results are, however, consistent for all the tree topologies examined, regardless of the values of the other parameters, although differences exist in the extent of accuracy achieved. The positive relationship between the value of $\kappa$ and phylogenetic accuracy is stronger when the evolutionary rate, $r$, is high and thus, contributing to saturation of the pairwise genetic distances among the sequences in the alignment. Our results also showed that all the four inference methods performed equally
well under substantial saturation (high \( r \) and high \( \kappa \)), while there were significant
differences in accuracy among them at high \( r \) and low \( \kappa \). Our findings suggest that highly
divergent datasets are still usable, as the phylogenetic information is often not completely
lost, and may be retrieved using sites that have experienced more transversions.

**Material & Methods**

**Computer Simulation**

Nucleotide sequence alignments were generated using the computer program
Dawg version 1.1.2 (Cartwright 2005) for four ultrametric, 16-taxon topologies (Fig. 1),
obtained from Ogden and Rosenberg (2006). Simulations were also performed using
non-ultrametric 16-taxon topologies as in Fig. 1 (not shown). DNA evolution was
simulated using only nucleotide substitution events under the HKY model (Hasegawa,
Kishino and Yano 1985), while systematically varying the following sequence
parameters in a fully factorial manner: sequence length (\( l \)), nucleotide base frequencies
(expressed as the G+C content, \( \theta \)), rate of nucleotide substitution (\( r \)) as a multiplier, the
transition to transversion rate ratio (\( \kappa \)), and the shape parameter of the gamma
distribution that describes the rate variation among-sites. The values of the sequence
parameters used in the simulations are given in Table 1. All other options in the
simulation program were set to default during simulation.
Figure 1 – The model trees. The four ultrametric 16-taxon topologies obtained from Ogden and Rosenberg (2006) used as model trees for the simulations of DNA evolution, shown with relative branch lengths: (A) Balanced tree with equal branch lengths, (B) Balanced tree with random branch lengths, (C) Random tree generated under a pure birth (Yule) process, and (D) Pectinate tree. During simulation, the total number of substitutions to be made for a given branch, for a given parameter-combination, was obtained as the product of the branch length in the model tree, the rate multiplier and the sequence length. Values from the latter two parameters were obtained from Table 1.

Table 1. The sequence parameter values used in the simulations. The sequence length, $l$, is measured as the number of nucleotides, the nucleotide substitution rate, $r$ is a multiplier, that, when multiplied by a given branch length in the model tree and the sequence length, yields the expected number of substitutions to be introduced in that branch during simulation. The nucleotide base frequencies are expressed in terms of G+C content, $\theta$, $\kappa$ represents the transition to transversion rate ratio, and the shape parameter, $\alpha$ specifies the extent of rate heterogeneity among sites.
<table>
<thead>
<tr>
<th>Sequence Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence length, $l$</td>
<td>500, 2500</td>
</tr>
<tr>
<td>G+C content, $\theta$</td>
<td>0.20, 0.50, 0.80</td>
</tr>
<tr>
<td>Transition-transversion rate ratio, $\kappa$</td>
<td>1, 5, 10, 20</td>
</tr>
<tr>
<td>Gamma distribution shape parameter, $\alpha$</td>
<td>0.1, 0.5, 5.0, $\propto$ (infinity)</td>
</tr>
<tr>
<td>Rate of nucleotide substitution, $r$</td>
<td>0.025, 0.05, 0.1, 0.2, 0.4, 0.8</td>
</tr>
</tbody>
</table>

While the ranges of these parameters have been deliberately kept rather large in order to understand the full scope of their influence on phylogenetic accuracy, they are not unrealistic, and have in fact been obtained empirically from mammalian genes and used in earlier studies (Rosenberg and Kumar 2003; Gadagkar and Kumar 2005). In particular, such ranges, particularly for $r$ and $\kappa$, are seen in mitochondrial genes (Yoder, Vilgalys and Ruvolo 1996; Yang 1996c; Yang 1998; Yang and Yoder 1999), nuclear non-coding introns (Saitou and Ueda 1994), and even in some nuclear genes (Rosenberg and Kumar 2003). Each sequence parameter combination (total 576 “genes”) was used for the simulations along the four model trees, and 100 replicates were obtained for each gene, thus producing 230,400 datasets.

**Phylogenetic tree reconstruction**

After the simulations were done, the sequence alignments obtained were subjected to phylogenetic inference using Neighbor-Joining (NJ), Maximum Parsimony (MP), and Maximum likelihood (ML) methods as implemented in PAUP* version 4.0 b10 (Swofford 2003). In addition, likelihood analysis was also done using PhyML.
version 2.4.4 (Guindon and Gascuel 2003) because of its speed and efficiency in comparison to ML. HKY (Hasegawa, Kishino and Yano 1985) pair-wise distance estimates were used for the NJ analyses. In PhyML, the initial tree was built using BIONJ (Guindon and Gascuel 2003). The parameters of the HKY substitution model (the four base frequencies and the transition/transversion rate ratio) along with the proportion of invariable sites and the gamma distribution shape parameter were estimated from the simulated data using PAUP*. For the MP and ML analyses, a heuristic search was done using the stepwise addition algorithm for the provisional tree and subsequent branch swapping was done using the Nearest-Neighbor Interchange (NNI) method. When multiple trees were recovered, a strict consensus of these trees was taken to produce a single tree. All other settings were set to default in PAUP*, and PhyML program.

Assessing phylogenetic accuracy

The accuracy of the phylogenetic trees inferred was measured as the percentage of internal branches (or nodes) reconstructed correctly in the inferred tree, $P_C$, obtained as

$$P_C = \left[ 1 - \frac{d_T}{(2m-6)} \right]^{100},$$

where $m$ is the number of sequences in the phylogeny (16) and $d_T$ is the topological distance between the inferred tree and model tree (Robinson and Foulds 1981; Penny and Hendy 1985). $P_C$ values were averaged over all the (100) replicates for each parameter combination, to give $\bar{P}_C$ and is expressed in percent. For example, 60 percent accuracy means 60 percent of the internal branches are reconstructed correctly in the reconstructed (or inferred tree) when compared to the model tree.
Results

Overall Performance

We first examine the overall accuracy, $P_c$, of each inference method for the lowest and highest values of each parameter (Table 2). It can be seen that there is a large difference in accuracy between the two extreme values considered in this study for some parameters, and not for others. Furthermore, some inference methods appear to show a greater difference than others. When the accuracy is compared between the lowest and highest parameter values, the inference methods, in general, show an increase in accuracy for $l$, the sequence length, a decrease for $r$, the substitution rate, very slight to no change for $\theta$, the base composition, an increase for $\kappa$, the transition-transversion rate ratio, and a decrease for $\alpha$, the shape parameter, with the greatest difference in accuracy seen in the case of $r$, and the least in the case of $\theta$ (although a G+C content of 0.50 shows a slightly higher accuracy when compared to the two extreme values of 0.20 and 0.80; not shown). In general, most of these results are not novel, and have been shown before, although perhaps not in such detail. However, what is surprising is the behavior of $\kappa$, the transition-transversion rate ratio. This parameter has seldom been the focus in the literature, but available studies have generally attributed a negative relationship between phylogenetic accuracy and the value of $\kappa$ (Yang 1998; Rosenberg and Kumar 2003). The results of this study, on the other hand, show that the marginal effect of $\kappa$, when averaged over the other parameters, has a positive relationship with accuracy. This is dealt with at length below.
Among the inference methods, NJ results show the greatest difference between the lowest and highest values of each of the parameters, on average, while all the other methods show comparable values of accuracy between them. NJ also shows the least values for $\bar{P}_C$, whether at the lowest or highest parameter values, when compared to the other methods, all of which were somewhat comparable. As far as the topologies are concerned, the highest accuracy is seen for the balanced tree with equal branch lengths (Balanced tree A), followed closely by the random tree topology. Balanced tree B (with unequal branch lengths) has a lower accuracy, in general, than the above two topologies, while the pectinate tree does very poorly. These results are consistent among the inference methods.

The general trend in the association of the sequence parameters with phylogenetic accuracy is the same (as explained above) among all the inference methods and model tree topologies, except when the MP method is used on the datasets obtained from the pectinate tree. This shows a slight decrease in accuracy with increase in $\kappa$. The pectinate tree shape also appears more sensitive to changes in the values of sequence parameters and shows greater differences in accuracy for the lowest and highest parameter values, regardless of the inference method examined. However, the ML methods (PhyML and ML) yield more accurate results than MP or NJ from the pectinate tree datasets.

| Table 2: Marginal effects of low and high values of sequence parameters on phylogenetic accuracy. The overall accuracy, $\bar{P}_C$, of the four phylogenetic methods (PhyML, ML, MP, and NJ) for each model tree at the lowest and highest values of sequence length ($l$), substitution rate $\kappa$ and rate of branch length change ($\lambda$). |
(r), G+C content (θ), transition-transversion rate ratio (κ), and the shape parameter (α), when averaged over all the values of the other parameters (see text).

Since the effect of κ on phylogenetic accuracy is contrary to generally held views, it warrants closer scrutiny. After studying its marginal effects for each inference method and for each topology when taken over all the other parameter values (Table 2), we next studied the interaction between κ and each of the other parameters, taken one at a time (while averaging over the rest of the other parameters), while noting differences among

### Sequence Parameters

<table>
<thead>
<tr>
<th>Method</th>
<th>Model Tree</th>
<th>L</th>
<th>r</th>
<th>θ</th>
<th>κ</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhyML</td>
<td>Balanced A</td>
<td>96.10</td>
<td>96.91</td>
<td>99.46</td>
<td>85.40</td>
<td>96.47</td>
</tr>
<tr>
<td></td>
<td>Balanced B</td>
<td>87.62</td>
<td>91.93</td>
<td>93.46</td>
<td>89.61</td>
<td>89.71</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>94.36</td>
<td>96.72</td>
<td>98.62</td>
<td>84.00</td>
<td>95.48</td>
</tr>
<tr>
<td></td>
<td>Pectinate</td>
<td>68.44</td>
<td>89.64</td>
<td>83.43</td>
<td>70.05</td>
<td>78.45</td>
</tr>
<tr>
<td>ML</td>
<td>Balanced A</td>
<td>94.45</td>
<td>97.87</td>
<td>99.12</td>
<td>95.69</td>
<td>95.71</td>
</tr>
<tr>
<td></td>
<td>Balanced B</td>
<td>87.65</td>
<td>93.82</td>
<td>94.64</td>
<td>90.11</td>
<td>90.22</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>91.26</td>
<td>97.36</td>
<td>98.06</td>
<td>93.67</td>
<td>93.78</td>
</tr>
<tr>
<td></td>
<td>Pectinate</td>
<td>66.03</td>
<td>90.06</td>
<td>86.63</td>
<td>76.59</td>
<td>76.69</td>
</tr>
<tr>
<td>MP</td>
<td>Balanced A</td>
<td>93.65</td>
<td>98.96</td>
<td>98.78</td>
<td>96.41</td>
<td>96.37</td>
</tr>
<tr>
<td></td>
<td>Balanced B</td>
<td>85.51</td>
<td>92.29</td>
<td>93.05</td>
<td>89.11</td>
<td>89.20</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>91.72</td>
<td>98.60</td>
<td>97.81</td>
<td>95.06</td>
<td>95.04</td>
</tr>
<tr>
<td></td>
<td>Pectinate</td>
<td>56.99</td>
<td>79.50</td>
<td>80.22</td>
<td>67.05</td>
<td>66.89</td>
</tr>
<tr>
<td>NJ</td>
<td>Balanced A</td>
<td>88.13</td>
<td>94.22</td>
<td>98.74</td>
<td>90.52</td>
<td>90.57</td>
</tr>
<tr>
<td></td>
<td>Balanced B</td>
<td>80.95</td>
<td>89.28</td>
<td>92.93</td>
<td>84.40</td>
<td>84.45</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>85.03</td>
<td>93.98</td>
<td>97.48</td>
<td>88.73</td>
<td>88.81</td>
</tr>
<tr>
<td></td>
<td>Pectinate</td>
<td>58.17</td>
<td>83.32</td>
<td>81.84</td>
<td>69.21</td>
<td>69.20</td>
</tr>
</tbody>
</table>

189
the model trees and inference methods as well. Finally, we explain the behavior of $\kappa$ from the perspective of substitution saturation.

**Interaction of $\kappa$ with each of the other sequence parameters**

$\kappa$ and $l$: When $l$ and $\kappa$ are varied simultaneously, the difference in $\bar{P}_c$ is almost non-existent between the two values of sequence length, whereas the effect of changes in the value of $\kappa$ is quite obvious, with the accuracy of almost 80 percent when $\kappa = 1$ and reaching practically 100 percent when $\kappa = 20$, in the case of balanced and random tree topologies (not shown). The pectinate tree also shows an increase in accuracy with increase in kappa (except when inferred using MP method), although the extent of accuracy obtained is the least in comparison to other tree shapes for all inference methods, and irrespective of the sequence length ($l = 500$ or $l = 2500$). The percent increase in accuracy for each increase in $\kappa$ was maximum in the case of NJ, improving the accuracy by about 10-20%.

$\kappa$ and $\theta$: No significant trends were observed between Kappa and the base composition (not shown). The accuracy improved with increase in $\kappa$, irrespective of the GC content (not shown). No significant differences were observed among the GC content values, $G+C = 0.50$ yields a slightly better accuracy than extreme $G+C$ values (0.20 or 0.80). The results were consistent among all the inference methods and tree shapes, except again for MP in the case of the pectinate tree, where a slight decrease in accuracy was seen with increase in kappa at all GC content values (not shown).
\( \kappa \text{ and } r \): The interaction of \( \kappa \) with \( r \), shown in Figure 2, is the most important in affecting the accuracy of phylogenetic inference because an increase in the value of \( r \) adds to the effect that \( \kappa \) has on accuracy. Taken individually, phylogenetic accuracy decreases with increase in evolutionary rate and increases with increases in Kappa, regardless of the inference method or tree topology (Figure 2). However, when a high \( r \) is coupled with a high Kappa (e.g., when \( \kappa = 20 \) and \( r \geq 0.4 \)), this results in an increased accuracy of about 90 percent or greater. The decline in accuracy is seen mainly when \( r \geq 0.2 \), and \( \kappa \leq 10 \). At low substitution rates (\( r \leq 0.2 \)), the accuracy is almost close to 100 percent at all kappa values. The extent of accuracy achieved for each interaction of \( r \) and \( \kappa \), differs among the phylogenetic methods (in the order, PhyML > ML > MP > NJ), and with the topology of the model tree (Balanced A > Random > Balanced B > Pectinate tree). Interestingly, when \( \kappa \) is high (\( \kappa = 20 \)), all the inference methods, and tree shapes perform equally well and lead to a similar improved accuracy. The only exception is that when \( r \leq 0.4 \), the accuracy decreases slightly (~ 5%) with increase in Kappa, for MP trees associated with the pectinate topology (Figure 2D). When Balanced topology A was used, the percent increase in accuracy at high rates (\( r \geq 0.4 \)) from \( \kappa = 1 \) to \( \kappa = 20 \) was high, but differed considerably among the inference methods: 30% in ML, 35% in PhyML, 70% in ML, and 90% in NJ (Figure 2A). In case of pectinate tree, the accuracy from \( \kappa = 1 \) to \( \kappa = 20 \) decreases by 3% at low rates (\( r \leq 0.05 \)), and at higher rates (\( r \geq 0.4 \)) eventually increases by almost 50% (Figure 2D).
Figure 2: Joint effects of kappa and substitution rates on the phylogenetic accuracy. The Phylogenetic accuracy ($\bar{P}_C$) plotted against substitution rate, $r$, and transition-transversion rate ratio, $\kappa$, for the four model tree topologies of Fig. 1: (A) Balanced ‘A’, (B) Balanced ‘B’, (C) Random, and (D) Pectinate. $\bar{P}_C$ for the four inference methods, PhyML, ML, MP, and NJ, is represented by the following symbols: filled circle, open circle, cross, and open triangle, respectively. Each point in the graph represents an average over all the replicates for $l=2500$, $\theta = 0.50$, and $\alpha = \infty$. 

192
\( \kappa \) and \( \alpha \): The among-site rate variation, measured as alpha (\( \alpha \)) is an important determinant of phylogenetic accuracy (Yang 1996c). In our study, the results reported above regarding the interaction between Kappa and substitution rate hold true only when the rate is homogenous across the sites (\( \alpha = \infty \)). When the among-sites rate is heterogeneous (\( \alpha = 0.1 \)) the results are relatively constant across all rates, irrespective of the value of Kappa (Figure 3). Figure 3 shows the influence of \( \alpha \) for each combination of substitution rate and kappa, for the inference method PhyML, under the four model tree topologies. When \( \alpha = \infty \) (Fig. 3A), the accuracy is low when Kappa is low and \( r \) is high. However, as the value of Kappa increases, the accuracy dramatically increases. On the other hand, when the rate is heterogeneous among sites, in particular, when only fewer sites experience evolutionary changes (\( \alpha \leq 0.5 \)) the accuracy is high to begin with, and, as a consequence, does not change much with increase in Kappa, even at higher rates (Figure 3B). When the rate was homogeneous, and at high evolutionary rates (\( r = 0.8 \)), NJ showed the highest increase in accuracy with increase in kappa (\( \bar{P}_c \leq 10\% \) for \( \kappa = 1 \) and \( \bar{P}_c \geq 90\% \) when \( \kappa = 20 \); not shown), irrespective of the model tree. In PhyML or ML and MP the percent increase in accuracy from \( \kappa = 1 \) to \( \kappa = 20 \) is approximately 60\% and 40\%, respectively. This percent increase however, is different for Balanced B topology and Pectinate tree shape (e.g., see Fig.3 for PhyML).
Figure 3: Combined effect of Kappa and Substitution rate on phylogenetic accuracy, for the PhyML method. The results are shown for two values of the shape parameter: (A) $\alpha = \infty$ (rate homogeneity among sites), and (B) $\alpha = 0.1$ (rate heterogeneity among sites). The phylogenetic accuracy for the four tree topologies (Fig 1), Balanced ‘A’, Balanced ‘B’, Random, and Pectinate, is represented by the following symbols: filled circle, open circle, cross, and triangle, respectively. Each point in the graph represents an average over all the replicates for $l=2500$, and $\theta=0.50$.

**Nucleotide Substitution Saturation and Role of Transversions**

Distantly diverged gene sequences (or sequences with high substitution rates) often experience substantial substitution saturation, especially in the third codon position of protein-coding genes. This saturation can misrepresent the phylogenetic information contained in the sequences, leading to incorrect phylogenetic inference. Some simple
ways of handling such sequences include avoiding sequences with pair-wise evolutionary
distances larger than 1 (Nei and Kumar 2000) and plotting either number of substitutions
or the transition to transversion ratios against a corrected genetic distance (Xia and Xie
2001). A more sophisticated method is the entropy-based test of substitution saturation
as implemented in DAMBE (Xia et al. 2003). This test, when used on our datasets,
suggested significant saturation in the sequences obtained from simulations with higher
evolutionary rates ($r \geq 0.4$), regardless of the tree topology. The effect was more
profound in dataset with homogenous rate distribution, irrespective of kappa and other
parameters. At high substitution rates, and under rate homogeneity, it is likely that each
site in the sequence will be “hit” multiple times. If the transition rate is much higher
than transversion rate (as in $\kappa = 20$), than it is more likely for one transition to be followed
by another transition simply because another transition is more likely than a transversion.
Under these circumstances, a high transition to transversion rate ratio should lead to
saturation of transitions more than transversions. On the other hand, a transition to
transversion rate ratio of one ($\kappa = 1$) means the two rates are the same. The latter situation
has a relatively higher likelihood of saturation (when compared to $\kappa = 20$) of both the
substitution types, although now transversions are more likely than transitions (twice as
likely, to be exact). We believe that, at high rates ($r \geq 0.4$), the improved accuracy $\geq
90\%$ at high kappa ($\kappa = 20$) in comparison to the accuracy $\leq 40\%$ at low kappa ($\kappa = 1$), is
because there is little saturation of transversions at high kappa and saturation of both
types at low kappa. Phylogenetic trees inferred from the transition-only and transversion-
only sites confirmed this expectation. Phylogenetic trees reconstructed from
transversion-only sites (for dataset with \( r = 0.8, \kappa = 20, \alpha = \infty \)) yielded trees with most of the of the internal branches correct (90\%), while trees inferred using the same dataset but from transition-only sites failed to infer almost any of the internal nodes correctly (and showed an accuracy close to zero), showing saturation in the transitions. However, for datasets with \( r = 0.8, \kappa = 1, \) and \( \alpha = \infty \), neither transition-only nor transversion-only sites gave accurate trees, suggesting that both transitions and transversions had undergone saturation. These results hold for all the tree shapes investigated in this study.

**Discussion**

We have presented here the results of a simulation-based study undertaken to investigate the influence of the following sequence parameters: sequence length (\( l \)), nucleotide substitution rate (\( r \)), base-composition (\( \theta \)), transition-transversion rate ratio (\( \kappa \)) and the shape parameter (\( \alpha \)) (that specifies the extent of heterogeneity in the substitution rate across sites), individually and jointly, on the accuracy of phylogenetic reconstruction by four inference methods: Neighbor-Joining (NJ), Maximum Parsimony (MP) and Maximum Likelihood (ML and PhyML) methods. The model trees used were four 16-taxon tree topologies (Figure 1). All five parameters were varied and combined in a factorial manner (Table 1) and 100 replicates were generated for each of the 576 parameter combinations, for each model tree. The accuracy of phylogenetic reconstruction was measured as the number of correctly inferred internal branches, determined based on the topological distances (Robinson and Foulds 1981; Penny and
Hendy 1985), and averaged over the 100 replicates for a given parameter combination, and finally expressed as $\bar{P}_c$, the average percent of correct branches.

The marginal effects of $l$, $r$, $\theta$, and $\alpha$ held no surprises in their general trends, although the results were not entirely devoid of interest (Table 2). Clearly, however, the most interesting parameter was $\kappa$, the transition-transversion rate ratio. In particular, this parameter was influenced with the changing values of $r$. When $r \leq 0.2$ the phylogenetic accuracy is almost 100%, regardless of the kappa. When rate is high $r \geq 0.4$, however, the accuracy is clearly affected and is correlated to the value of kappa. Of the four inference methods, NJ appears particularly prone to the impact of the combination of these two parameters (Fig. 2). Overall, when the accuracy is assessed for all kappa at high substitution rate ($r = 0.8$) the likelihood methods outperform MP, which outperforms the NJ method. This trend is common to the balanced trees. For the random tree shape, MP is better than the PhyML and ML (at low kappa), and under pectinate tree topology, PhyML outperforms MP, which in turn, does better than ML, and which in turn, is better than NJ (Figure 2). Also, when the rate is heterogeneous among sites, PhyML performs the best and MP the worst, overall (data not shown). When $\kappa \geq 10$ all methods perform equally well, with high level of accuracy. These results are for Balanced A topology (Figure 1A) in particular. Comparison among the model tree topologies shows no difference in the trend observed for kappa (see Figure 2A-C) except for the pectinate tree (Figure 2D).

We also determined whether the effect of kappa is limited to the specifics of the model tree used in the study. Therefore, we examined factors such as the number of taxa
and ultrametric vs. non-ultrametric trees to confirm the consistency in the effect of kappa on the phylogenetic accuracy. The results were similar (not shown).

Here, it seems that high accuracy at high kappa is a result of the phylogenetic signal being present in the transversional substitutions. At high kappa, relying only on transitions results in an incorrect phylogeny and at low kappa both transitions and transversions contribute to incorrect phylogeny, while an inference done using only the transversion events at high kappa almost always produced an accurate tree. This is because of saturation of transitions and transversions at low kappa, (when the rate is high), and saturation of transition events relative to other mutations at high kappa.

Application to Phylogenetic Tree Reconstruction using Real Data

DNA sequences comprising vertebrate mitochondrial (mtDNA) COI sequences from *Masturus lanceolatus* (sunfish), *Homo sapiens* (human), *Bos taurus* (cow), *Balaenoptera musculus* (blue whale), *Pongo pygmaeus* (Bornean orangutan), *Pan troglodytes* (chimpanzee), *Gallus gallus* (chicken), and *Alligator mississippiensis* (American alligator) were obtained based on Xia (2000). As in all protein-coding gene sequences, the third codon position is the most variable and the second is the most conserved. The substitution saturation test of Xia et al. (2003) for the first, second and third codon positions of the mtDNA sequences indicated that the third codon positions had experienced substitution saturation and as a consequence, were not likely to be useful for phylogenetic inference. The third codon positions in vertebrate mitochondrial COI
sequence evolves extremely fast, and exhibits a high kappa, $\kappa \sim 50$ and alpha, $\alpha = 0.70$ (moderate rate homogeneity). Despite the signs of saturation in the third codon position (Xia and Xie 2001; Xia et al. 2003), phylogenetic analysis with this data resulted in a tree that was almost congruent to the (first + second) codon positions that are believed to be conserved and expected to produce an accurate tree (Figure 4a&b). When sites with transitions and transversions were analyzed separately for the third codon position, the two trees were significantly different from each other. In fact, the tree inferred from the transitions-only sites of the third codon positions (Figure 4c) is absurd, whereas the phylogenetic tree with the transversions-only sites of the third codon positions (Figure 4d) showed exactly the same topology as in Figures 4a and 4b. The transversions-only sites of the third codon position also resulted in the correct grouping of two taxa, *Bos taurus* and *Balaenoptera musculus* which are not grouped in the tree inferred from all sites of the third codon position (Figure 4b). Analyzing the transitions and transversions separately in this manner shows that transversions contain stronger phylogenetic signal than transitions, and are capable of masking the distorted signals coming from the saturated transitions sites that may be misleading. This however, may be applicable only under certain conditions (as in this simulation based study or the empirical example used), for instance, high transition to transversion rate ratio, increased evolutionary rate, less rate heterogeneity among sites. Thus, even though a sequence (non-coding or coding) or subsets of a sequence have undergone substantial substitution saturation, it does not mean that it cannot be used for phylogenetic inference. The phylogenetic signal
is likely to be present in the sites that have experienced transversions, which may be useful in the inference of an accurate phylogeny.

The results in this study show that substitution saturation itself may not be a problem, as it may be dealt with in terms of whether it is the transition or transversions sites that have undergone saturation. We have shown with simulated and empirical datasets, that a correct phylogeny can be obtained from saturated datasets when there is saturation among transversions in the sequences.

**Figure 4:** Phylogenetic tree reconstruction based on the vertebrate mitochondrial COI sequences. Maximum Likelihood trees were reconstructed using DNAML program (Felsenstein 1989) with default options, as implemented in DAMBE (Xia and Xie 2001) based on the vertebrate mitochondrial COI sequences using: (a) First and Second codon positions, (b) Third codon positions, (c) Transitions-only sites of third codon positions, and (d) Transversion-only sites of third codon positions.
These results also emphasize the importance of the substitution rate in impacting the accuracy of phylogenetic inference. A low substitution rate usually ensures accurate phylogenetic inference, irrespective of the values of the other parameters used in this study. The results also establish optimal values of parameter combinations for accurate phylogenetic inference, under the different simulation conditions employed. It is expected that these results will be useful in studies that do phylogenetic analyses with empirical DNA sequences.

**Conclusions**

This study has established optimal values of five sequence parameters, singly and in combination, for improving the accuracy of phylogenetic inference, under varying conditions of model tree topology and inference method. An important conclusion of this study is that substitution saturation need not render a dataset unsuitable for phylogenetic analysis. In addition, the results here suggest values of the transition-transversion rate ratio and the evolutionary rate that result in saturation of the signal from the transitions in the data, but where the transversions still carry sufficient signal to offset the distorted signal from the transitions, so as to yield accurately inferred trees.
Acknowledgements

We thank Ohio Supercomputer Resources (OSC) for providing the computational resources for the bulk of the analysis in this study. This research was supported by means of research funds from the University of Dayton: start-up funds for SRG and summer research funding from the Graduate School for BD.
CHAPTER 3.1

WHOLE GENOME COMPARISON OF H1N1 AND H3N2

INFLUENZA A VIRUS

This chapter is published in *Biochim Biophys Acta*, 2008. 1783(2):177-87.
Whole Genome Comparison of H1N1 and H3N2 Influenza A Virus

Bhakti Dwivedi¹ and Sudhindra R. Gadagkar¹*

¹Department of Biology, University of Dayton, Dayton, OH 45469-2320

*Corresponding author

E-mail: gadagkar@notes.udayton.edu
Abstract

Previous influenza A viral pandemics in humans have been attributed to human-specific strains of certain viral subtypes (e.g., H1N1 in 1918) and to genetic re-assortment between avian and human strains of the virus (e.g., H2N2 in 1957 and H3N2 in 1968). This research was undertaken to determine if certain segments (particularly HA and NA, which are critical in ensuring successful infection) were undergoing rapid change in the subtypes, H1N1 and H3N2. DNA sequences were obtained from NCBI and comparative sequence analysis done for each of the eight genomic segments for the following pairs of subtypes/strains (H = Human; A = Avian): H-H1N1 & A- H1N1, and H-H3N2 & A- H1N1. Furthermore, comparisons were also made between collections from two different time periods for the same strain to study any changes over time. The results show that the HA and NA segments are both undergoing purifying selection, with the $d_S/d_N$ ratio (ratio of amino acid changing substitutions to neutral substitutions) <1 in all the comparisons (range: 0.09-0.93). On the other hand, the PB1-F2 gene, which induces apoptosis, is under positive selection in the avian H1N1, with a $d_S/d_N$ ratio >1 (1.5-2.4), and the evolutionary (substitution) pattern is changing in segment PB2 over time, as determined by the Disparity Index test ($I_D= 0.25$). These results show that it may be necessary to keep a close watch on segments other than HA and NA as well, in order to prepare for a mutation that would render the avian strain capable of transmission among humans.
**Keywords:** Influenza virus, positive selection, base composition, evolutionary substitution pattern.
**Introduction**

Pandemic outbreaks of influenza among humans in the last century have been attributed to different subtypes of the influenza A virus, including H1N1 (1918), H2N2 (1957), and H3N2 (1968). The H1N1 virus in 1918 was uncommonly virulent, with approximately 50 million fatalities worldwide in a little more than a year (Crosby 1989). Both the 1957 and 1968 viruses were avian-human reassortants in which few avian gene segments were replaced with the then-circulating human-adapted virus (Scholtissek, Koennecke and Rott 1978; Kawaoka, Krauss and Webster 1989; Schafer *et al.* 1993; Webster, Sharp and Claas 1995). Although, the 1918 virus is believed to be an avian like human adapted virus rather than a human-avian reassortant virus (Reid, Taubenberger and Fanning 2004; Tumpey *et al.* 2005). In fact, two major descendent lineages of the 1918 H1N1 virus and two reassortant lineages of 1968 H3N2 virus, still persist in nature. However, none of these viral descendents approaches the same level of pathogenicity as 1918 virus. Today, the world is anxiously monitoring the effect of the avian H5N1 subtype on bird populations, with the fear that the virus might mutate to become capable of human-to-human transmission. This study focuses on two of the subtypes that have already caused pandemics earlier (H1N1 and H3N2). Whole-genome comparisons have been made within and between subtypes, in order to determine if certain parts of the viral genomes are undergoing rapid change.

Influenza A viruses are negative-stranded RNA viruses of the family Orthomyxoviridae (Figure 1). The total genome length is approximately 12000-13000 nts. They are categorized into subtypes based on two surface proteins: hemagglutinin
(HA) and neuraminidase (NA). There are 15 different HA subtypes and 9 different NA subtypes. While all of these subtypes have been found among influenza A viruses in wild birds, only a few of them are associated with human infection.

**Figure 1. Influenza A virus in detail.** The figure was obtained from the website-
http://www.vetscite.org/publish/articles/000041/img0002.jpg

The influenza A virus genome comprises eight RNA segments encoding at least ten proteins (Figure 1 and 2). Each segment is associated with several nucleoproteins, several molecules of the three subunits of its polymerase (see Figure 1). The heterotrimeric polymerase complex of proteins, PB2, PB1 and PA, are responsible for replication of the virus. Recently, an alternative reading frame of PB1 segment has also been identified that encodes for a peptide, PB1-F2 (shown in red within PB1 segment in Figure 2), which is believed to play a crucial role in virus-induced cell death (Chen *et al.*
The proteins, HA and NA are involved in the attachment and release of the virus from the host cell, while the nucleoprotein (NP) is the nucleocapsid structural protein. Two non-structural proteins, NS1 and NS2, are involved in regulating numerous aspects of the viral life cycle. NS1 protein is shown to be a type IFN antagonist that blocks the host’s ability to make interferon (Diebold et al. 2003). Finally, two membrane proteins, M1 and M2, are involved in nuclear export and maintenance of pH, respectively.

![Figure 2. Influenza A virus Eight RNA segments](image)

Novel influenza A viral strains emerge either by continuous gradual mutation or by genetic recombination between gene segments of two flu viruses. Earlier studies often associate novel HA and NA proteins with the virulence of influenza A virus (Goto and Kawaoka 1998; Steinhauer 1999; Gibbs, Armstrong and Gibbs 2001). In addition, many of the present vaccine and antiviral drugs target primarily HA and NA genes. However, our study shows that besides HA and NA there are other gene segments within the influenza A virus genome that are under the selection pressure and might have a significant role in the emergence of future virulent strains. Thus, any control measure that
targets only one or two gene segments of this virus might not be efficient in the influenza treatment and prevention. Therefore, it becomes important to understand the molecular changes that substantially differentiate the viruses based on their virulence and also the undergoing mutations that might be giving some directionality to an emergence of highly pathogenic viral strain.

Material and Methods

Sequence Collection

The RNA sequences for all eight segments were downloaded from the Influenza Virus Resource at NCBI (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) for human and avian, H1N1 and H3N2 subtypes, for each of several different years (as shown in the Figures 3-5). The available PB1 segments were manually translated to generate the alternative transcript PB1-F2. The PB1-F2 segment for a subtype for a given year was generated by aligning the PB1 segment of that year with the PB1-F2 segment of other year (for which it was available).

Pairwise Alignment

Each of the eight segments was aligned with the corresponding segment of the same subtype (human-H1N1, human-H3N2, avian-H1N1, avian-H3N2) from a different year. Nucleotide sequences were aligned based on the protein alignments using the ClustalW (Thompson, Higgins and Gibson 1994) as implemented in MEGA version 3.1 (Kumar, Tamura and Nei 1994)
Sequence Analysis

The viral segments were characterized by computing the following: G+C content, the relative rate of nonsynonymous to synonymous substitutions ($\omega$) as an indicator of positive selection, and the Disparity Index ($I_D$) test (Kumar and Gadagkar 2001) to determine if two viral genomes were evolving with heterogeneous substitution patterns in a given segment.

The proportion of GC base pairs was determined for each of the eight segments in both the H1N1 and H3N2 lineages over time using MEGA version 3.1 (Kumar, Tamura and Nei 1994). The $\omega$ ratio was computed using the PAML version 3.15 (Yang 1997) with the following parameter specifications under codeml: model=0; NSsites=0; runmode=-2; CodonFreq = 2. The $\omega$ ratio greater than 1.0 are considered positive selection; less than one as purifying selection and a $dN/dS$ of one indicates a neutral selection. Disparity index test ($I_D$) was done to measure the observed differences in evolutionary pattern of nucleotide substitution for each pair of gene segments following the Monte Carlo procedure employed in MEGA version 3.1 (Kumar, Tamura and Nei 1994). The $I_D$ test was done on nucleotide sequences at 4-fold degenerate sites and the 1$^{st}$ and 2$^{nd}$ codon positions (or translated sequences) to detect both the evolutionary and functional pressures respectively.
**Results**

The Comparative sequence analysis of the eight viral segments revealed a uniform GC content of 40-50% in all the subtypes being compared over time (Figure 3). It seems that M segment has the highest GC content in both H3N2 and H1N1. While segment NS1 has one of the highest G+C content in Human-H1N1, it has the lowest G+C content in Avian-H1N1. On the other hand, NP shows a rather uniform G+C content across subtypes and hosts. The base composition, as reflected in the G+C content, in general, does not appear to vary very much across the years for most segments.
However, the change in the base composition was apparently enough for the $I_D$ test (Figure 4) to detect significantly altered nucleotide substitution patterns for a few segments, over the years, both at the 4-fold degenerate sites and the 1$^{st}$ and 2$^{nd}$ codon positions, reflecting evolutionary and functional constraints, respectively. The disparity index analysis showed that the gene segments HA, PA and NP are evolving under different functional pressure (Figure 4). However, no significant differences were noticed in the segments of human-H1N1 and avian-H3N2 lineage in the recent comparisons.
Figure 4. **Disparity Index test ($I_D$).** A comparison of the nucleotide substitution pattern for each viral segment from the four subtypes over time. The eight RNA segments are represented as circles (PB2 = 1; PB1 (including PB1-F2) = 2; PA = 3; HA = 4; NP = 5; NA = 6; M (including M1 and M2) = 7; NS (including NS1 and NS2) = 8). Pink and yellow circles denote significant $I_D$ values at 4-fold degenerate sites and the combined 1<sup>st</sup> and 2<sup>nd</sup> codon positions, respectively.

A comparison of the non-synonymous with synonymous substitutions of the viral sequences with their putative ancestors (previous years) revealed that some segments, particularly PB1-F2, were undergoing positive selection (Figure 5), indicating that the changes in the sequence were significantly directional.
Discussion

A salient finding in this study is that the PB1-F2 segment is under extremely high positive selection pressure in both subtypes in both hosts, with a $\omega$ ratio $>>1$, although in a more recent comparison, it appears to be under purifying selection in the case of...
human-H1N1 (Fig. 3). Positive selection in PB1-F2 has also been recently reported by Obenauer et al. 2006. Perhaps this viral segment, which is involved in apoptosis in the host cells, ought to be monitored closely. In our analyses, the two segments that are usually suspect in this virus, HA and NA, are both under purifying selection, in all comparisons. This is perhaps because sequence data from the years prior to the problem years of pandemics is not available for the corresponding segments. Other segments that also have exhibited positive selection pressures in at least one comparison are NA, M, and NS. In general, our analyses show that some of the RNA segments have been subject to significant changes in both subtypes in both hosts. Whether these changes can result in the kind of virulence shown by the pandemic strains of the previous years is a matter of speculation. More sequence data and a keener understanding of the interaction among the segments would perhaps help understand the dynamics of this virus and its virulence better.
CHAPTER 3.2

MOLECULAR MIMICRY: STRUCTURAL CAMOUFLAGE
OF PROTEINS AND NUCLEIC ACIDS

This chapter is published in Biochim Biophys Acta, 2008. 1783(2):177-87.
Molecular Mimicry: Structural Camouflage of Proteins and Nucleic Acids

Panagiotis A. Tsonis\(^1\)\(^*\) and Bhakti Dwivedi\(^1\)

\(^1\)Department of Biology and Center for Tissue Regeneration and Engineering, University of Dayton, Dayton, OH 45469-2320

\(^*\)Corresponding author

E-mail: panagiotis.tsonis@notes.udayton.edu
Abstract

When it comes to protein specificity and function their three-dimensional structure is the ultimate determinant. Thus, sequences that participate in key parts, such as catalytic sites or DNA binding have been favored and maintained highly conserved during evolution. However, in a reversal of fortune, selection has favored conservation of shapes over sequence, especially when proteins look like nucleic acids. Proteins from pathogens evade the host’s defenses because they are shaped as DNA; others use such a disguise for transcriptional regulation. Several factors are tRNA look-alikes so that they can efficiently control the process of protein synthesis. Molecular mimicry among RNAs could result in a new unexplored level in gene regulation. This comprehensive review outlines this important area and aims to emphasize that molecular mimicry could in fact be more widespread than initially thought and eventually adds a new layer of genetic regulation.

Keywords: Molecular mimicry; Protein-DNA; RNA-tRNA; Protein-tR
An unsuspecting fish swims by a piece of tree bark, a fallen leaf and forest debris. It seems like a very natural environment for the Amazon and Orinoco river basins. But alas! It is the last scenery the fish will ever see! The tree bark is the shell, the leaf is the head and neck and the debris is the loose skin of the matamata turtle (Fig 1a), which feeds on this fish (Blum. 1999). It is called camouflage, but in reality the turtle’s body has been selected to mimic its environment in order to give it an advantage over prey or predators. Such an ‘environmental’ mimicry can be seen in many different animals, such as the Malaysian leaf frog (whose name implies that it looks like a leaf; Fig. 1b) or the walking stick (Fig 1c), an insect that looks like an ordinary twig and thus, fools its predators, which pass by and ignore it!

Figure 1. Examples of camouflage in animals. a: Matamata turtle (*Chelus Fimbriatus*), b: Malaysian leaf frog (*Megophrys Montana*) (Blum. 1999). c: Walking stick (*Diapheromera femorata*) Photo was obtained from website http://home.comcast.net/~zirschkyd/walkingstick.html.
The evolution of camouflage is certainly a very interesting but poorly studied phenomenon, which favors animals that in shape and colorization resemble their surrounding environment. So, it should not be surprising that such a strategy has been exploited at the molecular level as well. Molecular mimicry between proteins is a well documented event, where bacteria, viruses or parasites evade the immune response, agonists and antagonists bind to receptors, or when an autoimmune response is initiated because a pathogen shares sequence similarities with native proteins (Hall 1994; Davies 1997). Protein-protein mimicry results because similar amino acid sequences are present in different proteins and it should not be surprising that it exists. The topic of protein-protein mimicry is beyond the scope of this review. Here we are concerned with molecular mimicry between proteins and nucleic acids, which are completely different molecules. This mimicry relates to domains of proteins that have assumed a very similar three dimensional structure to nucleic acids, DNA or RNA (Nissen, Kjeldgaard and Nyborg 2000). Such mimicry as we will see serves a good purpose either for survival or for function during several processes, such as replication, transcription, splicing and protein synthesis. We will also cover mimicry among different RNA molecules because it can also illustrate quite nicely the advantage of the strategy as well as the impact that this might have in gene regulation.

**Proteins disguised as DNA**

Why a protein should evolve looking like DNA? The best guess would be because it fools other proteins that normally bind DNA. The advantage of this is obvious when
such a protein is encoded by an invading genome. The unique structural features of
double helix DNA include negatively charged phosphate groups joining two adjacent
nucleotides, diverse major and minor grooves, nucleotides that provide either polar or
hydrophobic nature. An effective DNA mimic usually presents a surface that is
complementary in shape and similar in chemical composition (charge) to DNA. These
DNA mimics aim towards DNA-binding proteins and thus act as DNA molecule(Dryden
2006; Dryden and Tock 2006).

How pathogens use mimicry to battle the host defenses and survive

Protecting the genome

One of the first examples of DNA mimicry by a protein was ocr, which stands for
‘overcome classical restriction.’ This is a bacteriophage T7 protein, which is expressed
after the phage infects E. coli. Bacteria usually protect themselves from foreign DNA by
means of restriction enzymes. In this context, the ocr protein competitively inhibits type
I DNA restriction enzymes from binding to their DNA target, thereby allowing a
successful phage propagation through the bacterial population and thus acts as anti-
restriction protein (Studier 1975; Dunn et al. 1981; Mark and Studier 1981;
Bandyopadhyay et al. 1985; Kruger et al. 1985; Bickle and Kruger 1993; Blackstock et
al. 2001; Sturrock et al. 2001; Atanasiu et al. 2002). In addition to type I restriction
enzymes, ocr also binds to E. coli RNA polymerase(Ratner 1974). The shape of ocr is an
elongated 26 KDa dimeric protein (Mark and Studier 1981; Atanasiu et al. 2001;
Blackstock et al. 2001), which suggest structural similarity to the DNA and thus complementarity to the binding groove on the target protein (Mernagh et al. 1998; Powell, Connolly and Dryden 1998; Powell, Dryden and Murray 1998). X-ray crystallographic structure of ocr reveals a remarkable molecular mimicry of 24 bp of bent B-form DNA (Walkinshaw et al. 2002). The distribution of carboxylic groups on the surface of the protein mimics the distribution of phosphate groups on target DNA (Fig 2a). Also, the distinctive bend in the ocr structure is similar to the bend induced in DNA upon binding to type I restriction enzyme. It has also been shown that ocr can inhibit all known families of type I restriction enzymes, regardless of their target DNA sequence (Walkinshaw et al. 2002). Besides ocr, other antirestriction proteins such as T3 phage SAMase protein and Ard family proteins also seem to be potential DNA mimics (Bickle and Kruger 1993; Belogurov and Delver 1995; Zavilgelsky 2000; Wilkins 2002; Tock and Dryden 2005).

The Bacillus subtilis bacteriophage genome contains uracil instead of thymine and upon infection its genome is targeted by the key DNA repair enzyme uracil-DNA glycosylase (UDG). To counter, the phage expresses uracil-DNA glycosylase inhibitor (Ugi), a protein that protects the uracil containing phage DNA by irreversibly inhibiting the host’s UDG. In comparing the three dimensional structures of UDG-bound DNA (Mol et al. 1995) (Fig 2b) and UDG-Ugi (Putnam et al. 1999) (Fig 2b’) we can see that DNA and Ugi bind the UDG’s active site. This suggests mimicry between DNA and Ugi. Thus, Ugi protects the phage’s DNA by competitively binding to UDG’s active site. True
enough when the β1 strand of Ugi is structurally superimposed with DNA the conserved acidic residues align with the phosphate groups (as shown in Fig 2c).

Another protein that could be a DNA mimic is HI1450 from *Heamophilus influenza*. The mimicry is largely inferred by its structural similarity to Ugi (Parsons, Yeh and Orban 2004). As we can see in Fig 2c the HI1450 β1 strand also aligns with the DNA backbone mimicking part of the minor groove. Recently, it has been shown that HI1450 binds to a dsDNA-binding protein HU-α and that this interaction is mediated by residues in the β1 strand (Parsons, Liu and Orban 2005). HU-α is a histone-like protein in *Haemophilus influenza* that is involved in gene regulation, DNA bending, stabilization and DNA repair (Azam *et al.* 1999; Dorman and Deighan 2003; van Noort *et al.* 2004). This evidence strongly supports the view that HI1450 may act as a DNA mimic and could have a major role in inhibiting or regulating the nucleoid protein. Similar distribution of glutamate and aspartate residues can be seen in Dinl protein from *E. coli* (Ramirez *et al.* 2000). DNA damage prevents replication by DNA polymerase, resulting in the formation of single-stranded DNA (ssDNA). The binding of the Rec protein to ssDNA induces a DNA repair process known as SOS response (Friedberg E.G., Walker and Siede 1995). DinI is a protein that down regulates the SOS response by mimicking ssDNA (Ramirez *et al.* 2000). Biochemical studies suggest that DinI can prevent ssDNA binding to RecA and can also stabilize the ssDNA-RecA cofilaments (Voloshin *et al.* 2001; Lusetti *et al.* 2004). DinI seems to regulate RecA by competing with ssDNA for the same binding site of RecA molecule (Ramirez *et al.* 2000; Yoshimasa *et al.* 2003).
Figure 2. Pathogen proteins that mimic DNA. a: Bacteriophage T7 Ocr. Superimposition of phosphate groups of 12 bp B-DNA onto the 12 carboxyl groups of Ocr dimer (in blue ribbon form with dimer interface shown as red line). The phosphate groups are colored yellow (phosphorus) and purple (oxygen). The carboxyl groups are colored red (oxygen) and black.
(carbon). The sugar backbones of DNA helix are shown in two green shades. The base pairs are not shown for clarity. The black lines represent the direction of fitted DNA (Walkinshaw et al. 2002). b: UDG-dsDNA and b’: UDG-Ugi complex: The *E. coli* UDG (blue) is bound to DNA (green sticks), in fig 2b and to Ugi (brown) in fig 2b’ (Putnam and Tainer 2005). The arrow points to the interacting β1 strand. c: Structural alignment of β1 strands of HI1450 (blue) and Ugi (red) with dsDNA. The conserved acidic residues are labeled and highlighted along the strand. The dsDNA is shown in yellow with the phosphate groups colored and labeled in green (Parsons, Yeh and Orban 2004). d: The proposed model of *E. coli* gyrase (GyrA59)-DNA complex (Cabral et al. 1997). The DNA duplex is shown in green and red ribbon and the active-site tyrosine is denoted by a yellow star. d’: The proposed model of *E. coli* gyrase (GyrA59)-MfpA complex (Hegde et al. 2005). MfpA is shown in green. In both figures, *E. coli* gyrase is represented as surface. Surfaces are colored according to electrostatic charge distribution, blue for positive and red for negative.

**Drug resistance**

One of the drugs that are used to fight tuberculosis is fluoroquinolone. This drug inhibits bacterial DNA replication by binding reversibly to the gyrase-DNA complex, thereby causing all sorts of topological problems. Gyrase is a bacterial type II topoisomerase, which mediates catenation and decatenation of double stranded DNA. As one would expect, the infectious bacteria often become resistant to drugs through mutations in the gyrase itself, weakening the binding of the drug while retaining enough gyrase activity for bacterial survival. MfpA protein, however, provides an alternative resistance mechanism to fluoroquinolone. It competes with DNA for the gyrase surface.
thus protecting gyrase from fluoroquinolone until the enzyme has completed its function. MfpA is a member of the penta-peptide repeat family of bacterial proteins, where every fifth amino acid is either leucine or phenylalanine (Bateman, Murzin and Teichmann 1998). The structure of MfpA is a dimer composed of a right handed β helix with a size, shape and charge distribution remarkably similar to B-form DNA (Hegde et al. 2005).

Even though, the exact mechanism of interaction between MfpA and gyrase is not known, studies imply that MfpA might be mimicking a 30 bp segment of B-form DNA and thus could be capable of a direct interaction with gyrase (Hegde et al. 2005) (as shown in Fig 2d & d’). It has been suggested that other proteins of the penta-peptide repeat family could function as inhibitors of DNA binding by a similar mechanism as MfpA inhibition of gyrase (Bateman et al. 2004).

**DNA mimics involved in transcription**

**Shuttling factors to the nucleus**

Selective nuclear transport is mediated by nuclear localization signals (NLSs) and transport factors as karyopherins or importins. The structure of karyopherin α is composed of ten tandem arm repeats, which form a twisted helical shape with a large groove (Conti et al. 1998) as shown in Figure 3a. The structure of karyopherin α indicates a DNA double helix mimic, however, it has a larger rise per helical turn. NLSs peptides of several other nuclear proteins are known to be involved in DNA binding in the nucleus, for example NLS of transcription factor LEF-1 specifically binds to the
major groove of cognate DNA (Love et al. 1995). Interestingly, NLS peptides can also
bind to several lysine and arginine residues of the large groove of karyopherin α (Conti et
al. 1998). Therefore, it has been proposed that the nuclear import factors such as
karyopherin α have NLS-binding sites that mimic DNA targets in the nucleus (Fig 3a).
DNA mimicry by karyopherin α might also have some role in importing functional
transcription factors into the nucleus by mimicking their DNA binding sites.
**Figure 3. Protein-DNA mimicry at transcriptional level.** a: Karyopherin α (Kapα50) in complex with SV40 T antigen NLS peptide. The conserved residues of H3 of Kapα50 arms are colored in orange. The SV40 T antigen peptides are shown in blue (Conti et al. 1998). b,b’: The positions of σ factor, before and after DNA binding (Mekler et al. 2002). The RNA polymerase subunits-αI, αII, β, β’ and ω are colored in light blue, dark blue, orange, green and gray respectively. The template and non-template DNA are shown in gray and pink respectively and σ factor subunits are shown in yellow ribbons and spheres. In b’ the β’ pincers are rotated 16° into the active center. White circles and diamonds indicate location of probes used in the study. c: TBP binding domain of Drosophila TAFII230 in the TBP-TAFII complex. TBP is shown as green ribbon and TAFII230 as surface. c’: TBP-binding surface of the TATA box (surface) in the TBA-TATA complex (Liu et al. 1998). d: Interaction between the Rho residues Arg-66 and Asp-78 with cytosine. d’: G-C interaction in dsDNA (Bogden et al. 1999). Cytosines are shown in blue and residues from Rho pairing with cytosines are in black (stereo images).

**RNA polymerase recruitment to the promoter**

The prokaryotic RNA polymerase (holoenzyme) is composed of five subunits (β, β’ two α and ω) along with the σ factor. The σ factor is responsible for directing the polymerase to the promoter region. Once transcription has become processive the σ factor dissociates from the complex. When the structure of RNA polymerase was examined it was found that the σ factor’s position was different in the holoenzyme (before DNA binding) and in the open complex (bound to DNA) as shown in Figure 3b & b’ respectively. In fact, in the holoenzyme the 1.1 subunit of the σ factor occupied the region within the active center cleft where DNA is placed in the open form (Mekler et al.)
In other words, the σ1.1 subunit is a DNA mimic and corresponds to the downstream location of the DNA in the open complex. In order for the open complex to be formed, σ1.1 must be displaced. This structural data explain why deletion of σ1.1 affects the kinetics but not the stability of the open complex (Wilson and Dombroski 1997; Vuthoori et al. 2001).

The eukaryotic RNA polymerase II is recruited to the promoter region with the orchestrated action of the basal apparatus that is composed of many different factors. The first to contact the TATA box in the promoter region is TFIID, which is composed of the TATA binding protein (TBP) and several TBP-associated factors (TAFs). In order for the TBP to become active it must be relieved from TAFs and this is mediated by the recruitment of TFIIB. TAFs have a characteristic structure that mimics DNA double helix thus blocking the binding of TBP to TATA box. The solution structure of Drosophila TBP-TAFII230 complex (Liu et al. 1998) shows remarkable similarities with the TBP-TATA box complex (Kim, Nikolov and Burley 1993; Kim and Burley 1994) (Fig 3c & c’). The arch-shaped surface of TAFII 230 is similar to partially unwound minor groove of TATA bound to TBP. Both TAFII 230 and the TATA box have an extensive hydrophobic surface. In addition, the negative charged side chains of TAFII 230 mimic the positions of phosphate groups of the TATA box. A significant difference between TATA and TAFII 230 is that the TATA box surface is highly symmetric, whereas the TAFII 230 surface is asymmetric. This asymmetry results in closer matching of surface charges and hydrophobicity between TAFII 230 and TBP. This implies a direct inhibition of TBP interaction with the TATA box and thus regulating gene expression. As in the
case of σ1.1 displacement to form the open complex described above, such DNA mimics provide an obvious competition for regulation by activators or repressors.

**Termination**

In prokaryotes, transcription ends by two different processes, the Rho-independent and the Rho-dependent one. As the RNA polymerase transcribes the end of the unit, the sequences at the 3’ end can form a stem and loop structure. In the Rho-independent termination this structure can induce by itself the opening of the flap and the transcript falls off the RNA polymerase. In other cases termination has to be mediated by Rho. Rho is a protein with ATPase and helicase activity. Also, it has three strong nucleic acid binding sites with a preference for pyrimidines. The three-dimensional structure of Rho bound to an oligo-C sequence reveals that the interaction of a cytosine with Rho’s amino acids mimics the interaction of C-G in DNA (Fig 3d & d’), explaining, thus, the preference of Rho for pyrimidines (Bogden *et al.* 1999).

**RNAs disguised as tRNAs**

tRNAs are key players in protein synthesis. The 3-D structure of tRNA is a clover leaf L-shape with an anticodon at one end and an amino acid acceptor at the other end. The acceptor site is recognized by the aminoacyl tRNA synthetases, which charge each tRNA with its cognate amino acid. This site interacts with the peptidyl transferase center of the ribosome allowing the formation of peptide bonds. The anticodon is the sequence that interacts with the codons of the mRNA an interaction that is crucial to the decoding
process. However, the tRNA structure has many look-alikes. They can be found in other RNA structures or even in proteins.

**Splicing**

Studies of mtRNA splicing in *Neurospora crassa* and in yeast have pinpointed aminoacyl-tRNA synthetases (aaRSs) and other RNA binding proteins that have adopted a function in group I splicing. Most likely these proteins recognize structures in group I introns that resemble their RNA binding sites. In the case of *Neurospora crassa* it seems that tyrosyl-tRNA synthetase is “fooled” and recognizes a tRNA-like structural motif in two (LSU and ND1) group I intron’s catalytic cores (Caprara et al. 1996). Obviously, the catalytic core of these ribozymes is folded in a similar fashion as a tRNA. Such an association has evolved as a splicing mechanism of these introns. When LSU and tRNA$_{Tyr}$ are superimposed it is evident that domain L9 aligns with the acceptor stem, domain P7 with the variable loop (V) and domains P4/P6 with the anticodon stem (Fig. 4a).
Figure 4. RNA-tRNA mimicry. a: LSU intron (red) is superimposed with tRNA\textsuperscript{TYr} (blue), with the sites protected by aaRS shown in yellow (Tsonis 2003). b: tRNA with its acceptor stem and anticodon indicated, original picture obtained from www.cytographica.com/animations/. c: Valylatable tymoviral TYMV tRNA-like structure (Fechter \textit{et al.} 2001). Note the pseudoknot at the acceptor site. d: BMV tRNA-like structure (Fechter \textit{et al.} 2001). The eight tRNA-like structural domains are labeled (A, B1, B2, B3, C, D, E, F). Domains A and B2-3 mimic the amino acid acceptor stem and anticodon respectively. e: TMV tRNA-like structure (Fechter \textit{et al.} 2001). The three tRNA-like domains are indicated as D1, D2 and D3, connected by a central core (labeled as C). The acceptor stem and anticodon are mimicked by domain D1 and D2, respectively.
Helping viruses to replicate

Another set of RNAs that resemble tRNA, the so-called tRNA-like structures (TLS) has been found at the 3’ of many genomic RNA plant viruses (for reviews see (Hall 1979; Haenni, Joshi and Chapeville 1982; Dreher and Hall 1988; Dreher 1999; Fechter et al. 2001)). The tRNA-binding domains can be charged with three different amino acids, valine, tyrosine, or histidine, a reaction very similar to the aminoacylation of canonical tRNAs. In addition, the folded tRNA-binding domains are structurally very similar to the canonical tRNA except the pseudo-knotted amino accepting stem, which is unique to all viral tRNA mimics (Figure 4b-e). This pseudoknot is equivalent to the amino acid acceptor branch of tRNA. The amino acid acceptor branch of canonical tRNAs is composed of 12bp (7 from acceptor stem and 5 from T-stem). However, in Turnip yellow mosaic virus (TYMV) RNA it is formed from three helical segments each with a specified number of nucleotides (Rietveld et al. 1982; Dumas et al. 1987)(Fig 4c). Unlike TYMV tRNA-like domain, Brome mosaic virus (BMV) and Tobacco mosaic virus (TMV) RNA (Figure 4d & e) do not show a clear L-shape cover leaf architecture of tRNA (Felden et al. 1994; Felden et al. 1996). It has been shown that TLS are required for virus viability (Skuzeski, Bozarth and Dreher 1996) and are also crucial for gene expression (Gallie and Walbot 1990; Gallie and Kobayashi 1994). Recently, it has been found that valylated TLS of TYMV entraps ribosomes and directs them to the internal initiation of translation (with the valine N-terminally incorporated) for protein synthesis (Barends et al. 2004). Removal of TLS\textsuperscript{TYMV} completely abolishes polypeptide synthesis. One of the major explanations why these RNAs have adapted a tRNA-like structure as
well as aminoacylation is because replication of these viruses’ RNA genome requires interactions with a translation elongation factor (Dreher, Tsai and Skuzeski 1996; Giege 1996). Support for this comes from results indicating that aminoacylated viral RNAs can interact with elongation factors EF-Tu and EF-1a and that EF-Tu is a part of the bacteriophage Qβ RNA replicase. The mimicry has been evolved to reflect the surface of the elongator tRNAs that normally bind to elongation factors, thus, allowing efficient replication of the viral genome.

**Proteins disguised as tRNA**

This kind of mimicry is perhaps the most widespread and striking. Since tRNAs are major players in proteins synthesis, their mimics are also involved in this process. The charged tRNA enters the A-site of the ribosome bound to EF-Tu. In there the tRNA has two tasks, one is the decoding, by the interaction of the anticodon and the mRNA’s codon, and the other is the peptide bonding at its acceptor site. After this cycle, the tRNA-peptide translocates to the P-site leaving the A-site empty for the next tRNA. The translocation is facilitated by another elongation factor, EF-G. When a stop codon is located at the A-site, signaling the end of translation, it is recognized by a new factor, the release factor (RF). This factor is able to release the polypeptide from the tRNA. After this and the removal of RF the ribosomal subunits must dissociate, because if stayed as is unspecific translation can resume. This is mediated by the ribosome recycling factor (RRF). From this short trip to protein synthesis we can see the A-site of the ribosome is visited and must be occupied (except by tRNA) by several proteins. The accommodation
of tRNAs as they move from the A- to the P-site is very specific, thus if any protein is to occupy the same place it must have the shape and the size of a tRNA. The result is the incredible molecular mimicry between tRNA, EF-G, RF and RRF (Czworkowski et al. 1994; Nissen et al. 1995; AlKaradaghi et al. 1996; Nissen et al. 1999; Selmer et al. 1999; Song et al. 2000) (Fig 5a). In figure 5a we can see how the EF-Tu-bound tRNA looks very similar in size and shape with EF-G and RF. Also, the structural similarity between tRNA and RRF is spectacular (see also below). The interesting case here is that all these different proteins do not share any significant sequence homologies with each other, they are unrelated proteins.
Figure 5. Protein-tRNA mimicry. a: *Thermus thermophilus* RRF, yeast Phe-tRNA^Phe_, *Thermus aquaticus* EF-Tu:GDP:Phe-tRNA^Phe_, *Thermus thermophilus* EF-G:GDP and human eRF1. Atoms are colored- red (oxygen), green (carbon), blue (nitrogen), and yellow (phosphorus or sulfur) and protein domains are shown with their relevant numbers (Nakamura 2001). b: The binding of paromomycin onto helix 44 (H44) and its role in H44 codon-anticodon interactions. c: Localization of ETA (orange) and H44 (blue) in the 40S subunit with respect to eEF2 (grey), c’: The interaction between phosphates of ETA-bound βTAD and eEF2 (DIPH; a modified
histidine) is very similar to the sugar phosphate backbone of A192 and A1493 of helix 44 (Jorgensen et al. 2005).

Molecular mimicry can also account for the action of antibiotics, whose one major target is protein synthesis (Carter et al. 2000). For decoding and recognition of the codon by the tRNA a significant conformational change must occur in helix 44 of the small subunit, which is at the heart of the decoding center. Two universally conserved amino acids A1492 and A1493 (E. coli numbering) must be flipped out. This costly rearrangement makes sure that the cognate tRNA will be recognized. Antibiotics bind that exact place and result in flipped A1492 and A1493 (see Fig 5b). Thus, the conformation for tRNA and codon recognition is available without cost, resulting in non-cognate tRNA recognition, which in turn results in incorporation of the wrong amino acids and mutated proteins. A similar strategy has been devised by exotoxin A (ETA) from Pseudomonas aeruginosa in order to achieve universal recognition of its substrate eEF2 (Fig. 5c,c’) (Jorgensen et al. 2005). Notably, the toxin-bound βTAB (a non-hydrolysable NAD⁺ analogue) presents its phosphates to its target residue of eEF2 in an orientation that coincides with the phosphate backbones of A1492 and A1493. Thus, this mimicry maximizes Pseudomonas’ survival chances because any resistance to its toxin by the target organism would entail mutations in eEF2, which is crucial for function.
Insights into the mimicry issues

This review outlines how structural molecular mimicry can be used as strategy for one particular regulation or another. The obvious conclusion for the use of such mimicry is that if a protein looks like, say tRNA, it must do something that a tRNA does (at least structurally). However, we must stress here that this does not necessarily have to hold at the functional level as well. A good example is the structure and function of RF and RRF. RFs contain a GGQ motif that is responsible for hydrolysis (the equivalent of CCA in tRNA) and a tripeptide (PA/VT in RF1 and SPF in RF2) that mimics the anticodon and are responsible for discriminating against the different stop codons. The distance of the tripeptide and the GGQ in the crystal structures of isolated RF1 and RF2, however, is not compatible with the distance between the decoding and the peptidyl transferase center (Vestergaard et al. 2001). When the structure of ribosome bound RF2 was visualized by cryo-EM reconstructions, it was shown that RF2 undergoes changes in relation to the isolated structure. This allowed for accommodation of the tripeptide in the decoding center of the 30S subunit (Klaholz et al. 2003; Rawat et al. 2003). However, more details were received when the structure of ribosome-bound RF1 and RF2 were solved via X-ray crystallography with a resolution that allowed the protein helices and the backbone of RNAs to be resolved. It was then found that the tip of α5 helix (not part of the tripeptide tRNA mimic but common to both factors) could act as a discriminator of nucleotides at position 1 and that the tripeptide-containing anticodon loop, which is specific to RF1 and RF2, could recognize bases 2 and 3 (Petry et al. 2005). The conclusion is that recognition of stop codons involves other elements apart the anticodon tripeptide mimic. Similar
situation can be argued with the mimicry of the ribosome recycling factor (RRF). Structural studies have shown that despite the similar structure and dimensions to tRNA, RRF’s orientation on the ribosome is different that those of tRNA and the tRNA mimicking domains of EF-G (Agrawal et al. 2004). The studies on RFs and RRF reveal that structural mimicry is one aspect of molecular mimicry and functional mimicry is another. This underscores the importance to understand the biological significance behind structural mimicry. In this respect we can have two layers of mimicry. One is purely structural and has a strict topological aspect. In other words, for a protein to occupy a space normally occupied by tRNA it has to look like it or at least should have very similar dimensions. Once this has been achieved the protein has to do a task. In the case of RF it has to recognize a codon and hydrolyze the peptide bond (as a tRNA does). This second layer is a functional mimicry and the protein components mediating this do not have to look structurally like tRNA. For mimics that are used for competition (such as or cot TBP), functional mimicry might not be that important, but for mimics that also perform a function (such as catalysis) it must involve other elements of the protein or the interacting environment. How do we approach these issues experimentally? In the case of tRNA mimics they have to interact with other ribosomal components. Detailed studies on intermolecular interactions as derived from X-ray crystallographic data of the different mimics could identify possible common threads. Or if these are unique will extend our knowledge about functional mimicry. It is also important to learn more about these interactions as a function of time, which can be achieved by other emerging technologies, such as hydrogen exchange coupled to mass spectrometry (MS). When a protein is placed
in deuterium solution, hydrogen atoms are exchanged with deuterium atoms. Protein mass increase is then measured with high-resolution MS and the place of deuterium incorporation is determined in peptide fragments. With such a method protein dynamics and movements can be studied after binding and over time (Wales and Engen 2006). This might provide global signatures that can be attributed to structural or functional mimics. Another insight can be obtained by comparing evolution of sequences of DNA or RNA mimics. Are there common motifs that might account for structural mimicry? Such studies have not really been attempted largely because obvious sequence similarities have not been noticed. However, it might be possible to identify common signatures. For example, it is well established that a small number of amino acids (such as arginine, lysine, asparagine, glutamic acid) are the main participants with DNA interactions. It could be possible to identify such signatures that might indicate mimicry, but we seriously doubt that this will be definite. After all, these amino acids are also specifically interacting with DNA mediating the function of DNA binding proteins aside from mimics as well (Tsonis 2003).

**Evolution of molecular mimicry**

Molecular mimicry must be an ancient mechanism that probably originated in the RNA world, where specific three dimensional structures conferred unique functions to different RNAs (Nakamura 2001). It is easy to imagine that similarly shaped RNAs would have common partners and such an association might have been maintained during the transition to RNA-protein and DNA-protein world.
Natural selection has resulted in animal camouflage for their advantage. As we discussed in the beginning of the review in many cases looking like something else is a good survival strategy. Molecular mimicry has evolved for several different reasons. One is the obvious survival benefit as it is highlighted in cases where pathogens use molecular mimicry to fool the host’s defenses and survive. Another reason should be the acquisition of more efficient function. This is reflected quite well in the molecular mimicry of factors involved in protein synthesis. In this process evolution has favored proteins that look like tRNAs so that they perform well within the constrains of size and shape that are imposed by the ribosome. In this case there is selection for shape and not necessarily for sequences. RFs, RRF and EF-G do not share sequence homology. Another very important reason, we believe, is competition in gene regulation. A good example here is the competition of TAFs for the TATA box, which is the main promoter regulatory element. Such mimicry for regulatory purposes could be much more widespread than we know or appreciate. RNAs can fold in many different structures due to complimentarity of nucleotide sequences. Thus, structures could be more accessible for mimicking. For example, ribosomal protein S15 recognizes both the 16S rRNA and mRNA. The recognition is possible by molecular mimicry of 16S rRNA and mRNA. The sites that are required for S15 binding assume similar conformations (Mathy et al. 2004). Similar mimicry could also be the case involving other ribosomal proteins as well as regulation of gene expression by threonyl-tRNA synthetase. Except from its enzymatic role in aminoacylation, threonyl-tRNA synthetase regulates its own mRNA expression by
binding to the operator. The operator has a conformation mimicking the anticodon arm structure (Caillet et al. 2003). A different example for mimicry-directed regulation is the spliceosome’s U1A protein, which binds to both U1 snRNA and its own 3’UTR (Jovine et al. 1996). Given the plethora of RNA structures, we anticipate that molecular mimicry must play a much more pivotal role in regulation than previously suspected. The examples of molecular mimicry that we have so far could only reflect the tip of an iceberg. As more structural studies become available, especially those involving interactions of proteins and RNA, molecular mimicry could take a more prominent role in genetic regulation and evolution.
SUMMARY

Molecular phylogenetic studies usually relate topological/phylogenetic accuracy to the accuracy of sequence alignments and tree reconstruction methods. However, there are very few studies that expressly relate the accuracy of phylogenetic inference to evolutionary features of the sequences themselves in the alignment rather than any source of error during the process of sequence alignment or due to the choice of the method of phylogenetic inference. The focus of the present study was to determine the relationship between the number of gaps and phylogenetic accuracy, when the gaps are introduced in an alignment to reflect indel (insertion/deletion) events during the evolution of DNA sequences. Other studies have studied the effect of true but “gappy” sequence alignments on phylogenetic analysis in the context of the proportion of incomplete taxa in an alignment, the availability of partial sequences such as ESTs, or extent of divergence among the sequences. In this study, we have investigated the relationship between the number of gaps and phylogenetic accuracy, when the gaps are introduced in an alignment to reflect indel (insertion/deletion) events during the evolution of DNA sequences, and thus always carry phylogenetic signal. DNA sequence alignments were generated using computer simulation, while varying several sequence parameters and introducing both
substitution and insertion/deletion events, along a 16-taxon model tree, and systematically varying the expected proportion of gapped sites. The resulting alignments were subjected to commonly used gap treatment methods and methods of phylogenetic inference. Our results show that, when the alignment gaps reflect indel events without error, and the number of gapped sites per sequence is ≤20 percent of the sequence length, all the inference methods used perform well in accurately inferring the phylogeny. However, when the number of gaps is large (≥80 percent), likelihood based methods (Bayesian and PhyML), and Maximum Parsimony clearly outperform the other inference methods when the gaps are treated as missing data. Within the MP and Bayesian methods, the inference of the phylogeny was significantly more accurate when each gapped site was treated as an additional binary character state than when the gaps are treated as missing data. These results suggest that these inference methods, in combination with the indel coding method of gap treatment may be more useful than the other methods employed in this study, when the sequences in an alignment contain large number of gaps, as in the case of highly diverged sequences. Finally, our results also show that it is more difficult to accurately infer the phylogeny from an alignment where a greater proportion of gaps reflect deletion events rather than insertion events in the evolutionary history of the sequences in the alignment.

In this study, our goal was to determine the relationship between the number of gaps and the accuracy of phylogenetic inference, using only true simulated alignments. Inaccuracies in alignment certainly are sources of uncertainty and errors in real data analysis and this must be taken into consideration in empirical studies. The present
findings, however, can be further used to establish relationship between the proportions of gaps and alignment accuracy. This can also help us to understand how alignment errors can influence topological accuracy when looking at the number of gaps in the sequence alignment.

In this simulation-based study, we also examined the individual and combined effects of some of the salient nucleotide sequence parameters on the accuracy of phylogenetic inference, using computer simulation. DNA sequences were generated using 16-taxon model trees and the HKY model of nucleotide substitution. Simulation was done while systematically varying the sequence length ($l$), substitution rate ($r$), base-composition (expressed as the G+C content, $\theta$), transition-transversion rate ratio ($\kappa$), and the rate heterogeneity among sites ($\alpha$). Phylogenetic reconstruction on each of these alignments was done using the most commonly used inference methods. Accuracy of the inferred trees was determined in terms of the number of branches incorrectly reconstructed in comparison to the model tree. Surprisingly, a positive correlation was obtained between $\kappa$ and phylogenetic accuracy, which contradicts other studies. These results, however, hold true only when the rate of sequence divergence is high. We explained the increased accuracy with increase in kappa as a result of transversion substitutions that have not undergone substantial substitution saturation.

This is a simulation-based study and is confined to certain specific parameters and methods of phylogenetic inference used in this study. However, we believe that the
results obtained from these studies are sufficiently general to be useful to the community of molecular phylogeneticists.


Escherichia coli threonyl-tRNA synthetase as both an enzyme and a regulator of gene expression." Molecular Microbiology 47(4): 961-974.


Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J.,
Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. and
Origin of Eukaryotes." Proceedings of the National Academy of Sciences of the
United States of America 86(17): 6661-6665.


for Protein-Coding DNA-Sequences." Molecular Biology and Evolution 11(5):
725-736.

"Evolution of a noncoding region of the chloroplast genome." Molecular
Phylogenetics and Evolution 2: 52-64.

by a human influenza A virus." Proceedings of the National Academy of Sciences
of the United States of America 95(17): 10224-10228.


Sinauer Associates, Inc.

Graybeal, A. (1998). "Is it better to add taxa or characters to a difficult phylogenetic


Lanciotti, R. S., Roehrig, J. T., Deubel, V., Smith, J., Parker, M., Steele, K., Crise, B., Volpe, K. E., Crabtree, M. B., Scherret, J. H., Hall, R. A., MacKenzie, J. S.,


ribosomal protein S15 relies on both mimicry and site differentiation." Molecular Microbiology 52(3): 661-675.


DNA glycosylase in complex with a protein inhibitor - protein mimicry of DNA.


273


281


Studier, F. W. (1975). "Gene 0.3 of bacteriophage T7 acts to overcome the DNA restriction system of the host." Journal of Molecular Biology 94: 283-295.


Takahashi, K. and Nei, M. (2000). "Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and
maximum likelihood when a large number of sequences are used." Molecular Biology and Evolution 17(8): 1251-1258.


