Using Structural Information in Modeling and Multiple Alignments for Phylogenetics

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Xueliang Pan, M.A.S., M.S.

* * * * *

The Ohio State University

2008

Dissertation Committee:

Dennis K. Pearl, Adviser
Laura Kubatko
Steven N. MacEachern

Approved by

Adviser

GRADUATE PROGRAM in BIOSTATISTICS
Phylogenetic studies are increasingly based on structural biological data and on statistical formalization. That leads to the study of improved models and of extracting the maximum information from sequence data. In this research, I have proposed to incorporate the structural information in two areas that relate to phylogenetic inference: one is to use a spatial dependent substitution model for likelihood calculation in phylogenetic inference; the other is to use a gap distance measure for MSA evaluation. While the first application is to using an improved substitution models in phylogenetic inference, the second one focuses on the quality of the MSA produced by different alignment procedures.

The proposed spatial dependent model was based on our observation that the amino acids close to the functional core region tend to be conservative and these on the periphery are likely subject to mutation. So we proposed a substitution model with its rate for each amino acid dependent on its distance to the catalytic active center, or the functional core of the protein. The SD model has been implemented in the framework of Bayesian hierarchical model, the posterior distribution of the model parameters and the phylogenetic inference was estimated simultaneously using the MCMC Metropolis-Hastings algorithm. The SD model has been applied to 11 enzymes that are primarily to central metabolism that are found in species from all
Kingdoms. The SD model is much better than the currently available substitution models in terms of fitness consistently for all examples.

Besides the modeling, we also use the structural information of the sequences for MSA evaluation. The fixed alignments used in phylogenetic studies are derived in advance of phylogenetic analysis. There are many different ways to construct these alignments. The gap measurement proposed here is based on the assumption of structural superposition, and it not only evaluates the alignment quality of those sequences with structural information, but also those sequences without structural information. This measurement can be used to select a better MSA for our phylogenetic analysis. Furthermore, it may lead to improvement of the sequence alignment.
To my daughter, my wife and my parents.
ACKNOWLEDGMENTS

I first thank God, who has shown His amazing grace to me throughout the years of my graduate studies at the Ohio State University.

I thank my advisor, Dr. Dennis K. Pearl, for his inspiring and encouraging advice to guide me to a deeper understanding of the research, and his invaluable comments during the dissertation work. I also give special thanks to Dr. J. Dennis Pollack for his long term collaboration and for providing the amino acid datasets.

I am very grateful to Dr. Laura Kubatko and Dr. Steven N. MacEachern. They have given me valuable feedback and suggestions on my research and dissertation.

Thanks to my friend Hui Cao at the Department of Computer Science and Engineering for his help in programming. I thank the OSU Mathematical Bioscience Institute for providing a super-computer account.

My gratitude also goes to my friends on and off OSU campus for their friendship and support.

Finally I thank my dearest wife, my best friend, Jun Liu. I would never have been able to finish this work without her help, support, and encouragement.
VITA

1993 ...........................................B.S. Environmental Engineering

2002 .............................................M.S. Civil Engineering

2002 .............................................Master of Applied Statistics

PUBLICATIONS

Research Publications


FIELDS OF STUDY

Major Field: Biostatistics

Studies in:

Phylogenetics  Prof. Dennis K. Pearl
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>VITA</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xii</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>1</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Amino Acid Sequence and Crystallographic Information</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Phylogeny Inference</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Protein Structure and the Phylogeny Inference</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>17</td>
</tr>
<tr>
<td>SPATIAL DEPENDENT SUBSTITUTION MODEL</td>
<td>17</td>
</tr>
<tr>
<td>2.1 Substitution Model</td>
<td>18</td>
</tr>
<tr>
<td>2.1.1 The Development of Substitution Models</td>
<td>18</td>
</tr>
<tr>
<td>2.1.2 Amino Acid Substitution Models</td>
<td>20</td>
</tr>
<tr>
<td>2.1.3 Spatial Dependent Substitution Model</td>
<td>27</td>
</tr>
<tr>
<td>2.2 Likelihood Calculation and Posterior Tree Distribution</td>
<td>34</td>
</tr>
<tr>
<td>2.3 Bayesian Inference with MCMC Method</td>
<td>38</td>
</tr>
<tr>
<td>2.4 Implementation of the SD Substitution Model</td>
<td>48</td>
</tr>
<tr>
<td>2.4.1 Parameter Estimation</td>
<td>48</td>
</tr>
<tr>
<td>2.4.2 MCMC Convergence Diagnosis</td>
<td>51</td>
</tr>
<tr>
<td>2.4.3 Model Comparison</td>
<td>58</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>A segment of an amino acid alignment of some PGK sequences</td>
<td>2</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Four distinct aspects of a protein's structure</td>
<td>3</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>Bifurcating phylogenetic tree with five external nodes (a) Unrooted tree (b) Rooted tree</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>The crystallographic structure of a PGK (1VPE) molecule</td>
<td>24</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Correlation of averaged distance from amino acid alpha carbons (Ca) to the anchor-atoms at their catalytic active centers (C/AC) vs the positional-ConSurf conservation score (PCCS)</td>
<td>26</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Some examples of rate factor f(r) as a function of r. All the f(r) functions have an overall mean of 1</td>
<td>29</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>Example of a rooted tree with five species (leaves) for deriving the likelihood</td>
<td>34</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>The local rearrangement transition strategy in Li, Pearl and Doss’s model</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Mau and Newton's proposal of changing to a new tree from the current tree in two steps</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2.7</td>
<td>Time series of log likelihood (LnL), tree length (TL), and gamma shape parameter (alpha)</td>
<td>52</td>
</tr>
<tr>
<td>Figure 2.8</td>
<td>Examples of log likelihood autocorrelation plot and triple distance plot</td>
<td>53</td>
</tr>
<tr>
<td>Figure 2.9</td>
<td>SQR Plot of for Markov chains of various lengths</td>
<td>54</td>
</tr>
</tbody>
</table>
Figure 2.10 Distance density diagnostic, applied to two chains of length 1 million (panels (A)-(C)) and two chains of length 5 million (panels (D)-(F)) ..........................56

Figure 2.11 An example of split support correlation plot on two independent runs.........57

Figure 3.1 Time series plot, and its auto-correlation function (ACF) of the Log posterior Density for the 5PGK dataset in two independent runs based on the Full (Gamma+ SS4) model ......................................................................................................................71

Figure 3.2 Tree Distance Density Diagnostic for the 5PGK data set based on the Full model .................................................................................................................................72

Figure 3.3 Tree Distance Density Diagnostic for the 87PGK data set based on the continuous SD model ..........................................................................................................................83

Figure 3.4 Split support correlation plot on two independent runs for the 87 PGK dataset based on the SD continuous model ................................................................................................84

Figure 3.5 Site rates estimated from the Gamma model, Discrete SD model, continuous SD model and the continuous SD approximation model vs. site distance for the 5 PGK dataset ..........................................................................................................................85

Figure 4.1: Two hypothetically aligned amino acid sequences: A and B………………..93

Figure 4.2: The 3D structures of two phospholipase sequences 1bp2 and 1poba and the alignment .................................................................................................................................94

Figure 4.3: GDS under the null hypothesis and the alternative hypothesis......................98

Figure 4.4: Box plot of potential gap distance vs. gap size for the Serine proteinase MSA .................................................................................................................................102

Figure 4.5: Scatter plot of gap size vs. gap distance (A) and gap size vs. GD_P (B) for the MSA Alpha amylase, catalytic domain, aligned using the 3D-Coffee ...........103

Figure 4.6: the location of the gaps with GD_P larger than 0.5 for the MSA of Alpha amylase, catalytic domain, aligned using the 3D-Coffee. These are potential poorly-aligned gaps ...........................................................................................................104

Figure 4.7: Gap distance score (GDS) for ten sequence sets .....................................107

Figure 4.8: Comparing the five alignment methods based on APDB and iRMSD ......108
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>The twenty amino acids and their abbreviation and corresponding codons</td>
<td>3</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>The 11 selected enzyme datasets</td>
<td>64</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Catalytic-active enzyme center (CAC) selection</td>
<td>65</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>The spearman correlation of the Euclidean distance measure among five PGK scaffolds</td>
<td>67</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Summaries of model fit and computational times for the five-sequence PGK dataset</td>
<td>73</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Summary of log likelihood (lnL), substitution rate at selected partitions and computational times for 11 enzyme datasets</td>
<td>77</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Summaries of continuous SD models for the five PGK dataset</td>
<td>80</td>
</tr>
<tr>
<td>Table 3.7</td>
<td>Comparison of the discrete and continuous SD models for the five PGK dataset</td>
<td>81</td>
</tr>
<tr>
<td>Table 3.8</td>
<td>Comparison of the discrete and continuous SD models for the 87 PGK dataset</td>
<td>82</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Multiple Alignment Sequences selected from Homstrad</td>
<td>100</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>The count of the gaps with GD-P larger than 0.5 according to their sequence name and its corresponding reference sequence</td>
<td>105</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Compare APDB, iRMSD, and GDS for all ten datasets aligned with five different methods</td>
<td>109</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Phylogenetic analysis based on molecular data, such as DNA sequences, RNA sequences, or amino acid sequences, is a widely used technique in biological studies. With the advancement of molecular biology, a tremendous amount of protein sequence data with 3D structures has also became available. Using 3D structural information in phylogenetics is an emerging research topic.

In this dissertation, we will present two novel approaches for using the tertiary structures of the amino acid sequences in phylogenetics research: one is to evaluate the multiple sequence alignments from different alignment software, and the other is to reconstruct phylogeny with a substitution model dependent on the 3D structure using a Bayesian model.

Basic knowledge and terminologies about amino acid structure, phylogenetics, and some commonly used methods are introduced in Chapter one. Chapter two describes the structural dependent substitution model and the Bayesian inference technique to exploit it for phylogenetics. The application of this new substitution model is illustrated in Chapter three. Chapter four focuses on the new evaluation method for multiple sequence alignments using structural information with application. Finally, we conclude
with the strengths and weaknesses of the two proposed approaches and future research
directions.

1.1 Amino Acid Sequence and Crystallographic Information

An *amino acid* is a molecule that contains both amine and carboxyl functional
groups, and used as the basic components of proteins. There are twenty "standard" amino
acids used by cells in protein biosynthesis and are specified by the general genetic code
(Table 1.1). An *amino acid sequence* is an ordered group of amino acid residues,
connected by peptide bonds, in the chain of a protein. It is generally ordered from the N-
terminal end containing a free amino group to the C-terminal end containing a free
carboxyl group. Sequences sharing a common ancestor are called *homologous sequences*.
The homologous sequences can be aligned to identify regions of similarity that may be a
consequence of functional, structural, or evolutionary relationships between the
sequences. These *aligned sequences* are typically represented as rows within a matrix
with inserted gaps between the residues. An example of aligned sequences is show in
Figure 1.1.

<table>
<thead>
<tr>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGK_SCHMA</td>
</tr>
<tr>
<td>tr_Q9GPM4</td>
</tr>
<tr>
<td>PGK1_RHINI</td>
</tr>
<tr>
<td>PGK_SCHPO</td>
</tr>
<tr>
<td>PGK_ASPOR</td>
</tr>
<tr>
<td>PGK_EMENI</td>
</tr>
<tr>
<td>PGK_PENCH</td>
</tr>
<tr>
<td>PGK_GLOMO</td>
</tr>
<tr>
<td>PGK_NEUCR</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.1:** A segment of an amino acid alignment of some PGK sequences.
<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbreviations</th>
<th>RNA Codon(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>A Ala</td>
<td>GCU, GCC, GCA, GCG</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C Cys</td>
<td>UGU, UGC</td>
</tr>
<tr>
<td>Aspartate</td>
<td>D Asp</td>
<td>GAU, GAC</td>
</tr>
<tr>
<td>Glutamate</td>
<td>E Glu</td>
<td>GAA, GAG</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F Phe</td>
<td>UUU, UUC</td>
</tr>
<tr>
<td>Glycine</td>
<td>G Gly</td>
<td>GGU, GGC, GGA, GGG</td>
</tr>
<tr>
<td>Histidine</td>
<td>H His</td>
<td>CAU, CAC</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I Ile</td>
<td>AUU, AUC, AUA</td>
</tr>
<tr>
<td>Lysine</td>
<td>K Lys</td>
<td>AAA, AAG</td>
</tr>
<tr>
<td>Leucine</td>
<td>L Leu</td>
<td>UUA, UUG, CUU, CUC, CUA, CUG</td>
</tr>
<tr>
<td>Methionine</td>
<td>M Met</td>
<td>AUG</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N Asn</td>
<td>AAU, AAC</td>
</tr>
<tr>
<td>Proline</td>
<td>P Pro</td>
<td>CCU, CCC, CCA, CCG</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q Gln</td>
<td>CAA, CAG</td>
</tr>
<tr>
<td>Arginine</td>
<td>R Arg</td>
<td>CGU, CGC, CGA, CGG, AGA, AGG</td>
</tr>
<tr>
<td>Serine</td>
<td>S Ser</td>
<td>UCU, UCC, UCA, UCG, AGU, AGC</td>
</tr>
<tr>
<td>Threonine</td>
<td>T Thr</td>
<td>ACU, ACC, ACA, ACG</td>
</tr>
<tr>
<td>Valine</td>
<td>V Val</td>
<td>GUU, GUC, GUA, GUG</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W Trp</td>
<td>UGG</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y Tyr</td>
<td>UAU, UAC</td>
</tr>
</tbody>
</table>

**Table 1.1:** The twenty amino acids, their abbreviations and corresponding codons
Figure 1.1 illustrates part of the aligned sequences of several phosphoglycerate kinase (PGK) enzymes. The left part of the figure lists the nine names of the PGK sequences. The right part of the figure is a matrix, with fifty columns and seven rows. Each letter represents an amino acid and the character ‘_’ represents a possible deletion, or alternatively an insertion at the same site in other sequences. The term *indel* is often used to emphasize that it is unknown which of these two possibilities actually occurred. Each row represents one ordered amino acid sequence of this enzyme, and each column, called *site*, represents the aligned amino acids at that site. More details regarding the PGK enzymes will be discussed in Chapter 3.

Proteins are essential parts of all living organisms and participate in every process within cells. They are made up of amino acids, arranged in a linear chain and folded into unique three-dimensional protein structures. Protein lengths range from the lowest 40-50 up to several thousands with average of around 300 amino acids in the chain. Biochemists describe the protein structure by focusing on four distinct aspects: primary, secondary, tertiary and quaternary structure (Figure 1.2). *Primary structure* expresses the ordered amino acid sequence. *Secondary structure* describes the amino acids as they arrange themselves into highly regular sub-structures, such as alpha helix or strands of beta sheet. Secondary structures are locally defined, meaning that there can be many different secondary substructures present in one single protein molecule. *Tertiary structure* (the focus of this dissertation) characterizes three-dimensional structure of a single protein molecule; a spatial arrangement of the amino acids and the secondary structures. *Quaternary structure* illustrates the complex of several protein molecules or
polypeptide chains, which function as part of a larger assembly or protein complex. This structure is only for very large aggregates formed from protein subunits, does not apply to a simple protein.

The Protein Data Bank (PDB) of the Research Collaboratory for Structural Bioinformatics (www.pdb.org) provides a variety of tools and resources for studying the structures of biological molecules and their relationships to sequence, function, and disease (Berman et al. 2002). It currently holds more than 35,000 identified protein structures, and over 40,000 total structures. The amount of data for the structures with known 3D structure has increased dramatically during the last decade (only 510 total in 1990, but 13,600 in 2000, and more than 35,000 in 2007). This enables us to study how these structures relate to the evolutionary process of many proteins, which will help to refine the models of evolution used to infer phylogeny.
Figure 1.2: Four distinct aspects of a protein’s structure
1.2 Phylogeny Inference

*A phylogenetic tree*, also called an evolutionary tree or a tree of life, is a tree showing the evolutionary interrelationships among various species that are believed to have a common ancestor based on the *aligned sequences* of molecular (such as sequences of amino acids or nucleotides or single nucleotide polymorphisms) or observable characteristics (such as morphology or behavioral data). There are two types of trees, rooted or unrooted trees. In a rooted tree, there exists a common ancestor of all species under study and there is a unique evolutionary path from the ancestor to any species. An unrooted tree is a tree that only specifies the relationships among the species but does not provide any information about their common ancestor and does not define the evolutionary path. While unrooted trees can always be generated from rooted ones by simply omitting the root, a root cannot be inferred from an unrooted tree since it has no information about the ancestry, and the evolution history.

A tree has three major parts: the *external nodes*, the *internal nodes* and the *branches* (or equivalently, the *edges*). The external nodes represent the observed data, and are often called leaves or Operational Taxonomic Units (OTUs). The internal nodes stand for those ancestral data that cannot be observed. Each external node is connected to an internal node with one branch (edge). The internal nodes are linked with other internal nodes or with external nodes by branches. If the number of nodes connected to each internal node is three, except the root node, which has two, then the tree is called a *bifurcating tree* (Figure 1.3). If there is one or more internal nodes connected to more than three nodes (more than two for the root node), a tree is called a *multifurcating tree*. 
A phylogenetic tree is typically described with two components: the topology, referring to the branch pattern linking the external nodes, and the branch lengths, giving the length of each of the individual branches. Figure 1.3 shows an unrooted and a rooted bifurcating tree. They have five leaves, or external nodes (labeled 1, 2, 3, 4, and 5), and three internal nodes (6, 7, 8). Node 9 in (b) is the root node for the rooted tree. There are 7 branches for the unrooted tree, and 8 branches for the rooted tree in the figure.

For a rooted bifurcating tree with \(N\) (\(N>3\)) external nodes, there are \((N-1)\) internal nodes, \((2N-2)\) branches, and a total of \((2N-3)!/(2^{N-2}(N-2)!)\) different possible labeled topologies (Felsenstein 2004). The number of possible topologies of a tree increases dramatically as the number of external nodes increase. For example, a tree with 10 external nodes has more than 34 millions topologies, and a tree with 20 external nodes has \(8.2 \times 10^{21}\) topologies. The number of possible trees is enormous \((2.7 \times 10^{76})\) when the external nodes increase to more than 50 (Felsenstein 2004, pp 42).

Figure 2: Bifurcating phylogenetic tree with five external nodes (a) Unrooted tree (b) Rooted tree.
The process to reconstruct a phylogenetic tree from the observed data is called
*phylogenetic inference*. The most commonly used methods to infer phylogenies can be
categorized as distance-based methods, parsimony methods, and likelihood-based
methods.

The distance-based method is based on pair-wise evolutionary distances between
the external nodes of a tree. There are several ways to calculate these distances. For
example, starting from those molecular sequences, the distances can be calculated to
approximate the number of substitutions expected under an explicit model of evolution
such as those described in chapter two. Next, phylogenies are constructed that best match
the resulting matrix of pair-wise distances using techniques such as the Unweighted
Paired-Group Method with Arithmetic mean (UPGMA) method (Sokal and Michener
1958) or the Neighbor-Joining (NJ) method (Saitou and Nei 1987).

The UPGMA is the simplest method of tree construction. It was originally
developed for constructing trees that reflect the phenotypic similarities between OTUs,
but it can also be used to construct phylogenetic trees if all lineages are evolving at the
same rate (although this is often times an unrealistic assumption). For this purpose the
number of observed nucleotide or amino-acid substitutions can be used. UPGMA
employs a sequential clustering algorithm, in which local topological relationships are
identified in the order of their similarity, and the phylogenetic tree is built stepwisely.
First, identify the two OTUs that are most similar to each other among all the OTUs and
then treat these two as a new single OTU, referred to as a composite OTU. Subsequently,
identify the pair with the highest similarity among the new group of OTUs, and so on,
until only two OTUs left. It produces a rooted tree. This algorithm is rarely used these
days due to its limitations. UPGMA has been replaced by the neighbor joining algorithm
as the most favored distance-based method (a method using a distance matrix). The
neighbor joining algorithm allows for unequal rates of evolution between sequences and
is designed specifically for molecular data.

The neighbor-joining method is a special case of the star decomposition method.
At first, a distance matrix is estimated from the raw data, and the initial tree is a star tree.
Then a modified distance matrix is constructed in which the separation between each pair
of nodes is adjusted on the basis of their average divergence from all other nodes. The
tree is constructed by linking the least-distant pair of nodes in this modified matrix. When
two nodes are linked, their common ancestral node is added to the tree and the terminal
nodes with their respective branches are removed from the tree. This pruning process
converts the newly added common ancestor into a terminal node on a tree of reduced size.
At each stage in the process two terminal nodes are replaced by one new node. The
process is complete when two nodes remain, separated by a single branch. The NJ
method does not require that all lineages have diverged by equal amounts. The PHYLIP
program implements both the neighbor-joining method of Saitou and Nei (1987) and the

The maximum parsimony method (MP) seeks to find a tree with the minimum
total numbers of evolutionary steps required to explain the observed data (Fitch 1971). In
contrast to the distance-based methods that only produce one tree based on their
algorithm, the maximum parsimony methods explore all possible trees for the one with
the fewest amount of evolutionary steps. It usually involves two separate steps: (1) search through the space of trees, and (2) for a given tree topology, find the minimum numbers of changes needed to explain the data. Fitch’s algorithm focuses on the second search step. With the assumption that the evolution of each position is independent, and a cost function of converting from one state to another, first, a set of possible states (e.g. nucleotides) at each internal node is determined starting from the leaves to root; and then pick ancestral states for internal nodes to find the minimum cost for a given tree topology from root to leaves. The cost function is originally defined as 1, if the two characters are different; and 0 if they are the same. The cost function can be expanded to assign different cost dependent on the value of the two characters. The parsimony method is typically used to only reconstruct the topology and ignore the branch length, since the cost function is independent of the branch length.

David Swofford’s program, PAUP*: Phylogenetic Analysis Using Parsimony (and other methods), is one of the most widely used software package for the inference of evolutionary trees using parsimony method (Swofford 2003). As demonstrated in 1978 by Felsenstein, maximum parsimony can be inconsistent under certain conditions, known as long-branch attraction. And he promoted to use the likelihood method.

Likelihood based method is to find a tree based on the likelihood of the observed data and the possible trees. The maximum likelihood and Bayesian methods are both likelihood based.

Felsenstein (1981) introduced the basic algorithm for computing the likelihood of observed data given a tree (topology and branch length). The maximum likelihood
method finds a tree with the maximum likelihood which is calculated based on a substitution model as detailed in Chapter two. Because of the complex nature of the likelihood function, even for the simplest models and the most advanced stochastic search algorithms (Lewis 1998; Salter and Pearl 2000), it is hard to find a global maximum of the likelihood. There are many good maximum likelihood program for phylogenetic inference, such as PHYLIP (Felsenstein 1989), PAUP*(Swofford 2003), GARLI (Zwickl 2006) and RAxML (Stamatakis 2006).

The Bayesian method attempts to find the posterior distribution of trees given the observed data and may summarize this distribution using a consensus tree. The Bayesian method has several advantages over the other methods in terms of easy interpretation of results, ability to incorporate prior information and some computational advantages, according to Larget and Simon (1999). However, it has some difficulties to evaluate whether the computational algorithms used have reached convergence to the desired posterior probability of tree topologies; for instance, whether all islands with high probability have been visited. Despite these pitfalls, the Bayesian approach has been increasingly applied in the literature and has gained many advocates who consider it to be the best approach in phylogenetic analysis (Huelsenbeck, Ronquist, Nielsen and Bollback 2001).

Both MP and likelihood-based methods need to find a best tree according to certain criteria from the tree space. With the fact that the tree space, which contains all possible tree topologies, is very large when the number of external nodes is large, an exhaustive search of the entire space is impossible. Numerous heuristic search methods
and algorithms have been proposed to search for the best tree. These typically include a
method to move from tree-to-tree in this space such as using the branch swapping
methods of nearest neighbor interchanges (NNI), subtree pruning and regrafting (SPR),
or tree bisection and reconnection (TBR). For unrooted trees, NNI effectively dissolves
an internal branch connecting four subtrees, and reconnects these subtrees in one of the
two alternative ways. For a tree with N external nodes and (N-3) internal edges, there are
2(N-3) neighboring topologies that can be found each time. SPR consists of removing a
branch from a tree with subtree attached to it, then reinserting the subtree in all possible
places. There are 2N-8 neighboring tree topologies generated at each internal branch and
2N-6 for an external branch. The total number of neighbors examined by SPR is then
4(N-3)(N-2). The set of neighboring trees that can be produced by a single NNI move is a
subset of that produced by SPR. TBR differs from SPR in that any branch of the subtree
may be reconnected to all possible places instead of just the branch of the subtree that
was cut. So the total number for each internal node will be (2N₁-3)(2N₂-3), where n₁ is
the number of external nodes in the subtree and N₂ = N-N₁. There is no general analytic
formula for the total number of neighboring trees that can be examined by TBR (though
it is of order N³). The set of possible neighboring trees examined by SPR is a subset of
that by TBR. NNI, SPR and TBR methods cannot exhaust the possible tree topologies
More details of these and other heuristic search methods and discussion can be found in
1.3 Protein Structure and the Phylogeny Inference

In this dissertation, the protein structure will be used in phylogenetic inference in two areas: (1) to evaluate multiple sequence alignment (MSA), and (2) to calculate the likelihood of the observed data and infer phylogeny using the Bayesian approach.

The phylogeny inference process starts with a group of nucleotide or AA sequences sharing the same ancestors (homologous sequences). These sequences need to be aligned properly by adding insertions or deletions to produce an MSA. The MSA is then the input for almost all phylogeny inference software, such as PHYLIP (Felsenstein 1989), PAUP* (Swofford 2003), RAxML (Stamatakis 2006), BAMBE (Simon and Larget 1999), and MrBayes (Ronquist and Huelsenbeck, 2003). The protein structure can be incorporated in both of these connected procedures.

Methods to generate the MSA generally fall into three different categories. (1) Most common alignment algorithm assign penalty to base substitution, a larger penalty for opening a gap and another penalty to continuous a gap, which is often optimized using the dynamic programming techniques. This method is only based on the linear structure of the sequences. CLUSTALW and its menu driven variant CLUSTALX (Thompson Higgins and Gibson 1994) is one of the most popular alignment software implemented with DP. (2) Algorithms using structural information to produce structural-based MSA, such as LSQMAN (Kleywegt and Jones 1994), look for optimal superposition where intermolecular distances are minimized. (3) Structural information and sequence similarity together, specifically, are used to generate a substitution model and/or gap penalty scheme based on the 3D structure, such as the 3D-Coffee (Notredame
et al. 2004). It is widely believed that structural information can improve the accuracy of the MSA of proteins (O’Sullivan et al. 2004), but to evaluate the MSA generated by various methods remains a challenge.

To calculate the likelihood of the data for a given tree, one needs to make some assumptions about the evolutionary process occurring at each node, such as the substitution model. The substitution model, defined by the substitution matrix, describes the rate of evolution from one state to another state. For nucleotide, it is a $4 \times 4$ matrix with one or more parameters to be estimated, in which each cell $(i,j)$ represents the change rate of evolving from state $i$ to state $j$. Some famous models include Jukes-Cantor model (1969), HYK model (Felsenstein 1981) and the HYK model (Hasegawa, Kishino and Yano, 1985).

For amino acids, it can be described as a $20 \times 20$ matrix. Most of the substitution models for amino acids are empirical models, such as the Dayhoff model (Dayhoff, Schwartz and Orcutt, 1978), the Mtrev model (Adachi and Hasegawa, 1996), the Mtam model (Cao et al., 1998; Yang, Nielsen, and Hasegawa, 1998), the WAG model (Wheland and Goldman, 2001), the Rtrev model (Dimmic et al., 2002), the Cprev model (Adachi et al., 2000), the Vt model (Muller and Vingron, 2000) and the Blosum62 model (Henikoff and Henikoff, 1992).

During the last decade, it is widely recognized that the substitution rates are not the same across all sites. Substitution models allowing for site variation have been explored in the molecular evolution literature as early as the 1970’s (Uzzell and Corbin, 1971, Jin and Nei, 1990) and have been used in phylogeny inference since the 1990s.
(Yang 1993, Mayrose, Friedman and Pupko, 2005). Models that allow site variation to be
dependent upon the 3D structural have been explored recently by Parisi and Echave
(2001), Robinson et al. (2003) and Rodrigue et al. (2005). However, there is no
substantial work incorporating the 3D structural dependent substitution model into
substitution models for phylogenetic inference (Rodrigue, Philippe and Lartillot 2006).
We are going to discuss more details about the substitution models in chapter 2.
CHAPTER 2

SPATIAL DEPENDENT SUBSTITUTION MODEL

The substitution model is the core for the calculation of the probability of the data for a given phylogenetic tree. Molecular evolution is reasonably well modeled by finite state continuous time Markov chains that results in the observed aligned sequences from each of many species. In the process of evolution, a substitution matrix describes the instantaneous rate at which one state (such as A C G T for DNA sequences) changes to another state (negative values will occur on the diagonal of this matrix indicating a lessening probability of seeing the same state as the length of time moves away from zero). The transition probability is the probability of changing from a given state to another state over a branch that has a length of time t. The transition matrix can be calculated from the substitution matrix directly with the assumption of stationarity.

With the assumption of the independent evolution among each branch at each site, the branch with two end nodes and length t is the building block for the estimation of the probability of the observed data given a tree. The probability of the observed data for a given tree can be calculated as the product of the likelihood of all branches at each site, which can be obtained directly from the transition probability matrix.
2.1 Substitution Model

The substitution model is represented by an $S \times S$ instantaneous rate matrix, $Q$, where $S$ is the number of states. The transition matrix $P(t)$ is also represented by an $S \times S$ matrix, whose elements give the probability of changing from a state $i$ to another state $j$ in an amount of time $t$. $P(t) = \text{Exp}(Qt)$. A majority of these evolution models are time reversible, i.e. $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$, where $\pi_i$ is the stationary probability of state $i$. There are no biological reasons for using time reversible models, but it is used since no information about the direction of time is contained in the data when a molecular clock is not assumed and an unrooted tree is being constructed. Substitution models can be classified by the size of the alphabet they support. For data based on the presence or absence of specific characteristics, a 2-letter alphabet is used. For nucleotide models, amino acid models, or codon models, there would be 4, 20 or 64 states respectively. Different models proposed during the last several decades make different assumptions about the evolutionary process. It is very important to select a proper substitution model. Models with assumptions that poorly fit the observations could lead to erroneous inferences (Kuhner and Felsenstein 1994; Huelsenbeck 1998; Felsenstein 2004).

2.1.1 The Development of Nucleotide Substitution Models

A number of evolution models have been developed to describe the evolutionary processes acting on DNA sequences, protein sequences and others (such
as codon). The rate matrix for DNA substitution models is a 4×4 matrix. The simplest model for DNA evolution is the Jukes-Cantor model (1969) which has only one parameter. The Jukes-Cantor model assumes that the rates of changes between any pair of distinct nucleotides are the same. Later on, it was recognized that transitions (mutation from a purine (A or G) to another purine nucleotide or from a pyrimidine (C or T) to another pyrimidine nucleotide), occur more frequently than transversions (the substitution of a purine for a pyrimidine or vice versa). Kimura (1980) then introduced a model allowing a transition/transversion inequality rate. The F84 model (Felsenstein 1981; Kishino and Hasegawa, 1989; Felsenstein and Churchill, 1996) and the HYK model (Hasegawa et al. 1985) further relaxed the assumption of equal base stationary probabilities, giving these latter models five total parameters. Tamura and Nei (1993) introduced a six-parameter model that allowed different transition rates for pyrimidines and purines. The F84 and HYK models are a special version of the Tamura and Nei model. While the F84 model assumed there was no difference in the transition rates for pyrimidine and purine, HYK model assumed that the ratio of the two rates was proportional to the ratio of the numbers of pyrimidine and purine. The general time reversible (GTR) model (Lanave et. al. 1984; Gu and Li 1996; Waddell and Steel 1997; Rodriguez et al. 2005) had nine parameters that allowed the rate to be different for each pair of nucleotides.
2.1.2 Amino Acid Substitution models

The earliest substitution model, Eck and Dayhoff’s model (1966), is not surprisingly for amino acid data since amino acid information was available far earlier than DNA or codon information. For amino acids, most of the evolution models are empirical, with fixed empirically estimated rates of change. Such models include the JTT model (Jones, Taylor and Thornton, 1992) and the WAG (Whelan and Goldman, 2001) model, and each model may be only suitable for some proteins. A GTR model (with 208 parameters) based on limited sequences is not recommended for maximum likelihood based phylogenetic inference since the model is based on too few observations to reliably estimate the entire parameter vector (Felsenstein, 2004). Similarly in Bayesian inference the GTR model for amino acids introduces un-needed extra variability to the inference.

These empirical evolution models, including the Dayhoff model, JTT model and WAG model (also see references in Felsenstein, 2004) assume all sites evolve independently at the same rate. This assumption has drawn criticism from biologists because it ignores the biological reality that the evolution processes are not at the same rate at all sites, neither are they likely to be site-to-site independent.

One of the most important recent advances in the reconstruction of phylogenetic trees is the consideration of the heterogeneity of evolutionary rates among sites (Liò and Goldman, 1998). The biological basis of heterogeneous mutation rate among sites includes a reflection of the influence of nearest neighbors on mutation
rate. Stacking energies along the molecule, helix configuration, supercoiling, and DNA intrinsic curvature (that is sequence dependent) change the solvent accessibility and thus base reactivity. The rate of evolution is also subject to the protein structure/function selection pressures.

Models proposed by Yang (1993, 1994) and others (Gu, Fu and Li 1995; Waddell, Penny and Moore 1997; Mayrose, Friedman and Pupko 2005) allow each site to evolve at different rates following some distribution. Yang (1994) shows that the discrete Gamma model, with as few as four categories of evolutionary rates chosen to approximate a Gamma distribution, performs very well compared to the continuous Gamma model. It is also considerably more practical computationally. The discrete Gamma model (Yang, 1994) is still one of the most commonly used models in likelihood-based phylogenetics. This model has significantly improved the model fitness. However, it still assumes that all sites evolve independently.

Gu, Fu and Li (1995) further suggested a gamma + invariant model: the site is either invariant with probability $P_0$ or the mutation rate at the site follows a gamma distribution. This model is generalized into a mixture of gamma distributions by Mayrose, Friedman and Pupko (2005). Both models are dealing with the fact that the rate may not just follow one gamma distribution throughout the whole sequence of DNA/RNA or amino acids. While Gu, Fu and Li suggested some sites may be invariant, Mayrose, Friedman and Pupko suggested some sites may follow another gamma distribution, which led to a mixed gamma model. While these models
improved the fitness over those “simple” models, it is not clear how these modifications relate to the process of protein evolution. Models that reflect the biological reality are more desirable.

Some sophisticated models that allow the rate of evolution to be dependent on the primary, secondary or tertiary structure were then proposed during the last decade for protein data. Yang (1995) and Felsenstein and Churchill (1996) proposed two similar Hidden Markov Models (HMM), in which the rate of a site is correlated to the rate of its adjacent sites in linear position. These models take into account the rate correlation along the linear region of the protein sequence. Goldman, Thorne & Jones (1996) proposed a model taking into account the secondary structure of the protein molecule. Goldman, Thorne and Jones used three secondary structures: α-helix, β-sheet and L-loop states. They assumed that the secondary structure at site \( i \) depends only on the secondary structure at site \( i-1 \). They estimated the probability of a residue in a given secondary structure category being followed by a residue in the three categories considered in this work for the amino acid sequences. They also estimated the \( Q \) matrix for each of the secondary structure from the aligned sequences with at least 15% of identical residuals and known secondary structures. The authors continued their study (Goldman, Thorne and Jones, 1998) and proposed a more sophisticated model. In the latter model, each site falls into one of the eight categories based on their accessibility (buried / exposed) and their secondary structure (α-helix, β-sheet, turn and coil). They found that the solvent accessibility status of a site had a
particularly strong association with the amino acid replacement. Furthermore, the association between the secondary structure and the amino acid replacement was proven significant. The resulting HMM allows the simultaneous inference of phylogeny and protein structure, using information about each to improve inference of the other. Despite the proven usefulness of these models, they admitted that the process of protein evolution was still not well understood, and they believed that identifying the physical constraints that effect protein evolution might be helpful to elucidate the process.

Rate dependence on the tertiary structure of the amino acid sequences was explored recently by Parisi and Echave (2001), Robinson et al. (2003) and Rodrigue et al. (2005, 2006). In these models, the rate of a site depends not only on its own state, but also on the states of the sites surrounding the site due to the structural constraints imposed by its neighbor sites. Initially, the goal of their research was to predict secondary structure, and it has opened new avenues for using secondary structure in the likelihood calculation. These models allow both variable rates and site dependence. Unfortunately, however, these complicated models cannot be implemented to infer phylogeny due to prohibitive computational costs.

To gain insight into protein evolution, Paris and Echave (2001) developed a structurally constrained protein evolution (SCPE) simulation. In their underlying model, a trial amino acid sequence was selected against departure from a reference 3D structure. This simulation was based on sequence structure distance $S_{dist}$ and a
parameter that measures the degree of structural divergence $S_{\text{div}}$. The simulation starts with a DNA sequence corresponding to a reference protein with known 3D structure. Based on a simple Jukes-Cantor model, a new trial of DNA sequence was generated with one restriction: it cannot mutate to a stop codon. Based on the new trial DNA sequence, the $S_{\text{dist}}$ between the generated sequence and the reference sequence was calculated using the PROSA potential program, which is based on the energies at each position. Finally, the generated sequence was accepted if $S_{\text{dist}}$ was below $S_{\text{div}}$. Although this simulation is highly computationally demanding, it can generate descendants from an ancestral sequence. It also demonstrates how proteins may evolve with the restrictions on the protein structure.

Based on the SCPE simulation results above, Robinson et al. (2003) developed an evolution model that was dependent on the protein structures. They proposed to use an instantaneous rate matrix $R$. Unlike that in Goldman’s original approach (Goldman, Throne and Jones, 1998), $R$ cannot be expressed by a series of conventional rate matrices $Q$ since it no longer assumes independent changes among codons. The dimension of the $R$ matrix is determined by the length of the sequence. The underlying hypothesis of this model is that the substitution rate was dependent on whether the amino acid replacements would stabilize the known and presumably fixed tertiary structure. The stability was evaluated using the compatibility measurement between the sequence and the tertiary structure at each site. The compatibility of the sequence and structure was assessed based on two components: one is the solvent accessibility.
(E_s) and the other is the pair-wise interactions between the amino acids close to each other in 3D space (E_p). They also introduced a parameter that reflected the external impact (ω) of the protein evolution. \( R_{i,j} \), the rate from codon \( i \) to \( j \), was set to 0 if codon \( i \) and \( j \) differed at more than one nucleotide or if codon \( j \) encoded a premature stop codon. For the other cases where codon \( i \) and \( j \) differed by exactly one nucleotide, \( R_{i,j} \) was defined as:

\[
R_{i,j} = \begin{cases} 
    u\pi_h & \text{for a synonymous transversion} \\
    u\pi_k & \text{for a synonymous transition} \\
    u\pi_h\omega e^{(E_s(i) - E_s(j))S + (E_p(i) - E_p(j))P} & \text{for a nonsynonymous transversion} \\
    u\pi_k\omega e^{(E_s(i) - E_s(j))S + (E_p(i) - E_p(j))P} & \text{for a nonsynonymous transition}
\end{cases}
\]

where \( s \) and \( p \) are parameters reflecting the contribution from the sequence structure and \( h \) denotes the nucleotide in codon \( j \). When \( s=p=0 \), and \( \omega=1 \), this model reduces to the conventional codon model. Calculations of the likelihood for two-sequence cases were implemented based on this model. Although the authors implied that the application of this model in multiple sequences should not be difficult, the demand of massive computation may be prohibitive for a large number of sequences.

Inspired by the above work, Rodrigue et al. (2005) generalized Robinson et al.’s technique from two sequences to 10 sequences, thus introducing the approach in a broader phylogenetic context. Besides that, they used a model at the amino acid level instead of the codon level. They worked under a fixed tree topology, which was estimated using the PhyML program (Guindon and Gascuel, 2003) based on the JTT model (Jones et al. 1992) and the gamma + invariant model (Gu et al. 1995). One
sequence with known 3D structure was treated as a reference structure and the sequences were aligned without gaps (removing all insertions/deletions). They constructed a contact map, where each element was an indicator for whether site $i$ and $j$ were in contact (1) or not (0). The pseudo-energy of a protein was calculated based on the contact map and the pair potential matrix. The rate matrix $R$ at each site was derived from the pseudo-energy of a protein. They used an MCMC sampling method, Metropolis-Hasting algorithm to obtain the posterior distribution of the model parameters, branch length and the substitution history. They found that $p$, the parameter that reflects the contribution from the sequence structure, converged to a positive value. This indicated that the selection preferred sequence substitution histories that maintain a good structural fitness, which is the hypothesis that the model was based on. The tree in this study was fixed, which made it difficult to use for inference of the phylogeny of the aligned sequences. They suggested allowing topology to be a free parameter to be estimated. However, a rearrangement of the current tree will result in incompatibility of the current substitution history and the new tree. The task of updating both parameters will likely be very computationally demanding, which makes it very challenging for inferring a phylogeny with even a moderate number (>15) of sequences.

In summary, it is widely recognized that the substitution rate for the amino acid is not site independent. The secondary and tertiary structures of the protein have a significant influence on the substitution rate of the amino acids. The models
dependent on the tertiary structure of the protein proposed so far have demonstrated this effect (Rodrigue et al. 2005; Rodrigue et al. 2005, 2006), however, none of the current models can be used effectively for phylogenetic inference.

2.1.3 Spatial Dependent Substitution Model

The recent studies on the amino acid sequence evolution models gave strong evidence that the amino acid evolution is dependent on the tertiary structure of the molecules (Rodrigue et al. 2005; Rodrigue, Philippe and Lartillot 2006). Some substitution models that allow rate dependence on the tertiary structure have been proposed (Robinson et al. 2003; Rodrigue et al. 2005). However, due to the complexity of these models, they have not been implemented for inferring phylogeny. It is practically impossible for these models to estimate even a small tree with less than 10 external nodes. The research goal here is to develop substitution models that account for crucial aspects of protein structure dependence, and can be implemented to infer phylogeny with a large number of leaves.

The studies of crystallographic structure and the conservation rate of amino acids on phosphoglycerate kinase (PGK) and other enzymes suggest that the evolution rate of each amino acid in these systems may depend on its location in the molecule (Pan, Pollack and Pearl, 2005). It is discovered that the most conserved amino acids tend to be near the functional core of the molecule while the least conserved amino acids are mainly found on the periphery (Figure 2.1).
Figure 2.1: The crystallographic structure of a PGK (1VPE) molecule with the metal ion ligand shown in bright orange at the center. The conservation rates have been divided into nine groups and color coded from blue (least conserved sites) to red (most conserved). The panels 1a to 1h show the amino acids added one group at a time in increasing conservation rate. Amino acids closer to the ligand associated metal ion are added later than those near the periphery are.
Figure 2.2: Correlation of averaged distance from amino acid alpha carbons (Cα) to the anchor-atoms at their catalytic active centers (C/AC) vs the positional-ConSurf conservation score (PCCS).
Further study by Pollack, Pan, and Pearl (2008) suggests that the relationship between the conservation rate of the amino acid and its distance to the function core is consistent for each of twelve enzymes that are of primary importance to central metabolism. Figure 2.2 illustrates the correlation between the conservation score calculated using Consurf program (Landau et al. 2005) vs. the distance for all twelve enzymes. The Consurf score is a relative measure of evolutionary conservation at each sequence site of the target chain. The lowest score represents the most conserved position in a protein. It can be seen that the correlations are consistently around 0.5 for all twelve enzymes. It was also demonstrated in that research that this relationship is also consistent for data from the three different Kingdoms of life.

Based on these results, we proposed a new substitution model that allows rate variation to depend on the spatial location of an amino acid according to the molecular tertiary structure in the form of:

\[ Q^* = f(r) \; Q \]

Where \( Q^* \) is a new rate matrix; \( Q \) is the rate matrix for any of the usual Markovian substitution models (such as the empirical models of Dayhoff or WAG); \( r \) is the distance measure of each site; and \( f(r) \) is a monotonically increasing function of the distance, called rate factor, with a restriction that \( \sum f(r) / K = 1 \) where \( K \) is the total number of sites. The distance measure is based on the distance from each amino acid’s alpha carbons (Cα) to the anchor-atoms at their catalytic active centers (CAC).
of the protein that can be obtained from its tertiary structure and function. We hypothesize that a larger distance measure correlates to a higher substitution rate. The rate factor $f(r)$ can be defined as a continuous function or a discrete function. The parameters in the $f(r)$ function are denoted as $\theta$. Figure 2.3 gives examples of functions that can be used in our substitution model. For the discrete function, each site is allocated into one of several groups according to their distance measurement, such as the core region, the middle region and the surface region. All these groups evolve independently with each other, and all the sites in the same group share the same rate factor, which needs to be estimated. Figure 2.3 shows a discrete function with four groups. For the continuous function, each site will have a specific rate factor according to the distance measurement, and all sites evolve independently. The parameters in the rate factor function will be estimated from the posterior distribution. The model we proposed can be easily extended to the gamma model by allowing the rate factors at the same group to follow a gamma distribution.
Figure 2.3: Some examples of rate factor $f(r)$ as a function of $r$. All the $f(r)$ functions have a mean value of 1.
In Yang’s Gamma model (1993), a rate factor is assigned to site $h$ while keeping the overall rate matrix unchanged. The model assumes that the rate factors are independently and identically distributed according to a Gamma distribution with one parameter $\alpha$ (set $\alpha=\beta$). A small $\alpha$ suggests that the rates differ significantly, while a very large $\alpha$ means roughly equal rates. Here we assign a rate factor $f(r)$ at each site which is a function of the distance measure for that site. We assume that each site is evolved independently given its distance measure. Our model modifies the rate factor of the Gamma model by using a function whose parameters can be estimated, instead of a random number from the gamma distribution. This will likely increase the computational efficiency. Thus, it is expected that the modified substitution model will be computationally efficient for data with a larger number of sequences.
2.2 Likelihood Calculation and Posterior Tree Distribution

To illustrate the tree likelihood calculation, we use the example tree in Figure 2.4. For this tree, there are five sequences $x_1$ to $x_5$, four internal nodes, $x_6, x_7, x_8, x_9$ and eight branch lengths, $t_1$ to $t_8$.

At a site $k$ ($k=1,\ldots,K$) with distance measure $r_k$, if a rate factor $f(r)$ and conventional rate matrix $Q$ are selected, we can obtain a new substitution matrix:

$$Q^* = f(r_k)Q.$$

Based on the $Q^*$, the transition probability from one amino acid $i$ to another one $j$ after split time $t$, $P_{ij}(t)$, can be calculated as:

$$P_{ij}(t) = \exp(Q^*t)$$
Given a stochastic variable $X(v)$ describing the evolution state of a site in the evolution process at time $v$, the Markov assumption asserts that

$$P_{ij}(t) = \Pr[X(s + t) = j|X(s) = i]$$

for any initial time $s \geq 0$. Informally, this means that $P_{ij}(t)$ is the probability change from state $i$ at time any time $s$, to any state $j$ by time $s + t$, it does not matter how the process reached state $i$ by time $s$ (the process is "memoryless").

The probabilities of transition from one base to another, $P_{ij}(t)$, can be written as a matrix $P(t)$, and then we can write:

$$P(t+dt) = P(t)(I + Q dt)$$

where $dt$ represents a small time, and $I$ is the identity matrix. The matrix $Q$ is known as the instantaneous rate matrix and has off-diagonal entries $Q_{ij}$ equal to the rates of substitution of $i$ by $j$. (The diagonal entries, $Q_{ii}$, are defined by a mathematical requirement that the row sums are all zero.) This equation can be solved as:

$$P(t) = e^{Qt} = I + Qt + (Qt)^2/2! + (Qt)^3/3! + \ldots$$

Using the spectral decomposition (also termed diagonalization) of $Q$, one can calculate the matrix $P(t)$ as:

$$P(t) = U \cdot \text{diag}\{e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\} \cdot U^{-1}$$

where the matrix $U$ contains the eigenvectors of $Q$, the $\lambda_i$ are the Eigen values of $Q$ and diag{} denotes the diagonal matrix of the elements contained in the braces. The components $P_{ij}(t)$ can be written as:

$$P_{ij}(t) = \sum_s C_{ijs} e^{\lambda_st}$$
where the sum is over $s = 1,\ldots,20$ for amino acids; $C_{ij}$ is a function of $U$ and $U^{-1}$. Here time $t$ is scaled to units of expected substitutions per site.

We assume that the evolutions at different branches are independent. The likelihood of the data ($D$) at a given tree topology ($T$), split time ($t$) and the internal nodes ($v$) at each site $k$ can be calculated as follows:

$$P_k(D | T, t, v, \theta) = P_k x_1 x_7(t_1) * P_k x_2 x_7(t_2) * P_k x_3 x_8(t_3) * P_k x_4 x_6(t_4) * \ldots P_k x_8 x_9(t_9)$$

The above calculation assumes that the states of the internal nodes are given. Felsenstein (1981) introduced a peeling algorithm for calculating the above likelihood given a tree (both topology and branch length) while the state of the internal nodes are unknown. At a particular internal node, such as $x_8$ (Figure 2.4), there are two descendent nodes: $x_3$, and $x_6$. These nodes can be external node (observed data, $x_3$) or internal node ($x_6$). If the conditional likelihood, $L_v(s)$, be the probability of observing node $v$ having state $s$, we can compute that using the following:

$$L_{x_8}(s) = (\sum_x p(x_3 = x | s, t_3) * L_{x_3}(x_3 = x)) * (\sum_y p(x_6 = y | s, t_6) * L_{x_6}(x_6 = y))$$

where $P(x=ij,t) = P_{ij}(t)$, and the value of $L_v(s)$ for the external nodes to have a value equal to one for corresponding observed value and 0 for all other values. The conditional likelihood of all internal nodes, including the root node can be estimated by following the tree topology from the external nodes back to root. The overall likelihood of that site given the tree can be computed as follows:
\[ P_k(D \mid T, t, \theta) = \sum_s \pi_s L_{xg}(s) \]

With the assumption that evolutions in different sites are independent at a given distance measure \( r \), we can then calculate the likelihood of the tree as:

\[ P(D \mid T, t, \theta) = \prod_{k=1}^K [P_k(D \mid T, t, \theta)] . \]

The joint posterior distribution of all the parameters can be calculated using Bayes’s Theorem as:

\[ P(T, t, \theta \mid D) = \frac{P(D \mid T, t, \theta) * P(T, t, \theta)}{P(D)} , \]

where \( P(T, t, \theta) \) is the prior distribution of the parameters, which can be obtained as follows, assuming the independence of \( P(\theta) \) and \( P(T, t) \):

\[ P(T, t, \theta) = P(\theta) * P(t \mid T) * P(T) \]

For example, we can choose \( P(T) \) as a flat prior, \( P(t \mid T) \) as an exponential (10) distribution, and \( P(\theta) \) as an uniform (0.01, 10).

Theoretically, \( P(D) \) can be calculated by summing over the entire possible tree space. However, due to the tremendous size of the tree space, it is computationally impossible to be estimated numerically when the number of external nodes is bigger than 15. Using (Monte Carlo Markov Chain) MCMC, which requires only ratios of posterior probabilities and thus is independent of the denominator, is then a promising way to estimate the posterior distribution.
2.3 Bayesian Inference with MCMC Method

Although the Bayesian method was introduced to the phylogenetics field as early as 1968 by Gomberg (also see Huelsenbeck et al 2001; Felsenstein 2004), a Bayesian inference of phylogeny was not possible due to the computational difficulties until 1996. Three groups of researchers, Li, Pearl and Doss (1996, 2000), Rannala and Yang (1996, 1997), and Mau, Newton and Larget (1997, 1999), all used Metropolis algorithms to implement Markov Chain Monte Carlo (MCMC) methods, essentially independently around the same time. Larget and Simon (1999) further developed a program BAMBE (Bayesian Analysis in Molecular Biology and Evolution), while Huelsenbeck and Ronquist (2001) started to build the more comprehensive MrBayes program.

Generally, we describe the phylogenetic tree in the following parts: topology (T), branch lengths (t), and internal nodes (v). The posterior distribution of the tree is estimated as follows:

\[ P(\text{tree}|\text{data}) = P(\text{data}|\text{tree}) \times P(\text{tree}) / P(\text{data}), \]

where the data is the aligned sequences of homologous gene(s) or protein(s). It combines the prior information of trees, P(tree), and the likelihood P(data|tree) to produce the posterior distribution. P(data) can be estimated as the summation of all the possible trees on P(data|tree) \times P(tree). Although the theorem seems simple, the Bayesian analysis cannot be evaluated analytically due to the difficulty of the
summation over all possible trees. However, the posterior tree distribution can be estimated using the MCMC method.

The idea of MCMC is to construct a Markov Chain that has information about state spaces and parameters of the models, and has a stationary distribution that converges to the posterior distribution. One of the most widely used MCMC methods is the Metropolis-Hastings algorithm (Hastings 1970). The Metropolis-Hastings algorithm has two essential steps:

(Step 1) Propose a new tree \( T_n \) from the current initial tree \( T_i \);

(Step 2) Let \( b = f(T_n)/f(T_i) \). If \( b >=1 \), \( T_{i+1} = T_n \); else, \( T_{i+1}=T_n \) with probability \( b \) and \( T_{i+1}= T_i \) with probability \((1-b)\).

By repeating the two steps, we can generate a Markov Chain with a tree distribution converging to the posterior tree distribution.

The most computationally costly step of the above algorithm is to estimate \( b \).

Let \( f(T_i) = P(T_i|\text{data}) \), then

\[
b = \frac{[P(\text{data}|T_n)P(T_n)]}{[P(\text{data}|T_i)P(T_i)]} \\
= [P(\text{data}|T_n)/P(\text{data}|T_i)] \times [P(T_n)/P(T_i)]
\]

The first part is the likelihood ratio of the two trees. The second part can be determined from the prior tree distribution. The likelihood of the tree given the data and tree has been discussed in the previous section:
\[
P(\text{Data} \mid T) = \prod_{k=1}^{K} P_k(D_{(k)} \mid T)
\]

where \(D_{(k)}\) is the data at the \(k\)th site. With the assumption that evolution in different lineages is independent, \(P(D_{(k)} \mid T)\) can be calculated using the peeling algorithm.

The three groups of researchers, Li, Pearl and Doss, Rannala and Yang, and Mau and Newton, developed slightly different Monte Carlo methods to estimate the posterior tree distribution independently around 1996-1997. The first two used MCMC Metropolis-Hastings approaches and the latter group using Monte Carlo integration. These three methods are different in three key aspects of their Bayesian methods: prior distribution, new tree proposal (a key component of the Metropolis algorithm), and the summary of the posterior distribution.

Li, Pearl and Doss (2000) used the following assumptions in their algorithm:

(1) Uniform prior on all possible topologies,

(2) Uniform prior on splitting time given a specific topology and

(3) DNA at different site changes independently following HKY model while the distribution of the root nodes is stationary.
To propose a new tree from the current tree, they adapted a local rearrangement strategy similar to an NNI branch swap. The rearrangement can be separated into three steps. First, an internal node was randomly selected (excluding the root nodes) as the target node, and child 1, child 2 and sibling branches were defined as in Figure 2.6. Secondly, one of three branches was selected as the new sibling branch to generate a new topology. Finally, the new split time for the target branch was generated according to the new topology while the internal nodes were also randomly generated (Figure 2.5).

To summarize the posterior distribution of the tree, they suggested using a tree that minimizes the triple distance radius required to capture 90% of the simulated trees.
Li, Pearl and Doss’s algorithm not only estimates the tree topology and branch lengths, but also tracks the sequences at the internal nodes of the tree. They also proposed several convergence diagnostic strategies to evaluate whether the MCMC has run long enough. The calculation at each step proposed here is faster, but the rate of mixing for this algorithm is slower than the one proposed by Mau et al. (1999).

Mau et al. (1999) used a slightly different prior distribution. They assumed that all these rooted histories were equally likely distributed (flat prior – but distinguishing between the ordering of the internal nodes). These are similar to Li, Pearl and Doss’s assumptions which assumed flat prior for all rooted tree topologies. They assumed internal node times came from a uniform distribution. The proposal for a new tree from the current one has two steps, called Q1 and Q2. First, the tree leaf was reordered by flipping a coin at each internal node (Q1). Then, the distance from the leaf to the internal nodes was simultaneously and independently modified by assigning a new distance from a uniform distribution ($d_{i-a}$, $d_{i+a}$), where $d_i$ is the current distance for internal nodes $i$, and $a$ is the tuning constant (Q2) (Figure 2.6).
Figure 2.6: Mau and Newton's proposal of changing to a new tree from the current tree in two steps: (1) the leaf nodes are randomly reordered, (2) the distance of each internal nodes are modified with a tuning parameter $a$. The second step may change the topology.

They chose the same HKY85 model as Li (1996) did for their DNA sequence. The likelihood calculation is more costly for this algorithm because the topology of the proposed tree may change substantially (depending on the tuning parameter) in comparison to the current tree. However, this allows much faster visits to other places in tree space than Li, Pearl and Doss's algorithm. Mau and Newton summarized the posterior distribution tree using “clade” probabilities.

Rannala and Yang (1996, 1997) used a uniform prior on all the tree topologies, while the prior distribution of the node times was specified by a birth-death process. They used a nearest-neighbor interchange (NNI) at a randomly selected internal node to obtain a new tree topology from the current topology. The likelihood of each topology was approximately estimated by sampling a large number of trees with
different splitting times under the restriction of the topology. Instead of estimating the posterior distribution of the phylogeny with topology and splitting time, they only estimated the posterior distribution of the topology. This method requires intensive calculation on every topology (sum over various branch length), as it is essentially impossible to accurately estimate the likelihood of each topology when the tree is even of moderate size. The tree that has the highest posterior probability was taken as the estimate of phylogeny, which is referred to as the MAP tree.

All these approaches described above have large computational demands for large trees (>30 sequences). Larget and Simon (1999) further extended the ability of the Bayesian approach to reconstruct bigger trees. In their algorithm, the parameter space includes the tree topology, branch length and the parameters of the evolution model. They proposed two steps to move from one parameter space to another. First, with a fixed current tree (topology and branch length), a new set of model parameters was obtained using the Metropolis algorithm. Then, with fixed model parameters, a new tree was updated using the Metropolis algorithm. The updated model parameters and the new tree formed the new parameter space. They used two different algorithms for proposing new trees. Each algorithm had two versions: one assumed a molecular clock, while the other did not. One algorithm, called GLOBAL, is similar to the one proposed by Mau et al. (1999); the other one is called LOCAL, which only changed a small portion of the tree. The mixture of these two algorithms allowed more rapid mixing, as well as an improved calculation speed.
In Larget and Simon’s analysis of their primate data using the F84 model, they assumed a uniform prior on all trees. First, they conducted several short runs to find the initial values for the model parameters and the tuning parameters. Then, with a random initial tree, they ran 2000 cycles with the GLOBAL followed by 2000 cycles with LOCAL to complete the burn-in. They repeated the same cycles for another 100K cycles for the tree and parameter inference. They summarized the MAP tree similar to Yang and Rannala (1997). The estimation of the model parameters was obtained using their software package, BAMBE (Bayesian Analysis in Molecular Biology and Evolution).

Huelsenbeck and Ronquist developed (and are still updating) a program package, called MrBayes, which has more options in terms of prior distribution selection and substitution model selection. In this program, they implemented not only the standard single-chain MCMC algorithm, but also a multi-chain variant MCMC called Metropolis Coupled MCMC, or (MC)$^3$. This algorithm runs $n$ chains at the same time with $(n-1)$ of them being heated. A heated chain has the steady state distribution $[f(\text{Tree}|\text{data})]^b$, where $b=1/(1+(i-1)*H)$ and $H$ is a heating parameter. A swap is attempted between two randomly chosen chains from the $n$ chains. If the swap is accepted, the two chains switch states. The inference is only based on the cold chain ($b=1$). The effect of the heated chain is to lower the likelihood peak thus make it easier to explore the space of the phylogeny tree. The mixing of the Markov chain is dramatically improved using (MC)$^3$. This program also has implemented most of the
substitution models for nucleotide, amino acid, and codon sequences. Beside these advantages, this software allows changing the prior distribution of the tree and specifying the initial tree input. This open source coded program has also implemented several methods for relaxing the assumption of equal rates across sites, such as the gamma model (Yang, 1994) and the invariance model (Gu, Fu and Li, 1995). It also provides some diagnostic tools to evaluate whether the chain has been sufficiently run to converge to the posterior distribution. MrBayes is one of the most advanced software that implemented the Bayesian approach in phylogeny inference.

Recently, the Parallel Metropolis coupled MCMC algorithm (P(MC)^3) has been implemented in the MrBayes3 (Altekar et al. 2004) to utilize the multi-processor computational technology for increasing the computation efficiency. We will further discuss this method in the next session.

After the MCMC program produced a good sample of trees from the posterior, there are two ways to summarize the posterior. One is to use clade probabilities, as used in MrBayes and BAMBE. For each clade, a support value is assigned as the proportion of trees that contain the clade in all sample trees. Li et al. (2000) used an alternative way. They summarized the posterior using the central tree from the 90% of all trees that closest to this central tree based on the triples distance measure.

As suggested by Huelsenbeck et al. (2001), the Bayesian inference of phylogeny represents a significant advance for a number of reasons. First, it is a
likelihood-based method, and it inherits the good quality of the maximum likelihood method, which outperforms other methods, such as distance method and Parsimony, under a range of parametric conditions. Second, the Bayesian method can incorporate the available information on the prior probability distribution of trees (although computationally efficient methods of doing this with sound biologic underpinning are not yet available). Third, MCMC provides a computational way of approximating posterior probability of trees and other model parameters at the same time. Furthermore, the Bayesian analysis provides an intuitive measure of the support for trees and a practical way to estimate large phylogenies using a statistical approach. However, more studies are needed to improve this approach.

1. It is not clear how the prior tree distribution should be chosen, and how the prior influences the results.

2. There are many unanswered questions related to the convergence and mixing. For example, how to evaluate whether the chain has converged and is mixing well in a way that is tied to the example under investigation?

3. It is unclear how to interpret the meaning of the support values obtained from the posterior probabilities.

4. Criteria should be established to evaluate whether a substitution model is better than another one, and to use such knowledge to choose an optimal model. (The calculation of the Bayes Factor provides a simple omnibus approach – but may be poorly tied to the problem at hand.)
2.4 Implementation of the SD Substitution Model

We used the Parallel Metropolis coupled MCMC algorithm (P(MC)\(^3\)), which is employed in the MrBayes3 software (Altekar et al. 2004), to implement our SD substitution model in phylogenetic inference and parameter estimation. The program MrBayes (version 3.1.2) is available at www.mrbayes.net and a modified MrBayes program which can be downloaded at www.stat.osu.edu/~xpan/diss.htm, is used to estimate the posterior joint distribution of all parameters and the tree topology. Two types of diagnostics are also considered in this project: one is the diagnostics for the MCMC convergence; the other one provides evaluation of the model fitness.

2.4.1. Parameter Estimation

The posterior joint distribution of tree topology, branch length, and model parameters are evaluated using a P(MC)\(^3\) based Metropolis-Hastings algorithm (Hastings 1970; Altekar et al. 2004).

The P(MC)\(^3\) is a variant of MCMC in which M chains are run. Each chain has a different heat value \(b_m\). The approach is described below:

1. Let \(\phi_i\) be the current state of the parameters of the Markov Chain \(i\). Each chain has a different \(\phi_i^{(k)}\).
   
2. For all M chains, run parallel:
   
   (a) randomly propose a new \(\phi_i'\) from \(\phi_i\);
   
   (b) accept \(\phi_i'\) with probability \(R_i\) where
       
       \[ R_i = \min[1, \frac{P(D|\phi_i')/P(D|\phi_i)}{P(\phi_i')/P(\phi_i)}^{b_m} \times \frac{q(\phi_i')}{q(\phi_i')}] \]
   
   (c) if \(\phi_i'\) is accepted, let \(\phi_i = \phi_i'\), else, \(\phi_i = \phi_i\);
(d) repeat (a)-(c) for a given number, such as 100

(3) Randomly choose two chains from the M chains (chain j, and chain k) after all M chains run a given number of iterations (100). Swap the states of the two chains with probability $R_2$ where

$$R_2 = \min[1, \frac{P(\phi^k|D)}{P(\phi^j|D)} \cdot \frac{P(\phi^j|D)}{P(\phi^k|D)} \cdot \frac{b_j}{b_k}]$$

(4) go back to (2).

Step (2) involves the calculation of the likelihood function, the most computationally intensive operation in the iteration. Because the subroutines in step (2) are run in parallel, the $P(MC)^3$ algorithm is expected to significantly reduce the computational time compared the $(MC)^3$ algorithm, and its computational demands should be similar to the standard MCMC algorithm (with one chain). Compared to the MCMC method, $P(MC)^3$ improves the mixing of the search algorithm that leads to rapid convergence of the Markov chain to its posterior.

As described in session 2.2, the likelihood function is the function of the transition probability from one state change to another after a certain time period, which is based here on the Spatial Dependent substitution model we propose. Two SD substitution models have been implemented in the current version of the program. One is the discrete spatial dependent function $f(r)$ with step function; the other is the continuous inverse gamma function. The discrete SD model is implemented as follows:

(1) Obtain the distance measure according to the crystallographic information from the PDB.
(2) Partition the sites according to the distance measurement. For example, each site can be assigned into each of the four partitions: one for those without distance measure, and three for “core,” “middle,” and “periphery” respectively.

(3) Run the MrBayes with partitions to obtain the posterior distribution of tree, the parameters and the rate for each partition.

The continuous \( f(r) \) model is implemented using a modified MrBayes where the rate at each site is based on an inverse gamma function which is dependent on the gamma parameter, and its distance measurement:

\[
f(r_k) = G^{-1}(\frac{[k]^{-1/2}}{N})
\]

where \( N \) is the total number of sites with distance measure; \([k]\) is the order index of distance for sites \( k \) among all distance measurement; and \( \alpha \) is the parameter for the gamma CDF \( (G^{-1}) \). There are two steps for the continuous \( f(r) \) model:

(1) Obtain the distance measure for all possible sites according to the crystallographic information from the PDB of the reference sequences. For these sites without distance measure, assign the median value of the distance measures. Use the distance vectors as input for the modified MrBayes program.

(2) Run the modified MrBayes program using gamma rate with categories number equal to the number of sites with distance measure.
2.4.2 MCMC Convergence Diagnosis

Although the theory of the Markov Chain guarantees that the stationary probability distribution can be reached (Karlin and Taylor 1975; Smith and Roberts 1993) and the Metropolis-Hastings algorithm converges under very general conditions with certain rate (Mengersen and Tweedie 1995), the most difficult problem for the MCMC method is that of deciding when to stop the algorithm with some degree of confidence that a state of equilibrium has been reached or convergence achieved. Cowles and Carlin (1996) reviewed a number of proposed methods for diagnosing convergence of a particular sampler, based upon the output they produce. However, Asmussen, Glynn and Thorisson (1992) show that there can exist no universally effective means of detecting stationarity, applicable to all stochastic simulations. They suggest to concentrate on particular classes of problems, and to take explicit advantage of the unique characteristics of those classes.

For the convergence of MCMC in the phylogenetic inference problem, we used six independent diagnostics to evaluate if the Markov chain has converged, including the log likelihood used by Mau et al.(1999), a lag loglikelihood plot, a scaled regeneration quantile (SRQ) plot, a distance-based diagnostic for chain-to-chain variation used by Li et al.(2000), the standard deviation of split frequency (Huelsenbeck and Rannala, 2004) and the potential scale reduction factor (PSRF) (Gelman and Rubin, 1992) implemented in MrBayes.
(1) Log likelihood and parameter time plots

When the Markov chain converges to the posterior distribution, the likelihood and parameter should also converge and stabilize. These plots are expected to reach equilibrium after the burn-in period as seen from the example in Figure 2.7.

Figure 2.7: Time series of log likelihood (LnL), tree length (TL), and gamma shape parameter (alpha). When the posterior distribution reaches stationary, one can expect that these values will fluctuate in a stabilized manner. Here, the LnL and alpha seem to have stabilized after 30,000 trees while the TL stabilized after 60,000. The burn-in period for this analysis will be determined as 60,000.
(2) Autocorrelation plot of likelihood plot

These plots are used to investigate the independence between trees. Both plots will reach equilibrium if the Markov chain converges. For example, in Figure 2.8, the Autocorrelation plot of LnL vs. the numbers of lags for two runs suggests that the autocorrelation decreasing as the lag increases, the solid curve descends more rapidly for the chain sampled 1000 apart than the dashed curve which was sampled 200 trees apart.

Figure 2.8: Examples of log likelihood autocorrelation plot and triple distance plot. The horizontal axel is the number of lags (Li et al. 2000). In the autocorrelation plot, the solid curve is for trees 200 apart in a chain of length 1 million following the burn-in period. The dashed curve is for trees 1,000 apart in a chain of length 5 million
(3) SRQ plot

SRQ plot can be used as a diagnostic for mixing. A SQR plot is a plot of $t_i/t_M$ versus $i/M$, where $t_i$ = the time in the chain when it returns to a specific topology or topology group, $i = 1, \ldots, M$. This plot is to see if the chain is mixing well or not. If the chain mixed well, then, this plot will form only random deviations off of a 45° line. Big deviations from the 45° line indication that estimates of the topologies studied are drastically different in a major part of the chain than in the full chain. Figure 2.9 shows an example of SQR plots. It can be seen that the tree mixing became better when the length of the Markov chain increases to 40,000. When length = 10,000, a nonnegligible portion of the chain differs substantially from the global estimate, and this implies that the chain is not long enough to produce stable estimates.

![Figure 2.9: SQR Plot of for Markov chains of various lengths (from Li et al. 2000)](image)
(4) Distance Density Plots

The distributions of distances for random pairs of trees can be drawn in three different ways:

i) both from within the first chain,
ii) both from within the second chain, and
iii) one from each chain.

Under the assumption that both chains reached the same stationary distribution, the three distance distributions will be similar, otherwise, we consider that the two chains did not converge to the same distribution. As in Figure 2.10, for chain length of 1 million in a problem with 27 taxa, the mean and standard deviation of the distances are different among the three drawing. Thus, since both chains are estimating the same posterior distribution, this is an indication that at least one of them did not converge. While for chain lengths of 5 million, the distribution, the mean and the standard deviation are similar, providing more confidence that the chains of length 5 million have converged to stationarity.
(5) Standard Deviation of Split Frequency

Partition support values of the phylogenetic trees from two different runs should be similar in an MCMC phylogenetic analysis if both runs reach stationarity and each run has collected enough samples from the posterior probability distribution to estimate it reliably. Figure 2.11 illustrates an example showing the two runs with similar partition support values. The points are concentrated on a 45 degree line. So, this is consistent with reasonably reliable estimates arising from two runs that each came to the same stationary distribution.
Figure 2.11: An example of split support correlation plot on two independent runs. If two runs reach the posterior stationarity distribution, then, it will display only random variation around the same split frequency at each split, so it will form a 45 degree line. Less variation around the line indicates more reliable estimates.

The standard deviation of split frequency is a measure of how similar the tree samples of the two independent runs are based on the two split support values. If the chain reaches stationary, the standard deviation of split frequency will be close to 0.

(6) Potential scale reduction factor (PSRF)

The PSRF was proposed by Gelman and Rubin (1992) and can be interpreted as a convergence diagnostic. When the PSRF for an estimated parameter close to 1, it suggests that the distribution of the parameter from the sample is close to the targeted stationary distribution. The PSRF applies to all those situations where inferences are summarized by posterior means and variances, even if the posterior distributions are
not themselves believed to be normally distributed. However, the assumptions that allow one to interpret it as a scale reduction factor may not met in the phylogenetic context. So it may be unreliable if you have a small number of samples.

2.4.3 Model Comparison

While many substitution models can be used in phylogenetic inference for the same data set, it is important to evaluate which model is the best among the proposed models. The likelihood ratio test (LRT) is one of the most common tests for model fitness between two nested models (Huelsenbeck and Rannale, 1997). It is used to test the hypothesis \( H_0 \, \theta = \theta_0 \) vs \( H_a \, \theta \neq \theta_0 \). To compare the two nested models, we have the approximate distributional result

\[
2\Delta \ell = 2 \left[ \ln L(\hat{\theta}) - \ln L(\theta_0) \right] \sim \chi^2_p,
\]

where \( p \) is the number of extra unconstrained parameters between the two nested models. Since asymptotically, \( 2\Delta \ell \) follows a \( \chi^2 \) distribution with \( p \) degrees of freedom, if it is greater than \( \chi^2_{p, 0.05} \), the null hypothesis is rejected, and the complex model is preferred. Otherwise, the simple model will be favored. A difficulty in using this technique in the phylogenetics context is that we are rarely able to calculate the maximum likelihood under each model to the degree of accuracy needed to apply the theory.
Yang, Goldman and Friday (1994) show how two models that differ in only one parameter can be compared by examining the estimated value and variance of that parameter.

To compare the fitness among the non-nested models, the Akaike information criterion (AIC) is often considered to be the most appropriate (Akaike 1974; Alfaro and Huelsenbeck 2006). The AIC can be computed as:

$$AIC_i = -2 \ln L_i + 2p_i$$

where $p_i$ is the number of parameters in model $i$. A model that minimizes AIC is considered the best. AIC is the most widely used criterion in the phylogeny related literature when comparing more than two models at the same time.

The Bayes factor (Gelman, et al.2003) is the dominant method of Bayesian model testing generally as in phylogenetics (Suchard, Weiss and Sinsheimer 2001, Liu 2006). If we have to choose between two models $M_1$ and $M_2$ based on a data $X$, the Bayes factor $BF$ is given by:

$$BF = \frac{P(X | M_1)}{P(X | M_2)}$$

where $P(X | M_i)$ is the marginal likelihood for model $i$. Similar to a likelihood-ratio test, instead of using the maxim likelihood, Bayesians average the likelihood over the parameters space.
The Bayes factor can be estimated using the marginal probability of the data from the output of the MCMC for the different models (Verdinelli and Wasserman 1995; Suchard et al. 2001; Liu 2006).

The most challenging part of all these methods is to estimate the lnL, or the marginal posterior probability of the data, \( f(D | \theta) \), for each model. The simplest ways to estimate this from the MCMC output are the arithmetic and the harmonic mean methods (Newton and Raftery, 1994). The arithmetic mean is simply defined as:

\[
\frac{1}{m} \sum_{i=1}^{m} f(D | \theta_i)
\]

where \( \theta_i \) is the \( i \)th sample from the posterior distribution of \( \theta \) given data.

For the typical MCMC algorithms used here, the arithmetic mean will converge to the true \( f(D) \) when \( m \to \infty \), and it will also typically satisfy the central limit theorem (Newton and Raftery, 1994). However, this estimate has large variance and is highly influenced by a few large likelihood values. It converges very slowly.

The harmonic mean is also called the weighted likelihood bootstrap estimate. The marginal probability of the data \( D \) with respect to the posterior distribution of the \( \theta \), \( f(\theta | D) \), can be written as:

\[
f(D) = \left[ \int_{\theta} \frac{1}{f(D | \theta)} f(\theta | D) d\theta \right]^{-1}
\]

So, the harmonic estimate of \( f(D) \) is
\[
\left[ \frac{1}{m} \sum_{i=1}^{m} \frac{1}{f(D|\theta_i)} \right]^{-1}
\]

Although this harmonic mean estimate converges to the true \( f(D) \) as \( m \to \infty \), it does not generally satisfy the assumptions of the central limit theorem because it usually has infinite variance. So this estimate may be unstable and sensitive to small likelihood values. Nonetheless, this approach often gives results accurate enough for interpretation on the logarithmic scale.

There are more sophisticated estimates proposed by Newton and Raftery (1994) that can be used to estimate the posterior likelihood. These estimates combine the harmonic mean and arithmetic mean to avoid the central limit theorem trap. Applying these techniques in the current problem is a topic for future research.

The AIC used in this study is based on the harmonic mean of the posterior log likelihood, we denote this as AIC*. 

CHAPTER 3
APPLICATION OF THE SD MODEL

To demonstrate the effectiveness and properties of our spatially dependent (SD) model, we applied the model to a small PGK dataset and eleven big enzyme datasets for phylogenetic inference. The eleven enzyme datasets were kindly provided by Dr. J. Dennis Pollack of the Department of Medical Microbiology and Immunology at the Ohio State University. The datasets were also used in another study on the relationship between the amino acid conservation vs. distance (Pan et al. 2005, Pollack et al. 2008).

3.1 The Data

Eleven soluble enzymes of primary importance for central metabolism were selected representing three of the six major enzymatic classes (Table 3.1). Five of them are transferases (phosphotransferases-kinases): 3-phosphoglycerate kinase (PGK), pyruvate kinase (PYK), adenylate kinase (ADK), nucleoside diphosphate kinase (NDK), and acetylglutamate kinase (ACGK). Three are oxidoreductases (dehydrogenases): alcohol dehydrogenase (ADH), L-lactate dehydrogenase (L-LDH), and glyceraldehyde-3-phosphate dehydrogenase (GAPDH). The other three are lyases
(carbon-carbon and carbon-oxygen): deoxyribose aldolase (DERA),
D-fructose-1,6-bisphosphate aldolase II (FBPA), and enolase (ENO).

These 11 enzymes are of primary importance to central metabolism, and the
sequences in each enzyme are from a variety of taxonomic species among the three
Kingdoms: Archaea, Bacteria and Eukaryota. These sequences were collected from
various sources following certain criteria used in previous studies (Pollack, Li and Pearl
2005; Pollack et al. 2008). The proportions of sequences in each Kingdom among the 11
enzymes’ datasets are (mean ± s.d.): Archaea 8.1% ± 4.5%, Bacteria 68.4% ± 17.4%, and
Eukaryota 23.6% ± 17.4%. For each enzyme, at least three special sequences with
structural information (Table 3.2) are chosen with preference for their membership in
different Kingdoms in order to have some biologic diversity. These are a total of 51
sequences selected with a defined 3D structure and we will refer to these sequences as
scaffolds.

The 11 sets of sequences were aligned by MUSCLE (Edgar 2004) separately.
These MSA files can be downloaded here: www.stat.osu.edu/~xpan/diss.htm in FASTA
format.
Table 3.1: The 11 selected enzyme datasets with various numbers of taxa and different length of the aligned sequence.

3.1.1 Catalytic-active enzyme center (CAC)

The CAC is described in this study as nucleophilic, exerting an effect on a residue or water involved in the catalytic mechanism. It stabilizes a proposed transition-state intermediate and affects a substrate or co-factor that aids catalysis. Depending on the mutual availability of a similar anchor-atom for the scaffolds in each enzyme group, we chose a metal ion ($\text{Mg}^{2+}$, $\text{Mn}^{2+}$, $\text{Zn}^{2+}$), an atom of an amino acid (LYS, HIS, GLY, TYR), or an atom of a ligand ($\text{NAD}^+$ or $\text{NADP}^+$), as the CAC. The CAC was selected by Dr. Pollack according to the Catalytic Site Atlas (CSA), the catalytic mechanism database (EzCatDB), and other additional literature reports. Table 3.2 lists all the information about the selected CAC and the anchor-atom for all 51 scaffolds.
<table>
<thead>
<tr>
<th>ENZYME</th>
<th>PDB</th>
<th>SOURCE</th>
<th>KING-DOM</th>
<th>CSA-ID</th>
<th>PDB-POSITION</th>
<th>PDB-ATOM</th>
<th>ATOM-ID</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DERA</td>
<td>1mzh</td>
<td><em>Aquifex aeolicus</em></td>
<td>B</td>
<td>LYS</td>
<td>279</td>
<td>1407</td>
<td>NZ</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>1p1x</td>
<td><em>Escherichia coli</em></td>
<td>B</td>
<td>LYS</td>
<td>167</td>
<td>1284</td>
<td>NZ</td>
<td>260</td>
</tr>
<tr>
<td></td>
<td>1ub3</td>
<td><em>Thermus thermophilus</em></td>
<td>B</td>
<td>LYS</td>
<td>151</td>
<td>1153</td>
<td>NZ</td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>1vcv</td>
<td><em>Pyrobaculum aerophilum</em></td>
<td>A</td>
<td>LYS</td>
<td>146</td>
<td>1129</td>
<td>NZ</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>2a4a</td>
<td><em>Plasmodium yoelii</em></td>
<td>E</td>
<td>LYS</td>
<td>171</td>
<td>1336</td>
<td>NZ</td>
<td>281</td>
</tr>
<tr>
<td>ENO</td>
<td>2akm</td>
<td>Human</td>
<td>E</td>
<td>Mg$^{2+}$</td>
<td>440</td>
<td>6553</td>
<td>Mg$^{2+}$</td>
<td>439</td>
</tr>
<tr>
<td></td>
<td>2al1</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>E</td>
<td>Mg$^{2+}$</td>
<td>438</td>
<td>6558</td>
<td>Mg$^{2+}$</td>
<td>436</td>
</tr>
<tr>
<td></td>
<td>1oep</td>
<td><em>Trypanosoma brucei</em></td>
<td>E</td>
<td>Zn$^{2+}$</td>
<td>500</td>
<td>3213</td>
<td>Zn$^{2+}$</td>
<td>432</td>
</tr>
<tr>
<td></td>
<td>1e9i</td>
<td><em>Escherichia coli</em></td>
<td>B</td>
<td>Mg$^{2+}$</td>
<td>1431</td>
<td>3153</td>
<td>Mg$^{2+}$</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>1iyx</td>
<td><em>Enterococcus hirae</em></td>
<td>B</td>
<td>Mg$^{2+}$</td>
<td>1342</td>
<td>6511</td>
<td>Mg$^{2+}$</td>
<td>432</td>
</tr>
<tr>
<td>FBP</td>
<td>1a5c</td>
<td><em>Plasmodium falciparum</em></td>
<td>E</td>
<td>LYS</td>
<td>236</td>
<td>1720</td>
<td>NZ</td>
<td>368</td>
</tr>
<tr>
<td></td>
<td>1ald</td>
<td>Human</td>
<td>E</td>
<td>LYS</td>
<td>229</td>
<td>1765</td>
<td>NZ</td>
<td>363</td>
</tr>
<tr>
<td></td>
<td>1fba</td>
<td><em>Drosophila melanogaster</em></td>
<td>E</td>
<td>LYS</td>
<td>229</td>
<td>1791</td>
<td>NZ</td>
<td>361</td>
</tr>
<tr>
<td></td>
<td>1ok4</td>
<td><em>Thermoproteus tenax</em></td>
<td>A</td>
<td>LYS</td>
<td>177</td>
<td>1358</td>
<td>NZ</td>
<td>263</td>
</tr>
<tr>
<td>ADH</td>
<td>1e3e</td>
<td><em>Mus musculus</em></td>
<td>E</td>
<td>Zn$^{2+}$</td>
<td>378</td>
<td>2845</td>
<td>Zn$^{2+}$</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>1h2b</td>
<td><em>Aeropyrum pernix</em></td>
<td>A</td>
<td>Zn$^{2+}$</td>
<td>1362</td>
<td>5388</td>
<td>Zn$^{2+}$</td>
<td>359</td>
</tr>
<tr>
<td></td>
<td>1jxb</td>
<td><em>Sulfobabus saffaricus</em></td>
<td>A</td>
<td>Zn$^{2+}$</td>
<td>500</td>
<td>2624</td>
<td>Zn$^{2+}$</td>
<td>347</td>
</tr>
<tr>
<td></td>
<td>11lu</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>B</td>
<td>Zn$^{2+}$</td>
<td>343</td>
<td>20129</td>
<td>Zn$^{2+}$</td>
<td>342</td>
</tr>
<tr>
<td>GAPDH</td>
<td>1dc6</td>
<td><em>Escherichia coli</em></td>
<td>B</td>
<td>NAD</td>
<td>351</td>
<td>5020</td>
<td>NC4</td>
<td>330</td>
</tr>
<tr>
<td></td>
<td>1gyq</td>
<td><em>Leishmania mexicana</em></td>
<td>E</td>
<td>NAD</td>
<td>361</td>
<td>10910</td>
<td>NC4</td>
<td>358</td>
</tr>
<tr>
<td></td>
<td>1hdg</td>
<td><em>Thermotoga maritima</em></td>
<td>B</td>
<td>NAD</td>
<td>336</td>
<td>2597</td>
<td>NC4</td>
<td>332</td>
</tr>
<tr>
<td></td>
<td>1u8f</td>
<td>Human</td>
<td>E</td>
<td>NAD</td>
<td>336</td>
<td>10223</td>
<td>NC4</td>
<td>335</td>
</tr>
<tr>
<td></td>
<td>1yw</td>
<td><em>Plasmodium falciparum</em></td>
<td>E</td>
<td>NAD</td>
<td>410</td>
<td>10210</td>
<td>NC4</td>
<td>337</td>
</tr>
<tr>
<td>OXIDOREDUCTASE (DEHYDROGENASE)</td>
<td>2ecz</td>
<td><em>Pyrococcus horikoshii</em></td>
<td>A</td>
<td>NAD</td>
<td>501</td>
<td>10545</td>
<td>NC4</td>
<td>334</td>
</tr>
<tr>
<td></td>
<td>110z</td>
<td>Human</td>
<td>E</td>
<td>NAI</td>
<td>401</td>
<td>5148</td>
<td>NC4</td>
<td>333</td>
</tr>
<tr>
<td></td>
<td>11th</td>
<td><em>Bifidobacterium longum</em></td>
<td>B</td>
<td>NAD</td>
<td>1</td>
<td>2393</td>
<td>NC4</td>
<td>319</td>
</tr>
<tr>
<td></td>
<td>1ez4</td>
<td><em>Lactobacillus pentosus</em></td>
<td>B</td>
<td>NAD</td>
<td>1352</td>
<td>9438</td>
<td>NC4</td>
<td>318</td>
</tr>
<tr>
<td></td>
<td>1pzh</td>
<td><em>Toxoplasma gondii</em></td>
<td>E</td>
<td>NAD</td>
<td>701</td>
<td>9950</td>
<td>NC4</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td>1ldg</td>
<td><em>Plasmodium falciparum</em></td>
<td>E</td>
<td>NAD</td>
<td>401</td>
<td>2427</td>
<td>NC4</td>
<td>316</td>
</tr>
</tbody>
</table>

Table 3.2: Catalytic-active enzyme center (CAC) selection.
<table>
<thead>
<tr>
<th>ENZYM-YME</th>
<th>PDB ¹</th>
<th>SOURCE</th>
<th>KING-</th>
<th>CSA-</th>
<th>PDB-</th>
<th>PDB-</th>
<th>ATOM-</th>
<th>Size ²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>DOM ²</td>
<td>ID ³</td>
<td>POSITION ⁴</td>
<td>ATOM NO. ⁵</td>
<td>ID ⁶</td>
<td></td>
</tr>
<tr>
<td>1jxv Human</td>
<td>E</td>
<td>HIS</td>
<td>118</td>
<td>899</td>
<td>Cα</td>
<td>152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1k44 Mycobacterium tuberculosis</td>
<td>B</td>
<td>HIS</td>
<td>117</td>
<td>867</td>
<td>Cα</td>
<td>136</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1nb2 Bacillus halodenitrificans</td>
<td>B</td>
<td>HIS</td>
<td>116</td>
<td>813</td>
<td>Cα</td>
<td>150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1pku Oryza sativa</td>
<td>E</td>
<td>HIS</td>
<td>115</td>
<td>915</td>
<td>Cα</td>
<td>150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2az1 Halobacterium salinarum</td>
<td>A</td>
<td>HIS</td>
<td>119</td>
<td>885</td>
<td>Cα</td>
<td>181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ltk Plasmodium falciparum</td>
<td>E</td>
<td>GLY</td>
<td>373</td>
<td>2888</td>
<td>Cα</td>
<td>425</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1php Bacillus stearothermophilus</td>
<td>B</td>
<td>GLY</td>
<td>351</td>
<td>2697</td>
<td>Cα</td>
<td>394</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1vpe Thermotoga maritima</td>
<td>B</td>
<td>GLY</td>
<td>354</td>
<td>2696</td>
<td>Cα</td>
<td>398</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1vjc Sus scrofa</td>
<td>E</td>
<td>GLY</td>
<td>373</td>
<td>2726</td>
<td>Cα</td>
<td>416</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2cun Pyrococcus horikoshii</td>
<td>A</td>
<td>GLY</td>
<td>359</td>
<td>2879</td>
<td>Cα</td>
<td>410</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a3w Saccharomyces cerevisiae</td>
<td>E</td>
<td>Mn²⁺</td>
<td>1001</td>
<td>3775</td>
<td>Mn²⁺</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1aqf Oryctolagus cuniculus</td>
<td>E</td>
<td>Mg²⁺</td>
<td>534</td>
<td>3992</td>
<td>Mg²⁺</td>
<td>530</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1liu Human</td>
<td>E</td>
<td>Mn²⁺</td>
<td>583</td>
<td>15545</td>
<td>Mn²⁺</td>
<td>528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1gs5 Escherichia coli</td>
<td>B</td>
<td>LYS</td>
<td>8</td>
<td>67</td>
<td>NZ</td>
<td>258</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2ap9 Mycobacterium tuberculosis</td>
<td>B</td>
<td>LYS</td>
<td>36</td>
<td>242</td>
<td>NZ</td>
<td>299</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2buf Pseudomonas aeruginosa</td>
<td>B</td>
<td>LYS</td>
<td>33</td>
<td>250</td>
<td>NZ</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1z83 Human</td>
<td>E</td>
<td>LYS</td>
<td>21</td>
<td>159</td>
<td>NZ</td>
<td>196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1zak Zea mays</td>
<td>E</td>
<td>LYS</td>
<td>18</td>
<td>106</td>
<td>NZ</td>
<td>222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1zin Bacillus stearothermophilus</td>
<td>B</td>
<td>LYS</td>
<td>19</td>
<td>131</td>
<td>NZ</td>
<td>217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2aky Saccharomyces cerevisiae</td>
<td>E</td>
<td>LYS</td>
<td>17</td>
<td>123</td>
<td>NZ</td>
<td>221</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2cdn Mycobacterium tuberculosis</td>
<td>B</td>
<td>LYS</td>
<td>13</td>
<td>120</td>
<td>NZ</td>
<td>201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4ake Escherichia coli</td>
<td>B</td>
<td>LYS</td>
<td>13</td>
<td>120</td>
<td>NZ</td>
<td>214</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹PDB = Protein Data Bank code ID.
²KINGDOM: A = Archaea, B = Bacteria, E = Eukaryota).
³CSA-ID = Identification of an amino acid or a ligand associated with the CAC, refer to a specific atom or group of atoms in the anchor-amino acid or non-metal ligand (NAI, NAD) or the metal ligand (Mg²⁺, Mn²⁺ and Zn²⁺) of a specific chain.
⁴PDB POSITION = Chain number for the anchor position in the PDB file.
⁵PDB ATOM NUMBER = A unique PDB serial number for the anchor atom.
⁶ATOM ID = Chemical identification of the chosen anchor-atom, NZ = N of NH₃⁺ at the C5 of lysine, COOH⁻ = C5 of glutamic acid, and NC4 = C4 of the planar nicotinamide moiety in both NAD⁺ and NAI.
⁷Size = Number of amino acids in the scaffold.
3.1.2 Site Distance Determination

The Euclidean distance of the $\alpha$-carbon (C\(_{\alpha}\)) of each amino acid to the anchor-atom of the CAC is calculated for all 51 sequences based on their spatial XYZ coordinates from the PDB using the Yasara program (http://www.yasara.org). For each enzyme MSA, the number of distance measurements is the same as the number of scaffolds available in that enzyme. For example, PGK has five scaffolds; thus, it has five distance measurements. Table 3.3 lists the Spearman correlation of the Euclidean distance among these five scaffolds. The distances are highly correlated between each pair of distance measures. Based on this fact, the site distance of the MSA of an enzyme is defined as the average of the available Euclidean distances of the aligned amino acids from the scaffolds at that site. If a site that does not have any amino acid from the scaffolds aligned with it because of indels, it does not have a distance measure either. On the other hand, if a site has more than one amino acid from the scaffolds, the distance for that site is the average distance of the amino acid(s) from each scaffold.

<table>
<thead>
<tr>
<th></th>
<th>1HDI</th>
<th>1PHP</th>
<th>3PGK</th>
<th>13PK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PHP</td>
<td>0.989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3PGK</td>
<td>0.964</td>
<td>0.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13PK</td>
<td>0.835</td>
<td>0.845</td>
<td>0.792</td>
<td></td>
</tr>
<tr>
<td>1VPE</td>
<td>0.916</td>
<td>0.93</td>
<td>0.876</td>
<td>0.971</td>
</tr>
</tbody>
</table>

Table 3.3: The spearman correlation of the Euclidean distance measure among five PGK scaffolds.
3.2 Discrete SD model Application

The application of the discrete SD model includes two consecutive steps: first, the sites were partitioned according to their distance, and secondly, the phylogenetic inference, including the estimation of the model parameters and conditional likelihood, was conducted using the MrBayes program, version 3.1.2 (Altekar, et. al. 2004). The statistical properties of the discrete SD model and the partition methods were first investigated using a small dataset, which only includes five PGK scaffolds. The best partition method was selected based on their results from the small dataset. We then applied the selected discrete SD model to the eleven datasets to demonstrate the effectiveness of this model.

3.2.1 Application to Small PGK Dataset

The goal of using the small dataset in this study is to exam whether the SD model can be used in phylogeny inference, and whether it provides a better model fitness in terms of AIC. In addition, we explored what was the best way to partition the sites. The PDB IDs of the five sequences in the small dataset are 1HDI, 3PGK, 1PHP, 1VPE, and 13PK. The MSA has 457 sites. Among them, 42 sites had no distance measure.

Four classes of models, namely, a Base model, Gamma model, SD model and Full model, were implemented in this study. The Base model used a proper empirical substitution model (WAG), and assumes that the rates are independent of site and at the same value. The Gamma model did not assume equal rates across all sites; rather, the rates follow a gamma distribution. We used the discrete gamma model as suggested by
Yang (1994) in this study. The *SD model* assumed that the rates across all sites are dependent on the distance measure according to a discrete function. The fourth model, denoted as *Full model*, assumed the rate at each site $k$ is not only dependent on the distance, but also follows a gamma distribution: $P_k = \text{Exp}(Q^* \cdot g_k \cdot t)$ where $g_k$ is a random rate following a gamma distribution with shape parameter $\alpha$.

We implemented three different partition methods for the discrete function:

(1) partition the sites according to distance: 0-15, 15-25, >25 Å, (core, middle and periphery, respectively) and unknown (SS4, four-category). It has roughly the same number of sites for the first three categories.

(2) according to quartile: 0-25%, 25-50%, 50-75%, 75-100% and unknown (SS5, five-category) and

(3) according to quintile: 0-20%, 20%-40%, ..., 80-100% and unknown (SS6, six-category).

The empirical substitution model for the base model was selected using the ProtTest software (Abascal, Zardoya and Posada 2005). Based on the AIC* and likelihood criteria, the WAG model (Whelan and Goldman 2001) was the most suitable substitution model for the PGK data set among the empirical models implemented in ProtTest.

The phylogenetic inference for the PGK dataset based on these models was conducted using the MrBayes program on a supercomputer cluster, Xgrid provided by the Mathematical Biosciences Institution (MBI) at The Ohio State University. The Xgrid
supercomputer cluster has one front-end node and fifty eight compute nodes. The front end node, Apple xserve, has two 2.3 GHz G5 processors, 6 GB of memory, and 160 GB of local disk capacity. Twenty nine compute nodes with Apple dual 2GHz G5 processors, 1 GB of memory, and another 29 compute nodes with iMac single 1.8GHz G5 processors, 1 GB of memory.

This Markov chain Monte Carlo (MCMC) analysis used 2 independent runs. Each run has four chains (one base chain and three hot chains). It started from a random tree, on four different Markov chains for 300,000 generations saving every 100th tree. With 100,000 burn-in samples to have 4000 samples to estimate the posterior probabilities such as branch supports, topology, and the model parameters. The prior distribution for the \( \alpha \) parameter of the Gamma distribution was uniform (0.1, 10) and the prior for the branch length was \( \exp(10) \).

The time series plot of the log-likelihood and parameters, the distance diagnostics and the PSRF (Gelman and Rubin 1992) estimated in MrBayes, suggested the MCMC gave adequate convergence after about burn-in periods of 100,000. Figure 3.1 shows the time series plot and the auto-correlation function of the log posterior density in two independent runs based on the full model (Gamma + SS4), both runs has reasonable mixing, and have converged to the same tree distribution with similar distance density diagnostics (Figure 3.2). The PSRFs for all parameters ranges from 1.000 to 1.001.

The harmonic means of the conditional log-likelihood (lnL) of the data, the number of parameters, computational times, and their AIC* are summarized in Table 3.4.
Figure 3.1: Time series plot, and its auto-correlation function (ACF) of the Log posterior density, Ln(L) in two independent runs based on the Full (Gamma+ SS4) model. When both runs converge to the same posterior density, both plots will be similar as show above.
Figure 3.2: Tree Distance Density Diagnostic for the 5PGK data set based on the Full model. If the two runs follow the same distribution after the burn in period, and the number of sampled trees is large enough, then the densities estimated in the three panels will be approximately the same.
Table 3.4: Summaries of model fit and computational times for the five-sequence PGK dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>ln L *</th>
<th>Number of Parameters</th>
<th>AIC*</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (WAG)</td>
<td>-3863</td>
<td>1</td>
<td>7728</td>
<td>635</td>
</tr>
<tr>
<td>Gamma</td>
<td>-3839</td>
<td>2</td>
<td>7682</td>
<td>2349</td>
</tr>
<tr>
<td>SS4</td>
<td>-3824</td>
<td>4</td>
<td>7656</td>
<td>884</td>
</tr>
<tr>
<td>SS5</td>
<td>-3834</td>
<td>5</td>
<td>7678</td>
<td>952</td>
</tr>
<tr>
<td>SS6</td>
<td>-3828</td>
<td>6</td>
<td>7668</td>
<td>1033</td>
</tr>
<tr>
<td>Gamma + SS4</td>
<td>-3813</td>
<td>5</td>
<td>7636</td>
<td>3240</td>
</tr>
<tr>
<td>Gamma + SS5</td>
<td>-3821</td>
<td>6</td>
<td>7654</td>
<td>3513</td>
</tr>
<tr>
<td>Gamma + SS6</td>
<td>-3815</td>
<td>7</td>
<td>7644</td>
<td>3840</td>
</tr>
</tbody>
</table>

* the harmonic mean of the posterior likelihood ln L, the estimated standard deviations of the harmonic mean of the Ln L range from 0.1-0.3 for all models.

According to this result, the best partition method among the three SD models is SS4, which had a higher AIC* by 22 compared to SS5, and 12 when compared to SS6. In addition, the SS4 method used the least computational time among the three partition methods. Consistently, the full model (Gamma + SS4) is the best among the three full models, which had a higher AIC* (by 18 than the Gamma + SS5 and 8 than the gamma + SS6 models, respectively). Furthermore, the Gamma + SS4 model used the least CPU time. Similar conclusions can be drawn according to the Bayes Factor and BIC. We therefore used the SS4 partition strategy for the rest of the SD models on other datasets. The sites were partitioned into four categories, namely, the unknown, the periphery (surface), the core, and the middle sites.
It was also observed that the discrete function may not only improve the base model, but also the gamma model. The SS4 improved the base model in terms of AIC* by 72 with only about a 40% CPU time increase, while the Gamma + SS4 model was better than the gamma model in terms of AIC* by 46, with an additional 44% CPU time increase.

The Gamma model improved the model fitness by 46 in terms of AIC*, with about 3.70 times of the CPU time compared to the base model. Similarly, the Gamma + SS4 model improved the base SS4 model by 20 in terms of AIC* with about 3.67 times the CPU time.

Compared to the Gamma model, the SS4 model used much less CPU time and had a much higher AIC*. The SS4 model used about 37% of the CPU time spent by the Gamma model and had an AIC 26 higher than that of the Gamma model.
3.2.2 Application to Large Datasets

The goal of this application is to investigate whether the discrete SD model can be used for other enzymes that have similar 3D structures. The sequences in each of the 11 enzyme datasets were carefully selected to ensure wide diversity across all three kingdoms of life. All these enzymes had roughly round structure according to the scaffolds. Since the number of sequences in the enzyme ranged from 79 to 226 it is also important to investigate whether the SD model can be implemented to these kind of large phylogenetic datasets.

If the value of the SD model can be generalized to data sets involving sequences, a majority of which do not have known 3D structure, then its applicability in phylogenetics will be greatly enhanced. This was then tested in the other enzymes, to check if the SD model would yield consistent results that significantly improve the model fitness in terms of Bayes Factor, AIC, BIC or other criteria.

Two models were implemented in this study, the Base model and the SD models with four categories. The site partition method used here is based on the results from 3.2.1, the SS4 method, which partitioned the sites with known distance into periphery, core and middle categories with roughly the same number of sites in each category. The computational procedures were similar to that described in section 3.2.1. We used the same computers and MrBayes program except that we used up to 8 parallel processes, generated up to 3,000,000 trees with a sampling frequency every 1000 trees and a burn-in of at least 1000,000 trees to ensure that the MCMC procedures converge properly.
These MCMC runs appear to have successfully obtained convergence for all 11 enzymes using the SD model. This result demonstrated that the discrete SD model can be used in datasets with a large number of sequences. A summary of the comparison between the base model and the SD model is presented in Table 3.5. The increase of the loglikelihood, \( \ln L \), of the models was equal to the estimation of the log value of the Bayes Factor for the SD model vs. the Base model. The AIC* improvement is, of course, twice that of the LnL increment minus the difference of the number of the parameters between the two models as defined in Chapter 2. For example, the log (Bayes factor) of the SD model vs. the Base model for DERA is 734, and the AIC improvement is \( 2 \times (734 - 3) = 1462 \).

The CPU time increase is summarized as the ratio of the CPU time spent by the SD model vs. that spent by the Base model. Since the computer implementation for some enzymes used eight parallel processes with the SD model and four processes with the Base model, the ratio of the CPU time for these enzymes was overestimated (could be up to double the true ratio corresponding to the case if the same number of processes were used, such as that for the LDH. Therefore, the true CPU time ratio should be less than what is recorded in Table 3.5.

The substitution rates are the average rates for all sites in the Core, Middle and Periphery partitions. It should be noted that the overall average rate of all sites in different partitions is 1.0, according to the definition of the discrete function.
According to the log-likelihood, lnL, in Table 3.5, the SD model consistently improved the Base model with log(Bayes Factor) ranging from 62 to 2130. The improvement of the SD model was consistent for all enzymes. Similar conclusions can be obtained using AIC or likelihood ratio tests. This appears to demonstrate the viability of the SD model to be generalized to apply to a wide range of datasets. The CPU time ratio

<table>
<thead>
<tr>
<th></th>
<th>Increase of lnL*</th>
<th>95% Credible Regions of the Substitution Rate for the SD Partitions</th>
<th>Ratio of CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS4/Base</td>
<td>Others</td>
<td>Core</td>
</tr>
<tr>
<td>DERA</td>
<td>734 ± 1.4</td>
<td>(0.90, 1.21)</td>
<td>(0.36, 0.43)</td>
</tr>
<tr>
<td>ENO</td>
<td>909 ± 1.2</td>
<td>(1.36, 1.91)</td>
<td>(0.26, 0.31)</td>
</tr>
<tr>
<td>FBPA</td>
<td>259 ± 1.7</td>
<td>(0.90, 1.12)</td>
<td>(0.66, 0.74)</td>
</tr>
<tr>
<td>ADH</td>
<td>62 ± 0.7</td>
<td>(0.88, 1.14)</td>
<td>(0.77, 0.86)</td>
</tr>
<tr>
<td>GAPDH</td>
<td>564 ± 1.3</td>
<td>(1.71, 3.14)</td>
<td>(0.31, 0.37)</td>
</tr>
<tr>
<td>LDH</td>
<td>568 ± 3.4</td>
<td>(0.88, 1.00)</td>
<td>(0.58, 0.65)</td>
</tr>
<tr>
<td>NDK</td>
<td>510 ± 2.5</td>
<td>(0.71, 1.26)</td>
<td>(0.58, 0.69)</td>
</tr>
<tr>
<td>PGK</td>
<td>552 ± 1.2</td>
<td>(1.48, 1.79)</td>
<td>(0.28, 0.35)</td>
</tr>
<tr>
<td>PYK</td>
<td>2130 ± 1.7</td>
<td>(1.25, 1.39)</td>
<td>(0.20, 0.24)</td>
</tr>
<tr>
<td>ACGK</td>
<td>1473 ± 3.8</td>
<td>(1.21, 1.40)</td>
<td>(0.38, 0.42)</td>
</tr>
<tr>
<td>ADYK</td>
<td>606 ± 2.1</td>
<td>(0.58, 1.23)</td>
<td>(0.58, 0.63)</td>
</tr>
</tbody>
</table>

* The difference of the Harmonic mean of ln(L) between the SS4 model and the base (WAG) model ± estimated std. dev. of the difference based on calculating $s_{SS4}$ and $s_{Base}$ from 10 sub-samples from each run for each model and taking std. dev. = $\sqrt{s_{SS4}^2 + s_{Base}^2}$

** Using the double the CPU time of the SD model (8 parallel processes). Other models using 4 processes

Table 3.5: Summary of log likelihood (lnL), substitution rate at selected partitions and computational times for 11 enzyme datasets.
ranges from 1.33 to 2.6 for all 11 enzyme datasets. The CPU time cost for the SD model was a reasonable improvement compared to the Gamma model, which has a cost of about a CPU time ratio of at least 4.0 for the discrete gamma model with 4 categories (Yang, 1994).

In addition, the substitution rates estimated at different regions of the enzyme were different and not overlapping for all 11 enzymes. Furthermore, the substitution rates at different regions were ordered from lowest (core) to highest (periphery) and were approximately linearly proportional to the distance. Since the rate at each region was estimated independently with the only constraint being that the overall mean rate is 1, this result confirmed that the amino acids close to the CAC, or having the smallest distance measurement, have the smallest rate. Similar results were reported in other studies that the conservation of amino acids is highly influenced by its distance to the CAC based on the diversity index at each site (Pollack et al. 2008).
3.3 Continuous SD model

The continuous SD model was implemented using a modified MrBayes program. In this program, each site was assigned a rate according to the inverse gamma function of the distance rate as detailed in session 2.4. Similar to the discrete SD model, we applied the constraint that the average rate for all sites was equal to 1; thus, we set the parameters for the inverse gamma function as $\alpha = \beta$. Only one parameter $\alpha$ was then estimated in the program. We used the short scaffold-only PGK dataset to demonstrate the property of the continuous gamma model and its improvement on the base model, and to compare this model to the discrete SD model and the gamma model in terms of the AIC and the corresponding CPU time increase.

Two models were implemented in this study, the Base model and the continuous SD models. The continuous model also has several approximation versions in which these sites with similar distance are put into one of the predefined group to save computational time. These approximation continuous SD models are different from the discrete SD models: these approximation continuous SD models only have 1 shape parameter while the discrete model has $c$ parameters (with $c-1$ degrees of freedom) where $c$ is the number of discrete categories.

We used the same computers and MrBayes program except that we used up to 8 parallel processes, generated up to 300,000 trees with a sampling frequency at every 100th tree and burn-in of at least 50,000 trees to allow the MCMC procedures to converge properly on the 5PGK dataset.
A total of 13 models with various numbers of categories have been run simultaneously. The model with one category is the same model as the simple base model. The model with 415 groups is the continuous SD model. The remaining models are approximations to the continuous model by pulling the sites with similar distances into different numbers of groups. The rate at each group can be obtained using an inverse Gamma CDF. The CPU time, harmonic Ln(L) and AIC* for all models are summarized in Table 3.6.

<table>
<thead>
<tr>
<th>Number of groups</th>
<th>CPU time (s)</th>
<th>-Ln(L)</th>
<th>AIC*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>722</td>
<td>3861</td>
<td>7724</td>
</tr>
<tr>
<td>4</td>
<td>932</td>
<td>3831</td>
<td>7666</td>
</tr>
<tr>
<td>6</td>
<td>1069</td>
<td>3828</td>
<td>7660</td>
</tr>
<tr>
<td>8</td>
<td>1186</td>
<td>3826</td>
<td>7656</td>
</tr>
<tr>
<td>10</td>
<td>1299</td>
<td>3825</td>
<td>7654</td>
</tr>
<tr>
<td>12</td>
<td>1396</td>
<td>3827</td>
<td>7658</td>
</tr>
<tr>
<td>15</td>
<td>1567</td>
<td>3825</td>
<td>7654</td>
</tr>
<tr>
<td>18</td>
<td>1726</td>
<td>3825</td>
<td>7654</td>
</tr>
<tr>
<td>20</td>
<td>1845</td>
<td>3826</td>
<td>7656</td>
</tr>
<tr>
<td>30</td>
<td>2390</td>
<td>3828</td>
<td>7659</td>
</tr>
<tr>
<td>40</td>
<td>2930</td>
<td>3825</td>
<td>7655</td>
</tr>
<tr>
<td>100</td>
<td>6057</td>
<td>3826</td>
<td>7656</td>
</tr>
<tr>
<td>415</td>
<td>21468</td>
<td>3826</td>
<td>7656</td>
</tr>
</tbody>
</table>

Table 3.6: Summaries of continuous SD models for the five PGK dataset. The standard deviation of the harmonic mean of the Ln(L) ranges from 0.1-0.5 for all models.
According to the Ln(L) and AIC*, it can be see that the continuous SD models, both the approximated models and the continuous model have improved the base model by AIC of 58 to 70. It should be noticed that the -Ln(L) stabilized around 3825 to 3827 after the number of groups increase to 8. On the other hand, the CPU times are roughly proportional to the number of groups used in the model. Every additional group added to the base model will increase about 8-10% of CPU time by the Base model. Thus, using the current algorithm, it is not computational efficient to use the continuous SD model in comparison with the approximated continuous SD models with at least 8 groups.

Compared with the base model, the gamma model, the discrete SD model (SS4), the continuous SD approximation model (10 groups) and the continuous SD model (Table 3.7), all SD models have similar AIC* and Ln(L) values, which are better than the gamma model and the base model. The SS4 model has the least CPU time cost among all three SD models. According to the computation ease and speed, as well as the model fitness, the SS4, the partition model appears to be preferred in this application.

<table>
<thead>
<tr>
<th>Model</th>
<th>ln L</th>
<th>Number of Parameters</th>
<th>AIC*</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (WAG)</td>
<td>-3863</td>
<td>1</td>
<td>7728</td>
<td>635</td>
</tr>
<tr>
<td>Gamma</td>
<td>-3839</td>
<td>2</td>
<td>7682</td>
<td>2349</td>
</tr>
<tr>
<td>SS4</td>
<td>-3824</td>
<td>4</td>
<td>7656</td>
<td>884</td>
</tr>
<tr>
<td>Cont SD 10</td>
<td>-3825</td>
<td>2</td>
<td>7654</td>
<td>1299</td>
</tr>
<tr>
<td>Cont SD</td>
<td>-3826</td>
<td>2</td>
<td>7656</td>
<td>21468</td>
</tr>
</tbody>
</table>

Table 3.7: Comparison of the discrete and continuous SD models for the 5 PGK dataset. The standard deviations of all models are no more than 0.3.
We also implemented the continuous model in the 87 PGK dataset. Figure 3.3 shows the tree distance diagnostics and Figure 3.4 shows the split support plot. Both suggest the MCMC has reached stationary. Table 3.8 lists the Ln(L), AIC* and the CPU time for the base model and three SD models, one SS4 model and two approximate continuous SD model. Again, among the three SD models, the SS4 model spent the least computational time and has the highest AIC*. Thus, SS4 is preferred to the continuous SD model here.

<table>
<thead>
<tr>
<th>Model</th>
<th>Harmonic Mean of Ln(L)</th>
<th>Number of Parameters</th>
<th>AIC*</th>
<th>CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (WAG)</td>
<td>-42,491</td>
<td>1</td>
<td>84980</td>
<td>31,210</td>
</tr>
<tr>
<td>SS4</td>
<td>-41,882</td>
<td>4</td>
<td>83756</td>
<td>82,938</td>
</tr>
<tr>
<td>Cont SD 10</td>
<td>-42,273</td>
<td>2</td>
<td>84542</td>
<td>96,135</td>
</tr>
<tr>
<td>Cont SD 40</td>
<td>-42,227</td>
<td>2</td>
<td>84458</td>
<td>246,923</td>
</tr>
</tbody>
</table>

**Table 3.8:** Comparison of the discrete and continuous SD models for the 87 PGK dataset. The estimated standard deviations of all Ln(L) values for these models range from 0.47 to 0.95.
Figure 3.3: Tree Distance Density Diagnostic for the 87PGK data set based on the continuous SD model. If the two chains from the corresponding two runs follow the same distribution after the burn in period, then the densities estimated in the three panels will be approximately the same.
Figure 3.4: Split support correlation plot on two independent runs for the 87 PGK dataset based on the SD continuous model. If two runs do reach convergence, then, it will have the same split frequency plus random error at each split, so it will form a cloud near the 45 degree line.
Figure 3.5: Site rates estimated from the Gamma model, Discrete SD model, continuous SD model and the continuous SD approximation model vs. site distance for the 5 PGK dataset. The concentration of points near the top and bottom of the graph are due to the bounded nature of the maximum and minimum rates.

Besides the Ln(L) and CPU time, we also examined the correlation between the site rates estimated from different models and the distance of each site (Figure 3.5). The site rates estimated from different models show a similar pattern. For the Gamma model, the site rate and the site distance are moderately correlated with $r = 0.48$, and its corresponding linear regression line roughly parallels to the site rate from the SD models. The site rates based on the SD models can explain about 25% of the site rate variation using only a single or a few parameters. Since these SD models have a better model fitness, the site rate estimated from the SD models seems to better represent the amino
acid evolution in the PGK enzyme. The site rates estimated based on these three SD models are highly correlated, the difference in terms of rate variation among these three SD models are small. Therefore, the discrete SD model is recommended among all three SD models in this application.
CHAPTER 4

GAP DISTANCE MEASUREMENT AND ITS APPLICATION

Multiple sequence alignments (MSAs) are widely used as input for many bioinformatics applications, including phylogenetic tree reconstruction. A collection of homologous amino acid sequences can be obtained from many public databases, such as the Protein Data Bank (www.pdb.org). These homologous sequences can be aligned to form an MSA using many methods, including dynamic programming, progressive methods, and iterative methods. Duret and Abdeddaim (2000) and Notredame (2002) reviewed the wide variety of techniques that have been used to make multiple alignments. The MSAs obtained from different methods or different software are normally different. Since different alignments can lead to very different conclusions, it is critical to have the "right" MSA as the input for the bioinformatics application at hand. However, the accuracy of an MSA remains difficult to judge, due to both computational and biological reasons. Current methods generally rely on two uncertain factors: (1) the selection of a reference MSA that can be used as a “gold standard” to compare with the MSAs produced by various methods; (2) the criterion to measure the “closeness” between the “reference” MSA and the proposed MSA.
In this chapter, we propose a new method, using a gap distance measurement, to evaluate the quality of MSAs. The gap distance measurement overcomes issues with the two factors stated above since it does not need a reference MSA. We will use the tertiary structural information of the sequences to evaluate the quality of the overall MSA and the local alignment. The evaluation of the quality of the alignment in a local region aims to further improve the future alignment quality while the overall alignment evaluation can be used to select the best MSA.

It is widely believed that protein structure information can be helpful in determining the alignment of amino acids (Chothia and Lesk 1986; Godzik 1996; Ginalska 2003; O’Sullivan et al. 2004). Most of these structural alignments depend on the availability of structural information, so they can only be used for sequences whose corresponding structures are known.

Although the number of the sequences with known structure is continuously growing, many sequence structures still have not been identified. For example, the PGK enzyme sequences we collected in the previous study have 87 sequences, and only a few sequences have known structures. Nevertheless, the current best available MSA evaluation criteria, APDB (O’Sullivan et al. 2003) and iRMSD (Armougom et al. 2006), can evaluate the alignment quality of the sequences with structural information only, while ignoring the quality of the alignment connecting the remaining sequences in a data set that are without structural information.

Under the assumption that a protein’s 3D structure is highly conserved, the gap distance measurement proposed here is able to evaluate the MSA for all sequences, not just
those with structural information. It only requires that there are some sequences in the MSA with structural information and that there are gaps in the reference MSA to evaluate.

4.1 Multiple Sequence Alignment and Quality Evaluation

Amino acid sequences can be obtained from various public resources, including SRS (http://srs.ebi.ac.uk/), TIGR (http://cmr.tigr.org/), ExPASy (http://www.expasy.org/), and PDB (http://www.pdb.org/). These primary sequences with the same ancestor (homologous sequences) need to be aligned to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between them. There are various ways to align the sequences. Typically, they can be classified into the following methods according to their computation style: (1) direct methods using dynamic programming, (2) progressive alignment construction techniques, and (3) iterative methods. For amino acid sequences, the direct dynamic programming is based on two set of parameters: an indel/gap penalty and a substitution model that assigns scores for each possible pair of amino acid based on their similarity and mutation probability. This method is the most computationally costly one, and it can only be used to align small numbers of sequences. Progressive alignment methods begin by aligning the two most closely related sequences first and then aligns the next most closely related sequence to the alignment produced in the previous step and continues in this way until the final alignment is produced. This method is based on a guide tree to determine the “closeness” of sequences. CLUSTAL (Thompson et al. 1994) and T-Coffee (Notredame et al. 2000) are two examples of popular progressive alignment software. The iterative method, such as the one
implemented in MUSCLE (Edgar 2004), works similarly to the progressive method but repeatedly realigns the initial sequences while adding new sequences to produce the MSA. Besides the above computationally-based alignment methods, there is another type of alignment method, called structural alignment, which is solely based on the comparison of shape and the tertiary structure of the sequences (Mizuguchi 1998). An example is the MUSTANG algorithm (Konagurthu et al. 2006). Incomplete lists of available programs for multiple sequence alignment and structural alignment can be found here:

(1) http://en.wikipedia.org/wiki/Structural_alignment_software
(2) http://en.wikipedia.org/wiki/Sequence_alignment_software.

In many situations, structural information is believed to be highly conserved and thus able to improve the accuracy of the alignment of proteins (Bryant and Lawrence 1993; Eidhammer et al. 2000; O’Sullivan et al. 2004). The secondary or tertiary structural information of the protein sequences found in PDB have been used in structure based alignment methods in at least two ways: (1) structure superposition methods (Eidhammer et al. 2000), and (2) energy function methods (Bryant and Lawrence 1993). While the former is based on the spatial location of the α-carbon (Cα) of all residuals only, the latter is based on the chemical and physical properties of the residual and its surrounding residuals. Most of these structural alignments depend on the availability of structural information for every sequence being aligned, so they can only be used for those sequences whose corresponding structures are known.

Since there are many alignment methods to produce different MSAs for amino acid sequences, it is important to be able to evaluate which method gives optimal alignment.

90
However, it is very challenging to compare the MSAs produced by these methods due to the two factors stated above.

One method of evaluation is to establish a reference MSA in a real data set to serve as a "gold standard." Currently, there are several collections of datasets that have been commonly used as reference MSAs, for example, BaliBase (Thompson et al. 2005) and Homstrad (Mizuguchi et al. 1998). Both of these collections of MSAs are assembled based on structural information. To compare the accuracy of a proposed MSA with the reference MSA, two criteria have been typically used (Karplus and Hu, 2001): the Sum of Pairs Score (SPS) and the Column Score (CS). The SPS is the proportion of the pair of aligned residuals in the reference MSA that also occurs in the proposed MSA. The CS is the percentage of columns in the reference MSA that also occur in the proposed MSA. Both of these evaluation criteria rely heavily on the correctness of the reference MSA, which could be arbitrary and biased due to its dependence on the individual specialists who created them and the structural superposition methods they selected.

To avoid the potential pitfalls of the selection of the reference MSA and the difficulty of the superposition methods, a score measurement, called “Analyze alignment with PDB (APDB),” was introduced by O’Sullivan et al. (2003). A new type of root mean square distance (iRMSD) was also proposed by Armougom et al. (2006). Both APDB and iRMSD evaluate the proposed MSA based on the structural information given by the sequences in the MSA, not against a reference MSA. The APDB score measures the proportion of pairs of the residuals that have similar residuals in its surrounding area for the sequences with known structural information. On the other hand, iRMSD evaluates the
average difference of the distance between the index pair of residuals to the residuals around them on the two aligned structure. For every MSA produced by different methods, an APDB score and an iRMSD can be calculated. The higher the APDB score, the better the quality of the MSA; while the lower the iRMSD, the better the quality of the MSA.

Both APDB and iRMSD provide a measurement of quality to any MSA even without a reference MSA. However, they can only be used to evaluate the alignment quality of the sequences with structural information. Although the numbers of sequences with known structural information is growing, the proportion of sequences with known structural is still relatively small and shrinking. Since the aligned sequences without structural information in the MSA are not evaluated in the APDB and the iRMSD scores, the information carried by these sequences has to be totally discarded.

A gap distance measurement, including gap distance (GD_P) and Gap Distance Score (GDS), is proposed here to evaluate MSAs based on the putative gaps of all sequences and their relation to those sequences with the structural information. Thus, it not only evaluates alignment quality of the aligned sequences with structural information, but also those aligned sequences without structural information.
4.2 Gap Distance Measure (GD_P) and Gap Distance Score (GDS)

The GD_P and GDS are the two proposed gap distance measurements for evaluating how the alignments agree with the superposition of the sequence structures. Let A and B be the two aligned sequences with gap or insertions and deletions (Indels), and sequence A have a known structure (Figure 4.1).

**Figure 4.1**: Two hypothetically aligned amino acid sequences: A and B

For the two aligned sequences A and B in Figure 4.1, there are two putative gaps on sequence B, one is between amino acids D and K, and the other, between A and L, with a size of 3 residuals and 6 residuals respectively. The position of each gap is defined as the position of the residual right before the gap, or 20 and 47 for the two gaps in sequence B respectively. The distance of each gap in B is defined as the distance in the 3D structure between the α-carbon (Cα) of the corresponding residuals on the reference sequence A, i.e., the distance between D and V and that between S and L in sequence A, denoted as d(B,1) and d(B,2).
The 3D structures of two phospholipase sequences 1bp2 and 1poba and their alignment are illustrated in Figure 4.2. It can be seen that these two enzymes are highly structurally conserved and there is one indel from their alignment. Clearly, we can see from this figure that the insertion in 1bp2 forms a loop so that the gap distance will be very small. The gap distance measurement is based on the structural superposition property of these conserved protein sequences.

**Figure 4.2:** The 3D structures of two phospholipase sequences 1bp2 and 1poba and the alignment. The circled area on 3D structure of the 1bp2 corresponding to the gap in the alignment.
4.2.1 Gap distance measure for each gap (GD_P)

For each gap, we define the following identities: gap sequence, reference sequence, gap location, gap size, and gap distance. The gap sequence is the sequence where the gap is located, and the reference sequence is the sequence with 3D structure and from which the gap distance is estimated. The gap location is the site where the gap starts and the gap size is the number of indels the gap represents in the alignment being evaluated. Here, we further define the gap distance notation:

\[ d(\text{seq1}, j, \text{seq2}, \text{loc}, s) \]

as the gap distance of the \( j \)th gap from sequence seq1, at location =loc, with size= s and according to reference sequence seq2. According to this definition, the distances mentioned above, i.e., \( d(B,1) \), \( d(B,2) \) are denoted as \( d(B,1,A,20,3) \) and \( d(B,2,A,47,6) \) respectively.

Under the assumption that the backbone of the basic 3D structure of the sequence B is the same as that of A (O'Sullivan et al. 2003; Armougom et al. 2006), and the gap is introduced by the insertion of a loop of amino acids, then the distance of the gap should be similar to the distance between the \( \alpha \)-carbons of the two neighbor-residuals in the reference sequence. Intuitively, a gap with large distance suggests a potentially bad alignment between the two sequences at that position. The challenge is to define “large” numerically. We know that the distance between the two \( \alpha \)-carbons of neighboring amino acids is around 4 Angstroms, and that the insertion/deletion of amino acids may change the structure of the protein somewhat, but not radically if it is to maintain its proper function. One overall rule of thumb suggests a distance larger than 10 Angstrom is too large.
(O’Sullivan et al. 2003). However, this rule of thumb neglects the influence of the gap size. Here we propose a new gap distance measure, GD_P.

Let X(s) be the distance vector of all possible amino acid pairs that are s positions away on the reference sequence, which we call the potential gap distance. For each gap with size s, we can obtain an estimated percentile of the observed gap distance according to the empirical distribution of the potential gap distances as follows:

\[ GD_P = p(seq1, j, seq2, loc, s) = \text{prob} (X(s) < d(seq1, j, seq2, loc, s)). \]

Ignoring edge effects, when a gap is inserted randomly on a sequence, this measurement is independent of the gap size since GD_P is comparing the gap distance with the same sized potential gaps. In particular, when the aligned sequences, seq1 and seq2, are aligned randomly and have a size s gap, then the gap distance GD_P will be a random number from X(s). Thus, GD_P would have the same distribution as the cumulative distribution of X(s), i.e., a discrete uniform (0,1) distribution. In addition, GD_P has a mean value at 0.50 and finite variance.

When the two sequences are properly aligned, the gap distance will be around 4 angstroms or a little larger, but should be smaller than most values in X(s). Thus, its GD_P value will be close to 0. Conversely, a large GD_P value indicates a potentially improper local alignment. When GD_P is large, it will thus be worthwhile to check the details of the putatively aligned sequences at the locations of the gap. Thus, by assigning a corresponding penalty for large GD_P values, one may improve local alignments leading to further improvement of the overall alignment algorithm.
4.2.2 Gap Distance Score for the MSA (GDS)

Under the null hypothesis that the two sequences have a similar overall 3D structure but are randomly aligned, the distance of the gap in seq1 will be independent of seq2, then GD_P will have the same distribution as the cumulative distribution of X(s), a discrete uniform (0,1) distribution.

**Theorem:** Let X be a random sample set with n samples from a distribution F(X). We define

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}
\]

where \( I_{X_i \leq x} \) is an indicator function, and define

\[
D_n = \frac{1}{n} \sum_{i=1}^{n} (F_n(x_i) - F(x_i)).
\]

Then, \( D_n \rightarrow 0 \) almost surely as \( n \rightarrow \infty \).

**Proof:** for a fixed x, \( F_n(x) \) is a random variable which converges to \( F(x) \) almost surely by the strong law of large numbers, that is, \( F_n(x) \) converges to \( F(x) \) pointwise. Glivenko and Cantelli strengthened this result by proving uniform convergence of \( F_n \) to \( F \). Thus, \( D_n \rightarrow 0 \) almost surely as \( n \rightarrow \infty \).

Here, let \( F(X) \) be a uniform distribution and \( X_i \) be GD_P, we define GDS as

\[
GDS = 2 \times Dn = 2 \times \frac{1}{n} \sum_{i=1}^{n} (F_n(GD_P) - GD_P)
\]
Then, GDS will have the range of (-1 to 1) and will converge to 0 if the null hypothesis is true. If the alternative hypothesis is true, most GD_P will be close to 0, thus, GDS will be close to 1 (Figure 4.3). A higher GDS suggests a better alignment quality. GDS can be used to evaluate an MSA with at least one known structure.

**Figure 4.3:** GDS under the null hypothesis and the alternative hypothesis. Under null hypothesis (—), the two sequences have a similar overall 3D structure but are randomly aligned and GD_P is a random sample from $U(0,1)$, so $D_n$ is the area above the diagonal line minus the area under the diagonal line, which will be close to 0. Similarly, under the alternative hypothesis (---), $D_n$ will be close to 0.5, so GDS will be close to 1.
4.3 Applications of GD_P and GDS

We applied our gap distance measure on ten reference datasets from the Homstrad database that have at least 15 sequences and at least 100 residuals, to illustrate the behavior of the gap distance measurement and to compare five different alignment methods.

Besides the structural alignments obtained from the Homstrad, four alignment packages, including T-Coffee version 5.31 (Notredame et al. 2000), CLUSTALX version 1.83 (Thompson et al. 1997), 3D-coffee (EXPRESSO) version 5.31 (O'Sullivan et al. 2004) and MUSCLE version 3.52 (Edgar 2004) were used to realign the ten datasets with default parameters. For each dataset, there are five alignments: the structural alignment (from Homstrad), and the four alignments according to the four software packages. Table 4.1 lists the ten selected datasets with the various MSA lengths obtained from these five alignment methods.

The property and behavior of the gap distance and GD_P are illustrated with the Alpha amylase, catalytic domain dataset. GDS was studied using all the ten datasets. We also compared GDS with other two different measurements: APDB and iRMSD.
<table>
<thead>
<tr>
<th>family name</th>
<th>number of sequences</th>
<th>average residual number</th>
<th>average %ID*</th>
<th>Structure Alignmen t length</th>
<th>ClustalX Alignmen t length</th>
<th>Muscle Alignmen t length</th>
<th>T_Coffee Alignmen t length</th>
<th>3D Coffee Alignmen t length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha amylase, catalytic domain (aa)</td>
<td>23</td>
<td>401</td>
<td>23</td>
<td>780</td>
<td>638</td>
<td>780</td>
<td>739</td>
<td>802</td>
</tr>
<tr>
<td>Alpha amylase, catalytic and C-terminal domains (aa_NC)</td>
<td>23</td>
<td>486</td>
<td>22</td>
<td>891</td>
<td>691</td>
<td>902</td>
<td>976</td>
<td>950</td>
</tr>
<tr>
<td>Fatty acid binding protein-like (fabp)</td>
<td>17</td>
<td>132</td>
<td>36</td>
<td>147</td>
<td>145</td>
<td>147</td>
<td>149</td>
<td>147</td>
</tr>
<tr>
<td>Globin (globin)</td>
<td>41</td>
<td>146</td>
<td>32</td>
<td>174</td>
<td>171</td>
<td>178</td>
<td>177</td>
<td>174</td>
</tr>
<tr>
<td>Immunoglobulin domain -- immunoglobulin heavy chain (igvar)</td>
<td>21</td>
<td>123</td>
<td>52</td>
<td>137</td>
<td>137</td>
<td>138</td>
<td>137</td>
<td>137</td>
</tr>
<tr>
<td>Serine / Threonine protein kinases (kinase)</td>
<td>15</td>
<td>320</td>
<td>26</td>
<td>488</td>
<td>444</td>
<td>454</td>
<td>486</td>
<td>490</td>
</tr>
<tr>
<td>Lipocalin family (lip)</td>
<td>15</td>
<td>161</td>
<td>20</td>
<td>210</td>
<td>200</td>
<td>202</td>
<td>209</td>
<td>205</td>
</tr>
<tr>
<td>Phospholipase A2 (phosp)</td>
<td>18</td>
<td>121</td>
<td>49</td>
<td>137</td>
<td>134</td>
<td>134</td>
<td>136</td>
<td>134</td>
</tr>
<tr>
<td>Proteasome A-type and B-type (prote)</td>
<td>17</td>
<td>223</td>
<td>22</td>
<td>309</td>
<td>289</td>
<td>296</td>
<td>314</td>
<td>335</td>
</tr>
<tr>
<td>Serine proteinase - eukaryotic (serine)</td>
<td>27</td>
<td>233</td>
<td>38</td>
<td>309</td>
<td>300</td>
<td>308</td>
<td>309</td>
<td>321</td>
</tr>
</tbody>
</table>

* Percentage of identical amino acids, a measure of the similarity among the sequences.

**Table 4.1:** Multiple Alignment Sequences selected from Homstrad (http://www-cryst.bioc.cam.ac.uk/~homstrad/)
4.3.1 GD_P evaluation

The gap distances, and X(s), of an MSA are estimated using a modified T-Coffee program that can be downloaded at www.stat.osu.edu/~xpan/dist.htm. The percentage GD-P is estimated based on these two outputs using a program in R that can also be downloaded at the above website.

Figure 4.4 shows comparative box plots of potential gap distance X(s) vs. the gap size for the Serine proteinase MSA. It is observed that the gap distance is highly influenced by the gap size. When the gap size is less than 12, the potential gap distance increases as the gap size increases; for those potential gaps with a size larger than 12, they share a similar distribution. Since GD-P is essentially independent of the gap size, it is easier to interpret than gap distance to evaluate the quality of the alignment based on the gap.

Figure 4.5 shows a scatter plot of gap size vs. gap distance (panel A) and gap size vs. GD_P (panel B) for the MSA of Alpha amylase, catalytic domain, aligned using the 3D-Coffee program. From Figure 4.5 (A), one can see that most of the gap distances are less than 10 angstroms (~84.5%), and only a small portion have a value larger than 20 Angstrom (~1.8%). According to Figure 4.5 (B), the majority of GD_P values are less than 0.10 (82.5%). In addition, for the gaps with a small gap size, although it has small distance value, its GD-P value is relatively larger after making the adjustment for the effect of the gap size. Those gaps with large GD-P values (>0.5 or about 5.6% of the total) were further examined by plotting them according to their sequence name and gap location, as shown in Table 4.2 and Figure 4.6. Table 4.2 suggests that the 1ehaa is the sequence that is most poorly aligned with the other sequences since it has 42 gaps with a GD_P greater than 0.5,
and these big gaps are observed among 20 of the 22 reference sequences with structural information. Figure 4.6 shows the gap locations of these gaps. Most of the gaps are in nine locations of the alignment. This indicates that the MSA has potential alignment problems at these nine locations, which need to be further studied.

**Figure 4.4**: Box plot of potential gap distance vs. gap size for the Serine proteinase MSA. When the gap size is less than 12, the potential gap distance increases as the gap size increases; for those potential gaps with a size larger than 12, the distributions appear to be similar.
Figure 4.5: Scatter plot of gap size vs. gap distance (A) and gap size vs. GD-P (B) for the MSA Alpha amylase, catalytic domain, aligned using the 3D-Coffee
**Figure 4.6:** the location of the gaps with GD-P larger than 0.5 for the MSA of Alpha amylase, catalytic domain, aligned using the 3D-Coffee. These are potential poorly-aligned gaps.
| Sequence | 1avaa | 1bag | 1bf2 | 1bvza | 1cgt | 1ciu | 1cyg | 1d3ca | 1ehaa | 1g5aa | 1g94a | 1gcya | 1giwa | 1hx0a | 1jace | 1qhpa | 1smaa | 1smdd | 1uok | 1vjs | 2aaa | 7taa | Grand Total |
|----------|-------|------|------|--------|-----|-----|------|-------|-------|------|------|------|------|------|-------|-------|-------|------|------|------|------|----------|
| 1avaa   | 1     | 2    | 2    | 2      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 13      |
| 1bag    | 1     | 1    | 2    | 2      | 1   | 1   | 2    | 1     | 1     | 1    | 1    | 2    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 17      |
| 1bf2    | 1     | 1    | 1    | 1      |     |     |     | 2     | 1     | 2    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 11      |
| 1bvza   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 6      |
| 1cgt    | 1     | 1    | 2    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 2    | 2    | 1     | 1     | 1     | 1     | 1     | 1     | 15      |
| 1ciu    | 1     | 1    | 2    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 2     | 2     | 1     | 1     | 1     | 1     | 15      |
| 1cyg    | 2     | 1    | 2    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 2    | 2     | 1     | 1     | 1     | 1     | 1     | 16      |
| 1d3ca   | 1     | 1    | 2    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 2    | 2     | 1     | 1     | 1     | 1     | 1     | 15      |
| 1ehaa   | 3     | 1    | 1    | 2      | 2   | 2   | 2    | 3     | 3     | 2    | 2    | 2    | 1    | 2    | 2     | 3     | 2     | 2     | 2     | 2     | 42      |
| 1g5aa   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 5       |
| 1g94a   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 1     | 3       |
| 1gcya   |       | 1    | 1    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 2     | 1     | 1     | 2     | 1     | 1     | 9       |
| 1giwa   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 2    | 2     | 1     | 1     | 1     | 2     | 2     | 13      |
| 1hx0a   |       |      |      |        |     |     | 1    | 1     |       | 1    | 2    | 2    | 2     |     | 1     | 2     |     | 2     | 2     | 1     | 19      |
| 1jace   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 13      |
| 1qhpa   |       |      |      |        |     |     | 1    | 1     |       | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 3       |
| 1smaa   | 2     | 1    | 1    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 2    | 2     | 2     | 2     | 2     | 2     | 2     | 21      |
| 1smdd   | 1     |      |      |        |     |     | 1    | 1     |       | 2    | 2    | 2     |     | 1     | 2     | 2     | 1     | 2     | 2     | 1     | 13      |
| 1uok    |       |      |      |        |     |     | 2    |       |       | 1    | 1    | 2     |     | 2     | 2     | 2     | 2     | 2     | 2     | 2     | 6       |
| 1vjs    | 1     | 1    | 1    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 9       |
| 2aaa    | 1     | 1    | 1    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 7       |
| 7taa    | 1     | 1    | 1    | 1      | 1   | 1   | 1    | 1     | 1     | 1    | 1    | 1    | 1    | 1    | 1     | 1     | 1     | 1     | 1     | 1     | 7       |
| Total   | 8     | 3    | 11   | 10      | 5   | 6   | 5    | 5     | 23    | 13   | 9    | 6    | 21   | 22   | 15    | 6     | 9     | 12    | 14    | 23    | 25    | 20    | 20    | 291     |

**Table 4.2:** The count of the gaps with GD-P larger than 0.5 according to their sequence name and its corresponding reference sequence (for the MSA of Alpha amylase, catalytic domain, aligned using the 3D-Coffee).
4.3.2. GDS and comparison with other evaluation methods

GDS was analyzed based on the gap distance outputs of the modified T-coffee program. An example of the R program and the datasets can be obtained at this website: www.stat.osu.edu/~xpan/diss.htm. The GDS is used to measure the overall quality of an MSA. This score can also be used to evaluate the quality of the MSA aligned by different alignment methods. Here, we have ten datasets with five alignments each as described in Section 4.3.1.

Figure 4.7 plots the GDS score for the ten MSAs. Each point represents the GDS for an MSA. The GDS for the MSA aligned with the same method were connected as a line in the figure. The quality of the alignment method is evaluated for each dataset. For example, for the sequences of Alpha amylase, catalytic domain (aa), the 3d-Coffee method has the highest GDS score while the Clustaw has the lowest. This suggests the MSA aligned by the 3D-coffee program is the best among all these five methods while the Clustaw performed the worst, according to their GDS scores. This pattern appears to be consistent over most of the sequences: 3D-Coffee provided the best alignment on 7 out of the ten datasets, while Clustaw performed poorest on 8 out of the ten. The structural alignment, used as reference alignment in many other MSA evaluation methods, such as column scores, did not perform superior to 3D-Coffee, T-Coffee or Muscle in terms of the GDS. It only beats the other methods on 3 of the ten data sets. This result challenges the traditional practice of using the structural alignment as a "gold standard" to evaluate the quality of the MSA by different alignment software.
**Figure 4.7:** Gap distance scores (GDS) for ten datasets aligned in five methods.

The validation of the GDS was also assessed by comparing with other evaluation methods such as the APDB and iRMSD. The GDS, APDB and iRMSD are shown in Table 4.3. Figure 4.8 (A) (B) are the plots of the APDB and iRMSD on these MSA. Similarly, it appears that Clustaw performed poorly using these alternate criteria on these datasets. In addition, both APDB and iRMSD suggested that 3D-Coffee was superior to the other three methods in general, and Muscle and T-Coffee were similar in performance. These findings are consistent with the results from the GDS evaluation, and are in agreement with recent literature (O’Sullivan et al. 2003; Armougom et al. 2006). However, APDB and iRMSD suggested that the structural alignment generated the best alignment.
Figure 4.8: Comparing the five alignment methods based on APDB (A) and iRMSD (B)
<table>
<thead>
<tr>
<th>MSA</th>
<th>method</th>
<th>APDB</th>
<th>iRMSD</th>
<th>GDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>Clustaw</td>
<td>55.71</td>
<td>1.12</td>
<td>0.688</td>
</tr>
<tr>
<td>aa</td>
<td>3DCoffee</td>
<td>61.6</td>
<td>0.98</td>
<td>0.841</td>
</tr>
<tr>
<td>aa</td>
<td>Muscle</td>
<td>57.87</td>
<td>1.08</td>
<td>0.774</td>
</tr>
<tr>
<td>aa</td>
<td>Structural</td>
<td>68.35</td>
<td>0.92</td>
<td>0.764</td>
</tr>
<tr>
<td>aa</td>
<td>TCoffee</td>
<td>60.37</td>
<td>1.02</td>
<td>0.757</td>
</tr>
<tr>
<td>aa</td>
<td>Clustaw</td>
<td>51.88</td>
<td>1.22</td>
<td>0.547</td>
</tr>
<tr>
<td>aa</td>
<td>3DCoffee</td>
<td>58.4</td>
<td>1.05</td>
<td>0.834</td>
</tr>
<tr>
<td>aa</td>
<td>Muscle</td>
<td>52.45</td>
<td>1.19</td>
<td>0.783</td>
</tr>
<tr>
<td>aa</td>
<td>Structural</td>
<td>63.42</td>
<td>0.95</td>
<td>0.733</td>
</tr>
<tr>
<td>aa</td>
<td>TCoffee</td>
<td>55.32</td>
<td>1.14</td>
<td>0.768</td>
</tr>
<tr>
<td>fabp</td>
<td>Clustaw</td>
<td>74.25</td>
<td>0.79</td>
<td>0.304</td>
</tr>
<tr>
<td>fabp</td>
<td>3DCoffee</td>
<td>80.82</td>
<td>0.66</td>
<td>0.871</td>
</tr>
<tr>
<td>fabp</td>
<td>Muscle</td>
<td>81.01</td>
<td>0.67</td>
<td>0.600</td>
</tr>
<tr>
<td>fabp</td>
<td>Structural</td>
<td>81.28</td>
<td>0.66</td>
<td>0.631</td>
</tr>
<tr>
<td>fabp</td>
<td>TCoffee</td>
<td>80.21</td>
<td>0.68</td>
<td>0.658</td>
</tr>
<tr>
<td>globin</td>
<td>Clustaw</td>
<td>56.4</td>
<td>0.95</td>
<td>0.292</td>
</tr>
<tr>
<td>globin</td>
<td>3DCoffee</td>
<td>56.7</td>
<td>0.93</td>
<td>0.479</td>
</tr>
<tr>
<td>globin</td>
<td>Muscle</td>
<td>57.02</td>
<td>0.93</td>
<td>0.545</td>
</tr>
<tr>
<td>globin</td>
<td>Structural</td>
<td>57.1</td>
<td>0.92</td>
<td>0.430</td>
</tr>
<tr>
<td>globin</td>
<td>TCoffee</td>
<td>57.39</td>
<td>0.93</td>
<td>0.341</td>
</tr>
<tr>
<td>igvar</td>
<td>Clustaw</td>
<td>67.22</td>
<td>1.06</td>
<td>0.104</td>
</tr>
<tr>
<td>igvar</td>
<td>3DCoffee</td>
<td>67.58</td>
<td>1.06</td>
<td>0.835</td>
</tr>
<tr>
<td>igvar</td>
<td>Muscle</td>
<td>67.22</td>
<td>1.06</td>
<td>0.892</td>
</tr>
<tr>
<td>igvar</td>
<td>Structural</td>
<td>67.22</td>
<td>1.06</td>
<td>0.950</td>
</tr>
<tr>
<td>igvar</td>
<td>TCoffee</td>
<td>67.58</td>
<td>1.06</td>
<td>0.593</td>
</tr>
<tr>
<td>kinase</td>
<td>Clustaw</td>
<td>67.16</td>
<td>0.9</td>
<td>0.683</td>
</tr>
<tr>
<td>kinase</td>
<td>3DCoffee</td>
<td>67.08</td>
<td>0.9</td>
<td>0.838</td>
</tr>
<tr>
<td>kinase</td>
<td>Muscle</td>
<td>68.74</td>
<td>0.88</td>
<td>0.681</td>
</tr>
<tr>
<td>kinase</td>
<td>Structural</td>
<td>68.38</td>
<td>0.87</td>
<td>0.788</td>
</tr>
<tr>
<td>kinase</td>
<td>TCoffee</td>
<td>69.72</td>
<td>0.87</td>
<td>0.668</td>
</tr>
<tr>
<td>lip</td>
<td>Clustaw</td>
<td>53.64</td>
<td>1.17</td>
<td>0.654</td>
</tr>
<tr>
<td>lip</td>
<td>3DCoffee</td>
<td>67.4</td>
<td>0.92</td>
<td>0.893</td>
</tr>
<tr>
<td>lip</td>
<td>Muscle</td>
<td>57.71</td>
<td>1.09</td>
<td>0.717</td>
</tr>
<tr>
<td>lip</td>
<td>Structural</td>
<td>70.61</td>
<td>0.86</td>
<td>0.696</td>
</tr>
<tr>
<td>lip</td>
<td>TCoffee</td>
<td>58.3</td>
<td>1.07</td>
<td>0.494</td>
</tr>
<tr>
<td>phosph</td>
<td>Clustaw</td>
<td>79.51</td>
<td>0.69</td>
<td>0.915</td>
</tr>
<tr>
<td>phosph</td>
<td>3DCoffee</td>
<td>81.78</td>
<td>0.63</td>
<td>0.737</td>
</tr>
<tr>
<td>phosph</td>
<td>Muscle</td>
<td>79.18</td>
<td>0.68</td>
<td>0.763</td>
</tr>
<tr>
<td>phosph</td>
<td>Structural</td>
<td>82.79</td>
<td>0.6</td>
<td>0.907</td>
</tr>
<tr>
<td>phosph</td>
<td>TCoffee</td>
<td>80.09</td>
<td>0.68</td>
<td>0.725</td>
</tr>
<tr>
<td>prote</td>
<td>Clustaw</td>
<td>60.04</td>
<td>1.03</td>
<td>0.675</td>
</tr>
<tr>
<td>prote</td>
<td>3DCoffee</td>
<td>64.75</td>
<td>0.92</td>
<td>0.873</td>
</tr>
<tr>
<td>prote</td>
<td>Muscle</td>
<td>62.26</td>
<td>0.97</td>
<td>0.718</td>
</tr>
<tr>
<td>prote</td>
<td>Structural</td>
<td>66.43</td>
<td>0.88</td>
<td>0.698</td>
</tr>
<tr>
<td>prote</td>
<td>TCoffee</td>
<td>64.67</td>
<td>0.93</td>
<td>0.754</td>
</tr>
<tr>
<td>serine</td>
<td>Clustaw</td>
<td>80.96</td>
<td>0.7</td>
<td>0.786</td>
</tr>
<tr>
<td>serine</td>
<td>3DCoffee</td>
<td>85.76</td>
<td>0.6</td>
<td>0.945</td>
</tr>
<tr>
<td>serine</td>
<td>Muscle</td>
<td>84.05</td>
<td>0.63</td>
<td>0.884</td>
</tr>
<tr>
<td>serine</td>
<td>Structural</td>
<td>85.73</td>
<td>0.6</td>
<td>0.945</td>
</tr>
<tr>
<td>serine</td>
<td>TCoffee</td>
<td>84.37</td>
<td>0.63</td>
<td>0.909</td>
</tr>
</tbody>
</table>

**Table 4.3:** APDB, iRMSD, and GDS for all ten datasets aligned with five different methods
It should be noticed that the GDS is more sensitive to the quality of the alignment sometime. For example, the APDB and iRMSD show little differences among the five alignments for igvar and globin, while the GDS still detect the differences among the five alignments. At the same time, while the GDS cannot distinguish well the differences among T-Coffee, Muscle and structural alignments for enzyme aa, the APDB and iRMSD show the difference among the three.

Overall, GDS is a good evaluation of the quality of MSA, and it is different from APDB and iRMSD. The differences among the three evaluation systems are not surprising since GDS evaluate the MSA based on different criteria than APDB and iRMSD.
CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

Phylogenetic studies are probably going to be increasingly based on structural biological data and on statistical formalization (Liò and Goldman 1998). That leads to the study of improved models and of extracting the maximum information from sequence data. In this research, we have successfully incorporated a component of the structural information into two areas related to phylogenetic inference: one is to use a spatially dependent (SD) substitution model for likelihood calculations in phylogenetic inference; the other is to use a Gap distance measure for MSA evaluation. While the first application is to improve the models for phylogenetic inference, the second one focuses on the quality of the MSA produced by different alignment procedures.

5.1 Spatially Dependent Model

The SD substitution model developed here utilizes the protein structural information and the MSA. This improved substitution model has been implemented in the Bayesian Inference framework to reconstruct Phylogeny. This model is inspired by the observation that the amino acid close to the CAC tend to be conserved while these on
the periphery mutate frequently (Pan et al. 2005; Pollack et al. 2008). In this model, the substitution rate of each site is dependent on its distance to the CAC.

During the last few years, several structurally dependent substitution models were proposed (Parisi and Echave 2001; Robinson et al. 2003; Rodrigue et al. 2005). These models have provided significant improvement on the understanding of the rate variation among the amino acids. However, these models are too complicated to be used to estimate the phylogeny relating the sequences in an MSA. Our model can incorporate the spatial location of amino acids in a protein to reconstruct phylogeny and improve the model fitness tremendously. This model can also be use for phylogenetic inference for datasets with more than 50 sequences with small computational cost.

Two sets of SD models, discrete SD models and continuous SD models, were presented in this study. The discrete SD models were implemented in MrBayes, a program for Bayesian phylogeny inference. We found the four-category discrete SD model, SS4, performed well and achieved the best likelihood with the least CPU time increase, and meaningful parameter interpretation. This model partitioned the amino acids into core, middle and peripheral regions according to their distance to the CAC. The model has been applied to 11 large enzyme datasets containing sequences from all three Kingdoms of life and with roughly round structure for phylogenetic inference, and had significant model fitness improvement with log likelihood increases ranging from 62 to 2130 compare to the simple constant rate model. The rates for different partitions are monotonically increasing from core to periphery.
The continuous SD model implemented here uses an inverse gamma continuous SD model. In this model, the substitution rate at each site is dependent on its relative distance measure among all sites according to the inverse Gamma function. The prior distribution of the parameter of the gamma function can be selected either as uniform or exponential. The posterior distribution of the parameter can be estimated from the MCMC procedure as implemented in a modified MrBayes program. This model can also improve the likelihood with reasonable CPU time increase.

A limitation of our investigations of this model is that, the enzymes in this study have roughly round structures. Future work should develop a robust distance measure that can be used for enzymes and other proteins with irregular structures. The definition of the distance of each site requires a thorough understanding of the enzyme function and structure, which is the key to the application of this model. In this dissertation, we did not explore this topic in detail. Furthermore, the inverse gamma distribution is selected arbitrarily in this study based on our observation, other continuous SD functions, not limited to those mentioned in Chapter 2, may be adopted in the future. The discrete SD model selected here can be extended to the maximum likelihood phylogenetic inference. It is also worthwhile to implement this model on some ML program in the future.
5.2 Gap Measurement

Most phylogenetic studies are based on a fixed alignment, derived in advance of phylogenetic analysis and subsequently assumed to be correct, which is not the case in reality. The same set of homologue sequences can be aligned many different ways to form an MSA. The gap distance percentile (GD_P) and the gap distance score (GDS) proposed in this study can be used to evaluate how well the MSA fits the assumption of a conserved protein structure with insertions and deletions leaving the basic backbone unchanged. GD_P evaluates the quality of each gap. A higher value of GD_P indicates a potential mismatch between the sequence where the gap located and reference sequence where the distance value came from, while a lower value indicates a good match. In this study, we evaluated gaps with size greater than three. The cut off value for the GD_P was 0.5. Based on the location and the sequence name of these gaps with large GD_P values, one may identify sequences and locations as having a potential erroneous local alignment. A future direction is to assign a proper penalty for the gap based on its GD_P value and incorporate that into the alignment algorithm, thus improving the overall MSA quality.

GDS is a score evaluating the overall quality of an MSA. Numerically, the GDS value can be any number between -1 and 1. When an MSA aligned totally independent of structural superposition, the GDS value is expected to be around 0; and when the alignment is fully fitting the structural superposition, then GDS will be close to 1. Most MSAs are reasonably correlated to the superposition, so its GDS is expected to be between 0 and 1. A higher GDS value indicates a better MSA. In this study, we evaluated the quality of MSAs by five different alignment methods on ten datasets using GDS and
other MSA evaluation criteria: APDB and iRMSD. The results derived from GDS were comparable with those derived from APDB and iRMSD. We also found that the GDS can sensitively pick up the difference between MSAs in some situations where other evaluation systems cannot.

The strength of the gap distance measurement is that it not only evaluates the quality of the sequences with structural information, but also those without structural information. However, the gap distance measurement is dependent on the existence of gaps and the empirical knowledge of gap distance distribution of the potential gaps. Thus, it is important for the MSA evaluation method to be applicable to large numbers of sequences and long lengths. In this study, we applied the gap distance measurement to MSAs with at least 15 sequences and lengths larger than 100 sites. It will be helpful to study and address the issue of the requirement of the number and the length of sequences in the future.
BIBLIOGRAPHY


Felsenstein, J. (1978), “Cases in which parsimony or compatibility methods will be positively misleading”, Systematic Zoology, 27:401-410


117


Gomberg, D. (1968), ““Bayesian” postdiction in an evolution process”, Unpublished manuscript, Instituto di Genetica, University of Pavia, Italy


Li, S. (1996), *Phylogenetic tree construction using markov chain monte carlo*. Ph.D Dissertation, the Ohio State University, Columbus


Liu, L. (2006), *Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions*. Ph.D Dissertation, the Ohio State University, Columbus
dendrograms using Markov chain Monte Carlo”, *Journal of Computational and

Markov chain Monte Carlo methods”, *Biometrics*, 55:1-12

alignment based on multiple-structure alignment”, *Proteins Structure Function and

accounts for among site rate heterogeneity”, *Bioinformatics*, 21: Suppl 2:i151-i158.

Metropolis algorithms”, *Annals of Statistics*, 24, 101-121

Computational Biology*, 7:761-776

Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J. P. (1998), “HOMSTRAD:
a database of protein structure alignments for homologous families”, *Protein Science*, 7,
2469–2471.

multiple sequence alignments”, *Journal of Molecular Biology*, Vol 302, pp205-217

*Pharmacogenomics*, 3, 131–144

O'Sullivan, O., Zehnder, M., Higgins, D., Bucher, P., Grosdidier, A., Notredame, C.
reference alignments”, *Bioinformatics*, 19, Suppl. 1, i215–i221

“3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence
Alignments”, *Journal of Molecular Biology*, Vol: 340 Issue: 2, pp. 385-395,

amino acid substitution model that allows for rate variation to depend on spatial
location?”, Presentation at the Workshop on Phylogeography and Phylogenetics by
Mathematical and Biosciences Institute, Columbus, OH. September 26-30.


