Université de Montréal

Phylogenetic Shadowing Using a Model Selection Process

par
Mahshid Shakiba

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Juillet, 2006

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Phylogenetic Shadowing Using a Model Selection Process

présenté par:

Mahshid Shakiba

a été évalué par un jury composé des personnes suivantes:

Alain Tapp
président-rapporteur

Miklós Csűrös
directeur de recherche

Damian Labuda
membre du jury

Mémoire accepté le 1er août 2006

RÉSUMÉ

Le génome humain et les génomes de certains autres primates ont été récemment séquencés ou sont en cours de séquençage. Les primates sont d'excellents modèles pour étudier la biologie de l'humain.

La comparaison du génome humain à ceux d'autres primates est d'un très haut intérêt. Cependant, dû à la grande similarité qui existe au niveau des nucléotides entre ces derniers, l'interprétation des résultats des comparaisons entre génomes voisins constitue encore un grand défi. La méthode de "shadowing phylogénétique" a été largement utilisée dans la prédiction de la fonction des régions non codantes. Cette méthode utilise principalement l'approche de la fenêtre coulissante ou bien un modèle de Markov caché qui permettent tous les deux de détecter les régions sous sélection négative.

Ce mémoire présente une nouvelle approche dans la prédiction de régions fonctionnelles dans trois génomes voisins. Dans cette approche nous ne faisons aucune hypothèse quant à la distribution des régions conservées dans le génome. Nous utilisons le principe de la "description de longueur minimale" (MDL) provenant de la théorie de l'information. Cette stratégie permet, non seulement, la prédiction de régions du génome qui se trouvent être sous la sélection négative, mais aussi celles sous la sélection positive. Cela peut s'avérer très utile puisque ces dernières régions définissent souvent les traits biologiques particulier. Notre approche a été testée en utilisant les données de simulation et les alignements multiples des trois séquences génomiques de l'humain, du chimpanzé et du babouin.

Mots clés : Génomique Comparative, Ensemble de Segments de Plus Hauts Scores, Sélection de Modèles, Shadowing Phylogénétique.

ABSTRACT

The genomes of human and a few nonhuman primates have been sequenced and more genomes of primates will be completed in the near future. Nonhuman primates are the most pertinent organisms to comprehend human biology.

There has been a considerable interest in comparing the human genome with the nonhuman primates. However, due to the high degree of similarity between primates at the nucleotide level, interpreting the results between closely related genomes is very challenging. Phylogenetic shadowing has been a widely utilized method in predicting functionality in non coding genomic regions. The main method in phylogenetic shadowing is either a sliding window or a Hidden Markov Model that can detect the regions under negative selection in closely related genomes.

This thesis presents a novel approach to predict functional regions in three closely related genomes. This method does not make any assumptions about the underlying distribution of conserved regions. We use instead an information theoretic approach based on minimum description length. In addition to predicting negative selection, this strategy is used to identify regions under positive selection. Regions under positive selection are likely to determine unique biological traits of species. This approach is tested both on simulated data and on a multiple alignment of human, chimpanzee and baboon genomic sequences.

Keywords: Comparative Genomics, Maximum-Scoring Segment Set, Model Selection, Phylogenetic Shadowing.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Adenine
AIC	Akaike Information Criterion
AR	Ancestral Repeat
bp	Base Pair
BIC	Bayesian Information Criterion
C	Cytosine
DD	Delete-Delete Transition in Tree-HMM
DM	Delete-Match Transition in Tree-HMM
DNA	DeoxyriboNucleic Acid
G	Guanine
GFF	Gene-Finding Format or General Feature Format
HMM	Hiden Markov Model
LLR	Log-Likelihood Ratio
MD	Match-Delete Transition in Tree-HMM
MDL	Minimum Description Length
ML	Maximum Likelihood
MM	Match-Match Transition in Tree-HMM
My	Million Years
ORF	Open Reading Frame
RNA	RiboNucleic Acid
T	Thymine

NOTATION

n	Number of taxa
X	Alignment of DNA sequences
N	Length of alignment
X_i	ith column of alignment
$\Pr(a)$	Probability of a
T	Topology of the phylogenetic tree
$\hat{T}$	Maximum likelihood estimate of phylogenetic tree
t_j	Number of mutations along the branch connecting jth sequence to its ancestor
x_j^i	Nucleotide base at position i in jth sequence of alignment
Q	Instantaneous substitution rate matrix
μ	Mean instantaneous substitution rate
π_A	Equilibrium frequencies of Adenine
π_C	Equilibrium frequencies of Cytosine
π_G	Equilibrium frequencies of Guanine
π_T	Equilibrium frequencies of Thymine
$P(t)$	Substitution probability matrix
R	Ratio of transition to transversion
$E^i(x)$	Emission of sequence x at position i
$M^i(x)$	Transition of sequence x from the match state or * if x does not use match state at position i
$D^i(x)$	Transition of sequence x from the delete state or * if x does not use delete state at position i

π_{MM}	Prior probability of Match-Match transition
π_{MD}	Prior probability of Match-Delete transition
π_{DD}	Prior probability of Delete-Delete transition
π_{DM}	Prior probability of Delete-Match transition
r	Rate constant in match-transition matrix
u	Rate constant in delete-transition matrix
S	A segment
ϕ	A partition
$\|\phi\|$	Size of the partition ϕ
$w(\phi)$	Score of the partition ϕ
$\hat{w}(\phi)$	Complexity-penalized score of the partition ϕ
Z	Labels of each column of alignment after segmentation
C	Number of classes
z_i^c	Indicator variable of column i for class c
α_+	Mutation rate of the regions under positive selection
α_-	Mutation rate of the regions under negative selection
t^c	Divergence time vector for class c
w_i^c	Log-likelihood ratio of column i being in class c versus being in neutrally evolving class
d	Number of variables needed to specify a partition
λ	Penalization factor
$W^c(i)$	Score of the optimal partition for prefix $[1, i]$ that ends with a segment labeled c
Seg_{tol}	Tolerance parameter which controls the convergence of segmentation
$ftol$	Tolerance parameter which controls the convergence of Powell's method

To my beloved parents, Farangis and Sohrab

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisor Professor Miklós Csűrös for his valuable guidance, patience, constructive comments and friendly manner throughout the course of this research. This research would not have been possible without his supervision and support.

I am indebted to André Caron, my former boss at SignalGene Inc, for introducing me to field of bioinformatics.

On a personal basis, I am very grateful to my parents. No words can truly express my deepest and sincere appreciation for all they have done for me.

Last but not least, I would like to thank my husband, Behrouz, for his understanding, love and continuous moral support.

CHAPTER 1

INTRODUCTION

1.1 A Short Review of Biology

Biology is the science that studies living organisms, their structures, functions, origins, and evolution. The number of currently existing species is estimated to be in the order of 10^7 (Lynch 2006). All organisms are made of the same unit of life, the *cell* which has all the essential information and mechanism for its growth, maintenance and reproduction (Hunter 1993). Some organisms like yeast have a single cell. Multicellular organisms have different cell types. There are two major categories of organisms: *eukaryotes* and *prokaryotes*. Virtually every eukaryotic cell contains a nucleus, defined as an area of the cell that holds the genetic material. The nucleus is separated by membranes from the rest of the cell. Organelles such as mitochondria and chloroplasts are specialized cellular structures of eukaryotes. Prokaryotes do not have nuclei and organelles.

The genetic material is deoxyribonucleic acid, abbreviated as DNA. DNA can be linear (in eukaryotes and in some prokaryotes) or circular (in most prokaryotes) (Tamarin 1999). The genetic material of eukaryotes is organized into *chromosomes*. Each chromosome contains a double-stranded DNA molecule, along with a number of proteins. Sexually reproducing organisms have a paternal and a maternal chromosome for every chromosome. These organisms are called diploids. For instance, human cells are diploid, and contain 22 pairs of chromosomes (so-called autosomes), as well as two sex chromosomes.

DNA has a double helix molecular form like a twisted ladder (Watson and Crick 1953). Sugar and phosphate units make up the backbone of helices. Each backbone unit has a nitrogenous base and these bases make the "rungs" of the ladder. Normally, four types of bases are found in DNA: adenine, thymine, guanine and cytosine. Adenine and guanine nucleotides are *purines* and cytosine and thymine

are *pyrimidines*. Figure 1.1 illustrates the structure of the DNA.

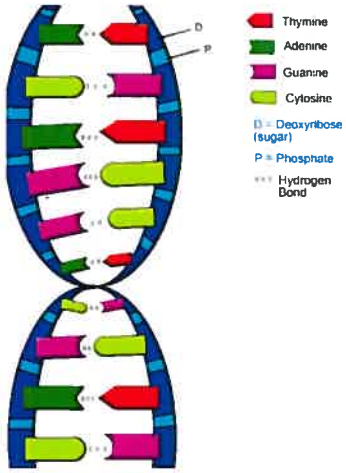


Figure 1.1: DNA structure (from BiologyCorner)

The four bases are abbreviated by the letters A, T, G and C and are concatenated into a sequence to represent a DNA molecule (e.g. CGGTTAC). There is no restriction on base types located on one strand. However, there is a restriction on the bases paired on opposing strands. If adenine is the base of one strand, the other must be thymine; if one base is cytosine, the other should be guanine. This relationship is called *complementarity* (Watson and Crick 1953). Complementary nature of base pairs makes the duplication of DNA an accurate process. The double helix "unzips" and each strand provides a template to create a new strand of DNA. This process results in two double helices exactly like the first. These base pairs are kept to each other by hydrogen bonds. The length of DNA is usually measured in the unit of base pairs (bp).

Functional tasks in a cell are mostly carried out by *proteins* (Hunter 1993). Human being is likely to have more than $100,000$ different kinds of proteins. Proteins are macromolecules that are made of the many combinations of *20 amino acids*. Long proteins may consist of up to 4500 amino acids. This makes the space of all possible proteins structures very large: 20^{4500} or 10^{5850}. However, most of proteins are 10 times smaller. Proteins fold up to make specific three-dimensional forms.

Biochemical functionality of proteins depends on their amino acid compositions and their three dimensional shape. Primary structure of a protein is the sequence of amino acids that makes a specific protein and is coded in DNA by sequence of nucleotide triplets. Each triplet of nucleotides is called a *codon* and corresponds to an amino acid. There are $4^3 = 64$ possible codons. Three of these codons specify the end of a protein sequence and are called stop codons. Other codons code for the 20 amino acids. Hence, the same amino acid can be encoded by several codons. For instance, alanine is represented by codons GCT, GCC, GCG and GCA. There are three possible places to start translating a strand of DNA sequence into amino acids. Each of these parsing is called a *reading frame*. If a reading frame is long enough and does not have a stop codon in the middle is called *open reading frame* or *ORF* and it can be translated into a protein. Since each strand of DNA can be parsed, therefore, there are six reading frames for every DNA sequence.

In most eukaryotes, the stretch of DNA sequence that codes a single protein has some non-coding sequences inserted into them. These non-coding sequences are called *introns* and are spliced out before a sequence serves as a template for protein synthesis (Gilbert 1978). The stretches of DNA that are translated into amino acids are called *exons*. In addition to the coding sequence of proteins, DNA encodes some other information. Every cell in the body has the same DNA but each cell type produces different set of proteins and in different quantities. The DNA signals specify where a protein should start and end; where splicing of intron should occur; how much of each protein should be synthesized. These signals are regulatory elements and are referred to as non-coding functional regions.

All the genetic material of an organism is called its *genome*. *Gene* is the discrete functional unit of genetic material and codes for some products (RNA or protein).

1.2 Evolution and Comparative Genomics

Evolution is the keystone paradigm in biology. Organisms reached their current state through evolution. The similarity of molecular mechanisms in living

organisms is explained by a common ancestry.

An evolutionary process has three elements: inheritance, variation and selection. Inheritance is the transfer of characteristics of parent to offspring. Almost all of the structure and function of an organism is passed by inheritance. The amount of variation between generations is limited and is related to the size of the population (Hunter 1993).

Variation in the inherited material is essential for evolution. Variation is defined as the process that make offspring different from their parents. When some of the bases in a genome are changed or a longer piece of a genome is duplicated or removed, a mutation occurs. Mutation in hereditary material is one of several possible sources of variation (Hunter 1993). Evolutionary changes by mutation are very slow because most of the mutations are deleterious or neutral. Sexual recombination is another source of variation.

Natural selection favors the organisms that have advantages and are better adapted to their environment. Therefore, if the generated variant has an advantage, then these changes propagate through the population with a certain probability (Kimura 1968). This probability is determined by the relationship between the size of the population and the effect of the mutation; small advantages in large populations do not tend to become fixed. In contrast, small disadvantages in small populations may become fixed.

The similarity between living things is the result of inheritance from a common ancestor; the variety comes from the variation and selection elements of evolution (Hunter 1993).

Recently, it has become possible to determine the genome sequence of species. *Genomics*, the most recent branch of biology, studies genomes. The size of the genome varies between organisms. The human genome consists of about 3×10^9 base pairs (Lander et al. 2001).

The focus of the next phase of the Human Genome Project is to find all the functional regions of genome sequences (Hardison 2003). Analysis of the individual genome sequences helps to understand the genome structure but it is not very

informative about genome functions (Miller et al. 2004). Comparative analysis of genome sequences has been and will be a major approach to identify functional regions of each genome (Miller et al. 2004).

It is known that functional sequences are subject to evolutionary selection (Miller et al. 2004). Mutation in functional regions has usually deleterious effect on the organism. Generally, mutations in non-functional regions do not have any effect on the procreative fitness of an organism and will accumulate over time (Kimura 1968). That is the reason functional sequences change more slowly than the non-functional sequences. These regions are referred to be under negative or purifying selection. Non-functional regions are sometimes referred to as neutral evolving regions. However, there might be a slight selective pressure on non-functional regions as well. It is estimated that about 5% of human genome is under purifying selection (Miller et al. 2004); within this subset, 1% to 2% encodes proteins (Margulies et al. 2003).

Positive selection, in contrast, causes sequences to change faster. These regions are often responsible for biological differences between organisms. Positive selection is sometimes referred to as Darwinian selection. One of the aims of comparative genomics is to identify these regions in different genomes. However, predictions about positive and negative selection regions need experimental tests to verify their importance and their functional roles (Miller et al. 2004).

1.3 Phylogeny and Evolutionary Trees

Based on the evolution theory, any set of species are related to each other. The more related two species are, the more recently they diverged from their common ancestor. Understanding the ancestry of the species compared and their relationship is central to many applications of comparative genomics.

Phylogeny studies the relationships between organisms and aims (1) to infer the evolutionary links between organisms and (2) to estimate the time when they shared a common ancestor (Durbin et al. 1998).

Comparative genomics employs phylogeny in order to understand the genomes of different species and to analyze their differences (Mount 2001). At the same time, as new genome sequences and methods to analyze these sequences become available, our understanding about phylogeny improves.

Traditionally, morphological characters such as beak shapes or number of legs have been used to infer the phylogeny. Recently, molecular data like DNA sequences and protein sequences are mostly used for this purpose (Mount 2001).

The evolutionary relationships of a group of organisms is represented by a phylogenetic tree (Eriksson 2004). A phylogenetic tree is a connected, acyclic graph, which directs all edges outward from a designated node, root. In other words, it is an arborescence. A phylogenetic tree is an unordered tree. Organisms under comparison are called taxa. A tree is composed of outer leaves representing the taxa or terminal nodes and nodes and branches representing the relationships between them (Mount 2001). The branch points within a tree are called internal nodes. These nodes are the hypothetical ancestral units and they are used to group existing units. The branching relationships between taxa show how they are related to each other.

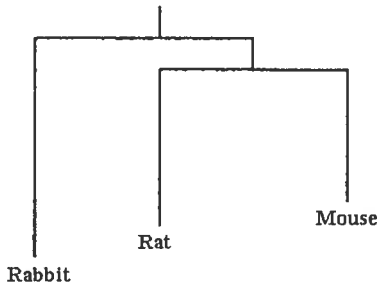
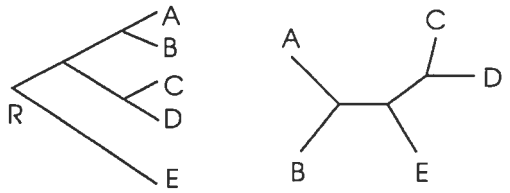


Figure 1.2: Phylogenetic tree for species rat, mouse and rabbit

The tree in Figure 1.2 shows two occurrences of speciation: first the lineage of rabbit had diverged, then the divergence between mouse and rat happened. Usually, the branch length represents the amount of time elapsed since the speciation from a common ancestor but in this figure the scale is not representative.

A real phylogenetic tree has a root, or a common ancestor of all taxa under study. Most phylogenetic inference methods are not informative about the position of the root. An example of a rooted tree is shown in Figure 1.3(a). In the rooted tree, an evolutionary path is defined as the path from root to a node. In the unrooted tree, relationships among taxa are specified but the evolutionary paths are not depicted (Fig. 1.3(b)).



(a) (b)

Figure 1.3: An example of phylogenetic tree; rooted tree (a), unrooted tree (b) (from Singh 1999).

The topology of a tree is the branching pattern of a tree and is denoted by symbol T. In a binary tree, every internal node has two offspring if it is a rooted tree, or 3 neighbors if it is an unrooted tree. In this thesis, trees are assumed to be binary. This assumption about the tree topology is not restrictive because every tree can be approximated by a binary tree with very short branches (Durbin et al. 1998).

A rooted tree with n leaves has $n-1$ internal nodes. This gives $2n-1$ nodes in total. Leaves are labeled. The total number of edges is $2n-2$. An unrooted tree with n taxa has $2n-2$ nodes and $2n-3$ edges. Any unrooted tree can be changed to a rooted tree by placing a root on any of its edges. Therefore, for a given number of n leaves, the number of rooted trees is $(2n-3)$ times the number of unrooted trees (Durbin et al. 1998). The total number of possible unrooted, labeled binary

trees with n leaves is (Felsenstein 2004):

$$B(n) = \prod_{i=3}^{n}(2i-5) \tag{1.1}$$

The length of each branch can represent the number of mutations that occurred in that branch or it can indicate the evolutionary time passed along the branch. Several methods are available in the literature to infer the phylogeny from a given set of sequences. The next chapter discusses phylogenetic inference methods in more details.

1.4 Sequence Alignments

A main use of sequence comparison is to investigate if sequences are related. This is usually done by first aligning the whole or parts of the sequences in question. Sequences can be aligned across their entire length (*global alignment*) or only in some regions (*local alignment*).

Sequence alignment algorithms are looking for proof that sequences under study have diverged from a common ancestor through mutation and selection processes (Durbin et al. 1998). If they have, they are defined as *homologs*. Homologous sequences are either *orthologs*, *paralogs*, or *xenologs*. Genes which are derived from a single gene in the last common ancestor of the sequences under study, are orthologs (Koonin 2005). Genes which are related by duplication event are paralogs. Two genes are xenologs if at least one of them is acquired by interspecies horizontal transfer of genetic material. In Figure 1.4, early globin gene is duplicated and α and β globin genes are formed. α globin genes in mouse, frog and chicken are orthologs. α and β globin genes in mouse are paralogs.

Substitutions, insertions and deletions are the basic mutational processes, which are considered in most alignment methods. Substitutions change nucleotides in a sequence; insertions and deletions add or remove nucleotides. Insertions and deletions are described as *gaps* (Durbin et al. 1998).

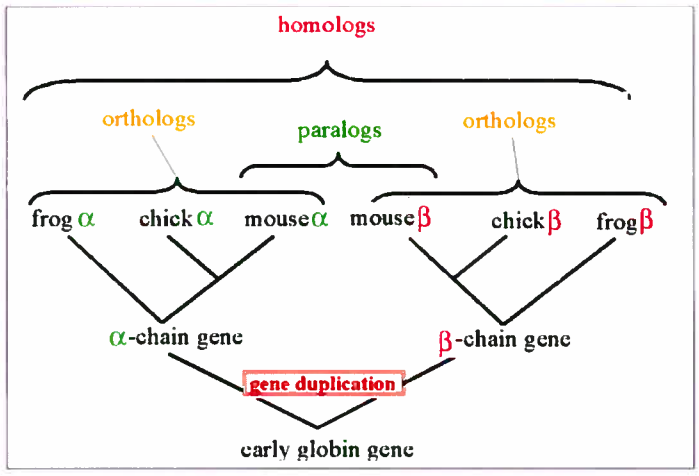


Figure 1.4: Examples of homolog, ortholog and paralog genes (from NCBI)

Here is an example of an alignment of four sequences. The columns in the alignment are also referred to as *sites*.

sequence1:	T	G	G	-	-	-	C	C	A	-	T	C	C
sequence2:	G	G	G	-	-	-	C	C	A	-	T	C	C
sequence3:	A	T	G	-	-	-	T	G	-	-	C	G	C
sequence4:	C	G	G	G	C	G	T	-	-	G	G	C	A

The quality of the resulted alignment is specified by an alignment score. Using this score, the next step is to decide whether that alignment is the result of sequences kinship, or it happened by chance (Durbin et al. 1998). The results of any comparative method depend on the quality of the underlying alignments used as inputs. To this end, the choices of alignment method and the scoring system used to evaluate the alignments are very important.

Two types of alignment algorithms have been generally used in sequence analysis (Yu et al. 2002). The first category looks for the *optimal* alignment (e.g. Smith-Waterman 1981 algorithm). The second type has a probabilistic nature (Yu and Hwa 2001) and searches for the most *likely* alignments (e.g. pairwise hidden Markov models).

In the optimal alignment algorithms, a cost is assigned to each kind of substitutions. Those mutations that are more common will get a smaller cost compared to the less frequent ones. For example, *transitions*, the substitution of purine-to-purine or pyrimidine-to-pyrimidine, are more frequent than *transversions*, which alter the type of nucleotides (Kimura 1981). Hence, transitions are penalized less than transversions. Biologically, deletions and insertions are more likely to occur as a consecutive group rather than to be scattered discretely and therefore, gaps are usually penalized using affine functions. This means that a cost proportional to the length of the gap is added to a cost for opening a gap. The optimal alignment of sequences seeks to minimize the total cost of nucleotide substitutions, insertions and deletions. For more information about these algorithms refer to Jones and Pevzner (2004).

The second type of alignment algorithms are based on a probabilistic framework. These methods assign probabilities to alignments and can be used to train a model on data and to obtain model parameters (Nielsen 2005). The alignment score is usually related to the likelihood of the alignment in a particular probabilistic framework. The resulting probabilistic model can be used to assess the quality of the alignment or to examine other possible alignments. By assigning probabilities to all alternative alignments, the similarity between sequences can be assessed without relying on any specific alignment (Durbin et al. 1998). Only few of these statistical alignment methods take into account the underlying phylogeny of sequences and the likelihood that these sequences have evolved from an unknown root is calculated. The most probable alignment along with the model parameters can be obtained using likelihood maximization or Bayesian techniques (see Durbin et al. 1998; Nielsen 2005).

There are several multiple alignment tools available. Among them are MLAGAN (Brudno et al. 2003), MAVID (Bray and Pachter 2004), DIALIGN (Morgenstern 1999), CLUSTALW (Thompson et al. 1994) and MUSCLE (Edgar 2004). MLAGAN and MUSCLE align DNA and protein sequences respectively. MAVID, DIALIGN and CLUSTALW can be used to align both DNA and protein sequences.

1.5 Comparative Genomics and Phylogeny

The first problem in comparative genomic studies is the choice of species for analysis (Nobrega and Pennacchio 2003) and it has two steps (Pardi and Goldman 2005). First, a range of species are chosen. This selection is based on the biology these species must share. The second step is to choose some of them for analysis and is usually based on maximizing the evolutionary distance between them. The less similar the sequences are, the easier is to discriminate functional conserved regions from neutrally evolving regions. Figure 1.5 shows the methods of comparing genomes at different phylogenetic distances.

Comparisons of distantly related genomes have been widely used to identify shared functionally conserved regions of genomes (Boffelli et al. 2003). For example, the evolutionary distance between human and mouse makes them good candidates for the identification of shared functionally conserved sequences. Since the last time they shared a common ancestor (about 80 million years ago), a large fraction of nucleotides have been changed. However, one can find similarity even between neutrally evolving sequences. These significant changes make it easy to identify functionally conserved regions (Boffelli et al. 2003).

Phylogenetic footprinting (Tagle et al. 1988) is a technique that uses multi-species sequence alignment to identify highly conserved elements in distant species. Since functional regions evolve much slower than nonfunctional sequences, the difference in mutation rates in functional and non-functional regions makes it possible to distinguish these regions from each other. This is achieved by comparing the orthologous regions of related distant species. If these regions have well conserved sequences, it is likely that they are functional (Blanchette et al. 2002).

Comparing distant species cannot identify the recent changes in DNA sequences responsible for primate biological traits (Boffelli et al. 2003). For instance, 20% of human functional elements do not have mouse orthologs (Nobrega and Pennacchio 2003). Therefore, comparison of the human sequence to that of other primates is needed. Figure 1.6 shows the phylogenetic tree of primates. Due to their short

divergence time (apes 6 to 14 My, Old World monkeys 25 My, New World monkeys 40 My), there is not enough sequence variation between human and other primates (Boffelli et al. 2003). More than 90% of human DNA is similar to that of primates (Ovcharenko et al. 2004).

In pairwise comparisons, the lack of sequence variation makes the discrimination of functional from nonfunctional sequences difficult (Boffelli et al. 2003). Considering more species and comparing the genomes of multiple primates can overcome this issue.

Phylogenetic shadowing analyzes the genomes of closely related species and considers the phylogenetic relationship of these species. This approach was first used by Boffelli et al. (2003) to identify coding and non-coding functional regions. Sequences from a set of 18 primates, which had a known exon, were used to estimate the mutation rate of the "conserved" and "non-conserved" regions. They found that the mutation rate for the non-coding regions was 7.3 times higher than the mutation rate of coding regions. They analyzed four genomic intervals with a known exon. Their results show that exon-containing segment has the smallest cross-species variation.

The interspecies comparison has the limitation that it can only be used to identify functional regions that are responsible for shared biological traits and it cannot reveal the features that are unique to a species. Boffelli et al. (2004) tested the use of population shadowing on sequences of the same species. However, this method needs a very large number of sequences from the individuals of the same species. This approach may be more feasible in the near future when large-scale resequencing becomes less costly.

Ovcharenko et al. (2004) developed eShadow, which is a computational tool for the identification of regions under negative selection through multiple sequence alignments of closely related genomes. eShadow applies phylogenetic shadowing and allows dynamic visualization of conservation profile of the genomes. eShadow can be applied to analyze distant genomes (e.g. human and mouse), as well as close genomes (e.g. two primates).

One of the prediction method used in eShadow is a two-state hidden Markov model. Ovcharenko et al. (2004) modeled the distribution of matches and mismatches in slow and neutral evolving regions with an HMM. The HMM parameters can either be obtained from a training dataset or be optimized using Baum-Welch algorithm (see Durbin et al. 1998).

eShadow, is the only existing phylogenetic shadowing tool, but it has a few drawbacks. First, it assumes that the distribution of slow and neutral evolving sites follows an HMM. However, there is no evidence in literature that supports the idea of modeling these regions with an HMM. Second, it can only identify the regions under negative selection, and positive selection regions cannot be identified. Positively selected regions are among the most interesting parts of a genome (Miller et al. 2004). These regions are likely responsible for unique traits of each species. It is with these considerations in mind that we have developed a probabilistic framework that allows the identification of regions under purifying selection, as well as positive selection in three closely related species. In this framework, no assumption is required about the distribution of slow, neutral and fast evolving regions. The model parameters along with the annotation of sequences are calculated by a maximum likelihood method.

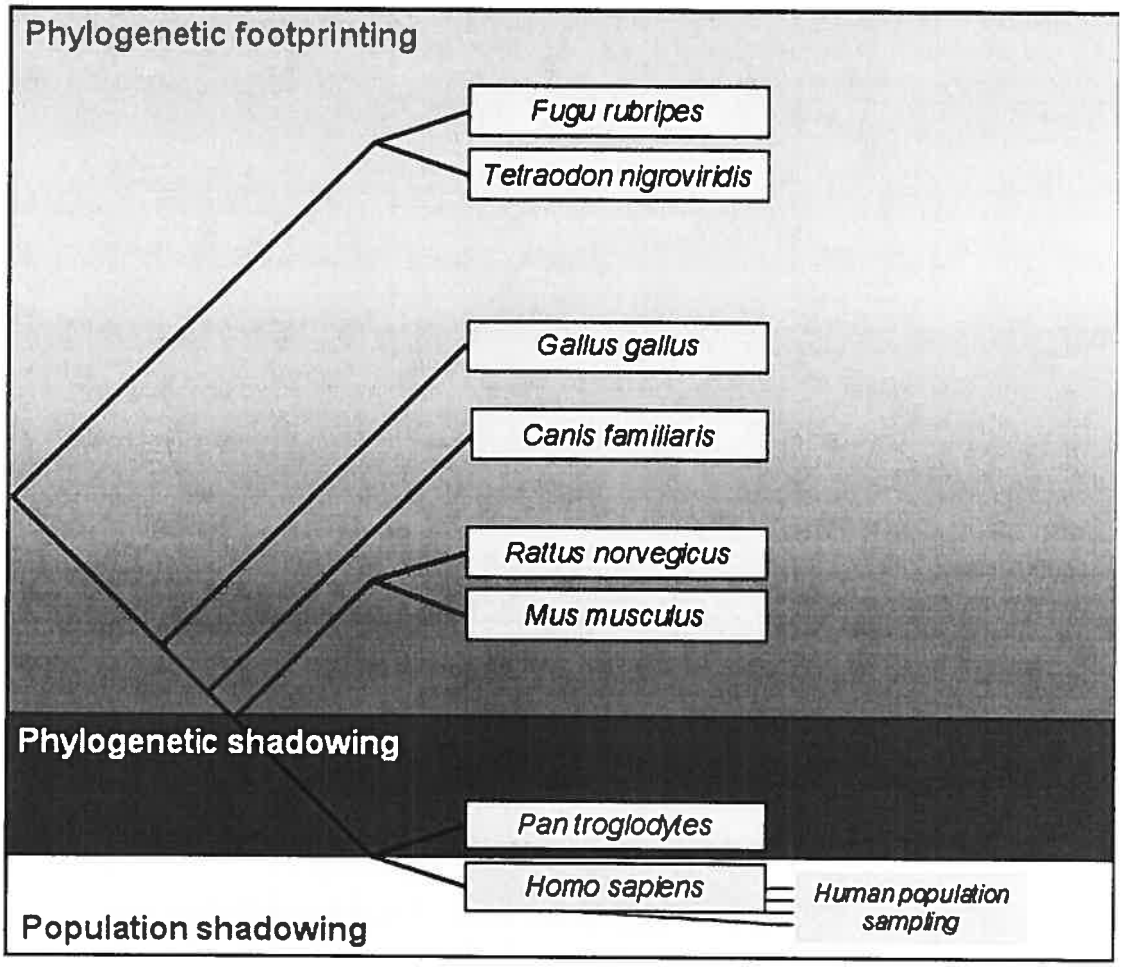


Figure 1.5: Use of genome comparisons at various evolutionary distances to annotate the human genome. Shaded areas representing different methods underlay a phylogenetic tree of selected vertebrates. In this figure, human (Homo sapiens) genome is compared with the chimpanzee (Pan troglodytes), mouse (Mus musculus), rat (Rattus norvegicus), dog (Canis familiaris), chicken (Gallus gallus), and fish (Fugu rubripes and Tetraodon nigroviridis) genomes (from Miller et al. 2004).

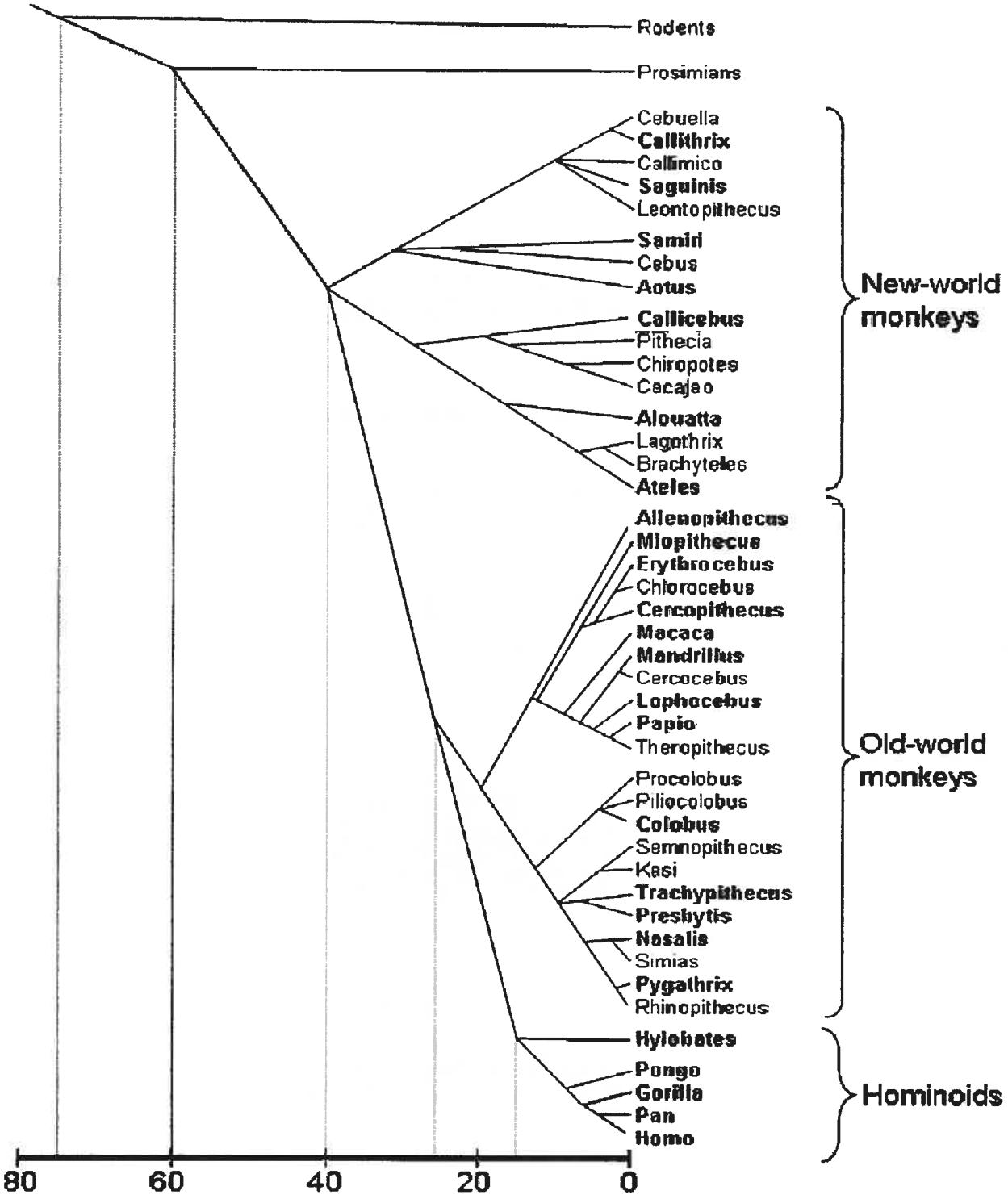


Figure 1.6: Primate phylogenetic tree. As a reference, prosimians' and rodents' age is also shown (from Boffelli et al. 2003)

CHAPTER 2

PHYLOGENY IN PROBABILISTIC FRAMEWORK

In many comparative genomics applications, as we emphasized, it is crucial to understand the evolutionary history of the compared species, or, in more general terms, their evolutionary relationships. To this end, it is important to infer the evolutionary tree topology and timelines (represented by the branch lengths) from the observed data.

Inferring a phylogeny is an approximation procedure which aims to provide the best estimate of an evolutionary history based on the information in the data (Hillis et al. 1996). The statistical and computational aspects of phylogeny reconstruction have been introduced about 40 years ago while phylogenies had been around for more than 140 years (Felsenstein 2004). There are several methods to infer the phylogeny of the sequences under study. Most of these methods use alignments computed in a preparatory step for phylogeny construction. They generally start with a multiple alignment of n sequences (representing n terminal taxa), and return one or more binary trees describing the evolutionary relationships among the sequences. The returned tree (or trees) is produced typically by maximizing some established objective function. The methods can be classified by the nature of their objective function into three main categories. Namely, *distance-based*, *parsimony* and *maximum likelihood* methods. Maximum likelihood approach is the one applied in this thesis. For more information about other methods used for inferring phylogenies see Hillis et al. (1996).

Most current approaches in sequence analysis treat alignment and phylogeny separately, although they are intimately linked (Nielsen 2005). Any error in the alignment can lead to a corresponding error in the identification of the tree (Mitchison 1999). At the same time, aligning DNA and protein sequences is based on the theory that aligned bases are derived from a common ancestor. For this reason, there are methods that perform alignment and tree-building simultaneously.

Notably, the models presented by Thorne, Kishino, and Felsenstein (1991, 1992) known as TKF1 and TKF2, and that of Mitchison and Durbin (1995), known as tree-HMM, estimate evolutionary history and alignment scenario at the same time (Nielsen 2005).

Maximum likelihood estimates, for sequence evolution and phylogenetic trees, are the basic mathematical tools in our methodology. The rest of this chapter is dedicated to introduce these concepts.

2.1 Maximum Likelihood

Maximum Likelihood is a statistical procedure which is used to estimate the parameters of the model that best describes a given data set (Nielsen 2005). For a given data, an analytical function is defined which is the probability of getting that particular set of data under a known model. Maximum likelihood estimates of model parameters is the set of parameter values that maximize the probability of the data.

Inferring phylogeny using maximum likelihood is achieved by evaluating different hypotheses about the evolutionary history of the underlying species. The probability that an explicit model of evolution and the hypothesized history would generate the observed data is calculated. It is assumed that a history with a higher probability to generate the observed data is better than the one with a lower probability of generating the observed data (Hillis et al. 1996).

Maximum likelihood estimation was first applied in phylogenetic inference by Cavalli-Sforza and Edwards (1967). Felsenstein (1981) applied maximum likelihood framework to DNA sequences, and developed the essential computational tools to infer the phylogeny of related species. Later, maximum likelihood was also applied to amino acid sequence data (Kishino et al. 1990; Adachi and Hasegawa 1992).

In order to apply the maximum likelihood method, a model of evolutionary process that accounts for the changes of one sequence into another is required. The model can be completely defined or it may have a few parameters left to

be estimated. A maximum likelihood approach evaluates the probability that the given evolutionary model and the tree topology generated the observed data; the tree that corresponds to the highest likelihood is defined to be the phylogenetic tree of the sequences under study (Hillis et al. 1996). The probability that the tree corresponding to the highest likelihood is the true topology of the taxa at hand increases as the length of alignment gets longer. This means that if sequences at terminal nodes are long enough, ML has a solution for the true tree topology of these terminal nodes; therefore the maximum likelihood method is statistically consistent (Chang 1996).

2.2 Models of Sequence Evolutions without Gaps

An explicit model of sequence evolution is needed to calculate the likelihood of a tree. The model gives the probability of various changes along the edges of the tree (Durbin et al. 1998).

Markov chains, which are stochastic processes, are mostly used to model molecular evolution (Nielsen 2005; Durbin et al. 1998). Usually a Markov model is defined by a set of 'states' and the 'transition probabilities' between states (Durbin et al. 1998). It is often assumed that the sequences evolve independently across different positions, and, thus mutations can be modeled at the single character level, where a Markov process is employed. In the context of DNA sequence evolution, states may be the base nucleotides. Evolution operates as a continuous-time Markov process on each edge of the tree; the processes branch at the tree nodes.

Markov chain models of sequence evolution assume that the probability of a mutation from state i to state j at a given site, does not depend on the history of the site before being in state i (Hillis et al. 1996). For example, if a sequence position at time t_0 has base C and at the later time t_1 has base G; knowing that at sometime prior to t_0 it is been in state A, is irrelevant in calculating the probability of change from C to G at this site.

One of the assumptions made about the Markov models used in sequence evo-

lution, is that as time passes without limits, the probability of being in each state j converges to a value which is non-zero and independent of the starting states. These values are called *equilibrium frequencies* for the base nucleotides (Hillis et al. 1996).

Different authors have described Markov models with different substitution models of evolution (e.g. Felsenstein 1981; Kishino et al. 1990). The substitution model is expressed as a matrix with elements as the probabilities of replacing one nucleotide by another nucleotide in the unit evolutionary time distance. Mathematically speaking, if the instantaneous transition matrix for the underlying Markov process is Q, then the substitution matrix is e^{Q}. For DNA sequences, the instantaneous substitution rate matrix, Q, is a 4×4 matrix and each element of Q, Q_{ij}, is the rate of variation from base i to base j in some infinitesimal time `dt` (Hillis et al. 1996). The most general instantaneous rate matrix is defined as:

$$\mathbf{Q}=\begin{pmatrix} -(a\mu\pi_C+b\mu\pi_G+c\mu\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ d\mu\pi_A & -(d\mu\pi_A+e\mu\pi_G+f\mu\pi_T) & e\mu\pi_G & f\mu\pi_T \\ g\mu\pi_A & h\mu\pi_C & -(g\mu\pi_A+h\mu\pi_C+i\mu\pi_T) & i\mu\pi_T \\ j\mu\pi_A & k\mu\pi_C & l\mu\pi_G & -(j\mu\pi_A+k\mu\pi_C+l\mu\pi_G) \end{pmatrix}$$

The rows and columns of this matrix correspond to the bases A, C, G and T, respectively. The factor μ is the *mean instantaneous substitution rate* and represents the expected number of changes per unit time. This value is different for each pair as it is being multiplied by the relative rate parameters a,b,c,...,l. By convention, μ is assigned to one and matrix Q is scaled in a way that the average rate of substitution at equilibrium is equal to 1 (Hillis et al. 1996). Therefore, timeline of each branch (represented by the branch length) is measured explicitly in number of mutations per site on that branch. π_A, π_C, π_G, π_T are the frequencies of base nucleotides A, C, G and T, respectively. It is assumed that these frequencies do not change over time.

As elements of Q show, the probability of the transition from one base to another

is proportional to the frequency of the target base and does not depend on the base frequency of the starting base. The Q matrix is constrained such that the sum of each row is equal to zero. The Q matrix can be decomposed into matrices R and Π as:

$$\mathbf{R} = \begin{pmatrix} -- & a\mu & b\mu & c\mu \\ d\mu & -- & e\mu & f\mu \\ g\mu & h\mu & -- & i\mu \\ j\mu & k\mu & l\mu & -- \end{pmatrix} \qquad \mathbf{\Pi} = \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}$$

Most of the substitution models reported in literature are special cases of matrix Q. It is generally assumed that the number of substitutions over time t, has a Poisson distribution with mean equals to μt where μ is the expected number of mutations per unit time (Bryant 2003).

The probability of a mutation along a branch of length t is defined as:

$$P(t) = e^{Qt} \tag{2.1}$$

$P(t)$ is referred to as the substitution probability matrix (Hillis et al. 1996). The element $P_{ij}(t)$ is the probability that base i changes to base j in evolution time t. Several important families of substitution matrices are assumed to be *time reversible* and *multiplicative* (Durbin et al. 1998). Substitution matrix is time reversible if the probability of the change from base i to base j is the same as the probability of the change from base j to base i in a given length of time.

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$$

A substitution matrix is multiplicative if

$$P(t)P(s) = P(t+s)$$

If a substitution matrix holds these two properties, then the likelihood of the phy-

logenetic tree does not depend on the position of the root (Durbin et al. 1998). It implies that searching for the best tree can be carried out on unrooted trees. Usually, the hypothetical root of all sequences is placed at an arbitrary position on the tree and the likelihood of this rooted tree is calculated (Hillis et al. 1996).

Most of the widely used substitution models use a simplified version of time reversible form of matrix Q where some constraints are imposed on parameters (Hillis et al. 1996). For example, some of these models, consider two types of substitutions: transversions
($A \leftrightarrow C$,$A \leftrightarrow T$, $G \leftrightarrow C$, $G \leftrightarrow T$) and transitions ($A \leftrightarrow G$, $C \leftrightarrow T$). Felsenstein (1984) defined a model with two types of substitutions: a general substitution which can produce all types of substitutions (transitions and transversions), and substitutions that do not change the type of the nucleotides (transitions). This model, referred to as F84, allows the base frequencies to be different (Hillis et al. 1996). An equivalent model was introduced by Hasegawa, Kishino and Yano (1985) known as HKY85, which only differs in the rate matrix's parametrization. The instantaneous substitution rate matrix for F84 model is defined as:

$$\mathbf{Q} = \begin{pmatrix} -- & \mu\pi_C & \mu\pi_G(1+R/\pi_U) & \mu\pi_T \\ \mu\pi_A & -- & \mu\pi_G & \mu\pi_T(1+R/\pi_Y) \\ \mu\pi_A(1+R/\pi_U) & \mu\pi_C & -- & \mu\pi_T \\ \mu\pi_A & \mu\pi_C(1+R/\pi_Y) & \mu\pi_G & -- \end{pmatrix} \tag{2.2}$$

R is the ratio of transition to transversion; $\pi_U = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. F84 model yields to the following substitution probability matrix:

$$P_{ij}(t) = \begin{cases} \pi_j + \pi_j(\frac{1}{\Pi_j} - 1)e^{-\mu t} + (1 - \frac{\pi_j}{\Pi_j})e^{-\mu t(R+1)} & \text{(i = j)} \\ \pi_j + \pi_j(\frac{1}{\Pi_j} - 1)e^{-\mu t} - (\frac{\pi_j}{\Pi_j})e^{-\mu t(R+1)} & \text{(i} \neq \text{j, transition)} \\ \pi_j(1 - e^{-\mu t}) & \text{(i} \neq \text{j, transversion)} \end{cases} \tag{2.3}$$

In the above equation, $\Pi_j = \pi_A + \pi_G$ if base j is a purine (A or G) and $\Pi_j = \pi_C + \pi_T$ if base j is a pyrimidine (C or T).

2.3 Likelihood of a Tree

Once a substitution model is defined, the likelihood of a given tree is calculated to examine its consistency with the data. Figure 2.1 shows the main steps of calculating the likelihood of a tree.

In Figure 2.1(A), sequences of four taxa with length N are aligned together and we want to calculate the probability of this alignment. Figure 2.1(B) is an example of one of the possible unrooted trees for these 4 taxa. As discussed earlier in this chapter, for the time-reversible models the position of the hypothetical root does not change the likelihood of the tree. Figure 2.1(C) shows an example of a rooted tree that is obtained by placing the root at an arbitrary position. It is assumed that the bases of a sequence evolve independently of each other. Hence, the likelihood of each column can be calculated separately and the likelihood of the tree is the product of the likelihoods for each column in the alignment (Fig. 2.1(E)).
In order to calculate the likelihood of column j, all possible scenarios that could generate the column j should be considered. Each of the hypothetical roots can be an A, a C, a G or a T. Since there are 2 hypothetical roots with 4 possibilities for each, there are $4 \times 4 = 16$ different scenarios that could result in the column j. Given that each of these 16 cases are possible, then the total probability of the column j is the sum of these probabilities (Fig. 2.1(D)).

The probability of alignment of n sequences which are related to each other by a phylogenetic tree, T, is equal to the likelihood of the tree and can be written as :

$$L(T) = \Pr(X|T) = \prod_{i=1}^{N} \Pr(X_i|T)$$

where X is the alignment of length N (Durbin et al. 1998). The probability of each site under a given phylogenetic tree is $\Pr(X_i|T) = \sum_{\mathcal{L}} \Pr(\mathcal{L}, X_i|T)$. In this equation, $\mathcal{L}$ is the state (i.e. A, C, G or T) of the $n-1$ hypothetical ancestral nodes of the tree.

Felsenstein proposed a method, based on dynamic programming, to calculate

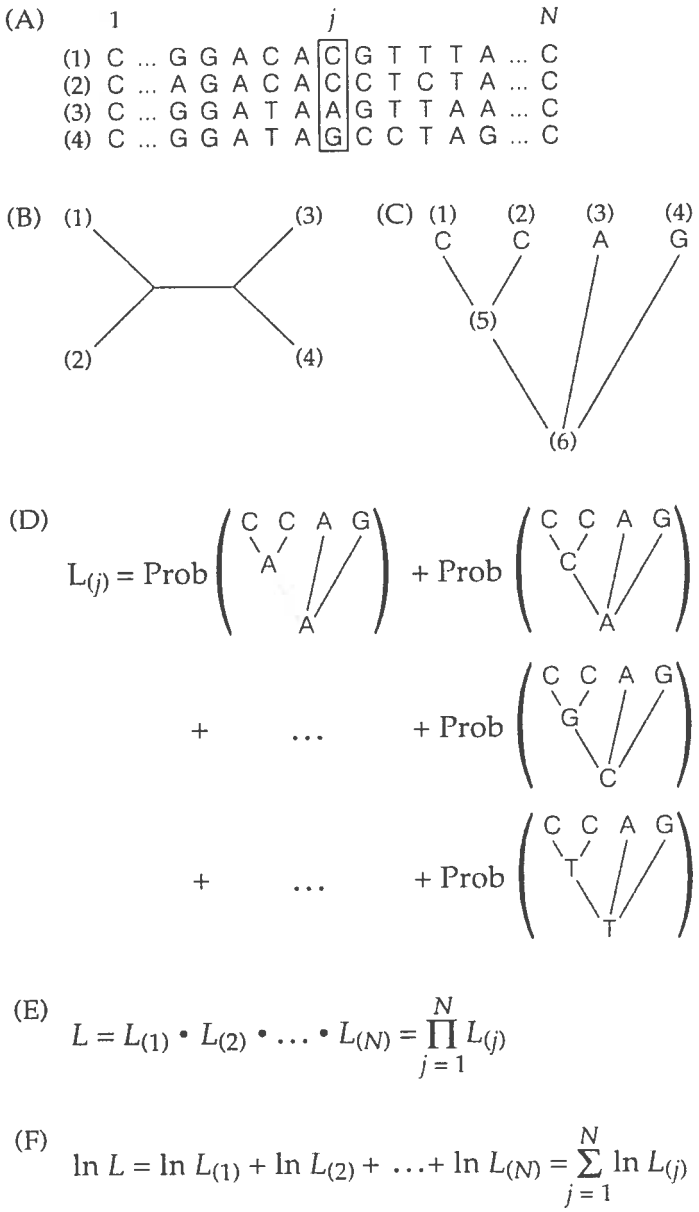


Figure 2.1: Overview of calculating the likelihood of a given phylogenetic tree (from Hillis et al. 1996)

this likelihood (Durbin et al. 1998). A short summary of this method is described next.

Suppose u is a node in T. If u is not a leaf, then let v, w be the children of node u. Suppose t_v and t_w are branch lengths that connect v and w to their parent. If we are given a model of evolution with its substitution probability matrix $P(t)$, then we can calculate the probability that base a changes to base b in time t. This probability is equal to $P_{ab}(t)$. If u is a leaf, $\Pr(L_u|a,T)$ represents the probability of having base a at node u. If u is not a leaf, $\Pr(L_u|a,T)$ is the probability of all the children of u when node u has base a. This probability is calculated from the probabilities $\Pr(L_v|b)$ and $\Pr(L_w|c)$ for all possible values of b and c. The recursions are given as:

$$\Pr(L_u|a,T) = \begin{cases} I(a = x_u) & \text{if } u \text{ is a leaf} \\ \left(\sum_b P_{ab}(t_v) \Pr(L_v|b,T)\right) & \\ \times \left(\sum_c P_{ac}(t_w) \Pr(L_w|c,T)\right) & \text{otherwise} \end{cases} \tag{2.4}$$

x_u is the base of X_i at node u. I is the indicator function. The probability of each column of alignment, X_i, is defined by $\Pr(X_i|T) = \sum_{a\in\{A,C,G,T\}} \pi_a \Pr(L_r|a,T)$, where r is the hypothetical ancestor of all sequences. This recursive procedure, implemented as a dynamic programming algorithm was introduced by Felsenstein (1981). Using this recursion, likelihood of a single tree with n leaves of length N and alphabet of size m, can be calculated in $O(nNm)$ time.

The maximum likelihood estimate of a phylogeny tree can be expressed as:

$$\hat{T} = \arg\max_T \Pr(X|T) \tag{2.5}$$

$\hat{T}$ can be obtained by calculating the $\Pr(X|T)$ for all the possible tree topologies. Finding the optimum tree is a NP-hard problem (Roch 2006; Chor and Tuller 2005). Therefore, for large number of taxa, heuristic search techniques are employed to obtain the near-optimal trees in reasonable computing time (Guindon and Gascuel 2003).

For three sequences, likelihood of column i of the alignment can be calculated as:

$$\Pr(X_i|T) = \sum_{a \in \{A,C,G,T\}} \pi_a P_{ax_1^i}(t_1) P_{ax_2^i}(t_2) P_{ax_3^i}(t_3) \tag{2.6}$$

where x_j^i is the base of jth sequence at column i; and t_j is number of mutations occurred along the branch connecting jth sequence to the common ancestor of all three.

2.3.1 Estimation of Model Parameters

From the above description, it can be seen that several parameters must be estimated from the data using maximum likelihood method. Q, an instantaneous substitution rate matrix; T, the tree topology; t, the vector of branch lengths and equilibrium base frequencies are those that are estimated.

In some simple cases (three or four taxa), the optimum can be found analytically, but in most cases heuristic optimization is necessary. Different optimization methods can be used to estimate the parameters of a given tree. There are two major families of algorithms for optimization (Press et al. 2002). The first family are gradient-based approaches which use the objective function and its derivatives to estimate the optimal values of each parameter. For instance, Newton's method, needs the first and second partial derivatives of the objective function with respect to each parameter (Hillis et al. 1996).

The second category, derivative-free optimization methods, do not need the derivative of the function and therefore are more practical. Brent's method (1973), for a single variable; and Powell's method (1964), for several variables; are examples of this category (Press et al. 2002).

Ideally, the best tree should be found, by searching over the n dimensional parameter space, and globally optimal values of these parameters should be reported at the end. This means that for every possible tree, all parameters should be optimized for that tree. The tree with the highest likelihood is selected as the best model (Hillis et al. 1996).

2.4 Evolutionary Models with Gaps

Alignment of most sequences contains gaps. Mutational changes like insertions, deletions, and rearrangement of genetic materials produce gaps in the alignment. Various methods are implemented in phylogenetic programs to treat the gaps, and each method has its advantages and disadvantages. Commonly, phylogenetic analysis programs ignore the columns of alignment that contain gaps (Siepel and Haussler 2004). This approach has the drawback that the alignment of divergent sequences may have only a small minority of columns without any gaps.

McGuire et al. (2001) suggested to treat the gap as an extra character in the evolutionary model. Using this approach, a multiple-site insertion or deletion is considered as a series of independent events. However, it is well known that gaps tend to be persistent and to occur as a consecutive group rather than to be scattered individually (Durbin et al. 1998). Therefore, this approach overweights multiple-site gaps when it comes to calculating the likelihood of a tree (McGuire et al. 2001).

Boffelli et al. (2003) replace the gaps in each column with the least frequently occurred base in that column. Global equilibrium base frequencies are used to break the ties. Since this thesis deals with three sequences, this approach is not appropriate.

In our program, two different approaches of treating gaps are implemented. Gaps can be treated either as missing data or they can be used to learn about their patterns through a tree-hidden Markov model (tree-HMM) architecture.

In the following sections, first I describe how gap can be treated as missing data. Next, the model proposed by Mitchison and Durbin (1995) is presented. This model, called tree-HMM, allows affine-type gap penalties to be learned and incorporated into the standard evolutionary models.

2.4.1 Phylogeny and Missing Data

Many phylogeny programs, including PHYLIP (Felsenstein 1993), treat gaps as missing data. In order to describe the method more precisely, the notation given by Siepel and Haussler (2004) is followed.

Suppose an alignment X is given. X_i is one of the columns of alignment with gaps. If gaps are replaced with a character from alphabet $\Sigma = \{A, C, G, T\}$, set M is obtained. Since every element of M could lead to have the column X_i, the probability of X_i is obtained by summing over all elements of M.

$$\Pr(X_i|T) = \sum_{y \in M} \Pr(y|T)$$

The elements of X_i that are gaps can be seen as wildcards. For instance, by denoting $*$ to gaps, $X_i = (A, C, G, *, A)^T$ would lead to
$M = \{(A, C, G, A, A)^T, (A, C, G, C, A)^T, (A, C, G, G, A)^T, (A, C, G, T, A)^T\}$. Felsenstein's formulas (E.q. 2.4) are extended to treat gaps as missing data (Felsenstein 2004). Equation 2.4 is generalized to

$$\Pr(L_u|a, T) = \begin{cases} I(a \text{ matches } x_u) & \text{if } u \text{ is a leaf} \\ \left(\sum_b P_{ab}(t_v) \Pr(L_v|b, T)\right) & \\ \times \left(\sum_c P_{ac}(t_w) \Pr(L_w|c, T)\right) & \text{otherwise} \end{cases} \tag{2.7}$$

Thus, if node u has a "*", $\Pr(L_u|a) = 1$ for every possible a. This can be seen as removing the branch connecting u to its parent (McGuire et al. 2001). Therefore, the gaps in the alignment neither remove nor add any information in inferring phylogenies. This approach has no additional cost in calculating the likelihood of the tree. Throughout this thesis, this model is referred to as standard phylogeny model.

In this model, each column of alignment is independent of other columns. This means that if the columns of alignment are shuffled, the likelihood of the resulted alignment is the same as the original alignment. Patterns of insertions and deletions

are not used in this model. Therefore, shared patterns of insertions and deletions which may disclose useful information about the relationship between sequences are ignored and are not taken into account (McGuire et al. 2001).

2.4.2 Tree-Hidden Markov Model

Tree-HMM takes into account the pattern of gaps in the alignment. This model is considered as the composition of an alignment model (i.e. a profile-HMM, Krogh et al. 1994) and a standard evolutionary model (Neyman 1971; Felsenstein 1981). In this model, alignment is regarded as a series of paths through an HMM-profile. Rearranging the columns of alignment would result in an alignment with different likelihood from the original alignment.

Tree-HMM has match (M) and delete (D) states. Insertions are not modeled with an explicit state; they happen when a sequence goes to a match state at a position while its ancestor goes to a delete state.

In order to describe the tree-HMM , we follow the description given by Mitchison (1999). A simple phylogenetic tree, T, is shown in Figure 2.2. This tree has a sequence y at the root and sequence x at the leaf. This tree indicates that sequence x has evolved from sequence y over time t. At each column of the alignment, each sequence has an emission and a transition; these two are referred to as the path of the sequence at that position. Figure 2.2 shows an example of the paths x and y take at four columns of the alignment. The differences in the paths can be either because they emit different nucleotides or they use different states and transitions. A probability is given to each path at every column; multiplying all these probabilities gives $\Pr(x|y,T)$. The probability of having such an alignment is $\Pr(x,y|t,T) = \Pr(y|T)\Pr(x|y,T)$, where $\Pr(y|T)$ is the prior probability of the root sequence.

At position 1, both sequences are in the Match (M) state; x emits A while y emits G. This can be seen as a tree with G at the root and A at the leaf and $\Pr(A|G,t,T) = P_{GA}(t)$ is computed from the substitution probability matrix of the phylogeny model (Eq. 2.1).

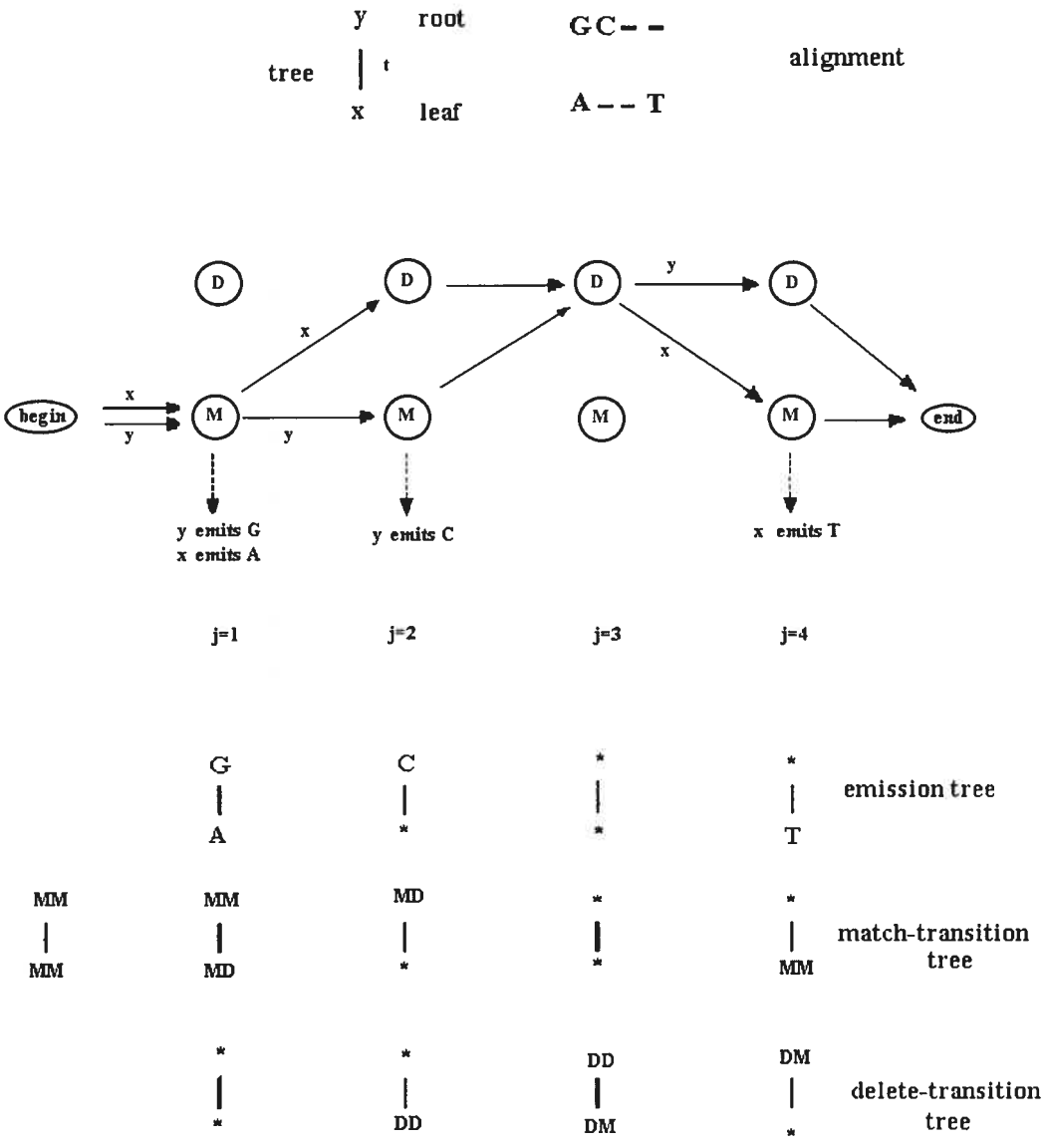


Figure 2.2: A short tree-HMM for a simple tree with two nodes (from Mitchison 1999)

The tree that depicts the substitution of nucleotides is called the *emission tree* in a tree-HMM. The probability of the column 1 in the emission tree is $\Pr(A, G|t, T) = \pi_G \Pr(A|G, t, T)$ where π_G is the base frequency of nucleotide G.

In going from position 1 to position 2, x has a transition from M state to D state, whereas y has a transition from M to M state. "MD" is used for the former transition and "MM" for the latter. This substitution of transitions is denoted by $\Pr(MD|MM, t, T)$. There are four possible substitutions of transitions from the match state and they form a 2×2 matrix:

$$\begin{pmatrix} \Pr(MM|MM, t) & \Pr(MD|MM, t) \\ \Pr(MM|MD, t) & \Pr(MD|MD, t) \end{pmatrix} \tag{2.8}$$

which is called the match-transition family matrix. The substitutions of transitions from match state define the *match-transition tree.* This tree has MM at the root and MD at the leaf at position 1. $\Pr(MD, MM|t, T)$ is equal to $\pi_{MM} \Pr(MD|MM, t, T)$. The prior probability of MM transition is denoted by π_{MM}.

At position 3, x has a DD transition and y a DM transition. The probability of this substitution of transition is denoted by $\Pr(DD, DM|t, T)$ and is equal to $\pi_{DM} \Pr(DD|DM, t, T)$. The delete-transition matrix is defined as:

$$\begin{pmatrix} \Pr(DD|DD, t) & \Pr(DM|DD, t) \\ \Pr(DD|DM, t) & \Pr(DM|DM, t) \end{pmatrix} \tag{2.9}$$

The substitutions of transitions from delete state define the *delete-transition tree.* In positions 1 and 3 both sequences are in the same state, but they are in different states in position 2 and 4; therefore, x can not be seen as being evolved from y (Mitchison 1999). Mitchison and Durbin (1995) proposed that at these positions the transitions of x and y should be regarded independent of each other. The symbol * is used to denote the missing ancestor or descendant sequences at such positions and is replaced by a sum over all possible transitions and emissions.

For example at position 2, delete-transition tree has * at the root and DD at the leaf.

$$\begin{aligned}\Pr(DD,*|t,T) &= \pi_{DM}\Pr(DD|DM,t,T)+\pi_{DD}\Pr(DM|DD,t,T)\\ &= \pi_{DD}\left(\Pr(DD|DD,t,T)+\Pr(DM|DD,t,T)\right)\\ &= \pi_{DD}\end{aligned}$$

The second step in the above calculation is based on the assumption that the delete-transition matrix is reversible. In the similar manner, at position 2, MD is at the root of match-transition tree and * is at the leaf, therefore

$$\begin{aligned}\Pr(*,MD|t,T) &= \pi_{MD}\Pr(MM|MD,t,T)+\pi_{MD}\Pr(MD|MD,t,T)\\ &= \pi_{MD}\end{aligned}$$

At the positions of the tree where ancestor and descendant at both root and leaf are missing, $\Pr(*,*|T)=1$.

$\Pr(x,y|T)$ is calculated by multiplying the probabilities of all transitions and emissions of x and y. Suppose $E^i(x)$ denotes the emission of x at position i. At position i, x can be either in match state or delete state. Let $M^i(x)$ be the transition from the match state or * if x does not use match state at position i. Suppose $D^i(x)$ denotes the transition from the delete state by sequence x at position i or * if delete state is not used. Then the joint probability of alignment of x and y can be expressed as:

$$\begin{aligned}\Pr(x,y|T,t) &= \prod_{i=1}^{N}\Pr(M^i(x),M^i(y))\times\Pr(D^i(x),D^i(y))\\ &\qquad\times\Pr(E^i(x),E^i(y))\end{aligned}$$

where N is the length of the alignment. Usually, the root sequence y is not known

and $\Pr(x|T)$ is calculated by summing over all possible paths of the root sequence:

$$\Pr(x|T) = \prod_{i=1}^{N} \Pr(M^i(x)|T) \Pr(D^i(x)|T) \Pr(E^i(x)|T)$$

$\Pr(E^i(x)|T)$, the probability of the emission of x at position i, is calculated by summing over all possible root residues. $\Pr(M^i(x)|T)$ is obtained by summing over all possible match transitions of the root. Similarly, $\Pr(D^i(x)|T)$ is defined by summing over all possible delete transitions of the root.

The probability of a tree T with n leaves modeled by a tree-HMM is:

$$\Pr(x_1, \ldots, x_n|T) = \prod_{i=1}^{N} \prod_{S} \Pr(S^i(x_1), \ldots, S^i(x_n)|T) \tag{2.10}$$

where S is E, M or D; and N is the number of columns in the alignment (Mitchison 1999). This probability can be calculated using the dynamic programming algorithm of Felsenstein. The above probability can be regarded as:

$$\begin{aligned} \Pr(x_1, \ldots, x_n|T) &= L(\text{emission tree}) \times L(\text{match-transition tree}) \\ &\quad \times L(\text{delete-transition tree}) \end{aligned} \tag{2.11}$$

where L is the likelihood. Since the columns of alignment are assumed to be independent of each other, likelihood of each tree is the product of the likelihoods for every column of alignment for that tree.

Mitchison used the following form for the match-transition matrix

$$\begin{pmatrix} a + (1-a)e^{-rt} & (1-a)(1-e^{-rt}) \\ a - ae^{-rt} & 1 - a + ae^{-rt} \end{pmatrix} \tag{2.12}$$

The rows and columns of this matrix correspond to the MM and MD transitions respectively. In this matrix, t represents the time elapsed since sequence x diverged from its ancestor; $r \geq 0$ is a rate constant; and $0 \leq a \leq 1$ is the equilibrium

probability of "MM" transition. The equilibrium frequencies are used as the priors: $\pi_{MM} = a$ and $\pi_{MD} = 1-a$. This form of matrix is time reversible and multiplicative (Mitchison 1999). Delete-transition matrix is also assumed to have the same form but with different parameters.

$$\begin{pmatrix} b+(1-b)e^{-ut} & (1-b)(1-e^{-ut}) \\ b-be^{-ut} & 1-b+be^{-ut} \end{pmatrix} \tag{2.13}$$

The rows and columns correspond to DD and DM respectively. Similarly $\pi_{DD} = b$ and $\pi_{DM} = 1-b$ and $u \geq 0$ is a rate constant.

As previously mentioned, in this thesis we are working with three sequences. Likelihood of column i in a tree-HMM for three sequences is calculated as:

$$\begin{aligned} \Pr(X_i|T) \;=\; & \left(\sum_{v\in\{A,C,G,T\}} \pi_v \left(\Pr(E^i(x_1)|v,t_1) \Pr(E^i(x_2)|v,t_2) \Pr(E^i(x_3)|v,t_3) \right) \right) \\ & \times \left(\sum_{m\in\{MM,MD\}} \pi_m \left(\Pr(M^i(x_1)|m,t_1) \Pr(M^i(x_2)|m,t_2) \Pr(M^i(x_3)|m,t_3) \right) \right) \\ & \times \left(\sum_{o\in\{DD,DM\}} \pi_o \left(\Pr(D^i(x_1)|o,t_1) \Pr(D^i(x_2)|o,t_2) \Pr(D^i(x_3)|o,t_3) \right) \right) \end{aligned} \tag{2.14}$$

where $[t_1, t_2, t_3]$ is the divergence time vector. In Equation 2.14, $\Pr(E^i(x_j)|v,t_j)$ is calculated from the substitution probability matrix $P(t)$ (Eq. 2.3); $\Pr(M^i(x_j)|m,t_j)$ is calculated from the match-transition family matrix (Eq. 2.12); and $\Pr(D^i(x_j)|o,t_j)$ from delete-transition family matrix (Eq. 2.13).

2.5 Phylogenetic Analysis Tools

Several phylogenetic analysis programs are available. The main ones are PHYLIP, phylogenetic inference package (Felsenstein 1989-1996); at `http://evolution.genetics.washington.edu/phylip.html` and PAUP,

phylogenetic analysis using parsimony; at `http://www.lms.si.edu/PAUP/`.
The three main methods of phylogenetic analysis (parsimony-based, distance-based and maximum likelihood) are implemented in these two packages. A comprehensive list of available packages and servers are listed at `http://evolution.genetics.washington.edu/phylip/software.html`.

There are also a few useful Web sites that provide information on phylogenetic relationships among species and organisms. Among them, tree of life at `http://tolweb.org/tree/` and Taxonomy browser at `http://pubmedexpress.nih.gov/Taxonomy/` can be mentioned.

CHAPTER 3

METHODOLOGY

The assumption about equal rate of evolution along a sequence is often unrealistic (Felsenstein and Churchill 1996). Mutational and selective pressures vary with nucleotide position in a genome (Nielsen 2005). The evolutionary forces affecting a nucleotide in a genomic sequence depend on miscellaneous factors, including local sequence context, and whether the nucleotide belongs to coding or non-coding region. The model of sequence evolution that imposes the same rate of evolution across all columns of a multiple alignment is used as *null hypothesis* in this thesis. Felsenstein (1981) and Neyman (1971) used this null hypothesis in maximum likelihood methods to infer phylogenies (Felsenstein and Churchill 1996).

Heterogeneous evolutionary rates at different sites in molecular sequences are addressed by different authors, including Yang (1993, 1994); Kelly and Rice (1996); Felsenstein and Churchill (1996).

We have employed heterogeneous evolutionary rates to reach our objective which is to find the signatures of negatively and positively selected regions in closely related genomes. Here, we limit the number of sequences under study to three. However, our method can be readily adapted to more than three sequences. The sequences under study can be aligned using any publicly available multiple alignment tool. The alignment of these sequences is used as the input of our program.

Three rate categories are defined: neutral, slow and fast. Neutral rate regions represent the positions in a genome that have accumulated mutations over time and have not been under selection. These regions are assumed to be non-functional.
Slow rate regions represent the positions in a genome that have been under negative selection and have undergone little changes since the separation from the species' most recent common ancestor. These regions are likely to be functional.
Fast rate regions represent the positions in a genome that have evolved faster than

neutrally evolving regions. These regions are likely to determine unique characteristics of the species.

Assuming three rate categories, there are 27 different ways of assigning them on three branches. We aimed at a reasonable compromise between statistical power (few column classes, few parameters) and model accuracy and considered only five of these column classes. We assumed that slow mutation rates operate on all three branches simultaneously. In contrast, fast mutation rate is assumed to affect a single branch only. Since three sequences are used and each of them may have a region under positive selection, fast evolving regions belong to one of three classes (i.e., one for each sequence). Each position in the alignment can be labeled either as neutral (i.e., all three are neutral), slow (i.e., all three are slow) or fast for a particular species. Therefore, each column is annotated by one of five classes.

The cases we did not cover are supposed to be rare or even implausible. One could also use many more rate categories and column classes. However, this would lead to a more complex model with many parameters which has less statistical power. This may not be desirable.

The problem of assigning labels to each position in the alignment is very similar to that of identifying homogeneous regions in DNA sequences. One generic feature of DNA sequences is that their statistical properties vary from position to position along the sequence (Sueoka 1962). This means that the density of any feature of interest, such as G+C content, the CpG dinucleotide content, fluctuates along the sequence. Usually these variations can be better explained by alternating homogeneous domains or segments than by random fluctuations in a homogeneous sequence (Li et al. 2002). Finding the borders between homogeneous regions in molecular sequences is an important task and has been extensively studied (Li 2001; Li et al. 2002).

Commonly, a moving window is used to see the changes of the features of interest along the sequence (Li 2001). By visualizing the variation, borders are usually specified in an ad hoc way.

Boffelli et al. (2003, 2004) have used this approach to identify the regions un-

der negative selection. The log likelihood ratio (LLR) under neutral versus slow-mutation for each column of alignment was calculated and was plotted. They showed that functional regions had the least amount of cross-species variation.

In order to identify homogeneous domains, different mathematical approaches are used. These approaches are known as "segmentation", "partitioning", or "change-point analysis" in different fields (Li 2001). Hidden Markov models, maximum likelihood estimation, and entropy based methods are among the various segmentation approaches (Csűrös 2004).

In this thesis, a novel method, based on work of Csűrös (2004), is used to segment the alignment into regions with different evolutionary rates.

3.1 Maximum Likelihood Estimate of Segments

Suppose $X = X_1, \ldots, X_N$ is an alignment of three sequences with length N. Each X_i represents the ith column of the alignment and is a letter from alphabet $\{\sigma_1, \sigma_2, \ldots, \sigma_{125}\} = \Sigma^3 \setminus \{(-,-,-)\}$ where $\Sigma = \{A, C, G, T, -\}$.

A *segment* S is defined as an interval $S = [a, b] = \{a, a+1, \ldots, b\}$ where $1 \leq a, b \leq N$, and all X_i, $i \in [a, b]$, belong to the same column class. This implies that base nucleotides at every position in this interval are under the same selective pressure for each sequence (i.e., they are homogeneous). A label is assigned to segment S which represents the class of every column of S.

When standard phylogeny model is applied, we are interested in a probabilistic model where X depends on an unobserved random sequence $Z = Z_1, \ldots, Z_N$ in such a way that each X_i depends solely on Z_i. The Z_i are called *segment indicators* and take values in a finite set $\mathcal{L}$. In the context of phylogenetic shadowing, $\mathcal{L}$ is the set of column classes which represent the different levels of selection operating along the sequence. We are in particular considering the situation where $\mathcal{L}$ has five elements: neutral class, negative selection (on all three branches), and three lineage-specific positive selection classes. The categories are parametrized using two rate factors: α_- for negative selection, and α_+ for positive selection. For the case that a column

is under no pressure, the divergence time vector is $t^0 = [t_1, t_2, t_3]$. Neutral class is depicted by label "0". If the first sequence in the alignment is under Darwinian selection at a given column and is evolved at a higher rate, then this column is labeled by class "1". The divergence time vector is $t^1 = [(\alpha_+)t_1, t_2, t_3]$ for class "1". If second or third sequences are under positive selection, then the column is labeled by class "2" or "3", respectively. The divergence time vectors for classes "2" and "3" are $t^2 = [t_1, (\alpha_+)t_2, t_3]$ and $t^3 = [t_1, t_2, (\alpha_+)t_3]$, respectively. If all the three sequences are under negative selection at a given column, the column is labeled by class "-1"; the divergence time vector is defined as $t^{-1} = [(\alpha_-)t_1, (\alpha_-)t_2, (\alpha_-)t_3]$. Table 3.1 summarizes the definitions of five classes.

The segmentation task is to form a hypothesis about Z, after seeing X only. For

Table 3.1: Labels and divergence time vectors for column classes

Class Label	Divergence Time Vector	Description
0	$t^0 = [t_1, t_2, t_3]$	All three under no selection
1	$t^1 = [(\alpha_+)t_1, t_2, t_3]$	First sequence under positive selection
2	$t^2 = [t_1, (\alpha_+)t_2, t_3]$	Second sequence under positive selection
3	$t^3 = [t_1, t_2, (\alpha_+)t_3]$	Third sequence under positive selection
-1	$t^{-1} = [(\alpha_-)t_1, (\alpha_-)t_2, (\alpha_-)t_3]$	All three under negative selection

fixed parameters, we choose Z based on a score that combines likelihood with a complexity penalty. In order to estimate the parameters, as well, we use a heuristic likelihood maximization method embedded in an expectation maximization framework, where in alternating steps, a hypothesis Z is chosen, then parameters are estimated using a chosen hypothesis.

When tree-HMM is applied, the probabilistic model is similar to standard phylogeny, described above, except that there is a Markov-dependence between X_i and X_{i-1}.

The likelihood of the alignment X under the hypothesis Z can be written as:

$$\Pr(X = x|Z = z) = \prod_{i=1}^{N} \Pr(X_i = x_i|\, Z_i = z_i)$$

where x_i is the observed sequence at position i of random variable X, and z_i is the hypothesis for the random variable Z at position i. The log-likelihood of the alignment under the hypothesis $Z = z$ can be written as:

$$\begin{aligned} \log \Pr(X = x|Z = z) &= \sum_{i=1}^{N} \log \Pr(X_i = x_i|\, Z_i = z_i) \\ &= \sum_{i=1}^{N} \sum_{c \in \{0,1,2,3,-1\}} \log \Pr(X_i = x_i|\, Z_i = c)\; z_i^c \end{aligned} \tag{3.1}$$

$$z_i^c = \begin{cases} 1 & \text{If } c = z_i \\ 0 & \text{If } c \neq z_i \end{cases}$$

Equation 3.1 can be written as:

$$\begin{aligned} \log \Pr(X = x|Z = z) &= \sum_{i=1}^{N} \sum_{c \in \{0,1,2,3,-1\}} \log \Pr(X_i = x_i|\, Z_i = c)\; z_i^c \\ &\quad + \sum_{i=1}^{N} \log \Pr(X_i = x_i|\, Z_i = 0) - \sum_{i=1}^{N} \log \Pr(X_i = x_i|\, Z_i = 0) \end{aligned} \tag{3.2}$$

A score is defined for each column of alignment i and for every class c such as:

$$w_i^c = \log \Pr(X_i = x_i|\, Z_i = c) - \log \Pr(X_i = x_i|\, Z_i = 0) \tag{3.3}$$

Using these scores, Equation 3.2 can be written as:

$$\log \Pr(X = x|Z = z) = \sum_{i=1}^{N} \log \Pr(X_i = x_i|\, Z_i = 0) + \sum_{i=1}^{N} w_i^0 z_i^0 + \cdots + \sum_{i=1}^{N} w_i^3 z_i^3 + \sum_{i=1}^{N} w_i^{-1} z_i^{-1} \quad (3.4)$$

The first term of Equation 3.4 is the log-likelihood of the null hypothesis assuming all columns evolved with the same rate of evolution and there were no selective pressure on any of them. The remaining terms form the LLR of the hypothesis $Z = z$ versus the null hypothesis. We are interested to find a hypothesis with the highest likelihood. Therefore, $\log \Pr(X = x|Z = z)$ should be maximized. Maximizing $\log \Pr(X = x|Z = z)$ leads to maximizing the LLR of the hypothesis z versus null hypothesis.

Csűrös (2004) introduced a concept which can be used in finding hypothesis Z. A *partition* ϕ is defined as a set of non-overlapping segments that span the whole sequence. A partition is thus a set of segments $\phi = \{S_1, S_2, \ldots, S_k\}$, along with the classes assigned to each $S \in \phi$. It is assumed that neighboring segments belong to different classes. If each column of $S = [a, b]$ has a score w_i, $i \in [a, b]$, the score of segment S will be $w(S) = \sum_{i \in S} w_i$. The score of the partition ϕ, is the sum of the segment scores: $w(\phi) = \sum_{S \in \phi} w(S)$.

Clearly, every segmentation z defines a partition and vice versa. If the score of each segment is defined to be the LLR of the segment belonging to the assigned class versus being in neutrally evolving, the highest-scoring partition corresponds to a segmentation z, which is the hypothesis with the highest likelihood.

As a first step toward finding z, the score of each column for every class should be calculated. If standard phylogeny model is used, the likelihood of a column of alignment is calculated using Equation 2.6. The divergence time vector for each class defined in Table 3.1, is used in each calculation. If tree-HMM is used as the model, the divergence time vector for each class is used in Equation 2.14 to obtain the likelihood of each column being in each class. Once likelihood for each column

and for every class is calculated, log likelihood ratio for each column being in each class versus being in neutrally evolving class is calculated from Equation 3.3.

It can be seen that the alignment would have the highest likelihood if the labels are chosen as follows: for each column the label corresponding to the class with the highest log likelihood ratio is used to annotate the column. However, the result of such a segmentation would be a partition with too many segments and is not representative for any meaningful pattern in the data. It has been suggested by different authors to view the segmentation as a model selection process (Csűrös 2004; Li 2001). For more information about model selection framework, see Burnham and Anderson (2002).

Regarding segmentation as a model selection process requires that a measure of the "merit" of a segmentation should be specified (Li 2001). Csűrös (2004) suggested to penalize the number of segments (i.e. partition size) and to use it for defining the "merit" of a segmentation. The score of a partition ϕ is modified as $\hat{w}(\phi) = w(\phi) - r(|\phi|)$. In this term, $r : \mathbb{N} \mapsto [0, \infty)$ is a monotone increasing penalty function and $|\phi|$ is the partition size. This score is called *complexity-penalized score* of a partition ϕ. The optimal partition is defined as the partition which has maximum complexity-penalized score.

Commonly used penalty functions are: Minimum Description Length (MDL), $d \log N$; Bayesian Information Criterion (BIC), $\frac{d}{2} \log N$; and Akaike Information Criterion (AIC), d; where N is the sample size and d is the number of variables needed to specify a partition (Csűrös 2004). d is directly related to the partition size (i.e., k). If penalty function is linear, then $\hat{w}(\phi) = w(\phi) - \lambda\, k$.

In Csűrös (2004), different algorithms are introduced to find the optimal partition when there are only two classes. If the complexity penalty function is linear, a dynamic programming algorithm is given that runs in $O(N)$ time. This algorithm is like the Viterbi algorithm in a two-state HMM model. For details about Viterbi algorithm refer to Durbin et al. 1998.

The algorithm MAXCOVER-C is implemented in this thesis to extend the dynamic programming algorithm from two classes to C classes (Cheng 2006). This

algorithm runs in $O(NC)$ time. The key to understand this algorithm is in the definition of $W^c(i)$ which is the score of optimal partition (i.e., $\hat{w}(\phi)$) for prefix $[1, i]$ that ends with a segment labeled c. At each column i, $W^c(i)$ is calculated for each class from $W^j(i-1)$ where $j = 1, \ldots, C$ and LLR of column i for class c (i.e. w_i^c). label is a $C \times N$ matrix. label$[c, i]$ is the class label of column $i-1$, knowing optimal partition of prefix $[1, i]$ ends in a segment labeled c. Once label$[c, i]$ is calculated for every $i \in [2, N]$ and $c \in [1, C]$, optimal partition is obtained by back tracking the label matrix.

In MAXCOVER-C, w_i^c is calculated from Equation 3.3. In our application, we have five distinct classes (i.e. $C = 5$). In MAXCOVER-C, the labels are from $1, \ldots, C$ for C classes; class "1" is the class associated with the null hypothesis. In our application, we have started labeling the classes from -1 instead of 1. However, class "0" is the class associated with the null hypothesis. This results in having labels 1, 2 and 3 for classes when first, second and third sequence are under positive selection respectively.

Algorithm 1 MAXCOVER-C

Definition: For all $i \in [1, N]$, $W^c(i)$ is the score of optimal partition for prefix $[1, i]$ that ends with a segment labeled c.

Input: w_i^c where $i = 1, \ldots, N; c = 1, \ldots, C; \lambda > 0$ penalization factor

Output: $Z = Z_1, \ldots, Z_N$ class label for the maximal score segment

Initialize $W^c(1) = w_1^c$ where $c = 1, \ldots, C$
for $i \leftarrow 2, \ldots, N$ **do**
    $W^j(i-1) \leftarrow \max\{W^1(i-1), \ldots, W^c(i-1)\}$
    for $c \leftarrow 1, \ldots, C$ **do**
        $W^c(i) \leftarrow w_i^c + \max\{W^j(i-1) - \lambda, W^c(i-1)\}$
        $\mathsf{label}[c, i] \leftarrow \begin{cases} j & \text{if } W^j(i-1) - \lambda > W^c(i-1) \\ c & \text{otherwise} \end{cases}$
    end for
end for
$W^{\acute{j}}(N) \leftarrow \max\{W^1(N), \ldots, W^c(N)\}$
$Z_N \leftarrow \acute{j}$
for $i \leftarrow N, N-1, \ldots, 2$ **do**
    $\acute{j} \leftarrow \mathsf{label}[\acute{j}, i]$
    $Z_{i-1} \leftarrow \acute{j}$
end for
return Z

3.2 Model Parameters

The parameters which have to be estimated depend on the applied model. In the standard phylogeny model, the divergence time vector $t = [t_1, t_2, t_3]$; the substitution model (Q); and the rate of mutations at fast and slow evolving regions (i.e., α_+ and α_-) should be estimated. The substitution model (Q) defined by Equation 2.2, has a parameter for the ratio of transition to transversion (R) in addition to the base frequencies ($\pi_A, \pi_C, \pi_G, \pi_T$).

In tree-HMM model, four more parameters are to be estimated in addition to the parameters used in standard phylogeny model; match-transition matrix has π_{MM} and r; delete-transition matrix has π_{DD} and u as parameters.
Each of these parameters has a default range which is specified in the application.

3.3 Estimating the Model Parameters

The model parameters are estimated using maximum likelihood method. Matrix Q is defined by equilibrium base frequencies and ratio of transition to transversion. The equilibrium base frequencies (π_A, π_C, π_G, π_T) are estimated directly from the data by counting the frequency of each base. These frequencies are considered fixed parameters in the rest of calculation. Based on the applied model (standard or tree-HMM), the vector [t_1, t_2, t_3, R, α_+, α_-] or [t_1, t_2, t_3, R, α_+, α_-, π_{MM}, π_{DD}, r, u] represents the set of parameters, respectively. The following steps describe the procedure to estimate these parameters.

1. Each parameter is randomly initialized in its default range.

2. The LLR of each column of alignment is calculated for each of five classes (Eq. 3.3). This implies calculating the log likelihood ratio of each column with each of the following divergence time vectors: [t_1, t_2, t_3], [$\alpha_+ t_1$, t_2, t_3], [t_1, $\alpha_+ t_2$, t_3], [t_1, t_2, $\alpha_+ t_3$] and [$\alpha_- t_1$, $\alpha_- t_2$, $\alpha_- t_3$].

3. Maximum-scoring segment set or optimal partition is obtained by MAXCOVER-C algorithm with $C = 5$. The penalty function used in our application is linear

(i.e., $r(|\phi|) = \lambda k$).

4. Likelihood of the alignment with respect to the partition obtained in step 3, is the objective function for optimization. The next section details the optimization procedures.

5. If convergence is achieved, maximum likelihood estimates of the parameters and the column labels are returned; if not, the process continues from step 2 with the new set of parameters obtained in step 4. Convergence is achieved when the following two conditions are met: the column labels obtained in two consecutive iterations are the same (i.e., same partition); and summed square of relative differences in estimated parameters for two consecutive iterations is less than a predefined threshold. This threshold is denoted by Seg_{tol}. The second condition means that the new set of parameters obtained in step 4 is very close to the set obtained in preceding iteration.

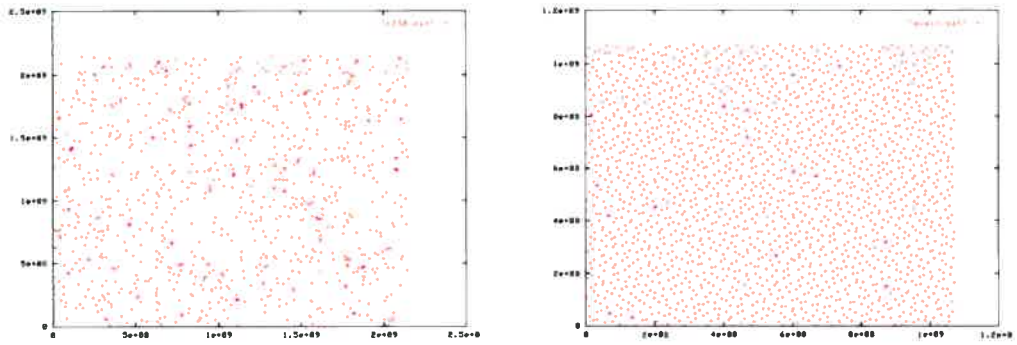
3.4 Optimization Method

In this thesis, the tree topology is known and parameters are selected so as to maximize the likelihood of the given tree under the best segmentation of the alignment.

The global extremum of a function is generally hard to find (Press et al. 2002). One heuristic approach is to find the local extrema by starting from different points in the space of parameters and then choose the most extreme of obtained local extrema. Since we are searching an n-dimensional space ($n = 6$ for standard phylogeny model or $n = 10$ for tree-HMM model) for a point where tree-likelihood is relatively maximum, it would be useful to choose the starting points far from each other. Quasi-random numbers are more suitable for this purpose than pseudo random numbers.

Quasi-random numbers are defined as sequences of n-tuples that spread throughout the n-space more uniformly than uncorrelated random points (Press et al. 2002). Figure 3.1(a) shows the distribution of a pair of sequences generated with

pseudo random number generator versus Figure 3.1(b) for the quasi-random sequences. It is clear that quasi-random numbers cover the space more uniformly than pseudo-random numbers. Halton's sequence is a simple example of quasi-



(a) Pair of pseudo-random numbers (b) Pair of quasi-random numbers

Figure 3.1: Scatter plot of the pair of sequences generated by pseudo-random generator (a), quasi-random generator (b) (from Taygeta Scientific Inc).

random numbers. To obtain jth number in the sequence, the following steps are performed: (1) j is written as a number in base b, where b is a prime. (2) The digits are reversed and a radix point is placed in front of the sequence. (3) The sequence is converted to base 10. For example, if $b = 2$ is used, the first five numbers of Halton's sequence are 0.5, 0.25, 0.75, 0.125 and 0.625. To obtain the Halton's sequence in n-dimension, usually the first n primes are used (Press et al. 2002).

3.4.1 Powell's Method

Powell's method is a derivative-free optimization algorithm (not gradient-based), which needs only the evaluation of the objective function f in the n dimensional space. Powell's method minimizes f iteratively in one dimension using a one-dimensional minimization method (i.e., Brent's method). For more information about Brent's method, see Press et al. (2002). Minimization is carried out along a set of directions which should be initialized. For simplicity, it can be set to the unit

vectors in each dimension of parameter space. After f is minimized in all set of directions for a given iteration, the following term is evaluated: $(f_{i-1} - f_i) < \frac{1}{2} * ftol * |f_{i-1} + f_i|$. f_i is the observed value of f at iteration i (current iteration); f_{i-1} is the function value at previous iteration and $ftol$ is a tolerance parameter.

If the convergence criterion is met, then the optimization ceases and the current minimum points in each direction are returned. If convergence criterion is not met, the algorithm proceeds with the updated set of directions. For details about this algorithm see Press et al. (2002).

The model parameters which are described in Section 3.2 are constrained. For instance, $\alpha_+ > 1$ and $\alpha_- < 1$. Since Powell's method is not a constrained optimization method, a pre-processing of the variables must be performed and variables should be transformed to unconstrained variables. First, constrained variable x is scaled between 0 and 1; then, inverse of sigmoid function is applied to transform the variable to an unconstrained variable y.

$$\begin{array}{lll} x_{min} < x < x_{max} & \Rightarrow \quad x' = \frac{x - x_{min}}{x_{max} - x_{min}} & 0 < x' < 1 \\ x' = \frac{1}{1+e^{-y}} & \text{Sigmoid Function} & \\ y = \log(\frac{x'}{1-x'}) & \Rightarrow \quad -\infty < y < \infty & \end{array}$$

The new unconstrained variables are optimized using Powell's method. The variables obtained by Powell's are post-processed to obtain the original variables.

$$x' = \frac{1}{1 + e^{-y}}$$

$$x = \frac{e^y * x_{max} + x_{min}}{1 + e^y}$$

3.5 False Positive Rates

In order to estimate the false discovery rate when real data is analyzed, a simulation is performed. The aim of this simulation is to find the percentage of misclassified columns if the sequence alignment was generated under the null model. Let $X = X_1, X_2, \ldots, X_N$ be the alignment of three sequences. After running the

application on this alignment, a set of segments predicted to be under negative and positive selection is obtained. The following simulation is performed to calculate the false discovery rate.

1. Use the estimated parameters (i.e., t_1, t_2, t_3, R) and the base frequencies observed in the alignment to generate an alignment with the same length. This alignment is generated under the null model and does not have any gap.

2. Using the estimated α_- and α_+, segmentation is performed. The task is to find the hypothesis Z conditional on the estimated values of model parameters. Since the alignment is generated under the null model, any column predicted as positive or negative is not correctly classified.

3. Repeat the steps 1-2 for 2000 times.

4. Calculate the average of misclassified sites (sites predicted to be under positive or negative selection).

5. P-value is calculated as the number of simulations with at least one misclassified site divided by the total number of simulations (i.e., 2000).

3.6 Application Description

The standalone application for this project is implemented in Java. This section briefly describes the specification of this application.

3.6.1 Input File

This application accepts the alignment of three sequences in Fasta (Pearson and Lipman 1988) format as input. The first line of the alignment for each sequence entry begins with a "greater than", (>) sign. Sequence name and description follows >. The alignment for the sequence starts the next line. An example of the alignment of three sequences in Fasta format is presented next.

>human

ttaggaactacctactatatacccaacaacttaatcactgtcaatatttacaataatgagatggctg----ttttttttt

catgtggtagaggcttggaaggcttggaa...

>chimp

ttaggaactacctactatatacccaacaacttaatcactgtcaatatttacaataatgagatggctgttttttttttttt

tcatgtggtagaggcttggaaggcttggaa...

>baboon

ttaggaactacctactatatacccaacaacttaaacactgtcaatatttacaataatgagatg-------ctttttttt

tcatgtggtagaggcttggaaggcttggaa...

3.6.2 Output File

This application generates one file per each sequence in a format based on GFF (i.e., Gene-Finding Format or General Feature Format). GFF is a format for describing genes and other characteristics related to DNA, and protein sequences. Files in GFF format are easy to parse and to process by different programs. This format can be used to report the predicted as well as experimentally confirmed features. Major genome browsers including UCSC genome browser and Ensemble can read GFF files and display annotation presented in the file. In a GFF file, each feature is described on a single line, and line order is not important. The obligatory fields in a GFF record are defined as follows.

<seqname> <source> <feature> <start> <end> <score> <strand> <frame>.

- <seqname> is the name of the sequence which is usually set to the identifier of the sequence in Fasta format. It can be set to the name specified by the user. If the chromosome on which the region resides is known, this field can be set to the chromosome identifier.
- <source> is the source of this feature and can be set to the program making the prediction. It is set to the name of our application.
- <feature> represents the type of the feature. Each file includes the regions pre-

dicted to be under positive and negative selection. Therefore, this field is either 'Positive' or 'Negative'. Regions which are predicted to be under no selection are not in the file.

- <start>, <end> are integers representing the starting and ending positions of the feature. <start> must be less than or equal to <end>.

- <score> is a floating point value for the score of the feature. When there is no score, '.' should be used instead. We have reported the log-likelihood ratio under slow versus neutral for regions under negative selection and fast versus neutral for regions under positive selection.

- <strand> is set to one of '+', '-' or '.'. If strand is not important, this field should be set to '.'. It is set to '.' by our program.

- <frame> is one of '0', '1', '2' or '.'. This field represents the phase of this frame from the first position of a codon. In our case, frame is not relevant and is set to '.'.

An additional field called "group" is needed to display a GFF record in the UCSC genome browser. All features with the same group are displayed on a single line. Each record in the output file has a unique "group" id. Features representing the positive and negative selection are labeled as Pos_XXX and Neg_XXX, respectively. XXX is a three digits number. For example, the first and the second region predicted to be under negative selection are labeled as Neg_001 and Neg_002. Similarly, Pos_001 and Pos_002 are the labels of the first and the second regions predicted to be under positive selection for a particular species.

All these fields should be separated by TAB character. For more information about GFF format, see `http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml`.

3.6.3 Options

This program has several options that can be set by the user. A list of these options is presented next.

- The model used for analysis should be specified. It can be either tree-HMM or

standard phylogeny model.

- The complexity penalty function which is used in segmentation algorithm can be set by the user.

- The path where output files are created, can be specified. Default path is the current directory.

- The name of the sequence for each species can be specified. This name is used in the <seqname> field of the output file. The defaults are "Sequence1", "Sequence2" and "Sequence3".

- For each sequence, an offset can be specified for the starting position of the sequence. The default value is "0".

- Number of starting points in the n dimensional space where optimization starts can be set. By starting the optimization from different points, it is more likely to get an optimum value close to the global extremum of the likelihood function.

- $ftol$ parameter which controls the convergence criterion of Powell's method can be set.

- Seg_{tol} which is the parameter used to control the convergence of segmentation can be set.

- Range of model parameters can be adjusted, if needed.

- α_- and α_+ can be given as constants. This means that they are not optimized.

3.6.4 User Interface

In this section, a complete scenario is presented to demonstrate how our application can be used. Suppose user has an alignment of three sequences (human, chimpanzee and baboon) in Fasta format and he/she wants to have the predicted signatures of positive and negative selection for this alignment. The human sequence is on chromosome 7 and starts at position 115597400. After launching the application, user clicks on the "Main" menu. Figure 3.2 shows the items in this menu. As illustrated in Figure 3.3, clicking on the "Run" menu opens a new window. User chooses the input file and the path for output files; enters "chr7" for the name of the first sequence and "115597400" for its starting position. In this

Figure 3.2: Screenshot of the main menu

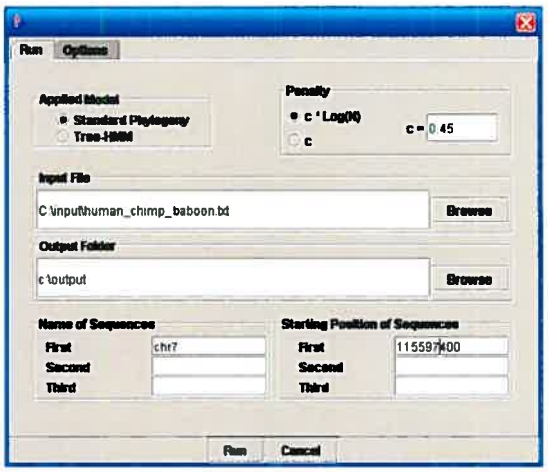


Figure 3.3: Screenshot of the "Run" dialog

example, the name and starting position of the second and third sequences are not set. Therefore, default values are used by the program. Standard phylogeny model is selected with the penalization factor equal to $0.45 * \log(N)$ where N is the length of the alignment.

Clicking on the "Options" tab, displays another window (Fig. 3.4), through which other options such as range of parameters can be customized. After setting the parameters, user clicks on the "Run" button and the process of finding the segmentation with the highest score starts. At the end of computations, user is informed by a message box.

Three GFF files are created in the output folder. At the end of each file, the like-

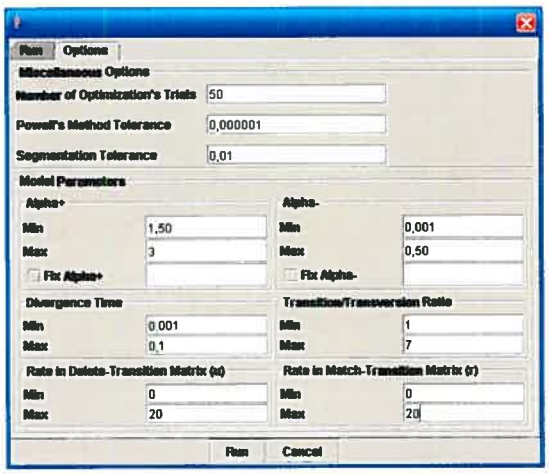


Figure 3.4: Screenshot of the "Options" dialog

lihood of the alignment under the obtained segmentation, the score of the resulted partition, and the estimated parameters are reported. Here is an example of the GFF file content for human sequence.

chr7 MDLShadow Negative 115597578 115598284 6.12 . . Neg_001
chr7 MDLShadow Negative 115624305 115625261 8.90 . . Neg_002
chr7 MDLShadow Negative 115653703 115653787 5.22 . . Neg_003
chr7 MDLShadow Negative 115656878 115656952 4.65 . . Neg_004
chr7 MDLShadow Positive 115656953 115657682 3.94 . . Pos_001
chr7 MDLShadow Negative 115657683 115657771 5.45 . . Neg_005
chr7 MDLShadow Positive 115659341 115660744 1.87 . . Pos_002
Likelihood = -318.81
Partition Score = 36.15
Parameters : t1 = 0.0042; t2 = 0.0044; t3 = 0.0378; R= 1.31; α_+ = 1.5; α_-=0.137

In order to display the result file in UCSC genome browser, GFF file should have headers that define the genome browser and annotation track display characteris-

tics. For example, the following two lines can be added to the beginning of human's GFF file.

browser position chr7:115597400-115660900
track name ="MDLShadow" description="MDLShadow Predictions" visibility=2 color=0,0,0

The first line specifies the genome position that the genome browser initially displays. In the above example, this position is set to the position of the human sequence on the chromosome. The second line specifies the track label, color of the track in the genome browser, and initial display mode of the annotation track. In this example, color is set to black; track label to "MDLShadow Predictions" and display mode to full (visibility 2 corresponds to full display mode). For more information about visualizing custom annotation file in UCSC genome browser see `http://genome.ucsc.edu/goldenPath/help/customTrack.html#TRACK`.
This output file can be loaded to genome browser as a user defined annotation file. Figure 3.5 shows UCSC genome browser with a custom track for the predicted regions given in human GFF file.

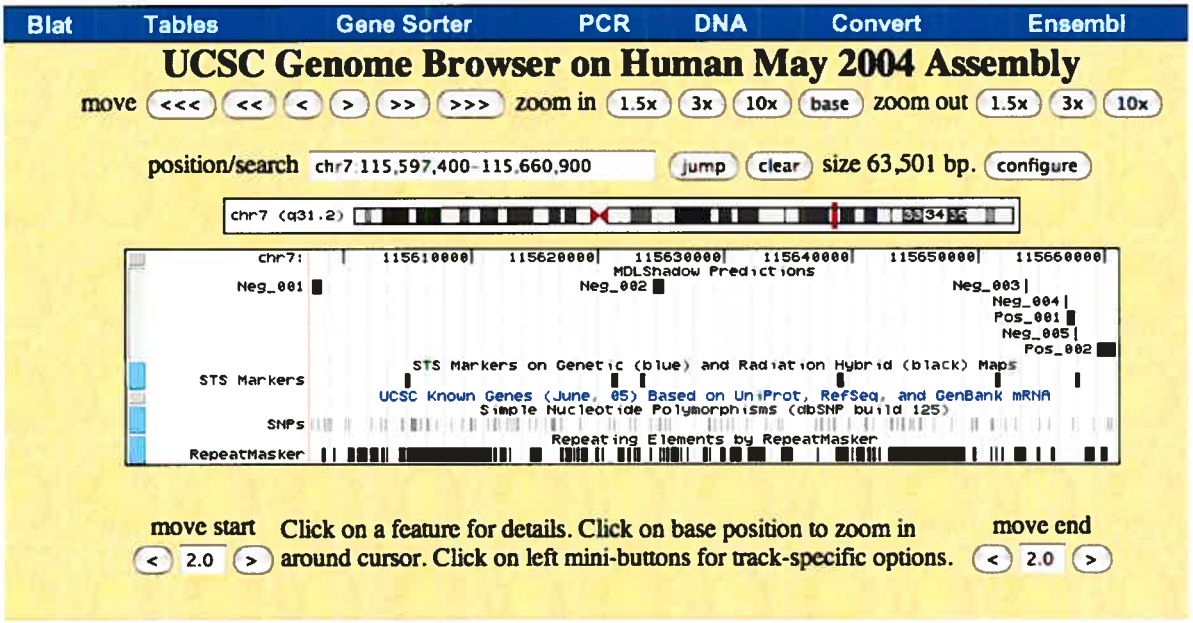


Figure 3.5: Screenshot of UCSC genome browser with the predicted regions as a custom track

CHAPTER 4

RESULTS

4.1 Simulation

The power of the presented methodology is evaluated on simulated data using standard phylogeny and tree-HMM model. The objective of simulations was to test the performance of methodology for sequences of different lengths and to choose the best penalty function accordingly. Simulations are performed with different penalty functions and for two alignment lengths, $10,000$ and $100,000$. The steps to generate the simulated data are presented next.

For standard phylogeny model:

1. Base frequencies are set to satisfy $\pi_A + \pi_C + \pi_G + \pi_T = 1$.
2. The length of alignment, N, is set.
3. Using the defined base frequencies, a sequence called $ParentSequence = x_1,\ldots,x_N$ is generated from alphabet $\Sigma = (A, C, G, T)$. This sequence is considered as the common ancestor of all three sequences.
4. The model parameters, $[t_1, t_2, t_3, R, \alpha_+, \alpha_-]$, are randomly generated in their default range.
5. The length of homogeneous segments is set to 2500, 1000, 500, 250, 100 and 50.
6. The length of alignment divided by the length of segments gives number of homogeneous segments. Segment classes are randomly selected from class labels {0,1,2,3,−1}. For instance, for the alignment of length $10,000$ with 10 homogeneous segments (i.e. segment length=1000), 10 labels are drawn from class labels {0,1,2,3,−1}. The first label is the class of the column 1

through 1000; the second label is the class of the column 10001 through 2000. The remaining labels are assigned to the columns of alignment in the same manner.

7. Three sequences are generated independently using the F84 model of substitution (Eq. 2.3). It is assumed that each of the three sequences evolved from *ParentSequence* in times t_1, t_2, t_3 respectively. The column label is used to determine which of the t_j, $\alpha_+ t_j$ or $\alpha_- t_j$, $j \in \{1,2,3\}$, should be used in the F84 model (see Table 3.1). These 3 sequences with the known column classes are considered as the input of our application.

8. Penalization factor, λ, is set to $0.3 * \log(N)$, $0.5 * \log(N)$, $0.7 * \log(N)$ and $\log(N)$. $\log(N)$ corresponds to the BIC penalization factor.

9. The model parameters are estimated along with the column classes as described in Section 3.3. This step is repeated 50 times with different initial values; the parameters and the column classes corresponding to the highest likelihood are reported at the end.

10. Segmentation error defined as the percentage of columns with the predicted class different from the defined class is calculated.

Steps 1 to 10 are repeated 100 times with different penalization factors and with different segment lengths. The length of segments is fixed in each simulation but varies between simulations.

To test the performance of the tree-HMM model, simulated data was also generated and analyzed under this model. The steps performed to generate the alignments under the tree-HMM model are as follows:

1. In addition to the base frequencies, frequencies of Match-Delete transition (π_{MD}) and Delete-Delete transition (π_{DD}) are set. The former is the probability of opening a gap and the latter is the probability of continuing the gap.

2. The length of alignment, N, is set.

3. Using the defined base frequencies, π_{MD} and π_{DD} a sequence called *ParentSequence* $= x_1,\ldots,x_N$ is generated from alphabet $\Sigma = (A, C, G, T,-)$. This sequence is considered as the common ancestor of all three sequences.

4. The model parameters, $[t_1, t_2, t_3, R, \alpha_+, \alpha_-, r, u]$, are randomly generated in their default range. Parameters $\pi_{DD} = \pi_{DM} = 0.5$ and $\pi_{MD} = 0.001$ (i.e. $\pi_{MM} = 0.999$) are fixed for all simulations.

5. The length of homogeneous segments is set.

6. Segment classes are randomly selected from $\{0,1,2,3,-1\}$.

7. Three sequences are generated independently using the F84 model of substitution (Eq. 2.3), match-transition matrix (Eq. 2.8) and delete-transition matrix (Eq. 2.9). Similar to the standard phylogeny model, column class determines which of the t_j, $\alpha_+ t_j$ or $\alpha_- t_j$, $j \in \{1,2,3\}$, should be used in Equations 2.3, 2.8 and 2.9. It is assumed that each sequence starts the transition from a dummy match state.

   The transition made by the *ParentSequence* at each position, and the base at the preceding position, determine which transition should be taken by the child sequence. Once the transition of the child sequence is specified, the base at the child sequence is set. Table 4.1 summarizes how transition and base of the child sequence at position i is determined. In this table "-" represents the gap character; "X", base nucleotide (i.e. A, C, G, T); y_i, base of the parent sequence at position i; x_{i-1} base of the child sequence at position $i-1$. Column "Pr(Transition)" specifies the transition probability of the child sequence. In the event of "MM" or "DM" transition, a base should be selected for position i of the child sequence. Column "Pr(Emission)" represents the probabilities that determine the base nucleotide of the child sequence when it should emit a character different from gap. When parent base is a gap character at position i, the equilibrium frequencies of nucleotides (i.e. π_A, π_C,

π_G and π_T) are used to determine the base nucleotide of the child sequence for this position. Cases where transition of the child sequence is independent from its parent are denoted by *.

For example, the first row corresponds to the case where parent sequence has "-" at position $i-1$; a nucleotide at position i and child sequence has a nucleotide at position $i-1$. Hence, parent sequence is in the delete state and child is in the match state at position $i-1$. This implies that delete-transition tree has "DM" at the root and a "*" at the leaf whereas match-transition tree has "*" at the root. Since child state is different from the state of its parent, the transition which child sequence is going to take is independent from its parent. Due to the presence of a nucleotide at the preceding position in the child sequence, child sequence transition is either "MM" or "MD" and is determined by equilibrium frequency of "MM" transition (i.e. π_{MM}). If "MM" is selected, then a nucleotide is generated for this position based on the F84 model of substitution (Eq. 2.3). If "MD" is the transition of child sequence, then x_i is a gap character.

Table 4.1: Transition and emission probabilities used in generating a sequence under the tree-HMM model

Parent Sequence			Child Sequence		
y_{i-1}	y_i	Transition	x_{i-1}	Pr(Transition)	Pr(Emission)
-	X	DM	X	π_{MM} *	F84 Model of Substitution
-	-	DD	X	π_{MM} *	Base Frequencies
X	-	MD	X	$\Pr(MM\|MD)$	Base Frequencies
X	X	MM	X	$\Pr(MM\|MM)$	F84 Model of Substitution
X	X	MM	-	π_{DM} *	F84 Model of Substitution
X	-	MD	-	π_{DM} *	Base Frequencies
-	X	DM	-	$\Pr(DM\|DM)$	F84 Model of Substitution
-	-	DD	-	$\Pr(DM\|DD)$	Base Frequencies

8. Penalization factor, λ is set.

9. The model parameters are estimated along with the column classes as described in Section 3.3.

10. Segmentation error is calculated.

4.1.1 Simulation Results

First, we describe the results of our simulations for alignment of length $10,000$. Table 4.2 provides the average of segmentation error for different penalization factors. It can be seen that for both models, the percentage of misclassified sites for

Table 4.2: Average of segmentation error ($N = 10,000$)

	Segmentation Error%	
Penalization Factor (λ)	Standard Model	Tree-HMM Model
0.3*log(N)=2.76	14.40	14.67
0.5*log(N)=4.6	4.25	5.26
0.7*log(N)=6.4	4.65	5.86
log(N)= 9.2	6.06	7.62

$0.5 * \log(N)$ is the least among all the four penalties but it is also very close to $0.7 * \log(N)$.

In order to evaluate the performance of the methodology in estimating the model parameters, a relative error is calculated for every variable in each simulation. If the true value of variable x is v and its estimated value is $\hat{v}$, the relative error is calculated as $\frac{|\hat{v} - v|}{v}$. The relative error of each variable calculated for each simulation is averaged over all simulations with the same λ. Table 4.3 shows the mean relative error of model parameters for standard phylogeny model. The relative error for branch length is calculated as the average of the relative error

Table 4.3: Relative error of model parameters with standard phylogeny model ($N = 10,000$)

	Relative Error %			
Penalization Factor (λ)	Branch Lengths	Ratio	α_+	α_-
0.3*log(N)=2.76	19	9	39	97
0.5*log(N)=4.6	6.62	4.5	6.1	70
0.7*log(N)=6.4	6.65	4.3	5.4	77.5
log(N)= 9.2	7.40	4.40	7.5	81.5

for t_1, t_2 and t_3. The segmentation error and error in estimated parameters is the highest for $0.3 * \log(N)$. The ratio of transition to transversion has the least amount of error among parameters. The error for parameter ratio is less than 4.5% for higher values of the penalization factor. The error for parameter α_+ is in the range of 5.4% to 7.5% for higher values of λ.

The parameter α_- has the highest error among all the model parameters for all the four penalties. Since the minimum error of α_- is 70%, it can be seen that this parameter can not be learned in training for sequences of this length. α_- is a parameter describing the sequences which have undergone little changes compared to neutrally evolving sequences. Obviously, α_- is hard to learn because it corresponds to very low mutation levels. Finding the exact estimation of α_+ and α_- is of secondary importance compared to finding the correct segmentation. Moreover, getting the right branch lengths is more crucial than obtaining the right value of α_- and α_+. It can be seen that the errors for branch lengths are small and less than 7.5% for higher values penalization factor. It should also be noted that we have assumed that α_- and α_+ are constants in each set of simulations. In reality, there are varying levels of selection in the sequences, which could only be modeled by using a set of α factors. Performing the simulations with this high amount of resolution is not an easy task.

Table 4.4 shows the relative errors of model parameters when tree-HMM model is applied. Since π_{MM} and π_{DD} are fixed at 0.999 and 0.5, the frequency of the gap

Table 4.4: Relative error of model parameters with tree-HMM model ($N = 10,000$)

	Relative Error %							
Penalization Factor (λ)	Branch Lengths	Ratio	α_+	α_-	π_{MM}	π_{DD}	r	u
0.3*log(N)=2.76	22	8.3	32.6	98.8	0.01	11.3	37.3	120
0.5*log(N)=4.6	7.8	4.4	7	48	0.01	11.4	31	120
0.7*log(N)=6.4	8.2	4.4	6.8	54	0.01	11.4	30.7	120
log(N)= 9.2	9.5	4.5	7.4	54.8	0.01	11.4	30	120

is very small in the simulated data. The parameters specific to tree-HMM model (i.e. π_{MM}, π_{DD}, r and u) are quite the same for all four penalties. Frequency of the Match-Match transition is estimated with an error as small as 0.01%. High error in estimating the rate constant (u) is probably a consequence of small samples with gap. Despite the high estimation error of u parameter, the average error of segmentation is slightly higher than that of standard model for all four penalties. Therefore, it may be concluded that tree-HMM model is not very sensitive to the ratio parameters in transition family matrices (i.e. r, u).

Comparison of results presented in Tables 4.3 and 4.4 shows that α_- is the only parameter which has less error with tree-HMM than with standard phylogeny model.

We have also performed simulations with sequences of length 100,000. The segmentation errors for alignment of this length is shown in Table 4.5. Similar to the results obtained for alignments of length $N = 10,000$, segmentation error is minimum with $\lambda = 0.5 * \log(N)$ for both models. However, the penalization factor $\log(N)$ leads to the highest segmentation error for this length. Segmentation errors calculated for both models are very similar with $\lambda \in \{0.3*\log(N), 0.5*\log(N), 0.7*\log(N)\}$.

The relative errors in estimating the model parameters are given in Table 4.6 when standard phylogeny model is applied.

Table 4.5: Average of segmentation error ($N = 100,000$)

	Segmentation Error%	
Penalization Factor (λ)	Standard Model	Tree-HMM Model
0.3*log(N)=3.45	5.23	5.72
0.5*log(N)=5.76	4.83	4.96
0.7*log(N)=8.06	6.09	6.07
log(N)= 11.51	7.32	8.14

Table 4.6: Relative errors of model parameters with standard phylogeny model ($N = 100,000$)

	Relative Error %			
Penalization Factor (λ)	Branch Lengths	Ratio	α_+	α_-
0.3*log(N)=3.45	2.60	3	8.30	10
0.5*log(N)=5.76	2.62	1.77	2.99	7.14
0.7*log(N)=8.06	3.46	1.68	3.44	8.48
log(N)= 11.51	4.40	1.80	3.90	8.60

It can be seen that errors in estimating the model parameters drop significantly compared to the results for alignments of length 10,000 (see Tables 4.3 and 4.6). Among the parameters, the estimated error for α_- is significantly dropped. For example, the estimated error of α_- dropped from 70% to 7.14% with $\lambda = 0.5 * \log(N)$.

The relative errors in estimating the model parameters analyzed with tree-HMM are given in Table 4.7 ($N = 100,000$). Comparing the results presented in Table 4.7 with Table 4.4, shows that errors in estimating model parameters drop significantly for alignments of length 100,000 compared to the alignments of length 10,000. We may postulate that as the sequence gets longer, more samples are available that can be used to obtain the maximum likelihood estimate of model parameters.

The error of rate parameter in delete-transition matrix (i.e. u) drops significantly from 120% to 48%. It can be seen that the errors of parameters specific to

tree-HMM model (i.e. π_{MM},π_{DD}, r, u) do not depend on λ but on the length of the alignment. The errors in estimating the equilibrium frequencies of Match-Match transition (i.e. π_{MM}) and Delete-Delete transition (i.e. π_{DD}) are 5×10^{-3}% and 3.3%, respectively.

We also observed that for alignments of length 10,000, the error of α_- is higher (except for $\lambda = 0.3 * \log(N)$) with standard phylogeny than with tree-HMM model. However, for alignment of length 100,000, the difference is not remarkable. With tree-HMM, α_- is learned from transition probabilities in addition to emission probabilities and that makes a difference for alignments of length 10,000. Though, for alignments of length 100,000, there are enough samples available to learn the α_- from emission probabilities and having additional training data (transition probabilities) does not improve the training.

Table 4.7: Relative errors of model parameters with tree-HMM model ($N = 100,000$)

	Relative Error %							
Penalization Factor (λ)	Branch Lengths	Ratio	α_+	α_-	π_{MM}	π_{DD}	r	u
0.3*log(N)=3.45	4.2	2.6	11	11.8	5×10^{-3}	3.3	10.8	48.3
0.5*log(N)=5.76	3.9	1.6	3.2	7.3	5×10^{-3}	3.3	10.1	45.2
0.7*log(N)=8.06	6.1	1.6	3.2	9.3	5×10^{-3}	3.3	10.2	45.4
log(N)= 11.51	7.4	1.9	5.2	13.4	5×10^{-3}	3.3	11.1	46.5

In order to see the effect of different segment lengths on the segmentation error, we plotted the mean of segmentation error for each segment length and each penalization factor (Fig. 4.1). These results correspond to the alignments of length 100,000 analyzed under standard phylogeny model. It can be seen that our approach has a better performance in detecting lengthy segments (more than 250bp). When the length of segments are short (50bp), $0.3 * \log(N)$ is more successful in predicting the correct class of the segments. In contrast, for the long segments (2500bp), $\log(N)$ outperforms the other penalization factors. However, it should

be noted that length of homogeneous segments in the alignment is usually unknown beforehand. Therefore, we selected the penalization factor that outperforms others for all segment lengths. Comparing the results of our simulation for sequences of length 10,000 and 100,000 demonstrates that the error corresponding to penalization factor $0.5*\log(N)$ is less than the corresponding error for penalties $0.3*\log(N)$ and $0.7*\log(n)$. Hence, $\lambda = 0.5*\log(N)$ is used when the same approach is tested on real data.

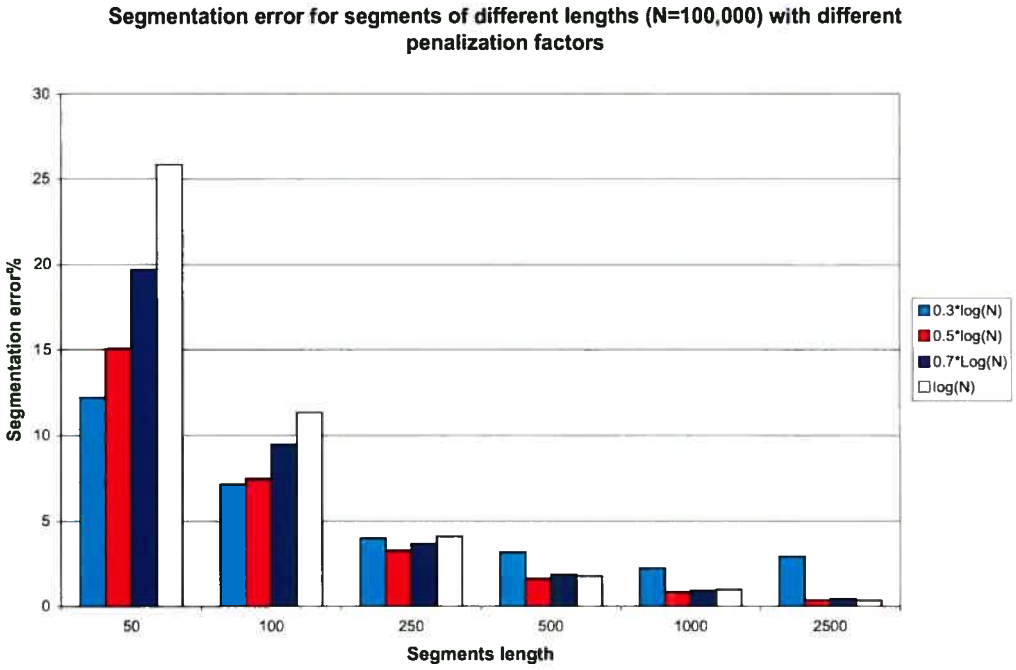


Figure 4.1: Segmentation error for segments of different lengths with different penalization factors

Since $\lambda = 0.5*\log(N)$ is used for analyzing real data, we further investigated categories of misclassified sites for this value of λ. Table 4.8 provides the segmen-

tation error for each class. These results correspond to the alignments of length 100,000 analyzed under standard phylogeny model. In this table, values obtained for fast classes, comprise the errors for all three classes of positive selection.
The highest error corresponds to the regions under positive selection classified as neutral (1.8%). This type of error is the result of over-estimating the branch length under positive selection. Since, there are three classes of positive selection, we may say that on average $\frac{1.8}{3} = 0.6\%$ of the columns under positive selection are erroneously labeled as neutral for each species. The second highest error is for the regions under no selection classified as being under positive selection (1.1%). This type of error happens when the branch length is under-estimated. Consequently, on average $\frac{1.1}{3} = 0.37\%$ of the neutrally evolving columns in each species are labeled as being under postive selection. The least amount of error corresponds to region under negative selection classified as positive (0.096%). The value at Fast/Fast entry (i.e. 0.83%), is for the regions under positive selection which erroneously labeled as being under positive selection but for different species.

Table 4.8: Average of segmentation error for each class using standard phylogeny model ($N = 100,000$)

	Predicted Class		
Actual Class	Neutral	Slow	Fast
Neutral		0.3%	1.1%
Slow	0.47%		0.096%
Fast	1.8%	0.2%	0.83%

In general, segmentation error and error in estimated parameters are slightly higher in tree-HMM than in standard model. However, with the number of simulations performed, this comparison is not appropriate since the simulated data for these two models are different. Further experiments are needed to compare the two models using same set of data.

4.2 Real Dataset (CFTR Region)

The result of our simulations confirmed that our approach has a better performance on sufficiently long alignments. Therefore, we analyzed 1.8 Mbp of the human sequence known as "greater CFTR region" (Thomas et al. 2003) for human and two of its primates, chimpanzee and baboon. This region is sequenced in different species and is widely analyzed in literature (Thomas et al. 2003; Margulies et al. 2003; Cooper et al. 2005).

Human sequence for this region corresponds to NCBI build 35, i.e. chromosome 7, 115404472-117281897 (Cooper et al. 2005). Baboon and chimpanzee sequences are shorter than human sequence (about 1.5 Mbp). The alignment of these three sequences were obtained from the multiple alignment of 13 species (available at `http://baboon.math.berkeley.edu/mavid/examples/zoo.target1/`). These sequences were aligned using MAVID (Bray and Pachter 2004). This alignment is 2074999 bp long and 44% of the columns contain either gap or ambiguous character (i.e. N). This region contains 10 known genes including the gene mutated in cystic fibrosis (i.e. CFTR) (Margulies et al. 2003).

Results of our simulation showed that the penalty function $\lambda = 0.5 * \log(N)$ is more appropriate. Therefore, the alignment was analyzed using standard phylogeny model with this penalty function. Ambiguous character (N) was treated like gap. Figure 4.2 shows the output of our program for this region on UCSC genome browser.

The predicted branch lengths are 0.0044 for human, 0.005 for chimpanzee and 0.042 for baboon. Other model parameters are estimated as $R = 1.3$, $\alpha_+ = 4.5$. α_- was fixed at 0.137, corresponding to the rate Boffelli et al. (2003) found for conserved region under the Hasegawa-Kishino-Yano (HKY85) model of evolution. They found that non-coding regions evolved 7.3 times faster than coding regions. Assigning a default rate of 1 to neutrally evolving regions, gives the value of $\frac{1}{7.3} = 0.137$ for α_-. Since HKY85 model of evolution is similar to the F84 model of Felsenstein (used in our program), this rate was used.

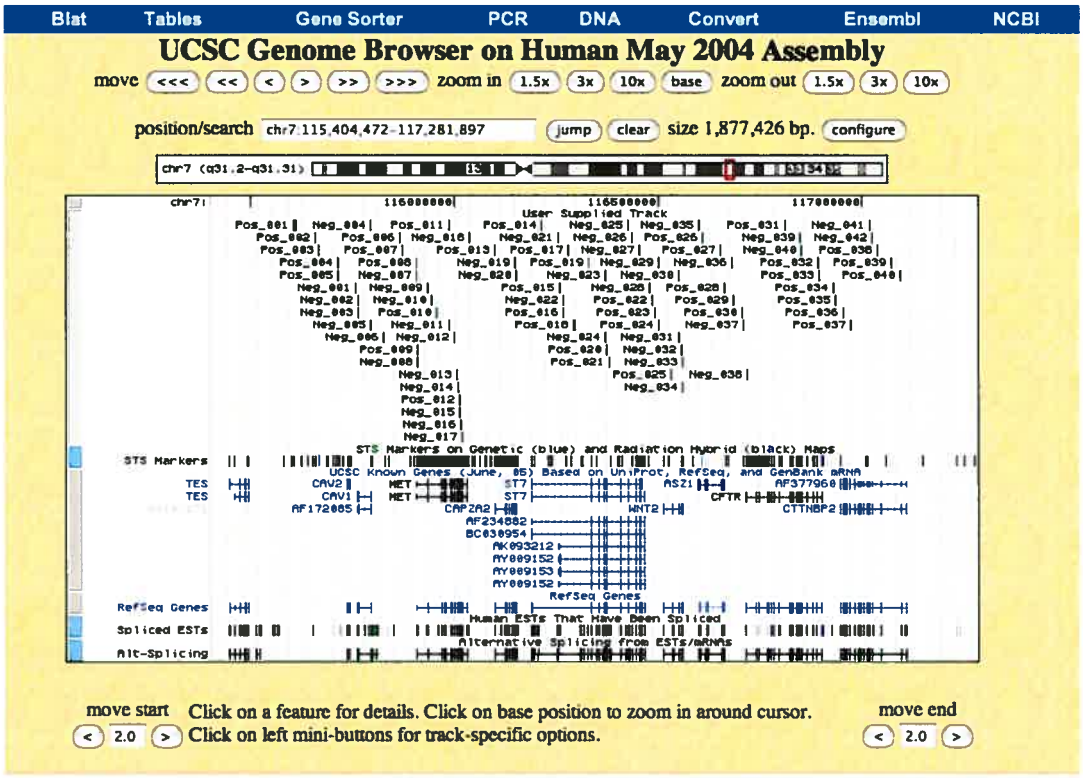


Figure 4.2: Annotation for CFTR region displayed as a user supplied track on UCSC genome browser.

The estimated parameters and segmentation converged after 15 iterations. Convergence is achieved if the column classes obtained in two consecutive iterations are the same (i.e. same annotation for the whole alignment); and sum square of the relative changes for all variables in two consecutive iterations is less than a predefined threshold ($Seg_{tol} = 0.00001$). Figure 4.3 shows the convergence of the annotation and parameters for the obtained partition. Iteration two is the starting point on the graph because this graph presents the differences between consecutive iterations. For instance, the values at iteration two correspond to the differences between iteration one and two.

There are 42 elements predicted to be under negative selection covering 1.25% of this region on human genome. 40 elements are predicted to be under positive selection. The percentage of the regions predicted to be under positive selection

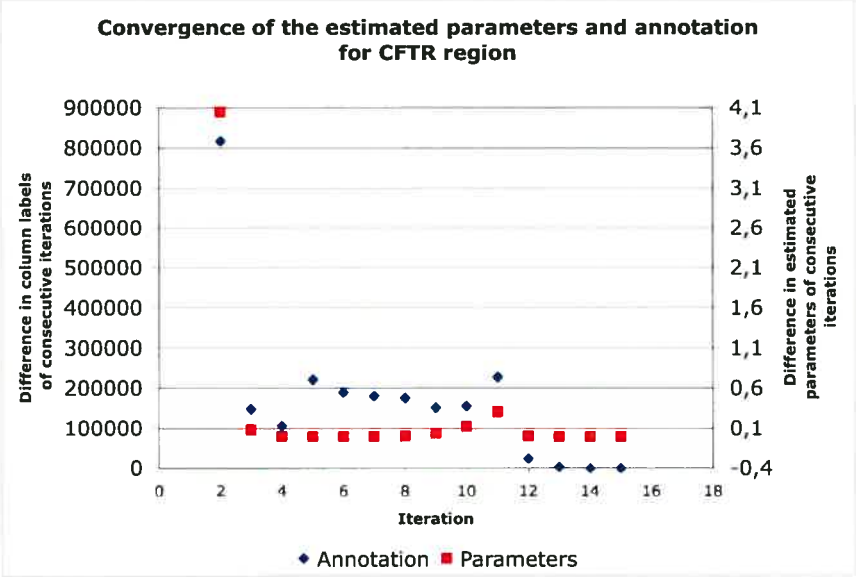


Figure 4.3: Convergence of the estimated parameters and segmentation for the CFTR region

is higher (i.e. 2.11%). However, 0.68% of these regions are located either at the beginning or at the end of the alignment which mostly contains gaps in the chimpanzee or the baboon. Table 4.9 shows several statistics about the length of the regions found under positive and negative selection. The shortest regions predicted to be under positive and negative selection are 11bp and 186bp, respectively.

Gene annotations for this region was download from UCSC genome browser (`http://genome.ucsc.edu/`). In order to find the predictions overlapping with the known annotations, Galaxy web server was used. Galaxy is a platform for interactive genomic analysis and can be found at `http://main.g2.bx.psu.edu/` (Giardine et al. 2005). Using this web server, one can perform different operations such as subtraction, intersection and union on genomic intervals.

In order the measure the false positive rates, simulation is performed as described

Table 4.9: Maximum, minimum, median and average length of the regions predicted to be under selection

Selective Pressure	Length (bps)			
	Maximum	Minimum	Median	Average
Under Negative Selection	1446	186	493	559
Under Positive Selection	7144	11	434	992

in Section 3.5. The number of misclassified sites is 1.6×10^{-6} and P-value $= 0.1$.

4.2.1 Regions under Purifying Selection

The composition of regions predicted to be under negative selection is presented in Figure 4.4. The majority of these regions are located in introns (52.1%). Introns account for about 28% of the conserved regions in Vertebrates (Siepel et al. 2005). Less then 25% of the regions predicted to be under negative selection are known coding exons and UTRs. Ancestral repeats (ARs) which are elements inserted to the common ancestor of most mammals consist about 7.4% of the regions. Generally, ARs are known to be nonfunctional. However, there are some evidence that these regions may have got functional roles during evolution (Cooper et al. 2005). Conserved regions which are not annotated (16.5%), may either represent non coding functional regions or non functional regions that did not accumulate enough mutations by chance.

A recent study by Bejerano et al. (2004) identified around 500 "ultraconserved" elements in human genome. Ultraconserved elements were defined as regions that have not changed over at least 200 bases in human, mouse and rat genomes. Cooper et al. (2005) have defined a metric for identifying the ultraconserved elements in mammalian genomic sequences. They analyzed an alignment of CFTR region from 29 mammals and reported 20 "ultraconserved" elements in this region. Comparison of our results with Cooper et al. (2005) shows that 12 regions significantly overlapped with these ultraconserved elements. Table 4.10 shows the locations, scores

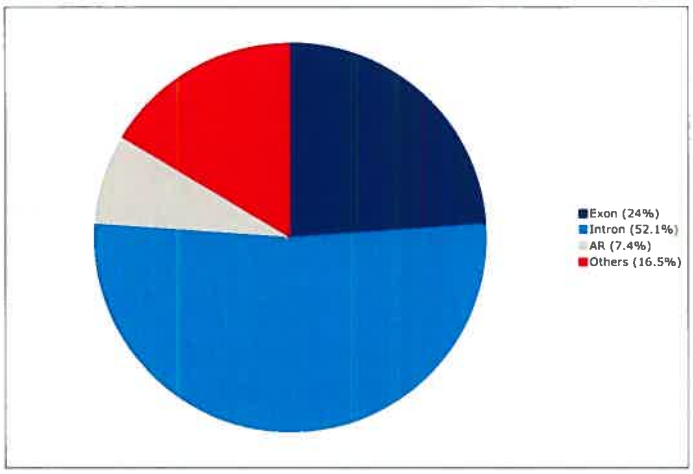


Figure 4.4: Composition of conserved elements by annotation types

of these 12 elements and the percentage of overlaps with the previously found ultraconserved regions. In the calculations, the first position of the human sequence for this region is set as position 0.

Column "Label" is the segment label in GFF file. "Score" is the LLR of the segment being in slowly evolving class versus being in neutrally evolving class. Column "Base Coverage" shows the percentage of the previously identified ultraconserved regions overlapped with MDLShadow predictions.

The segment which is located in the intronic region of ST7 (Neg_027) comprises a segment of more than 200 bps without any change in three primates analyzed. This region is not part of the reported ultraconserved element. We used the fasta3 program at `http://www.ebi.ac.uk/fasta33/nucleotide.html#` to find sequences in mammal database that are similar to this 202 bp sequence. Similar sequences are found in hominoids and old-world monkeys lineages with 100% identity. Other mammals including new-world monkeys have similar sequences but with less similarity (at most 96.5%). Figure 4.5 shows this sequence on the UCSC genome

Table 4.10: MDLShadow predictions overlapped with previously identified ultra-conserved regions

Label	Start	Length	Score	Type	Gene	Base Coverage %
Neg_004	388632	258	13.7	Exon	CAV1	54.1
Neg_009	528563	675	19.8	Exon	CMET	52.6
Neg_014	604428	356	13.8	Intronic/Exon	CMET	74.8
Neg_017	611464	733	18.4	Intronic/Exon	CMET	89.3
Neg_020	747227	1153	32.4	Exon	CAPZA2	80
Neg_023	963244	1180	36.5	Intronic	ST7	99.1
Neg_025	1019569	272	14.9	Intronic	ST7	27.8
Neg_027	1046542	1445	36	Intronic	ST7	100
Neg_031	1145375	933	30.1	Intronic	WNT2	90.7
Neg_032	1150823	617	29	Intronic	WNT2	89.4
Neg_041	1613773	381	21	Intronic/Exon	CORTBP2	83.7
Neg_042	1620651	814	17.7	Exon	CORTBP2	89.9

browser. The red bar shows the region identified by Siepel et al. (2005) as being removed from selection in rodent family. This segment might have some functionality in primates analyzed. These results imply that our program can be successfully used to detect highly conserved regions by comparing closely related genomes such as primates.

We were also interested to see if any region is identified as being under negative selection which is not previously annotated. To this end, we eliminated the segments that overlapped with known exons and the regions previously identified as ultraconserved and obtained 19 elements. The percentage of mutations per site in these 19 regions varies between 4.7% and 0.58%.

It is assumed that each species evolved independently after divergence from a common ancestor. For short branch lengths, the probability of mutation in each column can be approximated by the sum of branch lengths. Therefore, the probability of mutation in each column is $0.0044 + 0.005 + 0.042 = 0.0514$. This implies that if a region of the alignment evolved neutrally, we expect to see 5.14% variations in the

segment. All the regions predicted to be under negative selection have variations less than 5.14%.

Most of the regions which overlapped with known exons and ultraconserved elements are more than 98% similar (i.e. less than 2% variations). To further narrow down our predictions, the elements with mutations over 2% are excluded from the results. Nine regions qualify for this condition and are presented in Table 4.11. The column "Similarity" is the percentage of the sites where all three species have the same nucleotide. These regions are either intronic or intergenic. Intergenic regions are the DNA sequences located between genes and may have some functionality to control the genes nearby.

Table 4.11: Unannotated regions predicted to be under negative selection

Label	Start	Length	Score	Similarity%	Type	Gene
Neg_001	351707	395	15.97	98.98	Intergenic	CAV2-CAV1
Neg_008	502482	618	17.49	98.54	Intronic	MET
Neg_010	**539121**	**513**	**24.79**	**99.41**	**Intronic**	**MET**
Neg_011	581841	319	15.64	98.43	Intronic	MET
Neg_015	605261	330	7.71	98.18	Intronic	MET
Neg_018	630749	535	16.11	98.50	Intergenic	MET-CAPZA2
Neg_028	1071820	529	22.79	99.05	Intergenic	ST7-WNT2
Neg_034	1153430	782	22.27	98.21	Intergenic	WNT2-ASZ1
Neg_037	1301748	396	17.99	98.99	Intergenic	ASZ1-CFTR

The highest scoring segment (score =24.79), has a subregion over 325 bps without any mutation in these three primates. Since this region is not among the previously identified ultraconserved elements in mammals, it may be of functional importance in primates.

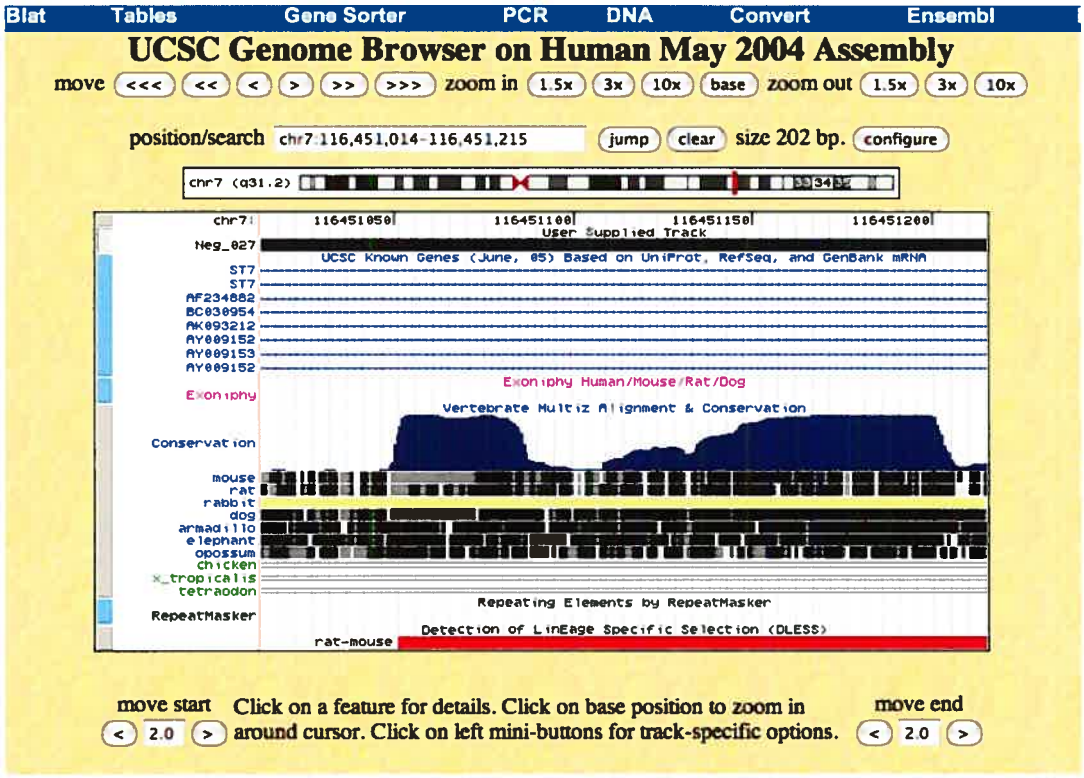


Figure 4.5: A conserved intronic region of ST7 displayed on UCSC genome browser

4.2.2 Regions under Positive Selection

It was noted that 40 elements were identified to be under positive selection. The composition of these elements by annotation type is illustrated in Figure 4.6. The majority of these regions are masked by RepeatMasker as ancestral repeats (63%). 47% of these repeats are located in introns (data is not shown here). Introns which do not overlap with ARs consist about 30% of the regions under positive selection. The remaining 7% are not annotated.

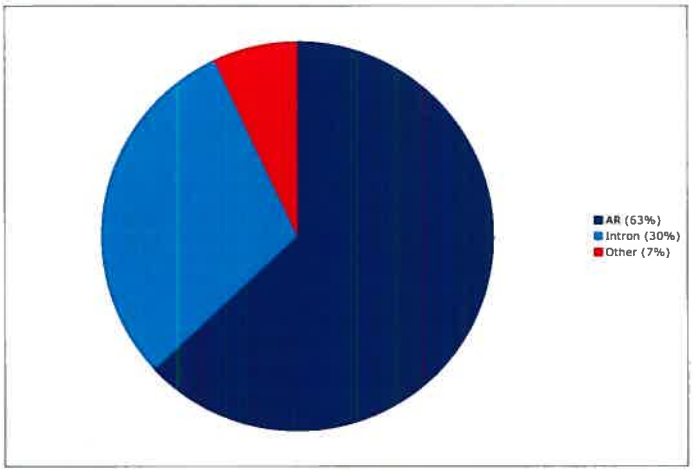


Figure 4.6: Composition of regions predicted to be under positive selection by annotation types

We investigated the alignment for the segments under positive selection. Columns where human base is different from chimpanzee and baboon bases are counted. The human sequence differs from other two primates by up to 90% in these 40 elements. Table 4.12 provides the list of the regions where mutation in human sequence exceeds 20%. Column "Mutation" gives the percentage of the columns where human base is different from chimpanzee and baboon bases. The length of these regions is relatively smaller than the regions predicted to be under negative selection. The

smallest region which is located in intronic region of CFTR gene is 11 bps long. Chimpanzee and baboon bases are identical for this region. Researchers at NHGRI have suggested that CFTR gene might be among the genes that has gone under selection over the past 250,000 years (NHGRI 2005). Among the listed regions, 3 reside in the intronic regions of CFTR.

It should be noted that the regions identified as being under positive selection can be alignment artifacts. Further research is needed to identify the predictions that are the results of point mutations and possibly putative signals of positive selection.

Table 4.12: Positive selection regions with more than 20% mutations in human sequence

Label	Start	Length	Score	Mutation%	Type	Gene
Pos_002	252480	67	17.33	33	Intergenic	TES-CAV2
Pos_006	460067	305	18.85	85	Intergenic	CAV1-MET
Pos_008	500822	60	29.25	29	Intergenic	CAV1-MET
Pos_015	850578	39	6.64	19	Intronic	ST7
Pos_016	860636	1634	26.63	25	Intronic	ST7
Pos_018	884970	6182	41.07	25	Intronic	ST7
Pos_022	1079721	44	19.43	50	Intergenic	ST7-WNT2
Pos_029	1276758	82	16.94	37	Intergenic	ASZ1-CFTR
Pos_030	1299284	359	20.84	22	Intergenic	ASZ1-CFTR
Pos_031	1405052	538	45.06	46	Intronic	CFTR
Pos_032	1485271	11	14.73	90	Intronic	CFTR
Pos_033	1487622	132	19.86	43	Intronic	CFTR
Pos_036	1546516	29	33.52	79	Intronic	CORTBP2
Pos_037	1565872	41	15.37	47	Intronic	CORTBP2

CHAPTER 5

SUMMARY AND CONCLUSION

5.1 Conclusion

The next phase of the Human Genome Project is to find all the functional regions of the genome and to determine their roles in human biology (Miller et al. 2004; Hardison 2003). Comparative genomics has a major role in this endeavour.

Nonhuman primates are the closest extant relatives of humans, and are thus the most pertinent organisms to understand human biology. However, due to the high degree of similarity between these primates at the nucleotide level, discriminating the functional from nonfunctional sequences is very difficult.

Recently, a strategy called phylogenetic shadowing was introduced to annotate genomes of closely related species, by analyzing the varying levels of variation along their multiple alignment (Boffelli et al. 2003, 2004). Phylogenetic shadowing is used to find genomic regions that are subject to non-neutral evolutionary selection, and are thus likely to be functional. The success of phylogenetic shadowing hinges upon the availability of orthologous sequences from several primate species. Often, the required data is not readily available and therefore this approach is of limited use at the moment. eShadow (Ovcharenko et al. 2004), the only publicly available phylogenetic shadowing tool, is based on a hidden Markov model that can be trained to detect regions under negative selection in closely related genomes. eShadow assumes that the distribution of columns with matches and mismatches follows a hidden Markov model. Moreover, eShadow's model does not account for positive selection.

In this thesis, we presented a new approach for finding the putative signals of purifying and positive selection in the alignment of three closely related species. In the presented methodology, no assumption is made about the distribution of the regions under selective pressure. We have applied an approach based on minimum

description length to find the regions subject to non-neutral evolution. Three evolutionary rate categories were defined: neutral, slow and fast. Each alignment column was labeled either as neutral (i.e. all three are neutral), slow (i.e. all three are slow) or fast for a particular species. Therefore, five distinct column classes were considered. Our method partitions the alignment into segments of fairly homogeneous mutation rates. The column classes were determined so as to maximize a score combining alignment likelihood with a complexity penalty.

Most alignments contain gaps. Two methods were implemented for handling gaps: the so-called standard phylogeny model which treats gaps as missing data, and the so-called tree-HMM model which takes into account the gap patterns.

The discussed methods were implemented in a Java program, called MDL-Shadow. The predictions of this program can be easily displayed on popular genome browsers like UCSC and Ensemble genome browser.

This methodology was tested on simulated data and on a genomic interval of the human sequence. Simulation was performed for alignments of length 10,000 and 100,000. The alignments consisted of several homogeneous segments and were generated with both standard phylogeny and tree-HMM models. Four different penalization factors were tested. We found that the penalization factor $0.5 * \log(N)$ led to the smallest segmentation error.

The segmentation and parameter estimation errors were considerably smaller for alignments of length 100,000 than for those of length 10,000. We concluded that our approach (as other alternatives) has a better performance on longer alignments because more samples are available to learn the model parameters. Another conclusion was that higher penalization factors are more successful in detecting long homogeneous segments. In contrast, smaller penalty functions would lead to smaller segmentation errors for short homogeneous segments, at the price of falsely annotating many regions as non-neutrally evolving.

We also analyzed an 1.8 Mbp long alignment of human, chimpanzee and baboon. This region, known as the greater CFTR region, was analyzed before by different groups (Thomas et al. 2003; Margulies et al. 2003; Cooper et al. 2005).

MDLShadow identified several segments as being under negative selection. A number of these segments are known exons. A previous study using 29 mammalian genomes had identified 20 ultraconserved elements in this region (Cooper et al. 2005). Among the predictions made by MDLShadow, 12 significantly overlapped with these ultraconserved elements. The majority of these elements are non-coding. We also identified unannotated regions which are well conserved in these three primates. These regions may represent non-coding functional regions or elements that have not accumulated enough mutations due to chance alone (albeit that chance is very low).

A few regions were also identified to be under positive selection in the human lineage after the human-chimpanzee split. These regions are either intronic or intergenic. A few of these segments are located in the intronic region of CFTR gene. It was suggested before that this gene might have been under positive selection in human population (NHGRI 2005). Nevertheless, further analysis is needed to see whether these regions are indeed under selection and to confirm their importance.

5.2 Future Work

The presented methodology can be extended to more than three species. For more than three taxa, different rate categories operate on subtrees of the phylogenetic tree. The number of classes should be specified beforehand based on the desired resolution and model complexity. A similar algorithm can then be applied to find the constraint elements as well as lineage-specific elements for each subtree.

Another important area for improvement is the optimization procedure. The Powell's method implemented occasionally reaches its maximum number of iterations without achieving convergence for some initial values. There are several alternative optimization methods that can be evaluated to find the more suitable ones. Since obtaining the right branch lengths are more important in partitioning alignments, one possible approach is to use Newton's method to optimize the branch lengths conditional on the current values of the other parameters and op-

timize all other parameters with derivative free optimization methods (similar to PAUP program). Further research is needed to see whether optimization of branch lengths and parameters at different steps can improve the results.

The penalty function used in the model selection process, was chosen based on the results obtained in our simulation experiments. We speculate that this penalty could be selected based on the expected length of the regions we are interested to find. If the objective is to capture short constraint elements, a smaller penalty function should be used.

Another interesting further research area is to "learn" the penalty function based on a training dataset for the species under study. For instance, by providing a set of constraint elements in the alignment, the program can search the possible space of penalization factors and report the factor capturing the constraint elements best. This factor can be subsequently used to identify other regions of alignment that are under selective pressure.

Another interesting extension is to modify the program to calculate the posterior probabilities of the predicted regions by posterior decoding (see Durbin et al. 1998). This extension will provide a means to eliminate predictions with small posterior probabilities and to concentrate on those corresponding to higher posterior probabilities. To calculate the posterior probabilities, a distribution over the partitions is required. In the spirit of algorithmic probabilities, the prior distribution can be selected proportional to 2^{-d} where d is the complexity penalty function (Li and Vitanyi 1997).

In the current method, no constraint is applied on the length of the segments estimated by the program. Our program can be extended to impose a minimum length on the identified segments (Cheng 2006; Csűrös 2004).

Segments identified in CFTR region as being under selection should be further analyzed. These regions can also be screened for potential non protein coding RNA genes. Sequences of different primates can be analyzed for the CFTR region to validate some of these predictions. Regions predicted to be under positive selection need more analysis because they might be alignment artifacts.

BIBLIOGRAPHY

BiologyCorner. http://www.biologycorner.com/bio1/DNA.html (as of February 1, 2006).

NCBI. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html (as of February 1, 2006).

Taygeta Scientific Inc. http://www.taygeta.com/rwalks/node3.html (as of February 1, 2006).

Adachi, J. and M. Hasegawa (1992). MOLPHY: Programs for molecular phylogenetics I–PROTML: Maximum likelihood inference of protein phylogeny. Technical Report 27, Institute of Statistical Mathematics, Tokyo.

Batzoglou, S. (2006). Lecture notes, stanford university. http://ai.stanford.edu/~serafim/CS262_2006/LectureNotes/ (as of June 1, 2006).

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler (2004). Ultraconserved elements in the human genome. *Science 304*(5675), 1321–1325.

Berg, J. M., J. L. Tymoczko, L. Stryer, and N. D. Clarke (2002). *Biochemistry.* New York, USA: W. H. Freeman.

Blanchette, M., B. Schwikowski, and M. Tompa (2002). Algorithms for phylogenetic footprinting. *Journal of Computational Biology 9*(2), 211–223.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science 299*(5611), 1391–1394.

Boffelli, D., C. V. Weer, L. Weng, K. D. Lewis, M. I. Shoukry, L. Pachter, D. N. Keys, and E. M. Rubin (2004). Intraspecies sequence comparisons for annotating genomes. *Genome Research 14*, 2406–2411.

Bray, N. and L. Pachter (2004). MAVID: Constrained ancestral alignment of multiple sequences. *Genome Research 14*, 693–699.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou (2003). LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research 13*(4), 721–731.

Bryant, D. (2003). Lecture notes, McGill Center for Bioinformatics. `http://www.mcb.mcgill.ca/\%7Ebryant/IHP/Paris1_1.pdf` (as of July 1, 2006).

Burnham, K. P. and D. R. Anderson (2002). *Model selection and multi-model inference*. New York, USA: Springer-Verlag.

Cavalli-Sforza, L. L. and A. W. F. Edwards (1967). Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics 19*(3), 233–257.

Chang, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences 137*, 51–73.

Cheng, M. T. (2006). Methods for multi-class segmentation of molecular sequences. Master's thesis, Université de Montréal.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson (2003). Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Research 31*(13), 3497–3500.

Chor, B. and T. Tuller (2005). Maximum likelihood of evolutionary trees is hard. In S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, and M. Waterman (Eds.), *RECOMB-05: Proc. of the 9th International Conference on Computational Molecular Biology*, Cambridge, England, pp. 296–310. ACM Press.

Collins, F. S., E. D. Green, A. E. Guttmacher, and M. S. Guyer (2003). A vision for the future of genomics research. *Nature 422*(6934), 835–847.

Cooper, G. M., E. A. Stone, G. Asimenos, NISC Comparative Sequencing Program, E. D. Green, S. Batzoglou, and A. Sidow (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research 15*, 901–913.

Csűrös, M. (2004). Maximum-scoring segment sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 1*(4), 139–150.

Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge, England: Cambridge University Press.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*(5), 1792–1797.

Eriksson, K. (2004). Statistical and combinatorial aspects of comparative genomics. *Scandinavian Journal of Statistics 31*(2), 203–216.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution 17*(6), 368–376.

Felsenstein, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Reviews in Genetics 22*, 521–565.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package), version 3.5c. Department of Genetics,University of Washington, Seattle.

Felsenstein, J. (2004). *Inferring phylogenies.* Sunderland, USA: Sinauer Associates, Inc.

Felsenstein, J. and G. A. Churchill (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*(1), 93–104.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and

A. Nekrutenko (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research 15*, 1451–1455.

Gilbert, W. (1978). Why genes in pieces? *Nature 271*, 501.

Guindon, S. and O. Gascuel (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology 52*(5), 696–704.

Hardison, R. C. (2003). Comparative genomics. *PLoS Biology 1*(2).

Hasegawa, M. and M. Fujiwara (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molecular Phylogenetics and Evolution 2*(1), 1–5.

Hillis, D. M., C. Moritz, and B. K. Mable (1996). *Molecular systematics.* Sunderland, USA: Sinauer Associates, Inc.

Huelsenbeck, J. P. (1995). The performance of phylogenetic methods in simulation. *Systematic Biology 44*(2), 17–48.

Hunter, L. (1993). *Artificial intelligence and molecular biology.* Menlo Park, USA: AAAI Press.

Jones, N. C. and P. A. Pevzner (2004). *An introduction to bioinformatics algorithms.* Cambridge, USA: MIT Press.

Kamvysselis, M. (2003). *Computational Comparative Genomics: Genes, Regulation, Evolution.* Ph. D. thesis, Massachusetts Institute of Technology.

Kelly, C. and J. Rice (1996). Modeling nucleotide evolution: a heterogeneous rate analysis. *Mathematical Biosciences 133*(1), 85–109.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature 217*, 624–626.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS 78*(1), 454–458.

Kishino, H., T. Miyata, , and M. Hasegawa (1990). Maximum likelihood inference of protein phylogeny and the origin of the chloroplasts. *Journal of Molecular Evolution 31*, 151–160.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics 39*, 309–338.

Krogh, A., M. Brown, I. Mian, K. Sjolander, and D. Haussler (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology 235*, 1501–1531.

Kuhner, M. K. and J. Felsenstein (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution 11*(3), 459–468.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. (2001). Initial sequencing and analysis of the human genome. *Nature 409*, 860–921.

Li, M. and P. Vitanyi (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. New York, USA: Springer-Verlag.

Li, W. (2001). DNA segmentation as a model selection process. In T. Lengauer, D. Sankoff, S. Istrail, P. Pevzner, and M. Waterman (Eds.), *RECOMB-01: Proc. of the 5th International Conference on Computational Molecular Biology*, New York, USA. ACM Press.

Li, W., P. Bernaola-Galván, F. Haghighi, and I. Grosse (2002). Applications of recursive segmentation to the analysis of DNA sequences. *Computers and Chemistry 26*, 491–510.

Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution 23*(2), 450–468.

Maes, F. (1998). *Segmentation and registration of multimodal medical images : from theory, implementation and validation to a useful tool in clinical practice.* Ph. D. thesis, K.U.Leuven.

Margulies, E. H., M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green (2003). Identification and characterization of multi-species conserved sequences. *Genome Research 13*, 2507–2518.

McGuire, G., M. C. Denham, and D. J. Balding (2001). Models of sequence evolution for DNA sequences containing gaps. *Molecular Biology and Evolution 18*(4), 481–490.

Miller, W., K. D. Makova, A. Nekrutenko, and R. C. Hardison (2004). Comparative genomics. *Annual Review of Genomics and Human Genetics 5*, 15–56.

Mitchison, G. J. (1999). A probabilistic treatment of phylogeny and sequence alignment. *Journal of Molecular Evolution 49*, 11–22.

Mitchison, G. J. and R. Durbin (1995). Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution 41*, 1139–1151.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics 15*, 211–218.

Mount, D. W. (2001). *Bioinformatics: sequence and genome analysis.* Cold Spring Harbor Laboratory Press.

Nei, M. and S. Kumar (2000). *Molecular evolution and phylogenetics.* Oxford, England: Oxford University Press.

Neyman, J. (1971). Molecular studies of evolution: A source of novel statistical problems. In *Statistical decision theory and related topics.*, pp. 1–27. New York, USA: Academic Press.

NHGRI (2005). New genome comparison finds chimps, humans very similar at DNA level. National Human Genome Research Institute at `http://www.genome.gov/15515096` (as of July 1, 2006).

Nielsen, R. (2005). *Statistical methods in molecular evolution.* New York, USA: Springer-Verlag.

Nobrega, M. A. and L. A. Pennacchio (2003). Comparative genomic analysis as a tool for biological discovery. *Journal of Physiology 554*, 31–39.

Ovcharenko, I., D. Boffelli, and G. G. Loots (2004). eShadow: A tool for comparing closely related sequences. *Genome Research 14*, 1191–1198.

Pardi, F. and N. Goldman (2005). Species choice for comparative genomics: being greedy works. *PLoS Genetics 1*(6).

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *PNAS 85*(8), 2444–2448.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2002). *Numerical recipes in C, The art of scientific computing.* Cambridge: Cambridge University Press.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 3*(1), 92–94.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research 15*, 1034–1050.

Siepel, A. and D. Haussler (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology 11*(2–3), 413–428.

Singh, M. (1999). Lecture notes, Princeton University. `http://www.cs.princeton.edu/~mona/Lecture/phylogeny.pdf` (as of February 1, 2006).

Stone, E. A., G. M. Cooper, and A. Sidow (2005). Trade-offs in detecting evolutionary constrained sequence by comparative genomics. *Annual Review of Genomics and Human Genetics 6*, 143–164.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *PNAS 48*(4), 582–592.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology 203*(2), 439–455.

Tamarin, R. H. (1999). *Principles of genetics.* Boston, USA: McGraw-Hill College.

Thomas, J. W., J. W. Touchman, R. W. Blakesley, G. G. Bouffard, S. M. Beckstrom-Sternberg, E. H. Margulies, M. Blanchette, A. C. Siepel, P. J. Thomas, J. C. McDowell, et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature 424*, 788–793.

Thompson, J. D., D. G. Higgins, and T. J. Gibsont (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research 22*(22), 4673–4680.

Thorne, J. L., H. Kishino, and J. Felsenstein (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution 33*, 114–124.

Thorne, J. L., H. Kishino, and J. Felsenstein (1992). Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution 34*, 3–16.

Watson, J. D. and F. H. C. Crick (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature 171*, 737–738.

Yang, Z. (1993). Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution 10*, 1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate method. *Journal of Molecular Evolution 39*, 306–314.

Yu, Y. K., R. Bundschuh, and T. Hwa (2002). Statistical significance and extreme ensemble of gapped local hybrid alignment. In M. Lässig and A. Valleriani (Eds.), *Biological Evolution and Statistical Physics*, pp. 3–21. Springer-Verlag, Berlin.

Yu, Y. K. and T. Hwa (2001). Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *Journal of Computational Biology 8*(3), 249–282.