A Likelihood Model of Gene Family Evolution

Lindsey Dubb

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2005

Program Authorized to Offer Degree: Genome Sciences

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Lindsey Dubb

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Joseph Felsenstein

Reading Committee:

Joseph Felsenstein

Elizabeth Thompson

Philip Green

Date:

Signature__________________________

Date______________________________

University of Washington

Abstract

A Likelihood Model of Gene Family Evolution

Lindsey Dubb

Chair of the Supervisory Committee:
Professor Joseph Felsenstein
Genome Sciences

The duplication of genes is a source of new genes, and thus can be of great interest. This thesis presents a simple model of gene duplication and gene loss, and a method by which the probability of a gene phylogeny can be computed for particular rates of these processes. This allows the likelihood of duplication and loss rates for a particular gene phylogeny to be determined. Using such calculations, I have examined the ability to estimate duplication and loss for a variety of different types of gene phylogenies.

Genetic data do not specify a single gene phylogeny with certainty. By integrating over many possible gene phylogenies via Markov chain Monte Carlo methods, uncertainty in the gene phylogeny is allowed. I have written a computer program to perform this calculation for DNA sequence data. With this, I have examined sources of variance in the estimation of gene duplication and loss rates.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

The author would like to express gratitude to Joe Felsenstein, my graduate advisor, for his thoughts, careful feedback on this thesis, and support. My thanks go also to the members of the laboratory for their help and example. I am also grateful to the members of my committee for their time and for their reading and thoughts on this thesis.

I would like to thank the Genetics and now Genome Sciences departments. In particular Brian Giebel in the Genome Sciences office has been of great assistance with the many forms and procedures on the way to graduation.

My sincere gratitude goes also to Brendan Stewart for her encouragement and support, and for her reading and editing of this document.

Chapter 1

INTRODUCTION

1.1 Importance of Gene Family Evolution; Purpose of Work

The evolution of gene families is an important aspect of molecular evolution. A central goal of the study of molecular evolution is an understanding of the relationships between genes. Therefore the form and parameters of those relationships are themselves notable. By finding out what we can about relationships between genes, we learn about the process by which new genes arise. In this way we can hope to better understand and predict relationships among genes in general.

These gene relationships can take many forms. Genes can be related due to duplication, which produces a new copy of a gene within a species, lateral transfer between species, or by conversion of part of a gene by part of a similar gene. Each of these relationships can potentially occur in a variety of ways. For tractability, however, it is necessary to narrow down the kinds of relationships under study. Here I will examine gene duplication and loss within a group of related species using one specific simple model.

The goal of this thesis is to present this model of gene family evolution, describe methods of its application to the analysis of data, and evaluate its behavior under a variety of circumstances. This distinguishes my work from previous studies of gene duplication primarily in that my approach uses an explicit probability model of duplication and loss. (These other gene duplication studies will be described in detail in chapter 2 of this thesis.) Clearly my model does not — and could not — allow for all natural processes that occur in gene evolution. However, with this model-based approach, the probability of data on a particular gene phylogeny can be calculated, and maximum likelihood estimates of the model's parameters can be inferred. This provides a means by which we can sum log likelihoods over multiple gene phylogenies, weighting according to the probability of each gene phylogeny. In addition, the assumptions of the model are explicit, and the effects of those assumptions on inference can be examined. Furthermore, by providing a simple model as a framework, it should be possible to extend the model in a variety of ways. Thus, though there are limitations to this simple model, they are explicit and can be analyzed, and it should be possible to add important complexities in further work.

Gene family evolution, though of interest in itself, is also important when inferring the relationships among species. In much phylogenetic research, the genes under study are in a family of related genes. (See, for example, Slowinski and Page 1999 for a review of this.) Though systematists often attempt to use only orthologous genes (genes related through speciations and not duplications), orthology is difficult to confirm. This also severely limits the genes available for study. In addition, many researchers choose the genes they study not for orthology but rather for a function of interest.

To allow for likelihood-based phylogenetic inference (including Bayesian methods) using genes which may be related through duplication, it must be possible to calculate the likelihood for the parameters of gene duplication and loss on a gene phylogeny. In turn, phylogenetic trees can be of interest not only for the species phylogeny in itself, but also for inference about traits of the organisms within the phylogeny. Likewise, for inference of these to be based on gene family information, a gene family model is needed.

1.2 Definition of Terms

To describe and discuss the evolution of genes within a gene family, it is first necessary to define the terminology being used. For the most part, I will use the standard terminology of the phylogenetics literature. However, though most of these terms are already familiar, some have multiple meanings depending on context, or are not clearly defined. Therefore, I will explain how these terms will be used within the thesis.

A “species” is taken in a historical, phylogenetic sense to mean a group of organisms which shared genetic information over a period of time.

A “species phylogeny” or “species tree” is defined here as a tree of the evolutionary relationships among a group of species. In equations and when using the term repeatedly, a specific species phylogeny will sometimes be referred to by with an uppercase S.

“Gene” will be used to mean a contiguous length of DNA in a single species. Its use does not imply that the DNA has a function.

A “gene family” is used to name a group of related genes. “Gene phylogeny” or “gene family tree” will be used to describe the tree of relationships among those related genes within a species phylogeny. “Gene phylogeny” is the term used by Goodman et al. (1979) in the paper that introduced the reconciliation of gene and species phylogenies. It is admittedly a somewhat unwieldy phrase. The shorter name “gene tree” has often been used for this in the gene family literature. However, “gene tree” has also been used in the discussion of the coalescent process to describe the relationship between different copies at a locus. To avoid confusion with other uses of “gene tree,” I have chosen to use the longer original terms, instead. (Perhaps at some future time, gene family trees will be found using multiple copies at each locus, and the distinction between gene trees and gene family trees will disappear. In this thesis, however, the two are different.) In equations, and when using the term repeatedly, a specific gene family tree will sometimes be referred to by an uppercase G.

Since much of this thesis is concerned with details of the structure of trees (both of species and of genes), they will need to be described clearly. Each vertex in a tree will be called a "node." Internal nodes in a species tree represent speciation events, while internal nodes in a gene family tree represent either gene duplications or the effect of a speciation event on a locus. Terminal nodes will be called "tips." A connection between nodes will be called an "internode." An internode and all nodes and internodes descendant of that internode will be referred to as a "subtree." A "child" node of a node N will be a node immediately descendant of N, while a "parent" of N will be the node immediately ancestral. See figure 1.2 for a graphical description of these terms.

The bottommost node at which two lineages come together will be called the "root" of the tree. The bottom of the branch below the root is called the "base" of the tree. This definition does not follow convention for the field, which in general does not consider the tree below its root. The base represents the tree before its earliest branching. Since it is possible that there will be gene duplication before the root of the species tree, it is necessary to consider this extra internode below the species tree root. In this paper, all the species trees will have a base and a root, and all gene family trees will have a root.

For consistency, the bottom of a tree will always represent its past, with time increasing toward the present at the top of the tree. Earlier points in a tree will sometimes be referred to as the "bottom" or "beginning," while a later points will be called the "top" or "end." (Note that this is far from standard in the literature. In particular, writings on graph theory of rooted trees

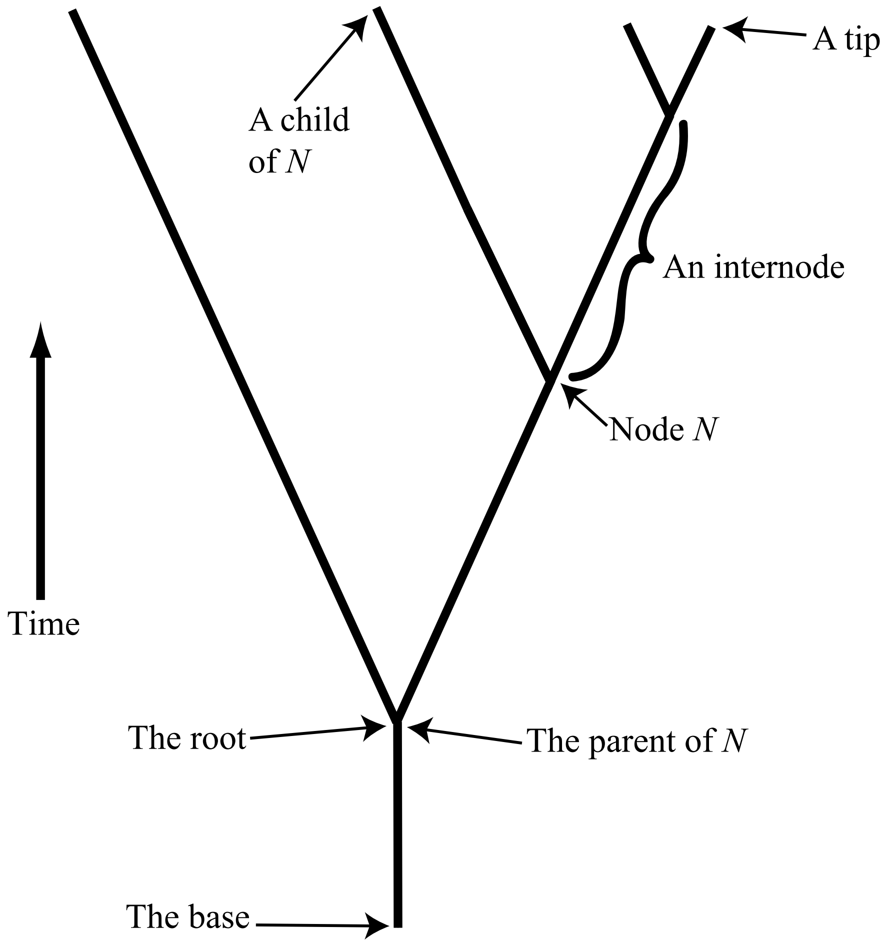

Figure 1.1: Names for nodes

customarily place the root at the top of the tree.)

An “exemplar” represents the original gene that gives rise to a new gene locus through duplication. In my model, the new gene and the exemplar cannot be distinguished using the data. However, there is some potential to infer this through means of models that consider population size or genomic context.

“Paralogs” are genes that share a common ancestor at a duplication event. “Orthologs” are genes in different species that share a common ancestor at a speciation event. These follow the original definition of terms by Fitch (1970). Paralogy and orthology are meaningful as a description of the common ancestor of two gene family members, and do not describe an entire tree. As a result there is some oddity to this terminology, since though genes A and B may be paralogs, gene C may be orthologous with both of them (see figure 1.3). This has been pointed out by a number of researchers — See, for example, Fitch (2000).

“Doomed lineages” are defined as lineages for which all descendants are lost before the time gene family members were sampled. “Surviving lineages” are the opposite: These are genes that have descendants among the sampled genes. “Known doomed lineages” are those doomed lineages that must necessarily have existed after a speciation event if a lineage survives in just one of the descendant species. (This is the result of a model assumption that a gene in an ancestral species will be inherited by both descendant species.)

A “duplication” is an event in which a single locus with a member of a gene family gives rise to two loci (the original exemplar and one new locus). In this thesis, there will be the simplifying assumption that two loci will be found throughout the population immediately after the duplication event. A “loss” is an event in which a single locus with a member of a gene family can no longer be found in a species. These definitions, their accompanying assumptions, and

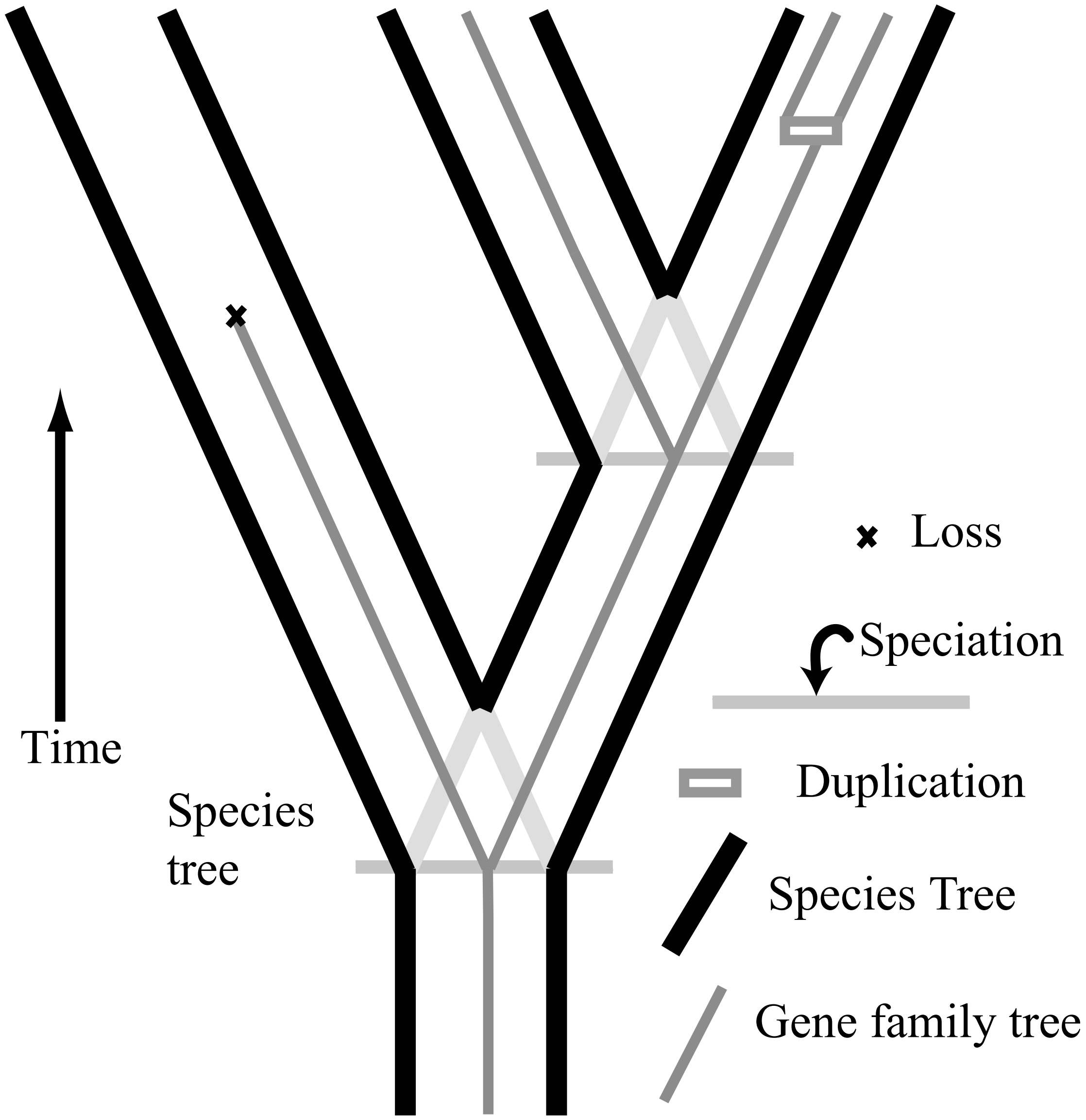

Figure 1.2: A gene family tree within a species tree, with the parts named

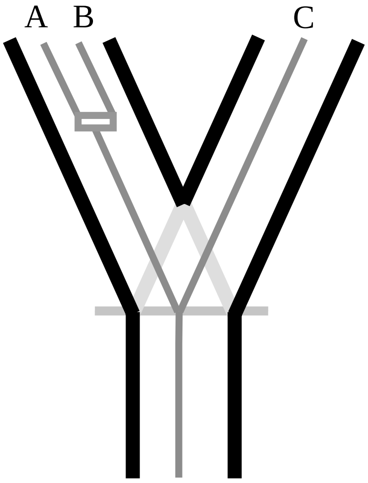

Figure 1.3: Paralogy and orthology: A and B are paralogs, as they are related through a duplication event. A and C (and B and C) are orthologs, because their common ancestor is at a speciation event.

their relation to biological events will be discussed in more detail in subsection 3.7.1.

“Lateral transfer” will refer to a historical event in which the gene from one species enters the genome of another.

“Gene conversion” refers to the replacement of part or all of one gene with part or all of another, similar gene in a nonreciprocal process. This does not change the length of the gene (as for unequal crossing-over).

“Blurp” and “Quilg” are the names of two computer programs I have written to implement the methods described in this thesis. Blurp calculates likelihoods when the gene family tree is known precisely, while Quilg does so with DNA data, without precise knowledge of the gene family tree.

Following standard practice in the literature of the birth-death process, the rate of duplication (births) will be referred to as λ, and the rate of losses as μ.

Chapter 2
RELATED RESEARCH

2.1 Analysis of Gene Family Members as Separate Species

There is a very extensive literature which examines families of genes. Most of these are concerned with the members of a single gene family, rather than with gene families in general. Many such studies use phylogenetic inference programs, treating each locus in the gene family as a separate species. This approach does not make use of information about the species in which each gene family member is found. Members of a gene family in different species can have a common ancestor only once they are in a species which is ancestral to both species in which those genes were found. When treating genes as separate species, genes can instead have a common ancestor at any time, implying that they are able to travel between species at will. Or from another perspective, the species phylogeny introduces constraints to the time of events on the gene family tree. Analysis of genes as separate species necessarily fails to consider these constraints. Furthermore, as gene duplications and losses are not considered to be stochastic events in these models, there is no change in the likelihood (or in the case of parsimony methods, penalty for the occurrence) of a tree with one or more such events.

The process by which gene families have evolved has also been examined for certain gene families, particularly in the study of immune proteins (see for example Ohta 1983; Gu and Nei 1999; Garrigan and Edwards 1999). Much of the recent literature has compared the relative importance of duplication/loss and gene conversion in the evolution of these molecules. These studies have largely been based on the degree to which members of a gene family in each species resemble one another more than the genes in another species.

2.2 Parsimony Based Approaches

The relationship between species phylogenies and gene phylogenies has led some researchers to define methods by which the phylogeny of the species can be used to help find the phylogeny of the genes in those species. These methods assume, as does my model, that a series of duplications and losses lead to the current set of genes. Unlike my method, the methods in this section are all based on parsimony criteria — the minimization of the number of certain kinds of events.

The first such method was introduced by Goodman et al. (1979). In that paper, the total number of gene duplications, “expression events,” and DNA substitutions were minimized for the (assumed) species phylogeny. Expression events were the gain or loss of the ability to identify the gene. The authors suggested that this might represent a deletion, but could also be considered any other change decreasing or increasing the ability to detect that gene. They used their described optimization criteria to evaluate different gene phylogenies of the myoglobin and α- and β- hemoglobin genes, and showed a result more in keeping with other data for the gene family trees than could be obtained from any of these groups of genes, alone. Although their method described criteria for the optimality of a gene family tree, it did not give an algorithm for the determination of such an optimal tree. Due to this and the lack of suitable data at the time of its publication, their method has rarely been used.

Fitch (1979) criticized this method as an arbitrary weighting of duplications, expression events, and DNA substitutions. This question of weighting was raised again along with possible weighting methods in Ronquist (2003).

Ronquist recommended as the best solution the selection of weights which provide the clearest phylogenetic signal as determined by resampling the character data.

The method of Page and Charleston (1997) is similar to that of Goodman et al. (1979), but is instead designed to find the unknown species tree from given gene trees. It also includes a heuristic algorithm to implement their method, and a computer program which uses this algorithm. Their method counts only duplications (or only duplications and losses), not expression events and DNA substitutions. The authors suggest that only duplications should be counted if the sampling of gene family members is incomplete, as otherwise their absence will be incorrectly attributed to a loss. If all the gene family members are fully sampled, however, the sum of duplications and losses is favored for its higher ability to resolve between gene phylogenies. Page and Charleston's method assumes that the topology of the gene phylogeny is precisely known. To avoid this, Page (2000) later suggests the step of allowing nearest neighbor interchanges at weakly supported nodes. The method has more recently been extended in Page and Cotton (2000) to resample the data to allow for a lack of certainty about the gene phylogeny. Page and Cotton (2002) also suggested the use of Bayesian credibility intervals as an alternative to bootstrap resampling. Page and Charleston's method has been improved by Hallett and Lagergren (2000), who provided a faster (polynomial-time) non-heuristic method of computing the optimal gene family tree for minimized duplications and losses.

A different method of parsimony-based inference of duplications was proposed by Zmasek and Eddy (2002) and made available in their RIO (Resampled Inference of Orthologs) package. This program takes a group of DNA sequences in a set of species along with a user provided species phylogeny. By resampling the DNA sequences, sets of gene family trees are generated. With each gene family tree produced, the program determines where duplications have occurred. By summing over all the bootstrap replicates, RIO is able to provide a percentage of those trees which support paralogy versus orthology for each pair of gene family members.

Another method, of Guigó, Muchnik and Smith (1996), instead treats any number of duplications in a species phylogeny's internode as caused by a single duplication episode, and suggests the minimization of these episodes. This third method allows the gene phylogeny and species phylogeny to be determined simultaneously, but the method is (according to the paper) very prone to finding a local rather than a global minimum number of duplication episodes. (To avoid this problem, Guigó, Muchnik and Smith used starting species phylogenies believed through separate analysis to be correct.) The problem of minimizing duplication episodes has since been shown by Fellows, Hallett and Stege (1998) to be NP-hard.

Simmons, Bailey and Nixon (2000) introduced a method to infer a species phylogeny without being mislead by gene duplications. This uninode coding method attempted to remove any paralogous comparisons in a phylogeny by first inferring the location of duplication events in the phylogeny, then treating each descendant subtree of the duplication separately. Each duplication (but not deletion) is considered to be a single event for purposes of minimization of events on the tree. No method is given for inferring the location of duplication events — It is assumed they can be found by inspection of the phylogenetic tree (in which all gene family members are treated as species). The number of DNA changes plus duplications is then minimized by comparing genes only within the subtrees. In the case of a gene phylogeny, this method correctly minimizes changes only if the locations of the duplication events are exactly correct. Thus it is likely to be vulnerable to gene losses and incomplete sampling, which can obscure the occurrence of duplications.

Each of the models described above minimizes a different quantity — gene duplications, expression events, and DNA substitutions (Goodman et al. 1979), duplications or duplications plus losses (Page and Charleston, 1997), duplication episodes (Guigó, Muchnik and Smith 1996), and duplications and character substitutions (Simmons, Bailey and Nixon 2000). Justification for the particular choice of events to minimize is rarely explicit in these articles.

Felsenstein (1981a) showed that parsimony-based inference of a phylogeny (without duplications or losses) is equivalent to likelihood based inference when the probability of one substitution is always greater than that of two substitutions — that is, when the probability of a change is extremely small. Using this logic, it can similarly be shown that the assumptions of duplications and losses made in the parsimony methods are justified under a likelihood model when rates of the events (duplications and losses, for example) are extremely small.

2.3 Probabalistic Treatments

Gu (2000) presents a way to find pairwise distances between species (and using distance-based tree inference methods, a species phylogeny) and provides formulas for the probability that a pair of species will both have members of the same gene family. This was limited by the use of a simple death process without births, and analysis of just one specific tree topology.

Gu (2000) suggests that this can easily be extended to a simple birth-death process. However, there are some problems with the method presented. The equation for calculating probabilities is given in the article on page 520, equation 16. There, Gu sums over all possible numbers of gene family members at every node in the gene family tree. This is correct, but is feasible only for trees with few nodes, and only in the case of a simple death model. The birth-death model presented immediately thereafter can theoretically lead to interior nodes with any number of gene family members.

More importantly, the method described considers only the number of gene family members present in each current species, and not the degree of similarity between members of the gene family. Thus it does not take into consideration the topology of the gene family tree, and thereby loses a significant source of information. On the other hand, in cases where it is extremely difficult to align and compare sequences, this loss might be less important, or even an advantage. In addition, this method could potentially be useful when employing a screening method (such as PCR) to identify the presence or absence of a gene family without attempting to sequence each gene family member.

Arvestad et al. (2003) used the results of Nee, May and Harvey (1994) (described in section 2.4 of this thesis) to calculate the probability of a birth-death process of gene evolution occurring within each species in a species tree. This was done through a Markov chain Monte Carlo (MCMC) method. Specifically, they begin with a gene family tree with a specified topology and a labelling of which tip of the gene family tree is found in each species. The branch lengths, however, are not given, but rather iteratively changed by MCMC. The algorithm then estimates the probability of each reconciliation. That is, given a known species tree and set of genes related by a known topology, they show how to estimate the probability of any particular set of branch lengths. This is similar to the calculation of the probability of a specific gene family tree in my own work. However, the method of Arvestad et al. (2003) does not permit duplications in the gene family tree prior to the root of the species tree. This method also requires the use of MCMC for the calculation. This requirement was removed by their next paper, described in the following paragraph.

In Arvestad et al. (2004), their method was changed in a number of ways. Among other work in the field, this is the most similar to the work presented in this thesis. Like my own methods, the authors determine the probability of the gene family tree by calculating the probability at each internode in the gene family tree, recursively calculating likelihoods toward the root of the species tree. Unlike my own methods, they do not explicitly calculate the likelihood of the tree above each node (i.e., of the subtree defined by the node) conditional on the number of doomed lineages at that node. Instead, they calculate probabilities directly from the process resulting from a "reconstructed" birth-death process that has had any lineages which are lost before the present removed.

Using this method of calculating the probability of a gene family tree in a known species tree, the authors use MCMC in a Bayesian framework to estimate the posterior distribution of gene family trees. They use this distribution to estimate the posterior probability that a particular pair of gene family members is related through orthology versus paralogy.

There is an implicit assumption in this method that the gene family tree begins at the base of the species tree. Ideally one would want to allow for duplications below the root of the species tree. Though my own method does allow for duplications prior to the root of the species tree, it, too, does not employ an ideal method to consider the root of the tree. This will be discussed in more detail in subsection 3.7.2, starting on page 42. It is also not possible to verify the formulas of Arvestad et al. (2004) numerically, as the program used to implement their method is not yet available.

2.4 Models of Duplication and Loss

Studies of the biology and population genetics of gene duplications and deletions are relevant to this work, as they suggest ways in which gene duplication models should be formulated and point to possible shortcomings in such models. The duplication and deletion of genes has been seen both *in vitro* and through examination of changes in DNA sequences. In bacteria, for example, as many as 1 in 10^4 new cells show duplications in certain genes (Watson et al. 1987). These are thought to arise primarily through crossing-over events between homologous regions of DNA during the replication process. Tandem duplications have often been observed as a result of selection for drug resistance in both bacteria and in cultured eucaryotic cells (Watson et al. 1987).

Deletions occur due to a number of different causes. In *E. Coli*, approximately 1 in 20 spontaneous mutations are deletion events (Clark 2005). They are believed to occur due to a variety of causes including unequal crossing over, radiation damage, and transpositions (Watson et al. 1987). Evidence for deletions has also been seen due to somatic mutation in cancerous tissues (see for example Montesano, Hollstein and Hainaut 1996 for a review). In addition, deletions have been seen as the result of chromosomal abnormalities in humans (K.-S. Chen et al. 1997).

The population genetics of a duplicated gene with the same function as the original were first studied by Haldane (1933) and Fisher (1935). This has more recently been modelled in more detail by a number of researchers. Christiansen and Frydenberg (1977) studied the process of loss of functionality in a two locus model with mutation to loss of function at each locus, and with selection against the mutations only when both loci had lost their functionality. They graphed the stable gene frequency levels on a two dimensional graph of frequency at locus A versus frequency at locus B. When the rate of mutation to loss of function is the same at each locus (as might be the case after a duplication event), their model predicts a hyperbola of stable frequency values along which the frequencies can drift. As their model does not allow a mutation to return to functionality, this unstable equilibrium will eventually drift to the loss of one or the other of the functional genes.

This was later examined by Clark (1994) and Nowak et al. (1997), who found that duplications could be maintained when deleterious mutations to the gene are common, as the presence of one functional copy could prevent the loss of fitness if the other copy were to mutate to become nonfunctional.

The genetics of duplicates of genes in the presence of changes in function has also been studied. Walsh (1995) looked at the dynamics of the duplication of genes with the addition of a process in which one of the pair of genes can gain a selectively advantageous new function, and compared the dynamics of such duplicates with those without new functions. With this model, Walsh found that the rate of fixation of advantageous duplicate genes was $S/(1-e^{-S})$, where $S = 4N_e s$, N_e is the effective population size, and s is the additional fitness of the advantageous gene in a heterozygote. Force et al. (1999) examined the fate of a duplicated gene when the new and original copies both lose parts of their function (referred to as subfunctionalization). Lynch, et al. (2001) further examined the fate of a new duplicate under a number of different circumstances using simulation. They again found a role for population size in the rate of fixation of duplicates with new functions. In addition, the authors found that the kind of new function established will also tend to depend on the population size; Larger populations tend to favor completely new functions, while smaller populations tend to acquire new genes through subfunctionalization. They also examined the possibility of linkage between gene duplicates, finding that in the case of linkage (only), the existence of a duplicate gene can be maintained due to its protection from a deleterious mutation in either of the copies. Otherwise, a duplication without a new function would only be selectively advantageous if two copies have a selective advantage over a single copy due to errors in intracellular processing.

2.5 Mathematical Properties of the Birth-Death Process

Kendall (1948) wrote much of the theory of the birth-death process. Thompson (1975) added to this by examining the tree of lineages from a birth-death process which survive to the present time. The author found the joint density of the number of current genes and the times and order of their branching.

Nee, May, and Harvey (1994) examined the birth-death process conditional on the survival until sampling of at least one lineage, which they called the "reconstructed process" (see figure 2.1). They found this conditional process to describe a pure birth process with a changing rate of births. Using this process, they determined the likelihood function of a reconstructed phylogeny. This was again done conditional on the survival of at least one lineage, but not conditional on the observed number of surviving lineages. Their equations counterintuitively demonstrate that rates of gene loss can indeed be inferred using just the tree of surviving lineages. They also provided equations for partial sampling of gene family members in the present, assuming a known probability that each would not be sampled.

Rannala and Yang (1996) conditioned instead on the number of genes at the time of sampling. In the context of their paper (which uses the birth-death process to model speciation and extinction), this was the number of species rather than the number of genes. They calculated the joint density of the order of branching and times of duplications conditional on a specific number of current species.

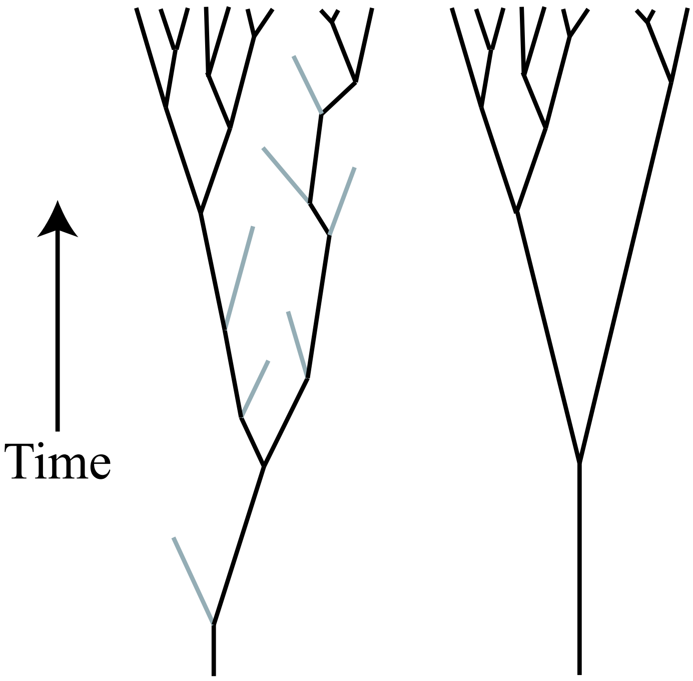

Figure 2.1: The tree on the left is a sketch of a simulated birth-death process with 10 tips. The bottom is the start of the process, and the top is the time of observation of the lineages. Losses are shown as lineages which stop before the top of the tree. On the right is the same tree, but showing only the relationships between lineages which survived to the time of observation. This tree on the right is the “reconstructed” birth-death process.

2.6 Detection of Paralogy

Much of the literature about gene duplications and losses appears to be concerned primarily with the distinction between paralogy and orthology — that is, whether the most recent common ancestor of two gene family members occurred at a duplication or a speciation event.

When a basis for interest in this question is given, it is most often because the researcher is asking another question for which the relationship between those gene family members is important. In the most common case, the researcher wants to know how multiple species are related, and would prefer to use orthologs rather than paralogs, as the gene family tree of the former will usually correspond more closely with the species phylogeny. Another similar example can be found when attempting to time speciation events via a molecular clock.

These are cases in which paralogy (i.e., gene duplications) make analysis more difficult. However, the problem here lies in algorithmic difficulty, not in any flaws with paralogous genes. Paralogs do contain information about species trees. In addition, there are problems introduced when discarding paralogous genes. In particular, when there are two genes which are paralogous to one another, but orthologous to a third gene (see figure 1.3, page 8), it is unclear which of the paralogs should be removed from the analysis. There is also a small danger of biasing the sample of genes, since only members of gene families are discarded. Unless there is some aspect of larger gene families which cause them to evolve in a different manner, however, this would not be a problem.

More importantly, it is not generally possible to determine with certainty whether genes are orthologs or paralogs. As a result, it is necessary to take account for this uncertainty in this estimate, rather than simply discard any genes which appear to be paralogous. This thesis will demonstrate the basis for this approach in the case of estimation of rates of duplication and loss.

There is, however, one instance in which the duplication event in and of itself could be important. If duplication is related to the acquisition of new gene functions, then orthologs will tend to have similar functions, while paralogs will have different roles. This supposition, however, cannot hold in all cases. This is made clear in the example where A and B are paralogous, but C is orthologous to both. If a duplication gives rise to new functions in the copied locus, then an ortholog to both will share the function of at most one of those copies, refuting the statement that orthologs share functions.

To avoid this pitfall, Zmasek and Eddy (2002) introduced a new concept called “super-orthology.” Super-orthologs are gene family members which not only share a common ancestor at a speciation event, but also were never duplicated between their most recent common ancestor and the present. This would appear to provide pairs of genes which will tend to share the same function. However, there remains the possibility that there has been a past duplication event for which the data has no direct evidence. For example, the researchers may not know of all members of the gene family, or a duplication might have occurred followed by the loss of one of the copies. One can avoid this problem by hypothesizing that new functions will only occur with duplications when both the original and the duplicate survive for some time after the duplication event. (If new functions tend to evolve only when selection is removed for the maintenance of the original gene function, then the loss of either copy could prevent the development of new functions.) However, this dependence on survival of both gene family members does not avoid the possibility of missing a gene family member when sampling. It also does not entirely avoid the question of gene deletion, since a gene might have been deleted a long time after a duplication event. Finally, it is also possible that new functions could evolve in a gene without its duplication.

If new functionality arises only in the new duplicate, and not in the original, then this implies a fundamental difference between the two copies. In this case the researcher is directly interested in paralogy versus orthology, since the event of duplication has a pronounced effect. One possibility is that the copy at a new location in the genome can have different regulation of its expression, and thus will be expressed in, and adapt to, different circumstances. This was examined by Lynch and Force (2000) in terms of expression in different animal tissues. If this is the case, it may be possible to look for differences between the original gene (called the exemplar) and the new copy.

There are three methods by which this could be approached. As the duplication will have taken place in a single individual, the most recent common ancestor of all versions at the new locus must be after the time of the duplication. Copies of the exemplar, on the other hand, may coalesce prior to the duplication event. Another source of information lies in the genomic context of the two copies — The sequences surrounding the exemplar may resemble the sequences surrounding exemplars in other species, but there may be less reason to expect similarities in the sequences surrounding different new copies. This was used by Sankoff (1999) to propose a method in which knowledge of the ancestral gene order in outgroups is used to determine the likely exemplar from each pair. On the other hand, a large duplication would copy some surrounding DNA, as well, making it more difficult to extract this kind of information. A third source of information could come in the form of functional assays of the different gene family members. However, it is in many cases difficult to model the evolution of gene function, or even to define functional similarity.

2.7 Genomic Rearrangement Models

The methods described above model the relation between members of a single family of genes. In this section, I will briefly discuss research covering the evolution of whole genomes. Clearly there is some relation between these kinds of models. Increases in ploidy, for example, will increase the number of chromosomes but also will create new copies of every gene. Rearrangements without duplication, on the other hand, might have no obvious effect on any gene families. However, by altering gene order, the probabilities of future events affecting gene family members near one another could be changed.

A number of researchers have attempted to infer trees of genomes derived under the assumption that all rearrangements are due to inversions, in which part of the genome is reversed in its order. For the most part, these models have attempted to minimize the number of inversion events. This is potentially a useful method of analysis, assuming that these inversions are very rare events. A polynomial time algorithm to determine the minimum number of inversions needed to explain differences between a pair of genomes was found by Hannenhalli and Pevzner (1999).

In Larget, Simon, and Kadane (2002), the authors present a probability model of inversions along with a method by which the probability of a tree with gene inversions can be calculated. Their method uses data indicating gene order and direction, but not DNA sequences. They model inversion events as occurring at exponentially distributed intervals in each internode of the species tree. The probability of each possible inversion (as defined by reversal operations on a signed ordering of genes) is assumed to be equal. They use Markov chain Monte Carlo to sample among possible histories of inversions. Their original prior distribution of trees (with all unrooted trees having equal prior probability) prevented sampling from any trees with high likelihood; However, this was changed in Larget, et al. (2004) to make use of prior knowledge about phylogenetic relationships at the phylum level.

Transpositions and translocations have been examined with genome rearrangement models, as well. Blanchette, Kunisawa, and Sankoff (1996) presented one algorithm to minimize the (weighted) number of inversions, transpositions, and translocations for a pair of genomes. This was improved and extended to a tree of multiple genomes by Bourque and Pevzner (2002).

For these models, gene duplication has largely been regarded as a nuisance, as it results in a pair of copies, either of which can be matched to the gene in the other genome. Sankoff (1999) proposed that only the original locus (the exemplar) from each duplication should be used in analysis, allowing these other methods to be applied. However, this discards information available from genes surrounding the new copy. Another approach has been taken by El-Mabrouk (2002), who describes an algorithm which minimizes duplications and reversals (inversions) on a phylogeny of multiple species. However, the algorithm is limited to gene families with at most two members.

Genome rearrangement has also been modelled more abstractly through the analysis of "breakpoints." This is a term introduced by Dobzhansky and Sturtevant (1938). Breakpoints are discontinuities in the linear order of a genome introduced by a variety of events, such as inversions, translocations, or deletions. In Blanchette, Bourque, and Sankoff (1997), the authors proposed minimizing these breakpoints between species as a criterion for optimizing the phylogeny of genomes. These breakpoints are not the same as counting inversion, translocation, or deletion events, but they are significantly easier to identify than those biological processes.

Another approach to genomic rearrangement has been attempted in the OrthoParaMap program by Cannon and Young (2003). This method incorporates both sequence data and genetic map information to both identify duplicated genes and determine amount of DNA duplicated in each event. OrthoParaMap does not use a model of genome rearrangement, but rather attempts to examine pairs of fully sequenced genomes, searching for long "diagonals" of similar genes using a deterministic heuristic method. Once these regions have been found, the program searches for gene family members (via sequence similarity using BLAST; Altschul et al. 1997) which are found in the same relative location in both of the matching DNA segments in a "diagonal." These gene pairs are then assumed to be orthologs. The last step looks for gene family members which are near each other in each subtree of the gene phylogeny. These are assumed to be paralogs produced through local duplications.

OrthoParaMap provides information unavailable to other current forms of analysis, but it does have substantial limitations. The heuristic nature of the search for diagonals does not provide a method for determining the accuracy of the estimates. Gene loss events are not considered by the model, and so the method can fail to infer speciation events even when one is logically required, if one of the resulting species does not contain an extant gene family member in that subtree (Cotton, 2005). The method also does not allow for small duplication events in which a gene is duplicated at a distant location from the original locus. That said, the examples shown in their paper do demonstrate that there is significant information available from looking for regions of duplications.

2.8 Empirical Studies of Duplication and Loss Rates

Nadeau and Sankoff (1997) used whole genome duplications to estimate rates of gene duplication and loss. Mice and humans were assumed to have undergone two whole genome duplications, leading to four genes from each original locus. They looked at three models with a single parameter representing that a new locus will be lost rather than diverge in function and be preserved by selection. They examined the distribution of the number of remaining members in different gene families, and attempted to infer rates of loss versus preservation as well as the time between the genome duplications. In their preferred models, they found rates of preservation to be the same order of magnitude as rates of gene loss. This is significantly greater than has been predicted by population genetics models of the loss of duplicate genes.

In Lynch and Conery (2000), the authors examined all known gene families in various fully (or mostly) sequenced organisms, subject to certain restrictions. With these gene families, the authors explored the relative rates of synonymous versus nonsynonymous substitutions, evaluated the possibility that gene duplicates tend to gain new functions, and estimated the rate of gene duplication. The latter was accomplished by examining the distribution of ages (i.e., number of substitutions) between pairs of duplicate genes. The authors attempted to exclude any suspected noncoding genes, as well as any genes from large multigene families (for this paper, any gene family with more than 5 members in an organism was considered large). Under the assumption that gene duplications occur at a constant rate overall, they estimated a gene loss rate of approximately half the genes lost per 0.053 substitutions/site for recently duplicated genes (i.e., genes with less than 0.25 expected substitutions per site). Using only very recent duplications in complete genomes, the authors also estimated the rate of gene duplication at 0.0023 duplicates per gene per 10^6 years in *D. melanogaster*, 0.0083 in *S. cerevisiae*, and 0.0208 in *C. elegans*.

Chapter 3

THE MODEL

This chapter will describe a model of gene family evolution based on the simple birth-death process. With this model, I will show how to calculate the probability of a particular gene family tree in a given species tree for values of the duplication and loss rates λ and μ. I will also discuss the assumptions made in the model.

3.1 Likelihood under the Model

Let the probability of the DNA sequences for a particular gene phylogeny G be $P(\text{Sequences}|G)$. "Sequences" will be used to refer not only to the list of bases at each site in the sequence, but also to the species in which each sequence was found. Let the probability of the gene phylogeny given model parameters λ (the rate of births, representing gene duplications) and μ (the rate of losses) be $P(G|\lambda, \mu)$. Then the likelihood of λ and μ over all possible G is

$$P(\text{Sequences}|\lambda, \mu) = \sum_{G} P(G|\lambda, \mu) P(\text{Sequences}|G) \tag{3.1}$$

This formulation divides the problem into three solvable parts.

$P(\text{Sequences}|G)$ is the likelihood of a particular gene phylogeny for the data set. This problem has been solved by a number of researchers beginning with Neyman (1971) with development for practical application to models of DNA sequence evolution in Felsenstein (1981b).

$P(G|\lambda, \mu)$ can be calculated to close approximation using the methods described below in section 3.4.

The summation over all possible gene phylogenies is essentially impossible for large gene families. However, this summation can be approximated using Markov chain Monte Carlo (MCMC) importance sampling. As the MCMC sampling time increases, the distribution of the sampled trees will approach proportionality to $P(G|\lambda,\mu)P(\text{Sequences}|G)$. This method will be described in detail in section 4.2 (page 51).

3.2 Doomed Lineages

Calculation of the probability of the data on a gene family tree is complicated by the possibility that some births may be obscured due to later deaths in the resulting genes. For example, if a gene is duplicated and one of the two resulting lineages is immediately lost, then there is little direct evidence indicating that either the duplication or the deletion occurred. (In fact there is the potential for some amount of evidence of this if the original gene is lost and the duplicate preserved, since that would mean all copies would necessarily coalesce by the time of the duplication event. However, my model does not incorporate population size or multiple copies at a locus.)

Though direct evidence for such events is largely lacking, it is nonetheless necessary to take these doomed genes into account. Otherwise, duplication events further in the past would be less likely to be counted than more recent duplications, as the resulting duplicates will have had more opportunity to be lost. Implicitly we would be assuming that the rate of births and deaths has been increasing toward the present. This can also be seen via graphs from the "reconstructed" birth death process, as described by Nee, May and Harvey (1994). In that paper, the distribution of existing lineages versus lineages which survive to the present is shown.

In this thesis, the calculation of $P(G|\lambda,\mu)$ is based on the bookkeeping of these doomed lineages with the gene phylogeny. By conditioning on the number of doomed lineages existing at each node N in the gene phylogeny, it is possible to peel the probability of the tree in the sense of Cannings, Thompson and Skolnik (1976) from the tips to the root. In other words, we can compute an array of likelihoods of the tree descended from N conditional on the number of doomed lineages present at that node. This array can be computed from the array of likelihoods at the immediately descendant node(s) in G. In this way, beginning with the tips, and proceeding to the base of the tree, the probability of G conditional on λ, μ, and the number of doomed lineages at the base of the tree can be determined.

A node N in a gene family tree G refers to any splitting in G corresponding with a divergence in the species phylogeny, as well as any duplication events. Loss events are handled implicitly, and do not result in nodes in G. The probability of the data on the tree will be calculated without the explicit calculation of the probability of all possible sets of events among doomed lineages. Instead, I will only consider the number of doomed lineages at each N in G, and use the transition probabilities of the birth death process to account for all scenarios for a particular number of doomed lineages at each N. This will be shown and explained below in section 3.4.

3.3 Description of the Model

The duplication and loss of loci is modelled as a simple birth-death process of gene family loci in each internode of a known phylogeny. That is, each existing gene will duplicate probabilistically with waiting time drawn from an exponential distribution with rate λ, and will be lost with a waiting time drawn from an exponential distribution with rate μ. In other words, the chance of a particular gene family member duplicating in time δt is $\lambda \delta t$, and the chance of the locus being lost in that time is $\mu\delta t$. (This assumes δt is sufficiently small that multiple events do not occur.) This results in a continuous time random walk in the number of gene family members. The rate of increase and decrease are proportional to the current number of genes. λ and μ represent the rate of formation (λ) and loss (μ) of loci, and are assumed to be constant throughout the tree. At each speciation event, all loci in the ancestral species are inherited by both resulting species. The genes at each locus are assumed to have a common ancestor at exactly the time of the speciation event.

As we are placing these events within a species tree, t in this model will not imply time, but rather a branch length on the species phylogeny. Thus, in a species with λ of 0.1, we expect an average of one duplication on a single lineage in a species branch of length 10.

The species phylogeny need not be clocklike, but for the purpose of the present likelihood computation is assumed to be known both in its pattern of branching and its branch lengths, with a known base. This base represents the time at which there are no doomed lineages that continue to the first node in the gene phylogeny. The gene phylogeny is assumed to be clocklike within internodes. That is, the rate of evolution of each member of the gene family in a particular species tree internode is assumed to be the same.

Genes are assumed to be related by duplication or speciation events, but not by gene conversion or lateral transfer between species. Duplication and loss events are assumed to involve exactly one whole locus. Such events could affect surrounding DNA, as well, so long as the DNA does not belong to another member of the same gene family.

There are a number of assumptions inherent to this model. The duplication and loss rates per locus are assumed to be the same throughout the history of the gene family, regardless of the number of gene family members in that species or of their chromosomal locations. That a speciation event gives rise to loci in both resulting species with a common ancestor at the time of speciation ignores the time to coalescence in the ancestral species. The phylogeny of the species is assumed to be known. These assumptions of the model will be examined in greater detail below in section 3.6.

3.4 Calculation of Probabilities

With this model, probabilities within an internode follow as per known properties of the birth-death process. These were found by Kendall (1948). Here I will be using the formulation employed by Bailey (1964). Let us first consider calculations within a single internode of a phylogeny, with a single lineage at its start.

From the probability generating function of the simple birth-death process, using two quantities, α and β, which will be defined below, the probability of a single gene dying out over time t (including the possibility that it has multiple descendants, all of which die out) is

$$p_{1\to 0}(t) = \alpha \tag{3.2}$$

and the probability of n genes ($n > 0$) resulting from 1 gene at the bottom of the internode after time t is

$$p_{1\to n}(t) = (1-\alpha)(1-\beta)\beta^{n-1} \tag{3.3}$$

where

$$\alpha = \mu \frac{e^{(\lambda-\mu)t} - 1}{\lambda e^{(\lambda-\mu)t} - \mu} \tag{3.4}$$

and

$$\beta = \lambda \frac{e^{(\lambda-\mu)t} - 1}{\lambda e^{(\lambda-\mu)t} - \mu} \tag{3.5}$$

When starting with a genes rather than a single gene, we have

$$p_{a\to 0}(t) = \alpha^a \tag{3.6}$$

and the probability of n genes resulting from a genes at the base of the internode after time t is

$$p_{a\to n}(t) = \sum_{j=0}^{\min(a,n)} \binom{a}{j}\binom{a+n-j-1}{a-1}\alpha^{a-j}\beta^{n-j}(1-\alpha-\beta)^j \tag{3.7}$$

That concludes the results here from Kendall (1948) and Bailey (1964). Let us now consider a gene phylogeny G for which the topology and branch lengths are known for all lineages that survive until the present. If we were able to determine a gene phylogeny from infinitely long DNA sequences, then the probability of the data for any trees with incorrect topology or branch lengths would approach 0. Thus the case of a known topology and branch lengths corresponds with the hypothetical situation if we had an infinite amount of DNA data at each sequence.

The calculations can be made most easily by beginning with the tips of the gene phylogeny, then calculating the probabilities for internal nodes for which the probabilities at the immediately descendant node(s) have already been calculated. This is continued backward in time on G until the base of S.

It is possible to calculate the likelihood for the pattern of gene duplication in the period between a duplication node N and the event at the immediately preceding node N' as follows. Let there be s surviving loci and an unknown number i doomed loci at N'. Let there be d doomed loci at N. Let $L_N(d)$ be the likelihood of the tree descendant from node N conditional on the existence of exactly d doomed loci at N. (See figure 3.1 for a visual description of the terms.)

We first calculate the probability of the internode between N and N', the immediately preceding duplication node. Within this internode, the final number of surviving genes is the same as the initial number. We then multiply

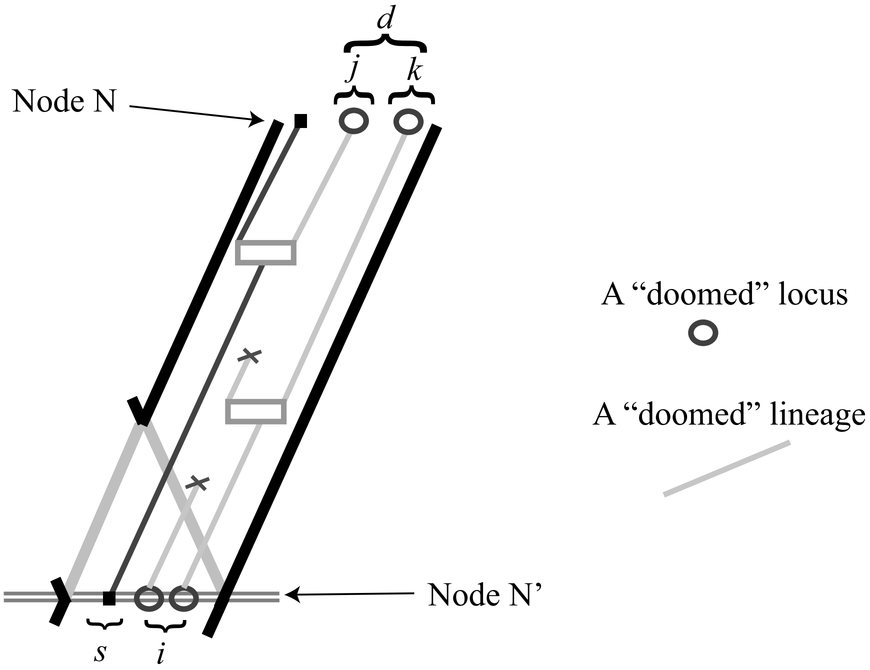

Figure 3.1: Terms used in the calculation of $L_{N'}(i)$: s is the number of surviving loci in the internode from N' to N. i is the number of doomed loci at N', while d is the number of doomed loci at N. j is the number of doomed loci at N which descend from surviving loci, and k is the number of doomed loci at N which descend from doomed loci at N'.

by the point probability density of the duplication event at N. For the simple birth-death process this is the duplication rate λ times the number of surviving loci s.

For an internode not ending in a tip, the probability of i doomed lineages at the start of the period conditional on d doomed lineages at the end is

$$L_{N'}(i) = \lambda s \sum_{j=0}^{\infty} \binom{s+j-1}{s-1} (1-\alpha)^s (1-\beta)^s \beta^j \sum_{k=0}^{\infty} p_{i\to k}(t) L_N(j+k) \tag{3.8}$$

The outer summation adds over the new doomed genes which arise from the surviving genes. The terms are derived from the probability $p_{1\to n}(t)$ (in equation 3.3), adding over all possible ways in which exactly j new lineages can come from s original lineages which themselves all survive until node N. Note that we cannot use equation 3.7 for this part of the calculation because no surviving lineage may be lost.

The inner summation adds over the number of genes which arise from doomed genes at N', which follows from the simple birth death-process (equation 3.7). The term $j+k$ is d, the total number of doomed genes at the end of the period. We do not know d, but as we are proceeding from the tips to the roots, we have already calculated the probability of the tree above this internode conditional on different values of d. Thus, we can sum over the possible values of d at the upper node, weighting each according to $L_{N'}(d)$, the probability of the tree above N' conditional on having d doomed lineages at that node.

Because of the two summations to infinity, this equation cannot be precisely calculated. However, as j or k grow larger than the number of surviving genes found in any species, the probability contributed by the resulting trees becomes vanishingly small. Therefore it is possible to cap the summations. The effect of these caps is examined in subsection 5.3.1 (page 87).

If the upper node is a tip, then there will be no doomed lineages at that tip. That is, $L_{\text{Tip}}(0) = 1$, and for all $x > 0$, $L_{\text{Tip}}(x) = 0$. This follows as any lineage which exists at the time of sampling cannot have been lost by the time of sampling. (However, in section 3.5, doomed lineages at a tip will be allowed as a way of representing unsampled loci.) In addition, as there is no duplication event at the tip, the probability of a duplication at the end of the internode should not be incorportated into the calculation. So, if the internode begins with a duplication and ends with a tip, the probability of the tree at or above the node N conditional on the number of doomed genes i simplifies from equation 3.8 to

$$L_N(i) = (1-\alpha)^s(1-\beta)^s\alpha^i \tag{3.9}$$

This is the probability that all s surviving lineages do survive (and do not give rise to any additional surviving lineages) times the probability that all i doomed lineages are lost.

Let us now consider the case of a speciation node in a gene phylogeny and an immediately preceding node. Here the calculations are very similar. However, a speciation event may result in a surviving gene in one species but a doomed gene in another. This allows for an additional possibility — a doomed gene which must exist at an interior node. For example, in the gene family tree in figure 1.2, page 7, there are genes sampled in the second and third species, but not in the first. This implies that there must have been a loss event at some point along the species tree internode leading to the first species.

Let i^* = the number of known doomed lineages in this descendant at the speciation event. Then

$$L_{N'}(i) = \sum_{j=0}^{\infty} \binom{s+j-1}{s-1} (1-\alpha)^s (1-\beta)^s \beta^j \sum_{k=0}^{\infty} p_{i+i^* \to k}(t) L_N(j+k) \tag{3.10}$$

The $L_{N'}(i)$ terms from the two descendants at a speciation event can be combined simply by multiplying the pairs for corresponding values of i.

These computations can be continued until the base of the tree. In my computer programs, the base is assumed to have no doomed genes. Thus the term $L_{\text{Base}}(0)$ at the base is the probability of the tree as a whole. This is substantially different from the method of Arvestad et al. (2004), for which the gene family tree is assumed to be rooted at the root of the species tree.

Note that if a gene phylogeny specified its duplication events only to a particular internode of the species phylogeny rather than with branch lengths, then this calculation could be made in fewer steps. Rather than calculate each period between tips, duplication, and speciation nodes, it would only be necessary to find the probabilities between tips and speciation events. Without the specification of branch lengths, however, such a treatment would lose any information on the duplication and loss rates which comes from the placement of the duplication nodes within internodes in the species tree.

3.5 Missing Members of the Gene Family

It is possible that a researcher might not have data from all existing members of a gene family. If we assume that, independently for each gene in the gene family, there is a probability b that the gene will be observed, then this can be included in the analysis. To do so, we need to change the calculation of conditional probabilities at the tips of the gene phylogeny. Specifically, as for the calculation of interior nodes:

$$L_{N'}(i) = \sum_{j=0}^{\infty} \binom{s+j-1}{s-1} (1-\alpha)^s (1-\beta)^s \beta^j \sum_{k=0}^{\infty} p_{i \to k}(t) L_{\mathrm{Tip}}(j+k) \tag{3.11}$$

Here, however, $j + k$ represents the number of unobserved genes, rather than doomed genes. $L_{\mathrm{Tip}}(j + k)$ then represents the probability of failing to observe $j + k$ genes. This can be calculated directly from b along with the observed number of genes g. That is,

$$L_{\mathrm{Tip}}(u \text{ unobserved genes}) = \binom{g+u}{g} b^g (1-b)^u \tag{3.12}$$

3.6 Assumptions of the Model

This model makes a number of assumptions about the process by which gene families evolve. This has been done both for computational tractability as well as conceptual simplicity. Since this is a probability model, the model can be extended so long as the probability of the events is properly accounted for. A simple model such as this may also act as a useful null model for other hypotheses. In the case of nested models, this would allow likelihood ratio tests to be used to test for rejection of the null hypothesis. In addition, it may turn out that this method is robust to certain violations of its assumptions. Robustness to some of these assumptions has been tested through simulations in section 5.3.

3.6.1 The simple birth-death process

The use of a simple birth-death process carries with it a number of implications. First, it is possible that the size of the population can become 0. Thus, according to the model, the gene family may die out in any or even all species in the tree. This raises the question of species and gene family ascertainment, which will be addressed below in subsection 3.7.4.

The assumption of a constant duplication and loss rate throughout the tree is suspect in some cases. In particular, though in this model all members of a gene family in a species can be lost by deletion, it is not plausible that a gene family necessary for survival would be eliminated by deletion. In such cases it would probably be more accurate to have a model in which the probability of deletion for the last member of the gene family in a species is 0. There is also evidence that some gene families such as those of MHC have undergone selection for a specific number of loci (see for example Takahata, Satta, and Klein 1992). Thus there are at least some gene families in which a simple birth-death model is not entirely appropriate.

There is evidence that rates of duplication vary depending on the chromosomal location of a gene. Specifically, genes near telomeres and centromeres tend to duplicate more quickly, as described for example in Horvath et al. (2001). This is not reflected in a simple birth-death process.

It is also possible that a single event could duplicate or delete more than one member of a gene family. These events clearly are not part of a simple birth-death process. This potentially introduces another way in which location of a gene could be important — Genes which are located close to one another will tend to be more likely to be lost or duplicated together. As a result, tandem duplications would be particularly likely to contradict the birth-death model in this case. In my model, such an event would need to be explained with multiple single events, exaggerating the frequency of such events, and neglecting larger scale occurrences. The duplication of an entire genome has been proposed as an explanation of some patterns of duplication, for example by Ohno (1970). This would likewise mislead inference using this model, though in this case proximity between the gene family members would not be important. This would also prevent the use of multiple gene families for estimation (as described below in subsection 3.7.3), as events in the different gene families would no longer be independent.

3.6.2 Same history for the entire gene

This model allows for duplication or loss of a gene, but does not permit duplication of part of a gene or gene conversion. Partial gene conversion could be expected to increase the similarity between two different genes in the same species, possibly resulting in the incorrect inference of a duplication event. Whole gene conversion could result in the replacement of one member of a gene family with another. This would be identical to a simultaneous birth and death in the gene family in a particular species. This would be a single event affecting two genes, and thus is not covered by the model.

3.6.3 Known phylogeny

The tree of the species which contains the gene phylogeny is assumed to be known throughout the analysis. The species tree must have a specified topology and branch lengths. This assumes there is sufficient information from other genes (or other sources of phylogenetic information) to perfectly determine the phylogeny of those species.

The effect of this assumption is unclear. Potentially one could resample the data used to estimate the phylogeny, and thereby represent the uncertainty of the phylogenetic estimate. However, the computational burden would be very high. A better approach would allow for different species phylogeny and gene phylogeny proposals using MCMC. (MCMC is described in more detail in section 4.2.) This would sample a variety of species trees as well as gene family trees. Such a method is beyond the scope of this thesis, but will be discussed in general terms in section 8.3, starting on page 110.

It is also assumed that we know the length of the base internode of the species tree. That is, we must specify a point at which we assume there is just a single member of the gene family — and no doomed genes with descendants which survive until the next node in the gene phylogeny. This cannot be done with certainty even with infinitely long sequences, and so represents a potential source of inaccuracy in the estimates. If we believe the birth-death process has been occurring far into the past, however, we can just choose a very large value for the length of the base internode. This will be discussed in more detail in subsection 3.7.2.

3.6.4 Molecular clock (or lack thereof)

This method does not assume that the species phylogeny or gene phylogeny follow a molecular clock. However, it does make the somewhat weaker assumption that all members of the gene family within a particular species internode evolve at the same rate. This assumption could run contrary to some recent models of gene family evolution, such as that of Walsh (1995), in which the authors argue that preserved duplicated genes will be expected to have a new selected function, and might therefore undergo a period of faster evolution soon after their duplication.

3.6.5 Lateral transfer

Lateral transfer could result in closely related genes in distantly related species. This might cause the incorrect inference of a high rate of gene loss as an explanation of the gene missing in any intermediate relatives.

3.6.6 Population size

The model in effect makes the assumption that the size of all species' populations is a single haploid genome. Birth and death events are considered to be instantaneous throughout a population. In a model which considers population size, the events would occur within a single genome in the population. This allows the possibility of multiple differing copies. Interestingly, in such a model the exemplar and new copy arising from a duplication would not be identical. Going forward in time, all copies derived from the new duplicate would necessarily coalesce (in the sense of Kingman 1982) at or after the time of duplication, while copies at the exemplar's locus might coalesce before the duplication event.

Another effect of this assumption is that coalescence between related gene family members after a speciation event is instantaneous. In a model with population sizes, the genes from the descendant populations would coalesce prior to speciation, with the time from coalescence to speciation distributed depending on the size of the population. It is even possible that the time of coalescence of one locus could be longer than the time of the duplication of another locus, resulting in paralogs more similar (and more closely related to the species tree) than orthologs from the same speciation event.

Population size might also have an impact on the rate of gene duplication and loss. In populations of varying size, this could result in birth and death rates which vary over time. Section 2.4 discusses a possible cause of this.

3.6.7 Model of DNA evolution

Lastly, this model inherits the assumptions of the phylogenetic likelihood inference it extends, as described in Felsenstein (1981b). Specifically, it models the DNA mutation process as Markovian, does not allow inversions or rearrangements, and assumes correct alignment of the DNA sequences.

3.7 Consideration of Specific Details

3.7.1 Relating the model to biological processes

The meaning of this model's birth (λ) and death (μ) rates can be defined in a number of ways, but it is important that the definition be used consistently. A birth represents an event which causes two gene family members to arise from one ancestral gene. As this could represent the duplication of a gene, these events are referred to as duplications throughout this thesis. A death represents an event which causes a gene to no longer be included in the gene phylogeny. This can represent the deletion of a gene, and is referred to in this thesis as a loss.

It is possible instead to apply a functional definition, such as the point at which the gene would be detected using a particular assay. In this case, a loss could represent a mutation which makes the gene no longer detectable with that assay. Caution is necessary, however, with the use of a functional definition. In particular, back mutation can result in the detection of a gene which was, by this definition, previously lost. But a simple birth-death model does not allow the possibility that a gene may be undeleted. Thus, the use of this model with an assay-based definition carries with it the assumption that no lost gene will later become detectable.

For example, if members of a gene family are identified using PCR, a base change within the primer region could result in a gene family member no longer being identified. If this single base mutation process can be modelled as occurring at a constant rate, then it not unreasonable to estimate a single rate throughout the gene family. As previously stated, however, this ignores the possibility of mutating back such that the gene can again be detected by PCR.

To take a second example — if genes are found using Southern hybridization, the situation is somewhat more complicated. In this case, multiple mutations will generally be necessary in order for a member of a gene family to no longer be identifiable. This suggests that the rate of loss would depend on the distance on the tree from the sequence used for hybridization. Thus the rate of loss through mutation to this assay would not be constant, and my simple birth-death model would be inappropriate.

3.7.2 Likelihood calculation at the root

In general, we will not know the point on the tree at which there is a single ancestral lineage. Thus the assumption that we will know the base of the species tree is a poor one. It would be preferable to make the much weaker assumption that all members of the gene family have a common ancestor at or before the earliest branching in the species tree.

A hidden difficulty lies here in the likelihood calculation at the base of the species tree. Specifically, the number of doomed lineages at this point is unknown. It is not correct to assume that there are no doomed loci at the base.

Any duplication event prior to this earliest node could give rise to a doomed locus, if one of the two resulting loci is lost after the base, but before the present. Though unobservable lineages at the base of the tree may seem irrelevant, this is not the case. Each doomed lineage must be lost prior to appearing at the tips of the tree. The probability of this occurring is determined by the parameters of the birth-death process, the topology and branch lengths of the species tree, and the time of the earliest duplication event before the root of the species tree. For example, if there are many doomed loci at the root, then intuitively it seems that higher loss rates will result in a higher likelihood, as all the doomed loci must be lost before the tips of the tree.

If we remove the explicit placement of the base of the species tree, we will instead be concerned with the number of doomed loci immediately prior to the root of the gene family tree. (This will also be immediately before the root of the species tree only if the earliest node in the gene family tree coincides with a speciation event.) At this node R there will be a single surviving lineage and an unknown number of doomed lineages. Using the methods described in 3.4, we will have an array of likelihoods $L_R(i)$ describing the probability of the tree starting at the root conditional on a number i doomed lineages at that node.

It is possible to attempt to address this using the Blurp and Quilg programs. By using a very long base internode on the species tree, we can attempt to approximate that there is an indefinite amount of time for new lineages to arise prior to the root of the species tree.

To exactly calculate the probability of each number i of doomed loci, we cannot just find the probability of $i+1$ lineages after an infinite period of time. For $\lambda > \mu$, the expectation for the number of lineages is infinite, and for $\mu > \lambda$, the expectation is 0. More importantly, the probability of any non 0 finite result is infinitesimal in both cases. (For $\lambda = \mu$, the expected number of lineages is 1, but the probability of 0 lineages counterintuitively approaches 1.)

However, if we condition on the survival of at least one lineage, then this becomes more tractable. This seems a sensible kind of conditioning, since we would not be examining the gene family if it had not survived. However, in the case of $\lambda \leq \mu$, it seems suspicious that we happen to be studying an infinitesimally probable event. If we are willing to ignore this, however, then for $\lambda < \mu$, the number of surviving lineages as $t \to \infty$ follows a geometric distribution with parameter λ/μ. This could be used to sum over the $L_N(i)$ results at the root of G, weighting in an appropriate manner. In the case of $\lambda \geq \mu$, every non-zero number of lineages approaches equal (infinitesimal) probability.

Really these results point to an oddity of the model as I have proposed it. It is not plausible that a simple birth-death process will describe a gene family over an infinite period of time. If such a model were correct, we would expect all the genes to die out, or for there to become an infinite number. Most likely there is some selection against too many or too few members of a gene family; I will discuss how such models could be handled in subsection 8.1.2 (page 104).

With a simple birth-death model, another promising approach is to calculate the probability of each number of doomed loci at the root of the gene family tree using the birth-death process considered backward in time. In Thompson (1975), followed by analysis by Rannala and Yang (1996) and Felsenstein (2004), the mathematics of this was worked out. The probability density of the birth-death process with n tips at the end and births at times t_i is

$$P(n, t_1, t_2, \ldots, t_{n-1} | \lambda, \mu, T) = p_{1\to 1}(T)\lambda^{n-1}(n-1)! \prod_{i=1}^{n-1} p_{1\to 1}(t_i) \tag{3.13}$$

Ideally we would want integrate this over possible times $t_1, t_2, \ldots, t+n-1$ and determine the probability for each n. However, this does not avoid the fundamental problem, since in each case we are multiplying by $p_{1\to 1}(T)$, which approaches either 0 or infinity as $T \to \infty$.

It is also interesting to consider the backward birth-death process conditional on the number of loci at the present, as shown by Felsenstein (2004). Because this conditions on a particular number of loci, the problems mentioned above do not arise. However, also because of this conditioning, it is not a useful result for computing the probability of possessing a certain number of loci at the end. Using symmetry arguments, counting the possible orders of duplication times, the probability density of duplication times t_i conditional on n final species after time T is (from Felsenstein 2004)

$$P(t_1, t_2, \ldots, t_n | n, \lambda, \mu, T) = \frac{\prod_{i=1}^{n-1} p_{1\to 1}(t_i) dt_i}{\int_0^T p_{1\to 1}(u) du} \tag{3.14}$$

Then choosing a time scale τ such that

$$\frac{d\tau}{dt} = p_{1\to 1}(t) \tag{3.15}$$

and integrating using equation 3.3 with $n = 1$, we get

$$\tau = \frac{e^{(\lambda-\mu)t} - 1}{\lambda e^{(\lambda-\mu)t} - \mu} \tag{3.16}$$

Though it does not provide the probabilities needed at the root of the gene family tree, as pointed out in Felsenstein (2004), this should lead to tests of fit to a birth-death model. This will be discussed more in section 8.4.1 (page 112).

3.7.3 Using multiple gene families

If we have data from multiple gene families and have reason to assume their duplication and loss rates are the same, then the data from those gene families can easily be combined. The likelihood of the duplication and loss rates is simply the product of their respective likelihoods for each gene family. However, as more loci are included in the analysis, it becomes more plausible that there may have been an event which duplicated or deleted two or more loci at once. As this is not incorporated in the model, it is a possible source of error.

Also, it does not generally seem reasonable to assume genes in every family have the same duplication and loss rates. However, following the logic used when integrating over possible evolutionary rates (see for example Felsenstein 2004 for a summary of rate variation methods among sites or Beerli and Felsenstein 1999 for rate variation among loci), then data from multiple gene families can still be combined. We assume a prior distribution such as a gamma from which the duplication and loss rates are drawn. If the gamma distribution is parameterized to have a mean of 1, then its density function can be described in terms of a "shape" parameter α as

$$f(r|\alpha) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\alpha r} \tag{3.17}$$

If there are h gene families, the shape parameter for the distribution of λ is α_λ and for the distribution of μ is α_μ, the likelihood of the parameters combined over all the gene families is

$$L = \prod_{i=1}^{h} \left[\int_0^\infty \int_0^\infty f(\lambda_i, \mu_i | \alpha_\lambda, \alpha_\mu) L_i(\lambda_i, \mu_i) d\lambda_i d\mu_i \right] \tag{3.18}$$

where $L_i(\lambda_i, \mu_i)$ is the probability of the data for gene family i when λ for that gene family is λ_i and μ is μ_i.

This would require the calculation of the likelihood for every pair of parameter values at each locus, which is not feasible. But in a manner analogous to the method of Yang (1994) for DNA sequence data, discrete approximations of the gamma distribution could be used to approximate the calculation of the integrals.

There is no reason to believe that duplication or loss rates vary among gene families according to a gamma distribution. Rather, the gamma family of distributions is a convenient arbitrary choice which varies between 0 and infinity and allows a wide variety of shapes according to the choice of α. It would be possible to use other families of distributions with similar properties, such as the lognormal distribution.

3.7.4 Ascertainment issues

In estimation of λ and μ, this model assumes the gene phylogeny is a random instance of a birth-death process in the phylogeny. If the researcher limits the gene families or species considered in a nonrandom way, inference of the parameters can be affected.

One particular form of conditioning is inevitable. No researcher is going to analyze a gene family in which all genes have been lost. The effect of this can be corrected with relative ease. The likelihood of λ and μ for a tree with a single doomed gene at the base and no surviving genes at the tips can be calculated for the phylogeny as described in section 3.4. The corrected probability of the data for the studied gene phylogeny is then

$$P^*(\text{Sequences}|G) = \frac{P(\text{Sequences}|G)}{1 - P(\text{No surviving genes}|\lambda, \mu)} \tag{3.19}$$

Note that the correction term $1 - P(\text{no surviving genes}|\lambda, \mu)$ depends only on the particular species phylogeny and the parameters, and does not need to be recalculated for each G.

As this situation is unavoidable, the correction has been implemented in all the analysis in this thesis. That is, it has been incorporated in Blurp, for calculation of probabilities of gene family trees in which the tree is known precisely, and in Quilg, for the similar calculation of probabilities when the gene family tree is unknown.

Other effects of ascertainment are may be easier to avoid, but are more difficult to deal with. Some are more a matter of psychology than necessity, and are much harder to model. For example, a researcher might choose to exclude from analysis any species without gene family members from the phylogeny, perhaps in a mistaken attempt to speed the computations. This would be ignoring a very good source of information about loss rates in particular, and would likely lead to an incorrectly low estimate of loss rates. The best solution to this would be reanalyze the data without removing those species.

A subtler problem might be that a researcher would not bother using gene family models if there is no more than a single gene family member in each species. It is possible to consider a model for this to correct for ascertainment. If there are n species under consideration, then there are 2^n possible kinds of trees (as defined by the number of gene family members at their tips) which contain either 0 or 1 gene family members in every species. By summing over all these trees, we could then correct as in equation 3.19. A problem with this situation is that it would be difficult in many cases for the researcher to judge if such a tree would have been excluded from analysis as a gene family.

Other specific kinds of species tree ascertainment could be similarly corrected, by summing over all trees which would be excluded from analysis. However, this could result in a great many trees which would be excluded, and the calculation of the ascertainment correction could therefore become very computationally difficult.

Chapter 4

COMPUTATIONAL METHODS

In this chapter, I will describe the manner in which the model in chapter 3 was implemented. Section 4.1 covers the implementation of Blurp, a program which calculates the probability of a particular gene family tree with known topology and branch lengths. Section 4.2 describes the methods used in the program Quilg, which estimates the likelihood surface for μ and λ when the topology and branch lengths of the gene family tree are unknown. Using data from DNA sequences, Markov chain Monte Carlo (MCMC) is used to sample possible gene family trees proportional to their likelihood. The methods for MCMC among gene family trees are nontrivial. Though computational methods may not initially seem as interesting as the mathematics used to describe the model, they are nonetheless essential for implementing and understanding the analysis.

These two approaches – of a known gene family tree and of sampling among gene family trees according to MCMC – are complimentary. Analysis of unknown gene family trees is useful, as this is the situation which will be faced by researchers examining real gene families. The use of known trees, on the other hand, allows a great many scenarios to be analyzed very quickly. (See chapter 5 for many examples of this.) It also provides a potential point of comparison with the use of unknown gene family trees and DNA data. Results from known gene family trees provide an ideal situation, showing the best possible results which can be expected. By comparing results from a known gene family tree with sampling among unknown gene family trees, we can also see the importance of our lack of precise knowledge about the gene family.

Results using the methods described in this chapter will be described and discussed in chapters 5 and 6.

4.1 Known Gene Phylogeny Case

Though it is possible to use MCMC to allow for uncertainty about the gene phylogeny, analysis with a known gene phylogeny is much faster. This allows a great many hypotheses to be examined quickly. For example, this makes it possible to consider the effect of many different factors, such as variability of estimation between different gene family trees, or of departure from the assumptions of the model. In addition, this analysis describes the most information which can possibly be obtained from the gene family tree, providing a comparison with the result with finite DNA sequences, which will always allow some uncertainty in the gene family tree. It might at first seem that precise knowledge of the gene phylogeny would give precise knowledge of the parameters of the process which produced the tree. This, however, is not correct, as each such tree contains only a finite number of duplication and loss events, and therefore can only give an estimate of the parameters which produced it. However, an infinitely large, perfectly known tree — or an infinitely large number of gene families, all with the same duplication and loss parameters, would be sufficient to determine the parameters exactly.

To conduct these tests, I wrote a computer program called Blurp, which simulates and analyzes perfectly known gene family trees, implementing the formulas for $P(G|\lambda, \mu)$ detailed in the model description in this dissertation (section 3.4, page 31). It also includes the ascertainment correction for the lack of observed gene phylogenies in which all genes are lost before the time of sampling (described in subsection 3.7.4).

More exactly, Blurp is given a user specified species tree along with the parameters λ and μ, a random number seed, and a specified number of gene phylogenies which are to be simulated. Using a simple birth-death process with the supplied parameters, and under the assumption that gene phylogeny branches will be the same length as those in the species tree, gene phylogenies are simulated within the species phylogeny. Gene phylogenies without any surviving members are discarded, and a new G is simulated in its place until the requested number of trees has been produced.

The program then calculates the sum of the log likelihoods of the resulting gene phylogenies for various combinations of λ and μ arranged in a grid. These are represented in a graph with contours plotted at the asymptotic 95% and 50% confidence intervals (using a χ^2 distribution with 2 degrees of freedom) from the determined maximum likelihood value. As the maximum likelihood estimate of the parameters is taken from the grid of tested parameter values, this will tend to be somewhat lower that the real maximum likelihood value.

As will be described in chapter 5, a number of variations were made in the simulation of gene phylogenies in order to assess the kind of information gained from different gene phylogenies, robustness to a variety of assumptions, and the effects of specific kinds of ascertainment.

4.2 DNA Sequence (Unknown Tree) Case

The Blurp program allows assessment of the ability to determine the likelihood surface of λ and μ when the gene phylogeny is known exactly. With real sequence data, however, there will inevitably be some uncertainty about the branch lengths and topology of the gene phylogeny. In addition, it is likely that the branch lengths of an inferred gene phylogeny — even one without duplication events — will not be an exact match to those of the phylogeny.

I will begin by restating formula 3.1 from page 27. Let G represent the set of all gene phylogenies which are consistent with the phylogeny. Then

$$P(\text{Sequences}|\lambda,\mu) = \sum_{G} P(G|\lambda,\mu)P(\text{Sequences}|G)$$

For a particular set of DNA sequences, there will be a great many different gene family trees G which can explain the data (though some will explain the data better than others). As a result, to examine the effect of finite sequence length on the inference of parameters of gene family evolution, it is necessary to sum over all G. The set of possible gene phylogeny topologies can be very large for phylogenies with many species and/or many genes. Even where it is possible to enumerate over gene phylogeny topologies, differing possible branch lengths make the calculation of $P(\text{Sequences}|G)$ for all Gs very difficult on such a tree.

The large set of tree topologies along with the many possible estimates of branch lengths likewise makes simple Monte Carlo sampling from $P(G|\lambda,\mu)$ infeasible. This is because the vast majority of trees have extremely low likelihoods, even if they are consistent with the set of DNA sequences. For example, in the case of the hemoglobin genes, a tree in which human α- and β-hemoglobin genes have a common ancestor 1,000 years ago would have an extremely low likelihood, as the sequences of the two genes are only distantly related.

Therefore it is necessary to use an importance sampling method to approximate this sum, choosing most often those gene phylogenies that provide a high likelihood. I have done this using Markov chain Monte Carlo (MCMC) with the Metropolis Hastings sampling algorithm, roughly according to the method described in Kuhner, Yamato and Felsenstein (1995) for estimation of effective population size. Here, however, we are using the method to sum over gene phylogenies, rather than over coalescent trees.

MCMC begins with an initial gene family G and proceeds by updating G with possible changes. These updates are accepted or rejected according to the method of Metropolis et al. (1953) and Hastings (1970). The sampling has been chosen to be proportional to the posterior probability of G

$$P(G|\text{Sequences}, \lambda_0, \mu_0) = P(\text{Sequences}|G)P(G|\lambda_0, \mu_0)/P(\text{Sequences}|\lambda_0, \mu_0) \tag{4.1}$$

where λ_0 and μ_0 are a specific pair of values of the parameters. It is not known how to calculate $P(\text{Sequences}|\lambda_0, \mu_0)$. However, this quantity is a constant for a particular λ_0 and μ_0. In the Metropolis Hastings method, only the ratio of these posterior probabilities will be used, and therefore its calculation is not needed.

This results in a Markov chain with gene phylogenies as states. The stationary probabilities of this Markov chain are

$$P(\text{Sequences}|G)P(G|\lambda, \mu)/P(\text{Sequences}|\lambda, \mu) \tag{4.2}$$

Following Hastings (1970), this can be achieved by accepting a proposed new G' with probability $\min(1, r)$, where r is given by

$$r = \frac{P(\text{Sequences}|G')P(G'|\lambda, \mu)}{P(\text{Sequences}|G)P(G|\lambda, \mu)} \frac{Q(G', G)}{Q(G, G')} \tag{4.3}$$

$Q(G, G')$ is the probability of proposing G' when the current tree is G, and $Q(G', G)$ is the converse — the probability of proposing G if the current tree were G'. The ratio of these two terms is known as the “Hastings ratio.” When the proposed tree is rejected, we retain the original G rather than replace it with G'.

$P(\text{Sequences}|G)$ is calculated as for the determination of likelihood for a particular phylogeny as described in Felsenstein (1981b). In my program I have used a Kimura two parameter model (Kimura 1980) of DNA evolution.

However, any other model of DNA evolution with calculable transition probabilities could have been used, instead.

Note that this method will accept any tree for which r in equation 4.3 is greater than 1. Even if the product is less than one, there is always some chance of accepting the new hypothesis so long as it is not completely inconsistent with the data. In this way, the space is explored with a tendency toward regions with higher $P(\text{Sequences}|G)P(G|\lambda,\mu)$, but even areas with lower likelihoods will be examined, though less often.

The method of choice of starting G and proposals G' will be described later in sections 4.5 and 4.6.

Note that this result is for a particular pair of values λ_0 and μ_0 of the parameters of the birth-death process. To find the relative likelihood for other parameter values, we have

$$\frac{L(\lambda,\mu)}{L(\lambda_0,\mu_0)} = \frac{1}{n}\sum_{G}\frac{P(G|\lambda,\mu)}{P(G|\lambda_0,\mu_0)} \tag{4.4}$$

By sampling from the posterior probability $P(G|D,\lambda_0,\mu_0)$, we tend to take G which have a greater contribution to the likelihood at that value of λ and μ. This will be most efficient when λ_0 and μ_0 are closer to the true λ and μ. To achieve this, my implementation uses multiple chains of G. After each chain, λ and μ are estimated from the most recent chain. The newly estimated λ and μ are then used as λ_0 and μ_0 for the following chain.

This only uses information from the most recent chain to estimate λ and μ. Ideally we could combine the information from trees in all the chains which have been run, and use that to estimate the parameters. Geyer (1991) showed a way in which this can be performed. However, due to its computational difficulty, I have not implemented this method.

4.3 Program Design

To implement this method for unknown gene family trees, I have written a program called Quilg. Quilg performs the following tasks:

- Simulation of gene family trees. As with the simpler Blurp program, the user specifies the species phylogeny (including base and branch lengths) as well as λ and μ. This as well as the next item (DNA simulation) are used for the simulation studies described later in chapter 6, and can be replaced by user supplied DNA sequences.

- Simulation of DNA data on a gene family tree. This is done for a user designated number of bases for the gene family members. The sequences are simulated according to the Kimura two parameter model (Kimura 1980) of DNA evolution (which specifies an overall rate of change and the ratio of DNA transitions and transversions) with user supplied parameters.

- Determination of an initial hypothesized gene family tree (described in section 4.5)

- Alteration of the current hypothesized tree into a new hypothesized tree (Section 4.6)

- Determination of the probability of a hypothesized tree under given values of λ and μ (described in section 3.4) and the likelihood of the gene family tree for the DNA sequences as described in Felsenstein (1981b).

- Determination of acceptance or rejection of new hypothesized trees (section 4.7)

- Storage of a subset of the hypothesized trees.

- Maximization of λ and μ for a particular set of the stored accepted hypothesized trees.

- Output of maximum likelihood values of λ and μ for the last set of stored accepted hypothesized trees along with a likelihood surface for the two parameters.

4.4 Maximum Likelihood Estimates and Chain Control

Though the sampling of gene phylogenies has been made from the posterior probability $P(G|\text{Sequences}, \lambda_0, \mu_0)$, we are interested in the probability of the data on sampled trees at other values of the parameters. This is shown above in equation 4.4. In addition, for computational reasons, it is not feasible to keep records of every tree from the MCMC. As a compromise, Quilg stores the tree after every tenth proposal.

This set of stored trees is used within Quilg to compute the maximum likelihood estimate of the parameters. This has been implemented with a simple hill-climbing method in which λ' and μ' are changed in alternating steps. λ_0 and μ_0 are used as the initial values. New parameter values are accepted when they result in a higher likelihood. When a parameter change results in a lower likelihood, the change is rejected, and another change to the same parameter is proposed by alternating between a reduction in the step size and a change in sign. In this way, new values for λ' are proposed until a maximum likelihood for that value of λ' (with a fixed value for μ') is found. Then, in a similar way, μ' is optimized to maximize the likelihood for that λ'. These steps are alternated until the likelihood cannot be improved by changes in either λ' or μ'. When the step size for both parameters becomes smaller than a predetermined constant, the resulting estimates are reported. This is not an especially efficient maximization method, and is not guaranteed to find a local maximum in the case of a narrow diagonal ridge of the likelihood surface for λ and μ. However, with these likelihood surfaces, this simpler method was able to find the maximum when more sophisticated maximization methods which depend on derivatives of the likelihood surface (such as the Newton-Raphson method) appeared to have difficulties.

The sampling of trees will be most efficient when λ_0 and μ_0 are close to the true λ and μ. To achieve this, Quilg runs multiple chains of gene family trees. After each chain, the maximum likelihood estimates of λ and μ are calculated. These estimates are then used as the new λ_0 and μ_0 in the following chain.

One special case occurs when the maximum likelihood values of λ or μ are found to be too close to 0. In this case, I artificially force the values to be high enough that we would expect one duplication and one loss event on a tree with total branch lengths equal to that of the most recently accepted tree. This is necessary because runs with $\lambda_0 = 0$ or $\mu_0 = 0$ will not accept any events of the corresponding type. This is a case of "fatal attraction" of the type pointed out by Kuhner, Yamato, and Felsenstein (2000).

In practice, I have used chains totalling 30,000 proposed gene phylogenies, using 10 chains each with 1000 proposed gene phylogenies followed by 2 chains of 10,000 proposed gene phylogenies. All estimates of the likelihood surface for output are made using only the final 10,000 tree chain. For all runs, only one tree in 10 is stored for later evaluation of parameter likelihoods. For the final run of 10,000 proposed trees, we are interested not only in the maximum likelihood estimates of the parameters, but also on the surface itself. In this case, Quilg finds the likelihood across a fine (logarithmic) grid of values for λ and μ, and reports the likelihood of each. As before, but beginning maximization with the maximum likelihood values of the parameters found among parameter pairs on the grid, the maximum likelihood parameter estimate is made.

Unlike for previous runs, the estimates are not adjusted upward even if they represent fewer than one duplication or loss per tree.

4.5 Choice of Starting Tree and Parameters

To begin the MCMC over gene phylogenies, it is necessary to use a starting tree and initial parameter values λ_0 and μ_0. So long as the sampling is run for sufficiently long, it will not matter which are chosen, so long as they have a finite likelihood. However, a starting tree and parameters which correspond well with the data will tend to take fewer steps to find other trees which contribute more to the likelihood.

To select a tree with which to begin the sampling, I have chosen a fairly arbitrary method which results in trees that correspond reasonably well with the DNA sequences of the genes. Specifically, the genes in the gene family are each treated as separate species, and their tree found using the UPGMA method (Sokal and Rohlf 1962) using the Kimura two parameter model of DNA evolution. UPGMA uses pairwise differences to progressively join together similar sequences into a clocklike tree. The pairwise differences are used to calculate the branch lengths in the tree. This UPGMA tree will be referred to as g.

This g is then reconciled with the phylogeny of the species. This reconciliation is in the sense described by Goodman et al. (1979). That is, it is possible that the topology of the gene phylogeny will not correspond with the topology of the species phylogeny. However, the trees can be reconciled by hypothetical duplications. In this case we must reconcile the topologies while retaining branch lengths which correspond reasonably well to the differences between the DNA sequences.

My procedure for doing so works as follows.

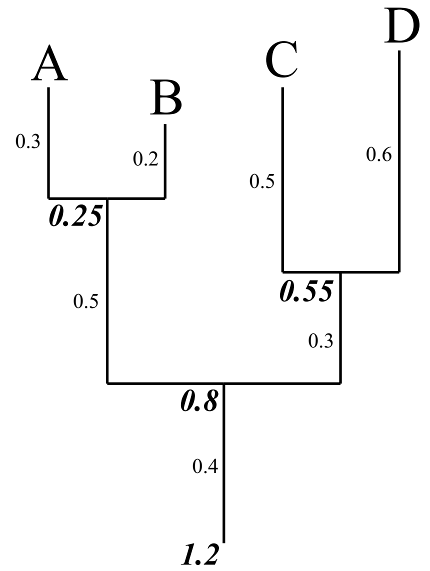

Figure 4.1: Calculation of average branch length to tips $t_s(M)$ on a species tree. Branch lengths are shown in small numbers to the left of each internode. Distances to tips $t_s(M)$ are shown in larger italic type just below and to the left of the corresponding nodes.

- Find the average branch length to tips — called $t_s(M)$ — for each node M in the species phylogeny. This is found recursively from the tips to the base, finding $t_s(M)$ for each node as the average of (t_s + branch length of internode) for the two nodes tipward from it. If the species tree is not clocklike, it is not impossible that M might have a lower t_s than a tipward node. For an example of the calculation of $t_s(M)$, see figure 4.1.

- Form a UPGMA phylogeny g of the genes, without regard to the species tree.

- Find the average branch length to tips — called $t_g(N)$ — for each node N in the UPGMA gene phylogeny. This is done as for the calculations of $t_s(M)$, above.

- Find the set of species R which contain one or more of the descendants of N. See figure 4.2.

- For each node N in g, find the most recent node M in the species phylogeny which is ancestral to all of R. This may also contain species not in R.

- If $t_g(N) < t_s(M)$, assign node N in the UPGMA gene phylogeny to the speciation event at node M in the species phylogeny. N may correspond with a duplication or a speciation event; See below for the determination of this.

- Otherwise, find the branch in the species phylogeny which corresponds with N. Beginning at node M, proceed down the species tree for branch length $t_g(N) - t_s(M)$. Set N as a duplication event at the corresponding location in S. If this procedure leads below the base of S, put the duplication event just above the base.

It is possible that multiple nodes in the gene family tree will be assigned to exactly the same location on the species phylogeny. If one of the nodes is ancestral to another, it is placed toward the base of the tree, with a tiny arbitrary branch length between it and its descendant node. When neither is ancestral, there is no need to determine which is placed tipward. If multiple nodes in the gene family tree are placed at a speciation event, the most tipward node(s) are assumed to come from the speciation event. Any other events at the speciation correspond with duplications.

At a speciation event, it has been assumed by my model of gene family evolution that both resulting species will have all genes from the ancestral species. However, it is possible with the assignments described above that

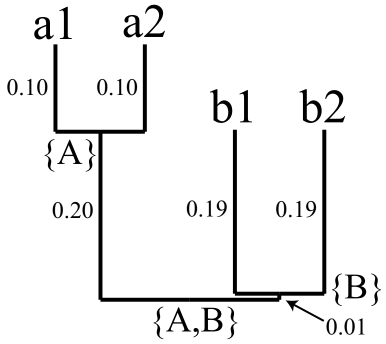

Figure 4.2: A tree g of four gene family members found in the four species shown in tree 4.1. Gene names consist of the species in which that gene was found followed by an identifying digit. Note that no genes were found in species C or D. Branch lengths are shown in small numbers to the left of each internode (where possible). The set of species R which contain one or more of the descendants of each node in g are shown in braces next to each node. In this example, the node in g labelled "{A,B}" corresponds with the speciation event at the more recent common ancestor of species A and B in figure 4.1.

a lineage in the gene phylogeny may exist at a speciation node but not have a descendant at the tips of all species which descend from that node. In the birth-death model, this must be caused by one or more losses. This is designated with a known doomed lineage. A starting tree with a known doomed lineage will not permit a loss rate of 0. However, as new trees are accepted as rearrangements (see section 4.6, below), these new trees may allow a loss rate of 0.

Depending on starting options, the user can provide the starting values λ_0 and μ_0, or they may be guessed by the program. In the latter case, this is done as follows. Once the starting tree has been found, the number of duplications in the starting tree are counted, and divided by the total branch length of the tree. This gives the initial duplication rate λ_0. The number of known gene losses is counted and divided by the the total branch length to give the initial loss rate μ_0. If either of these rates give an expectation of fewer than 1 event in the whole starting tree, the rate is adjusted upward to provide an expectation of 1 event. This method seems likely to provide an underestimate of the true rates (unless the true rates are very low), as it only includes observed events. However, it does remove the requirement that the user guess the rates.

Though this method makes use of concepts from Page (1994) in its initial assignment of gene family nodes to species nodes, it differs substantially in many details from his procedure. The most notable difference is that this method is not designed to minimize any quantity. Rather, it is intended to quickly provide a gene phylogeny which is consistent with the species phylogeny. It will hopefully find a gene phylogeny which is not too different in branch lengths from the UPGMA tree derived from the gene sequences. This is meant to give a tree with fairly high likelihood. However, it is not guaranteed to do so. If the tree does not have a high likelihood, the MCMC algorithm will search the space of trees, eventually finding trees with higher likelihood. Another notable difference is the specification of branch lengths; Page's method is concerned only with topology.

4.6 Tree Rearrangement

This section describes the proposal of new trees G' for use in the Metropolis Hastings updates to the Markov chain of trees. It also derives the probability of a particular proposed rearrangement $Q(G', G)$, allowing the Hastings ratio to be determined for a particular tree proposal. The general scheme is to take a node N, detach the subtree defined by that node from its parent node P, then reattach N anywhere up to two nodes ancestral to or a descendant from P. In other words, the method "slides" a subtree up or down at its current location, possibly past other nodes. This has some similarity to the rearrangement method of subtree pruning and regrafting (Swofford et al. 1996), but with substantial additional constraints due to the use of rooted trees and the restriction

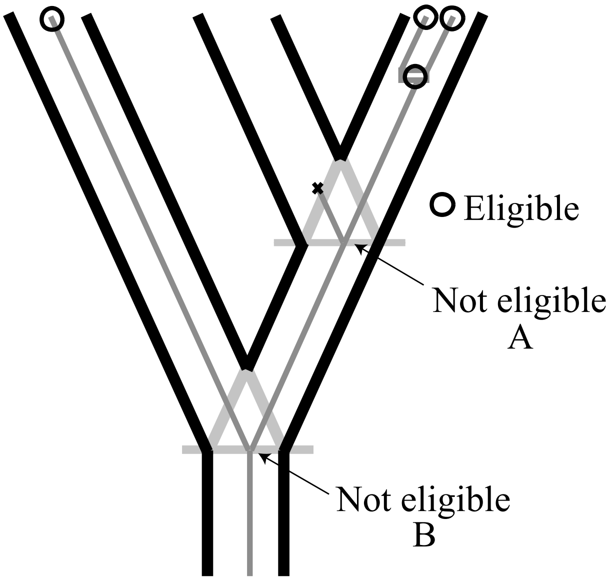

Figure 4.3: Nodes in G which can be chosen for rearrangement. Node A is ineligible because it is a speciation node which specifies a known doomed lineage. Node B is ineligible because it is one node above the base.

on allowable attachments introduced by the species tree, as well as limitation to relatively local rearrangements.

A node N in G is chosen at random, with equal probability given to each available node (See figure 4.3). This node may be a tip, a duplication event, or it may be due to a speciation event. However, speciation nodes which designate a known doomed lineage may not be chosen. Because of this limitation, the number of available nodes will remain the same in every G. In addition, the node connected to the base of G cannot be chosen. (Note, however, that a node connected to the *root* of the gene family tree is an allowable choice.) There will always be exactly $2n - 2$ available nodes, where n is the number of tips in G.

To produce G', the internode connecting N to its parent node P^* is erased. If P^* corresponds to a speciation event, the internode from P^* to N is replaced by a known doomed locus at P^*. If, as a result of this change, P^* now consists of a pair of known doomed loci, P^* is pruned back toward the root until it reaches a node at which at least one of the descendants is a surviving locus. This node is called P. Otherwise, P^* is used as P. See figure 4.4 for an example of pruning back from the original parent node, and figure 4.5 for an example in which the parent node is not pruned back. (The pruning step avoids many rearrangements which would result in duplications leading to a doomed lineage, which contribute nothing to the gene family tree. The choice of P can also be thought of in a different way: We are are only considering those nodes in G which lead to two surviving lineages.) If P is a duplication event, no additional changes are made at P. (This will result in a "duplication" node with just one descendant. This will affect the reattachment process, but will be removed after the rearrangement has been proposed.) N will be reattached either to a nearby ancestor of P, or to a nearby descendant, but never higher in the tree than the time of N, and not to a descendant of P where it is in a different species than N.

First we determine the number of possible reattachment locations for N. Reattachment is allowed anywhere from up to two nodes up from P to two nodes down from P, or at any intermediate point. However, this is constrained as follows:

- N can only reattach in its corresponding branch of the species tree, or in an ancestor of that branch. Otherwise, reattachment would cause a lateral transfer event.

- N cannot reattach to a node higher in the tree than N, or to an internode all of which is higher in the tree than N. This is because the ancestor cannot be more recent than the descendant.

- N cannot reattach at an existing duplication node, since duplications are assumed to give rise to only a single new gene family member.

- N can reattach at a node corresponding with a speciation event only if the node has a known doomed lineage in N's species. This is because the lineage must have an explicit ancestor in the gene family tree. Without a known doomed lineage to replace, it is unclear how N would be attached.

- N may not reattach at P. This is implemented only to reduce the necessary computations.

By travelling the tree two nodes upward and downward from P, following both loci when travelling upward through a duplication node, it is possible to make a list of all eligible reattachment nodes and internodes. The total number of these is designated v. One of v is then chosen at random with equal probability $1/v$. If an internode is chosen, the reattachment point is chosen uniformly between the bottom of the internode and the earlier of the top of the internode and the time of N.

This may seem to preclude the possibility that a node may be reattached above its current location, but that is not the case. If a child of N had been chosen for erasure rather than N, then the location of N could be changed upward.

The point of reattachment of N to the rest of G' will be referred to as P'. If P' is an existing speciation node, we remove the known doomed lineage at P'. Attachment in an internode results in a new duplication node at P'. When the internode from N to P' passes through a speciation event, a node is placed at each speciation. The descendant of the locus in the species which does not lead to N is marked as a new known doomed lineage.

The new tree G' formed by the reattachment is then examined for any unnecessary nodes. Specifically, any node in G' associated with a speciation event

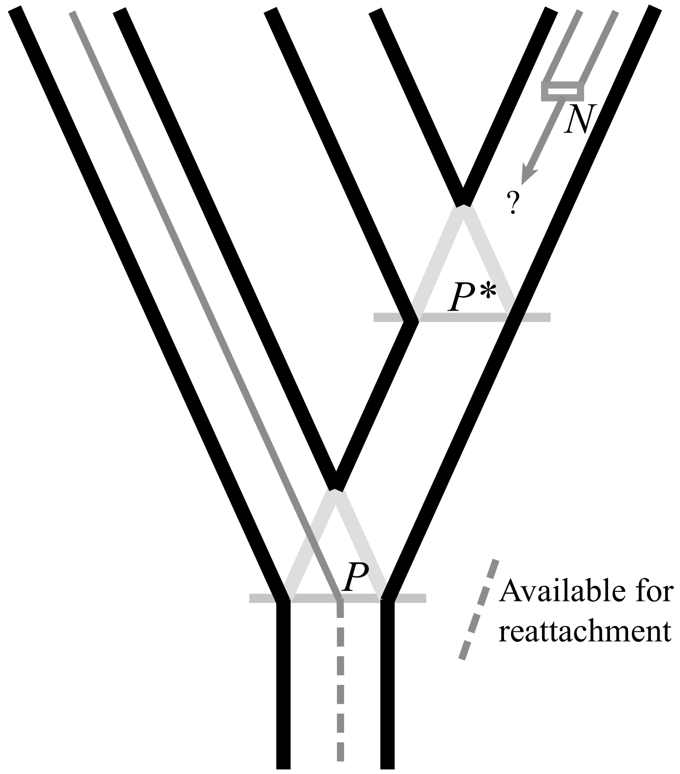

Figure 4.4: Reattachment, example 1: From the tree in figure 4.3, the duplication is selected as node N, and its parent node (the speciation on the right) is labelled here as P^*. After removing the internode to N, P^* leads only to doomed lineages, so it is pruned back until we reach the node P, which leads to another surviving lineage. The internodes going from N to P are removed, and N can be reattached up to two nodes away from P. In this tree, there is only one such lineage available — at the internode between the base and P. This will result in a duplication between the base and P, and known doomed lineages going to the left from the two speciation nodes.

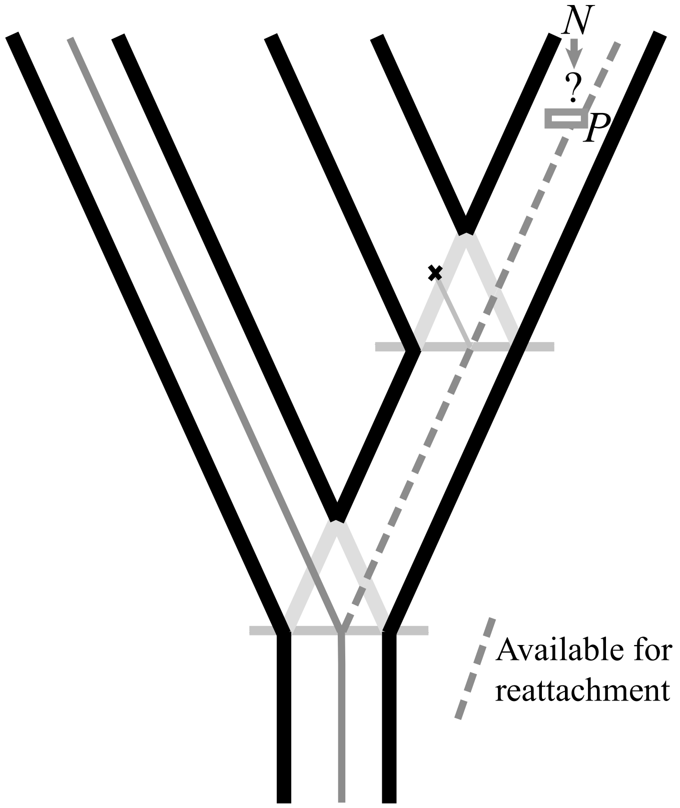

Figure 4.5: Reattachment, example 2: From the tree in figure 4.3, a tip is selected as node N, and its parent node (the duplication) is labelled here as P. N can be reattached two nodes away from P. Going upward from P there is one internode where N can reattach, and downward there are two. Thus, $v = 3$, and each of these internodes has a 1/3 chance of being chosen for reattachment. Though the speciation node below P does have a known doomed lineage, it is in the other child species, and therefore cannot be used for reattachment.

is removed if both descendants are now known doomed lineages. Any duplication node with just one descendant is removed, connecting its child and parent nodes. These steps are continued recursively until all such unnecessary nodes are removed.

Once G' has had these nodes removed, we then begin as for another rearrangement, this time choosing the same node N, rather than choosing a node at random. No proposal is made in this case. Rather, this is done to determine the number of rearrangements v' which are available from G' when detaching below N. This is used when computing the relative probability of a proposal of G' from G versus G from G', for use in calculating the Hastings ratio for the probability of acceptance of G'. When P is within an internode, v and v' for the other child node of P on G need to be calculated and added in, as well.

The choice to restrict the reattachment of N to two nodes above or below P is not entirely arbitrary. Two is the minimum node distance which can be used and guarantee that all possible trees can be achieved by rearrangements. By keeping possible rearrangements to a relatively small number, it is faster to count all allowable rearrangements. By keeping rearrangements fairly local, it is also more likely that the proposed rearrangement will not be rejected. However, there is potentially a drawback to this, as it is possible that two relatively likely gene family trees will be separated in the gene family tree Markov chain by trees with lower likelihoods, slowing the exploration of the space of likely trees.

4.7 Calculation of Acceptance Probabilities

As stated above in equation 4.3, each proposed change in the tree will be accepted with probability $\min(1, r)$, where r is determined from

$$r = \frac{P(D|G')P(G'|\lambda,\mu)}{P(D|G)P(G|\lambda,\mu)}\frac{Q(G',G)}{Q(G,G')}$$

The $P(D|G)$ and $P(D|G')$ terms are the likelihood of DNA data on a phylogeny, and can be calculated as described in Felsenstein (1981b). Calculation of $P(G|\lambda,\mu)$ and $P(G'|\lambda,\mu)$ has already been described in section 3.4 (starting on page 31). $Q(G,G')$ is the probability of proposing G' when the current tree is G, while $Q(G',G)$ is the probability of proposing G if the current tree were G'. Here, $Q(G,G')$ is the probability of choosing node N times the probability of choosing the particular point of reattachment. However, the probability of choosing a particular node is constant, and as it appears only in a ratio with an identical term, can be ignored. In the case of reattachment at a speciation node, the probability of reattachment is just $1/v$. When reattachment occurs within a branch of length l, the point probability density of that particular reattachment is δ_l/lv.

The reverse calculation $Q(G',G)$ can be made analogously, using $1/v'$ rather than $1/v$. When detaching at a speciation node and reattaching within a branch, we will have a probability density on just one side of the ratio of Qs. However, reattaching within a branch results in the introduction of a new duplication event. This duplication event also causes a probability density to be introduced into the term $P(G|\lambda,\mu)$ or $P(G'|\lambda,\mu)$ (see equation 3.8, page 33) on the opposite side of the ratio.

4.8 Verification of Results

As the steps performed by the program Quilg are nontrivial, it is important that they be verified. I have done so by testing in limited special cases in which the output can be measured against calculations or simpler simulations.

Quilg can accept DNA data for which the state at a particular base is unknown. From equation 4.2, the stationary probabilities of the Markov chain Monte Carlo are

$$P(\text{Sequences}|G)P(G|\lambda,\mu)/P(\text{Sequences}|\lambda,\mu)$$

When the DNA data are unknown, $P(\text{Sequences}|G)$ and $P(\text{Sequences}|\lambda,\mu)$ are both 1, and the stationary probabilities of the trees should approach proportionality with $P(G|\lambda,\mu)$. This quantity can be calculated for trees using the Blurp program. Thus by comparing the stationaries of Quilg with the calculations of $P(G|\lambda,\mu)$ from Blurp, I was able to verify that the rearrangement and acceptance schemes in Quilg were correct.

A less accurate but more general method of testing comes from simulation. I took a known species tree S and generated gene family trees G from S using known parameters λ and μ. The Gs were then used to simulate a specified length of DNA data at each tip using an evolution model (Kimura 2 parameter) with known parameters. This DNA data was then supplied to Quilg, which found the likelihood surface of λ and μ for each G. These estimates were then combined, and the result compared with the known supplied parameters. These results will be discussed below in chapter 6.

4.8.1 MCMC mixing

Although Markov chain Monte Carlo methods are guaranteed to eventually reach their stationary probabilities, this may not necessarily be achieved in a reasonable amount of time. This can be a particular problem if most proposed rearrangements are rejected, or if there are widely separated peaks in the likelihood surface. The former does not appear to be a problem with my rearrangement scheme, as (very roughly, and depending on the conditions of the run) 1/3 of rearrangements are accepted. The latter possibility, however, cannot be completely excluded.

To evaluate this potential problem, I ran chains until they appeared to reach a stable estimate of the parameters. However, this admittedly does not guarantee that mixing was sufficient. It is still possible that there is some unexplored area of the space of Gs which has a substantial contribution to the likelihood. Without a better understanding of the rearrangement method and its relationship with the tree space, it may not be possible to refute this with certainty.

In addition, I made small-scale checks of the effect of my method of choosing a starting tree and parameter values was resulting in a long period of "burn-in" before reaching convergence. To do this, I compared results with the algorithm's starting tree and parameters versus supplying the known correct tree and parameter values. Very inaccurate starting parameter values (off by a factor of 10) did require more starting chains before a good estimate of λ and μ were reached. The starting tree, however, appeared to have relatively little effect on the convergence time.

4.9 Computational Optimization

Due to the many steps required for good mixing, and the substantial calculations required to compute the probability of the data for each G as well as of the DNA sequences, Quilg is a slow program. However, it should be possible to substantially improve its current performance. Some speed improvement would likely come from more optimized organization of data structures and other algorithm implementation. Rewriting to allow chains to run on separate processors would likely speed up results quite a bit.

A promising theoretical improvement would be a heating scheme such as that described in Geyer and Thompson (1995), in which multiple chains are run at different "temperatures," speeding exploration through Gs with lower likelihoods. This would also work much faster if an implementation for multiple processors were available.

Chapter 5
ANALYSIS WITH KNOWN GENE PHYLOGENIES

5.1 Infinite DNA (Unknown Doomed Lineages)

It is computationally much easier to analyze cases where the gene family tree is known exactly than when only know the DNA sequences are known, and many different gene family trees are possible. Exact knowledge of the gene phylogeny corresponds to the hypothetical existence of infinitely long DNA sequences for each member of the gene family. Such infinitely long sequences would allow the gene tree to be known precisely in both topology and branch lengths. This can be thought of as a special case of equation 3.1 (page 27), in which all the likelihood comes from a single G, as all other Gs are in conflict with the data. In all the examples in this section, gene family trees will be simulated and then taken as known. Using these trees, the properties of the likelihood surface of λ and μ will be examined in a variety of cases.

Given a known λ and μ and species tree S, simulation of a gene family tree G^* within a species tree is straightforward. Starting at the base of the species tree (potentially prior to its root), and continuing to the tips, the waiting time to the next event is drawn from the exponential distribution $\frac{1}{l(\lambda+\mu)}e^{-l(\lambda+\mu)t}$, where l is the number of lineages. If the time drawn is less than the time to the next speciation event, then the type of event (duplication or loss) is determined and placed on G^* on a randomly chosen lineage in that internode, after which simulation continues within that same species internode. If the time drawn is instead past the next speciation event, no duplication or loss occurs. Instead, the genes are copied into each of the daughter species, and simulation continues anew in each of those daughters.

After this preliminary G^* is simulated, the observable G is determined by retaining only those lineages which survive until the tips, as well as all ancestors of those surviving lineages. All other lineages are trimmed off. Note that it is possible in a birth-death process for no lineages to survive until the tips. When this occurs, the tree is discarded and a new one is simulated. This is accounted for in the calculation of the probabilities using the method described in subsection 3.7.4 (page 47). With the resulting G, the likelihood of λ and μ parameter pairs is evaluated on a grid. By marking the maximum likelihood $\{\lambda,\ \mu\}$ point and showing the approximate contours of the likelihood surface, this allows the general shape of the likelihood surface to be examined.

It should be remembered throughout this section that the ability to know G exactly will have an effect on the results. In general, exact knowledge of the gene family will provide more information on birth and death rates than would be available with real data. Thus, their estimates will be more exact than they would realistically be.

5.2 Properties of the Model

5.2.1 Number of gene families

It is important to have an understanding of the amount of information available in a single gene family tree to infer rates of duplication and loss. In this subsection, the likelihood surface for single gene family trees will be shown. By repeating this for a number of different single trees, it is possible to visually examine the variability of estimates made with a single gene family tree.

The first examples will use the four tip species tree in figure 5.1, with arbitrary letters used as the species labels. Note that I have specified a branch length below the root of the species tree of 0.4.

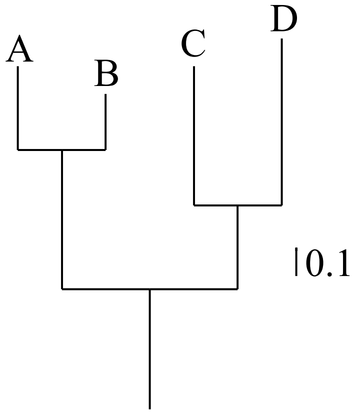

Figure 5.1: Species tree example: Likelihood estimates from single gene family trees

For the first example, I will examine a gene family tree with $\lambda = 1.2$ and $\mu = 0.8$. The simulated tree is shown in figure 5.2 within the species tree from figure 5.1.

The following figures show the likelihoods for various pairs of values of λ and μ. The dot represents the "true" values, while the contours show the 95% and 50% approximate confidence regions. These are determined using the asymptotic property of the likelihood ratio with $df = 2$. That is, the 95% confidence region boundary is at 0.5 x 5.9915 ($= \chi_2^2(0.95)$) from the maximum likelihood, while the 50% confidence region is at 0.5 x 1.3863 ($= \chi_2^2(0.5)$) from the maximum. It should be noted, however, that I have used the maximum likelihood value among the parameter pairs tested on a grid. This will in general be less than the true maximum likelihood. As a result, the approximate confidence regions will be slightly larger than would be provided by the true maximum. We can expect the effect of this to be greatest when the likelihood curve is steeper, as occurs when data from many trees are combined.

It may be apparent from these figures that the "true" values of the parame-

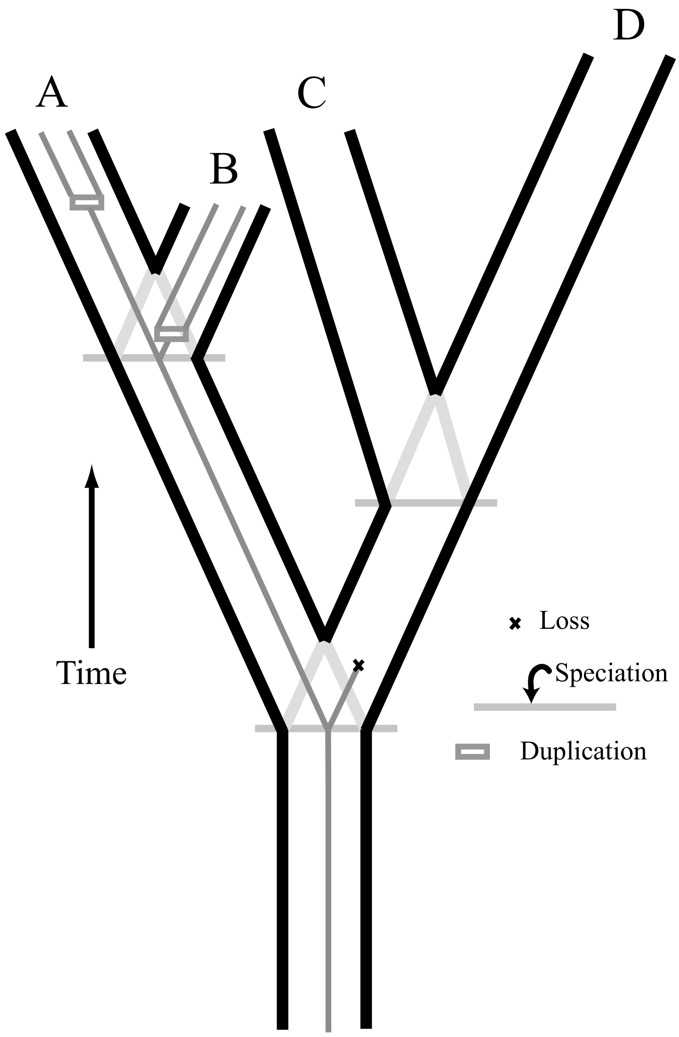

Figure 5.2: One gene tree simulated on species tree 5.1. (The duplication nearest to the tip with species B was moved somewhat upward to make the diagram easier to read.)

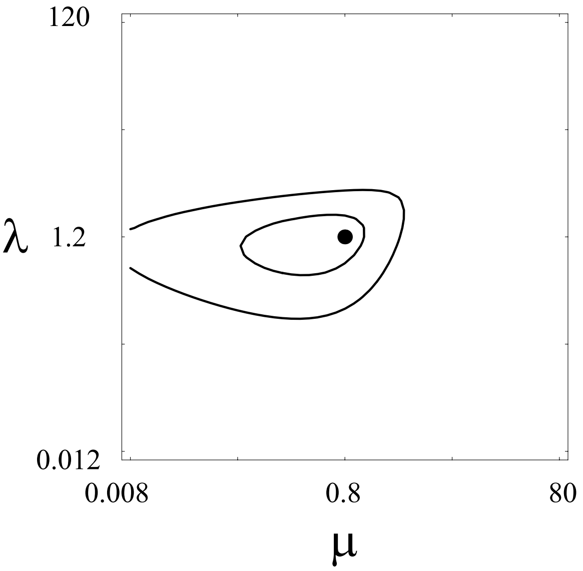

Figure 5.3: Likelihood surface for the single gene phylogeny shown in figure 5.2. Note that the axes are on a logarithmic scale. The dot in the center shows the “true” values of the parameters used to simulate the tree.

ters are only rarely found in the 50% approximate confidence region. I have confirmed this through simulation. In a sample of 100 gene family trees (using $\lambda = 1.2$, $\mu = 0.8$, and the species phylogeny in figure 5.1), the true λ and μ were only within the 50% confidence region 27 times. It should be remembered that the confidence regions here are approximate, as they are derived from the asymptotic properties of the likelihood ratio test. Thus the regions would be expected to approach true confidence regions as the sample size increases. Here, however, the sample size is unclear, since each graph comes from a single tree, on each of which multiple duplication and loss events are possible.

There are a number of observations which can be made from these trees. The likelihood surfaces vary a great deal from tree to tree, but generally follow certain kinds of patterns. Those trees that show a duplication have more evidence about the duplication rate. In particular, $\lambda = 0$ is inconsistent with such

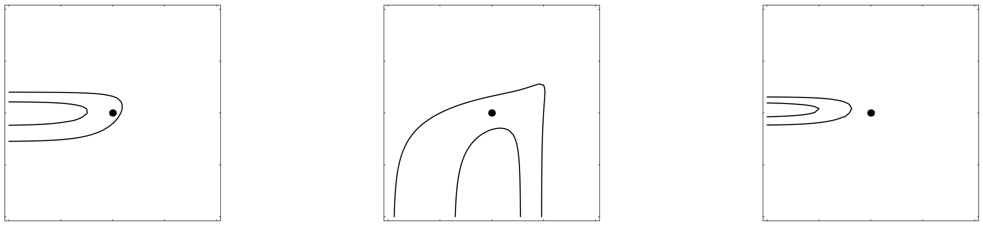

Figure 5.4: Likelihood surfaces for 12 different gene family trees produced with the species tree in figure 5.1 and parameters $\lambda = 1.2$ and $\mu = 0.8$; The first three surfaces correspond with trees A, B, and C (as referred to in table 5.1). The scales are the same as for figure 5.3.

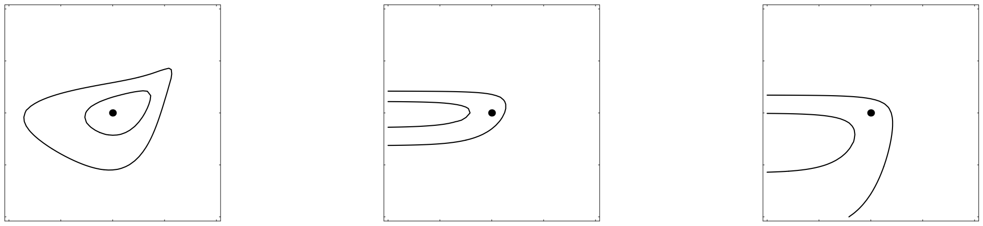

Figure 5.4: (continued) Trees D, E, and F

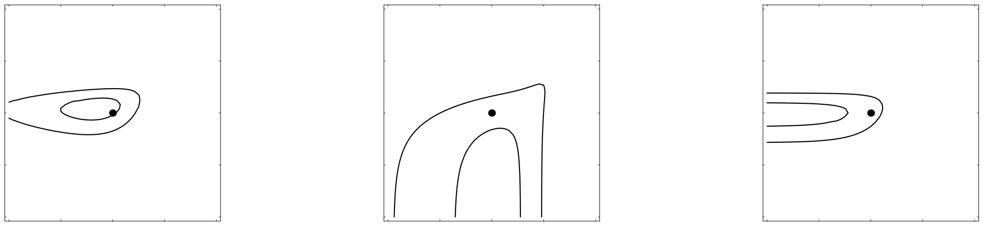

Figure 5.4: (continued) Trees G, H, and I

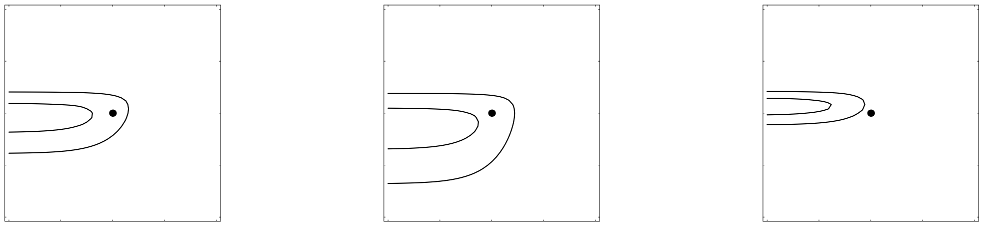

Figure 5.4: (continued) Trees J, K, and L

Table 5.1: Types of gene phylogenies which gave rise to the likelihood surfaces in figure 5.4

	Has Duplications	No Duplications
Known Deletions	D	BH
No Known Deletions	ACEGIJKL	F

a tree, and very low values of λ similarly tend to have a low likelihood. Those trees that require a loss event (that is, which have one or more losses resulting in a species without any gene family members) are particularly informative about the loss rate, and are incompatible with $\mu = 0$. Note that with surviving genes from a birth-death process in a single species, we would never observe a tree in which a deletion is required. This would be a gene family tree in which all gene family members have been lost, which of course cannot be observed. Thus the birth-death process of genes in a tree of species can be much more informative than the birth-death process in a single species.

From these two types of events — Trees with duplication events and trees with necessary loss events — we can classify four general kinds of trees. These are A) trees with duplications and known losses, B) trees with duplications but no known losses, C) trees without duplications and with known losses, and D) trees without duplications or known losses. For a summary of the types of topologies which resulted in the likelihood surfaces in figure 5.4, see table 5.1.

We also can get a rough idea of the ability to infer parameters from a small tree like this. As each axis shows a factor of 10^4 for its parameter, the likelihood surface is less precise than it may seem from these pictures. Nonetheless, even a small gene family tree tends to have enough information to reject extremely high values, and usually also very low values of one or both of the parameters.

It is also interesting to observe that there appears to be a small correlation between the estimates of λ and μ. Intuitively, one might expect that there would be an ability to estimate $\lambda - \mu$ or λ/μ, but that determining the individual parameters might be difficult. This question also arises from the properties of the birth-death process. Using the probability that a single lineage will go extinct after time t from equation 3.2, page 31, the probability of survival is $1 - P(\text{extinction})$, which is $1 - \alpha$, where α is given in equation 3.4 as

$$\alpha = \mu(e^{(\lambda-\mu)t} - 1)/(\lambda e^{(\lambda-\mu)t} - \mu)$$

As t grows large, in the case where $\lambda > \mu$, we see that

$$P(\text{nonextinction}) \to 1 - \mu/\lambda \text{ as } t \to \infty \tag{5.1}$$

However, more informative are the properties of the birth-death process when considering only surviving lineages. This is shown in Harvey, May, and Nee (1994) when examining the “reconciled” birth-death process with $\lambda > \mu$ in a single species. (Their paper is actually framed in terms of a speciation-extinction process rather than as duplications and losses in a gene family.) The authors showed that the rate of increase of number of lineages increases exponentially at the rate λ when near the time of sampling, and as $\lambda - \mu$ near the first time of duplication. Using this, they demonstrated that both λ and μ can be estimated when only the surviving lineages from a birth-death process are observed. It should be noted, however, that there are additional considerations in the present case, as there are multiple species branches, each with their own birth-death processes, and the number of gene family members at the beginning of a species internode is not known with certainty.

As shown above, the information available differs greatly even among gene family trees on the same species tree. Hereafter, in order to summarize information from multiple gene families, I will make use of 25 gene families for each case, and add their log likelihoods together for each parameter pair. This

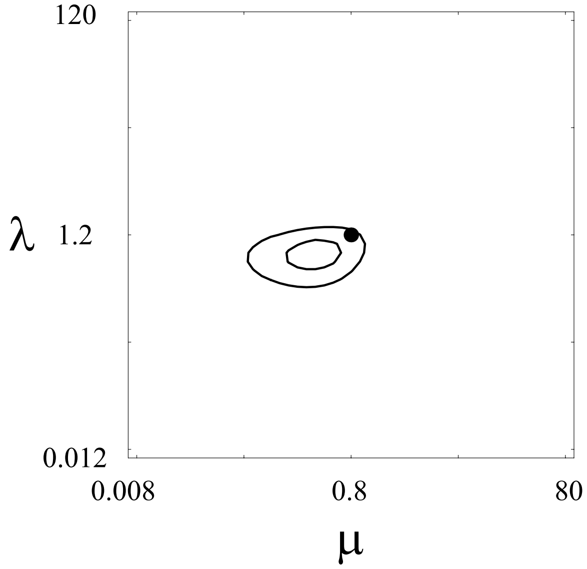

Figure 5.5: Likelihood surface from 10 gene families with the same parameters

corresponds to the assumption that there are multiple gene families, all with the same λ and μ (see subsection 3.7.3, page 45). It is important to remember, however, that the accuracy and confidence intervals will be substantially better than those which could be derived from a single gene family. For example, figure 5.5 shows the likelihood surface when combining likelihoods from 10 gene families on species tree 5.1.

5.2.2 Effect of the species tree and of base branch length

As the process of gene duplication and loss occurs in every internode of the species tree, the number of such internodes can affect the ability to estimate duplication and loss rates. In this section, I will show a variety of different species trees, then show the likelihood surfaces for λ and μ in each. As in the last section, the true parameter values here are $\lambda = 1.2$ and $\mu = 0.8$. In this section, all figures will show the combined result from 25 simulated gene

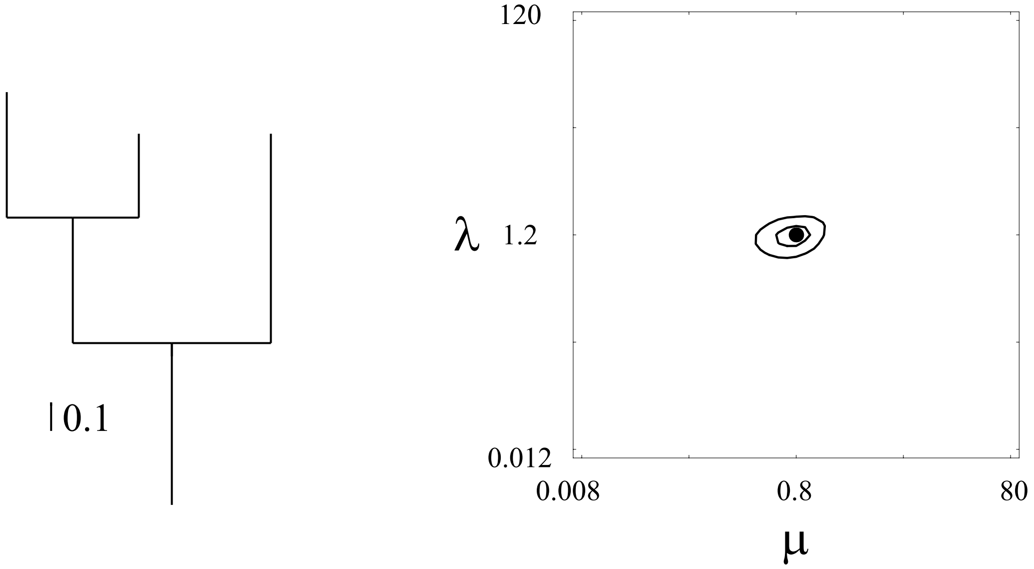

Figure 5.6: Species tree example: With base branch

family trees, rather than the result of a single tree.

I will begin with a three tip tree, shown in figure 5.6. This is the same as the previous tree (figure 5.1), with one branch removed.

The next tree is exactly the same, but with a 0 length base branch (figure 5.7).

There appears to be little difference between these likelihood surfaces. Note that the calculations of the likelihood take account for the different lengths of the base branch in each case. Thus, this is not a test of whether it is important to analyze the base of the tree, but rather of the effect of the base branch on the ability to infer birth-death parameters.

Now we will examine some more complex trees. Figure 5.8 shows a fairly balanced tree with 10 tips, while figure 5.9 uses an unbalanced 10-tip tree.

Though more balanced trees may give a slightly greater amount of information, the differences among species trees with the same number of branches is subtle. From observation of other sets of 25 gene family trees produced on

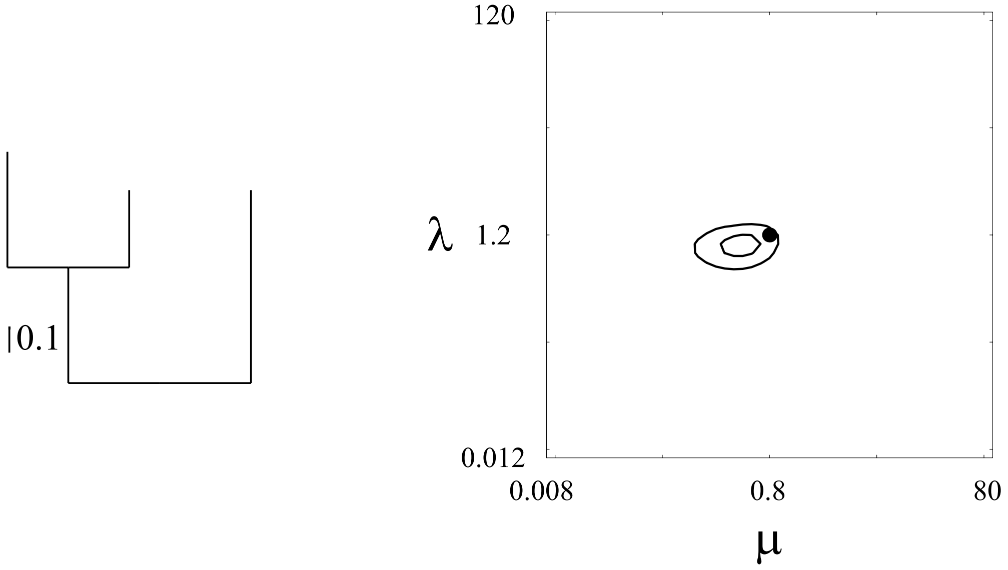

Figure 5.7: Species tree example: No base branch

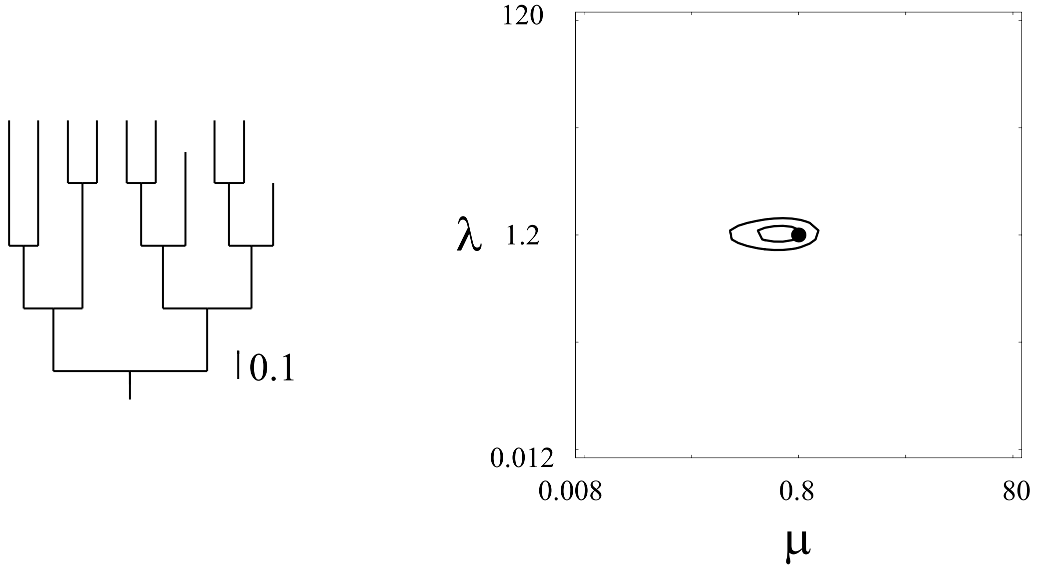

Figure 5.8: Likelihood surface of a larger tree

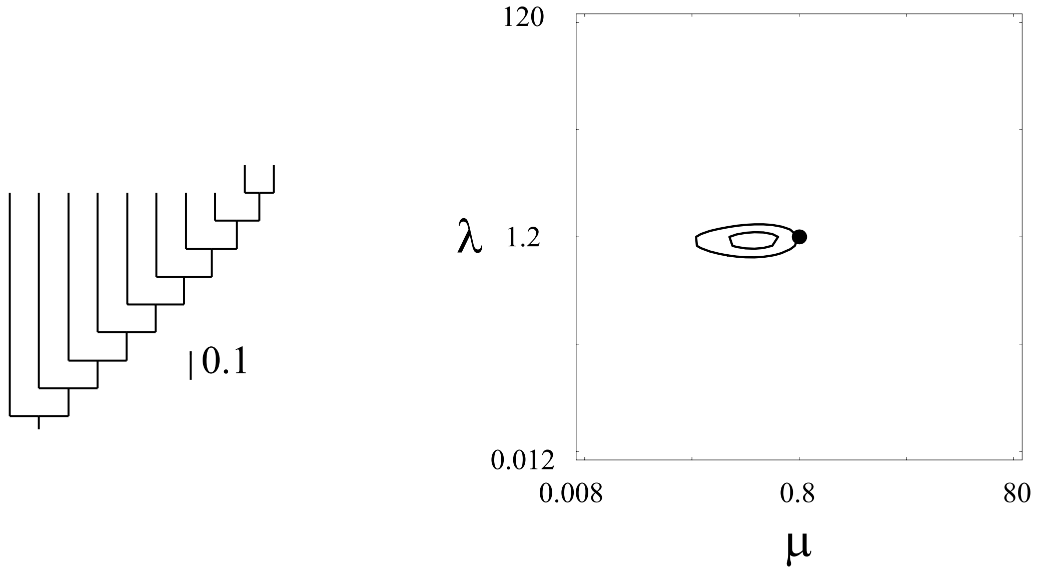

Figure 5.9: Likelihood surface of a large, less “balanced” tree

these species trees, it appears that the difference between these likelihood surfaces does not represent a strong pattern; The gene family trees used in 5.9 contained unusually few which excluded low estimates of μ. However, trees with more tips do allow a more reliable estimate of λ than is available from trees with fewer branches, as can be seen by the flatter confidence regions in figures 5.8 and 5.9, as compared with figures 5.6 and 5.7.

5.2.3 Rates of duplication and loss

It is also interesting to ask how the magnitude of the duplication and loss rates affects their inference. (Since there is no DNA in the analysis in this section, there is no reason to look at changes to the overall branch lengths of a tree, since that would be equivalent to rescaling λ and μ.) In general, we would expect that higher parameter values will be easier to detect, since their events will be more common. However, as duplications and losses become very common, our ability to infer them may eventually decline, as many of the

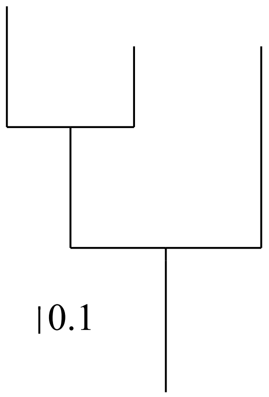

Figure 5.10: Species tree example: Effect of scaling parameters

duplication events will be unseen due to losses.

Let us take the species tree in figure 5.10. The effect of different “true” values of λ and μ are first illustrated in figure 5.11. On the left we see the situation when the rates are both doubled. For comparison, see the likelihood surface in figure 5.6. With increased rates, the ability to infer λ is improved, but inference of μ is largely unaffected.

On the right we see a tree with $\mu > \lambda$. This tends to result in a likelihood surface which shows more of a diagonal, suggesting that the two parameter estimates are correlated. As discussed above, this is not unexpected. It also tends to produce somewhat larger confidence regions than are seen when $\lambda > \mu$.

With only the duplication rate high, $\lambda = 1.8$ and $\mu = 0.1$ (see figure 5.12), we are close to a pure birth process. Here inference of the duplication rate is very accurate, but inference of the low loss rate is difficult.

When only the loss rate is high, with $\lambda = 0.1$ and $\mu = 1.8$, inference is difficult for both parameters. Note that this figure once again shows the tendency toward a diagonal likelihood surface, though it is somewhat less pronounced than in figure 5.11.

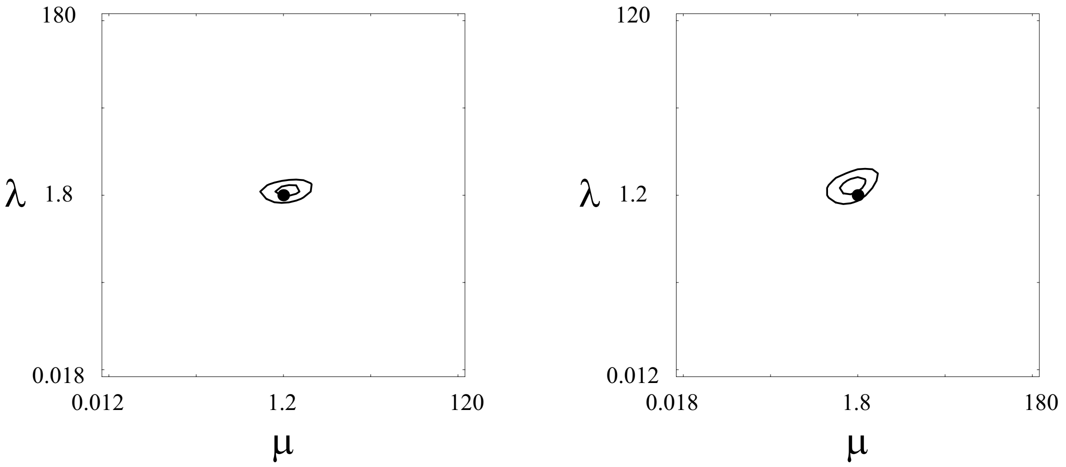

Figure 5.11: On the left: $\lambda = 1.8$, $\mu = 1.2$; On the right: $\lambda = 1.2$, $\mu = 1.8$

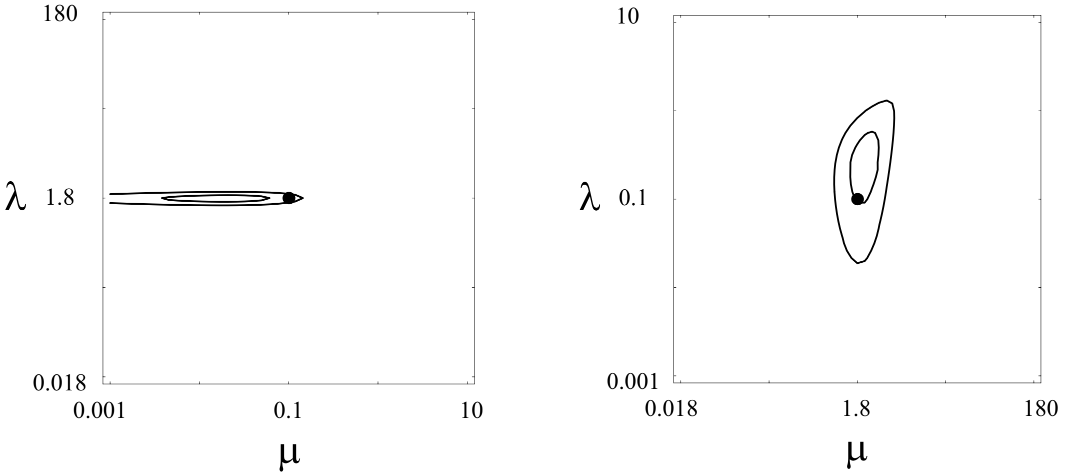

Figure 5.12: On the left: $\lambda = 1.8$, $\mu = 0.1$; On the right: $\lambda = 0.1$, $\mu = 1.8$

5.3 Robustness to Assumptions

5.3.1 Effect of approximating the likelihood calculation

In my calculation of the probability of the gene family tree, there are a pair of summations (in equation 3.8, page 33) which theoretically need to be made to infinity. In any case tried, however, it has sufficed to use a cutoff value of at most twice the maximum number of gene family members in any species. This can be confirmed by checking the $L_N(i)$ values calculated by Blurp for higher values of i. Here I will instead graphically examine different cutoff values to look at their effect. The cutoff in the summations affects only the likelihood calculation and not the simulation of trees. To make the comparison more accurate, in each set of graphs the exact same set of 25 gene family trees will be used.

In general, we would expect the cutoff to become important when high numbers of doomed lineages are fairly likely. This tends to occur in larger trees, and with higher rates of λ and μ. See figure 5.14 for results from a 10-tip tree (figure 5.13) using $\lambda = 2.4$ and $\mu = 1.6$. Though results are only shown for a cap of 1 and of 5 doomed lineages, even when just 2 lineages is used as the cap, results are very similar to the result on the right.

5.3.2 Gene family ascertainment

It is interesting to consider the reasons why a researcher will want to examine the likelihood of duplication and loss rates in a gene family. It is possible, for example, that gene families with more members in each species will be analyzed disproportionately more often. But if we exclude (or reduce the frequency of) analysis of certain gene families, this will mislead estimation of their parameters.

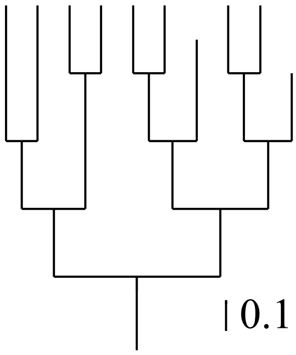

Figure 5.13: Species tree example: Effect of approximation of sums

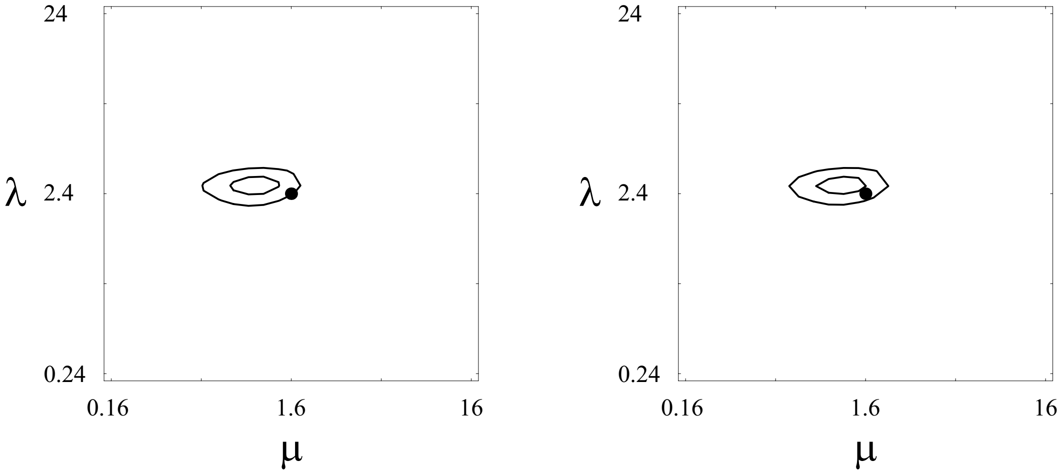

Figure 5.14: Likelihood surface for an identical set of 25 gene family trees simulated with $\lambda = 2.4$ and $\mu = 1.6$ on tree 5.13. Note that the scale here is of only a factor of 100 for each parameter, unlike other graphs in this chapter which use a factor of 10,000. On the left is the surface with a cap of 1 doomed gene family member on the summation in equation 3.8; on the right with a cap of 5 doomed gene family members. Surfaces with a cap of 3 or more appear identical to the graph on the right.

Unfortunately, it is in most cases extremely difficult to identify what kind of ascertainment is occurring. However, we can at least look at a few plausible specific kinds of ascertainment, and evaluate what effect they would have on estimation.

As was discussed in subsection 3.7.4 (page 47), it is always necessary to condition on observation of at least one gene family member in one of the species. If no members of the gene family exist, clearly the gene family would not be studied. This kind of ascertainment correction has been made throughout the thesis, including in the results in this chapter as well as in chapter 6.

One avoidable but still plausible scenario is that the researcher would only examine a gene family if it has more than one gene family member in at least one species. This can be examined by simulating gene family trees as described before, but discarding any which do not have at least one species with multiple gene family loci. The likelihood for each $\{\lambda, \mu\}$ pair is then analyzed only for these gene families. As before, 25 accepted gene families are analyzed for each figure. In this subsection, I will be using the 7-tip species tree in figure 5.15. These examples will again use $\lambda = 1.2$ and $\mu = 0.8$.

In figure 5.16, it is easy to see the effect of this. When gene families which have at most one gene in each species are not analyzed, estimation of λ is biased upward, and estimation of μ is biased substantially downward.

5.3.3 Species ascertainment

It is also possible that a researcher might analyze only those species which have one or more members of the gene family. This is a more easily avoidable kind of ascertainment, but it is still worth taking a look at its effects.

I will continue to use the same 7-tip tree from the previous section (figure 5.15) with $\lambda = 1.2$ and $\mu = 0.8$. In this case, all existing gene family members

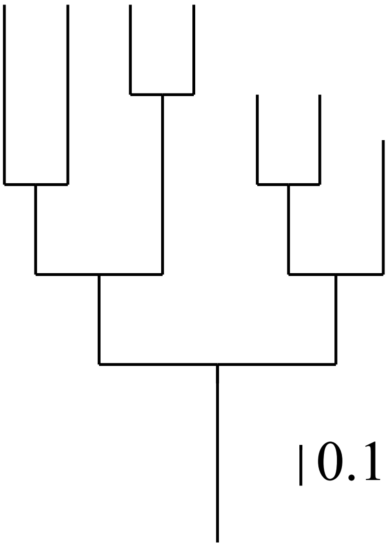

Figure 5.15: Species tree example: Effect of ascertainment

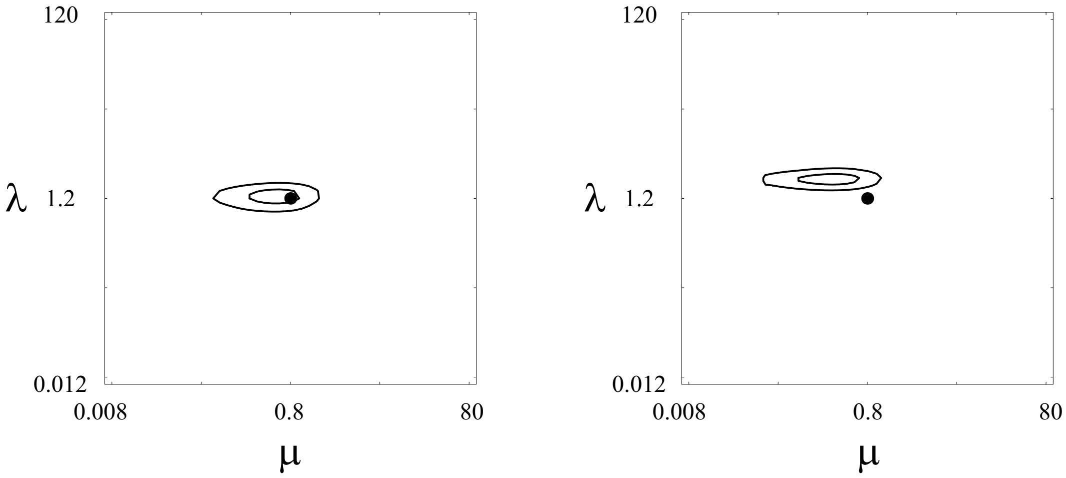

Figure 5.16: Likelihood surfaces for species tree 5.15 with and without gene family ascertainment; Left: Using all gene family trees; Right: Using only those trees with two gene family members in at least one species

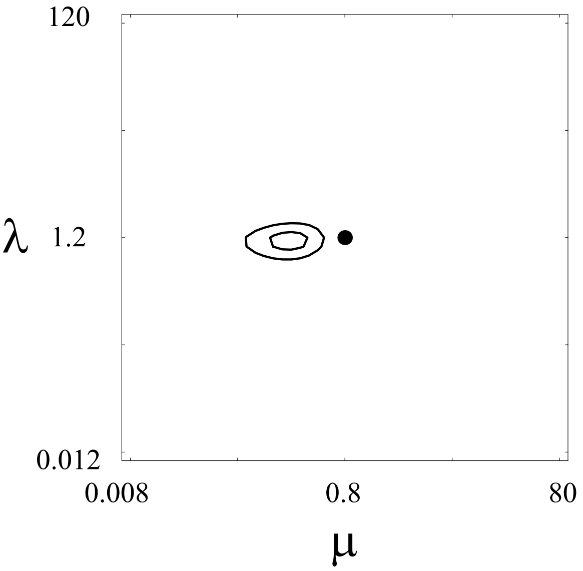

Figure 5.17: Likelihood surface for species tree 5.15 when species without gene family members are ignored.

are observed, but any species branch without any gene family members will be removed from the analysis. As with all these examples, if the gene family does not have members in any of the species, then it is rejected and replaced. The results are shown in figure 5.17. Here estimation if μ is biased downward, but estimation of λ appears to be fairly robust to this kind of ascertainment.

5.3.4 Effect of missing gene family members

There are now a large number of species for which nearly the complete genome has been sequenced. But there are still a great many organisms in which finding gene family members is much more difficult. In such a case, the researcher may well not be able to include all members of the gene family. As was previously discussed, this can be corrected in the case where the researcher has an estimate of the chance they might be missing a gene family member. I have implemented this as a single fixed chance across all species, but it could just

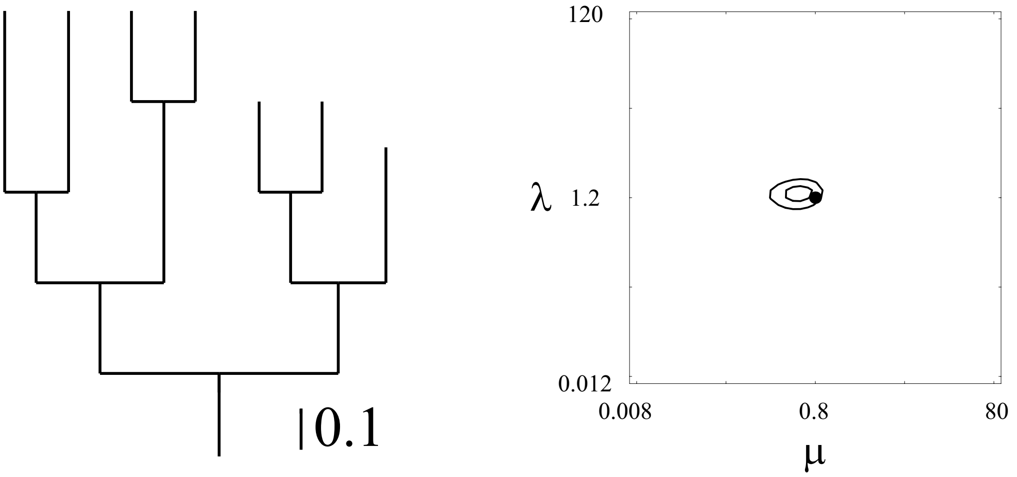

Figure 5.18: On the left is the species tree used in the following three graphs. On the right is the likelihood surface obtained with the tree when not missing any gene family members.

as easily be applied as a separate chance in each species.

To see whether this correction is necessary, I simulated 25 gene family trees as described in the rest of this chapter, but removed a fixed (50%) fraction of the loci (and any nodes and internodes which lead only to removed tips) from analysis before calculating the likelihoods. The resulting likelihood surfaces are shown in figures 5.18 and 5.19, without removing the tips, when removing the tips without adjusting the calculations, and with adjustment of the likelihood calculation, as described in section 3.5 (page 36). Here $\lambda = 1.2$ and $\mu = 0.8$. A tree with no observed gene family members (whether because they were all lost, or in figure 5.19 because they were not observed) was rejected and simulated again. The graphs in figures 5.19 and 5.18 represent the same 25 trees; The graph in figure 5.18 would have contained the same trees as were used in figure 5.19, plus any trees which were rejected due to a lack of observation of genes, but in this case no such trees were simulated.

Here we see that the confidence regions are larger when the sequences may

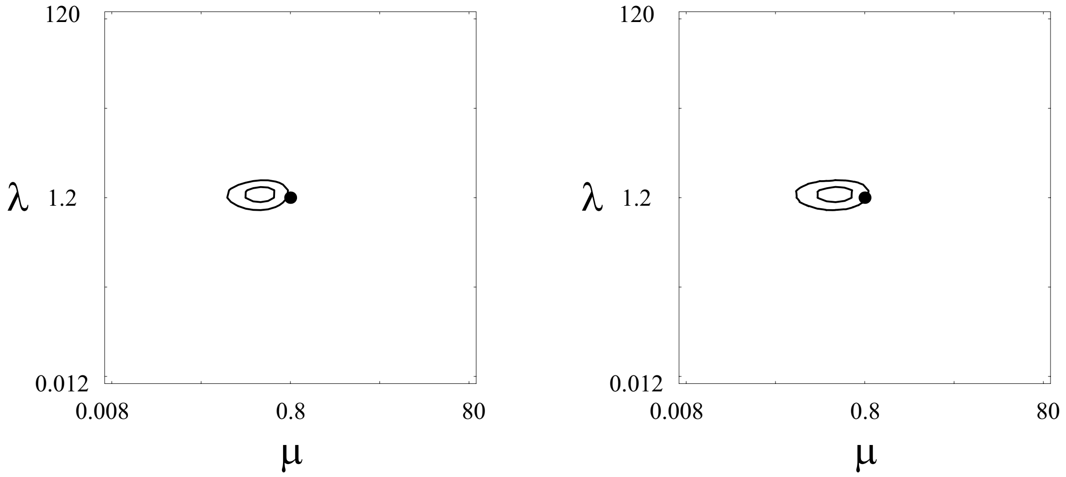

Figure 5.19: On the left is the likelihood surface from the tree in figure 5.18 when each gene has a 50% chance of being missed from the analysis, but when assuming that all genes have been sampled. On the right is the surface obtained when the same genes are missing, but the likelihoods are correctly adjusted for a 50% chance of missing each.

be missing, if we correct for the missing data. If we fail to adjust for missing data, μ in particular is underestimated. This is not an extremely large effect (though the logarithmic scale of the figures should be kept in mind), but still suggests that confidence region estimates will be more accurate if the possibility of missing genes is correctly incorporated in the analysis.

Chapter 6
ANALYSIS OF DNA SEQUENCES

6.1 Accuracy on Simulated Data

As described in section 4.8, one way to confirm that methods and code are working is to set up simulations of data under the conditions of the model, and see if the inferred likelihood surfaces correspond well with the "true" parameter values. As both the simulations and the inference of the rates are stochastic[1], there will be some variability in the estimates. However, it will be informative to see the kinds of variance and bias which are present in the values which are inferred.

For my simulated data runs, I supplied a known species tree (figure 6.1), known parameters λ and μ of the birth-death process, and 2.0 for the transition/transversion ratio of the Kimura two parameter model of DNA evolution.

From the species tree and birth-death parameters, 100 gene family trees G were simulated. Any G which has no surviving lineages was discarded and simulated again. (Quilg corrects for these missing Gs as described in subsection 3.7.4, page 47.) For each G, the specified number of bases (200 or 500) were simulated from the root to the tips using the DNA evolution model, starting with random sequence with an equal chance of each base at the root. The sequences were labelled with their corresponding species, as is required by Quilg for input.

Quilg was run once for each set of simulated data. The starting tree was

[1] As of November 2005, according to the definition of stochastic on the web site Answers.com, this means I am conducting "stochastic stimulations [*sic*]."

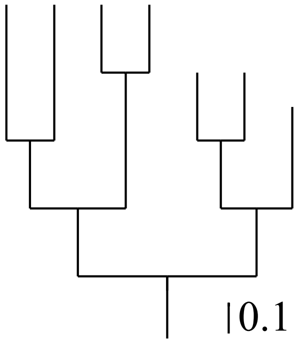

Figure 6.1: Species tree used with simulation study

Table 6.1: Means and SDs of estimated parameters ("true" λ = 1.2, μ = 0.8) from simulated DNA

Sequence Length	Mean $\hat{\lambda}$	SD $\hat{\lambda}$	Mean $\hat{\mu}$	SD $\hat{\mu}$
200 bases	1.0946	0.3435	0.5223	0.4216
500 bases	1.0740	0.2962	0.5511	0.4137

inferred by the program, but the starting parameters λ_0 and μ_0 were supplied from the "true" parameter values of $\lambda = 1.2$ and $\mu = 0.8$. With each run of Quilg, the maximum likelihood estimates of $\hat{\lambda}$ and $\hat{\mu}$ were recorded. As described in section 4.4, Quilg was run in multiple chains for a total of 30,000 proposed trees. One in ten trees were stored for calculation of likelihoods at the end of the runs. 10,000 proposed trees (1,000 stored trees) were used in the final MCMC chain, used to calculate the output maximum likelihood values for λ and μ. The results for these simulations are summarized in table 6.1.

6.2 ANOVA Methods

Though the results in table 6.1 give a general indication of the accuracy of estimation of λ and μ, it is worth examining this in more detail. Even under conditions ideal for the model, there are a number of potential sources of inaccuracy of estimation in this process. Some specific factors which might affect the estimates are

- The underlying species tree topology and branch lengths
- The different topologies and branch lengths of the gene family trees
- Variability of DNA evolution
- Variability between runs of the MCMC process of estimation

To evaluate the relative importance of each of these potential effects on estimation of duplication and loss rates, I conducted a nested ANOVA (analysis of variance) as follows.

- For each of 500 replicates, a species tree was simulated.
- For each species tree, two gene family trees were simulated
- For each gene family tree, two sets of DNA data were simulated
- For each set of DNA sequences, Quilg was run two times to estimate $\hat{\lambda}$ and $\hat{\mu}$.

It is necessary to use a nested ANOVA because of the relationship between the levels of the experimental design. The gene family tree depends on the species tree in which it is simulated; The DNA sequences depend on the gene family tree. The MCMC run depends on all three — the species tree, the gene family tree, and the DNA sequences. As there are two quantities being estimated — λ and μ — I will run this analysis separately on the data for each.

That is, using the data for $\hat{\lambda}$ and $\hat{\mu}$ from the same runs, I will analyze the causes of variation of each. The factors will be designated as S for the species tree, G for the gene family tree, and D for the sequence data. Following the methods in Sokal and Rohlf (1969), the equation for decomposing the variates of each estimate can be given as

$$Y_{ijkl} = \text{mean} + S_i + G_{ij} + D_{ijk} + \epsilon_{ijkl} \tag{6.1}$$

where Y_{ijkl} is the parameter estimate from lth MCMC run with the kth data set with the jth gene family tree of the ith species tree, "mean" is the parametric mean of the population, and S_i, G_{ij}, and D_{ijk} are the random contributions of the species tree, gene family tree, and sequence data, respectively. Finally, ϵ_{ijkl} is the error term attributable to each Quilg run. For the initial analysis, I will assume that S_i, G_{ij}, D_{ijk}, and ϵ_{ijkl} are normally distributed.

Species trees were simulated using the general method described in Kuhner and Felsenstein (1994). That is, we use a pure birth ("Yule") process until the specified number of tips are present, then simulate until one additional birth event is about to occur, at which point the tree is ended. Unlike the simulations of Kuhner and Felsenstein, these simulations begin with a single species, rather than with the first speciation event. The time to the first speciation event is taken as the length of the base internode of the species tree. This produces clocklike species trees. As has been noted, however, clocklike species trees are not necessary for analysis by this method. For the purposes of this study, the rate of branching was set to give an expected branch length of 0.3 mutations/base, and simulation was stopped just before the first branching event after five species were present in the phylogeny. The last event serves just to provide the endpoint of the tree. No further speciations occur after the fifth species has been produced.

Table 6.2: Sources of variation in $\hat{\lambda}$: df is degrees of freedom; SS is sum of squares; MS is mean square

Source of variation in $\hat{\lambda}$	*df*	SS	MS	F_s	Components
Among Ss	499	891.2683	1.7861	2.1081	17.12%
Among Gs within Ss	500	581.6525	1.1633	4.5330	47.25%
Among D within Gs	1000	303.8537	0.3039	5.2338	31.19%
Within D	2000	40.3802	0.0202		4.44%

Gene family trees were simulated as described before in section 4.1, page 50. For purposes of these simulations, the "true" parameter values were $\lambda = 1.2$ and $\mu = 0.8$.

For each gene family tree, 500 bases of DNA were simulated (twice) using a transition/transversion ratio of 2.0, and starting with random DNA at the root of the gene family tree.

Quilg was run twice on each set of DNA data. It was supplied with the correct known parameters of DNA evolution. The starting trees and starting values λ_0 and μ_0 were determined by Quilg based on the data (see section 4.5, page 58).

6.3 Sources of Variance in Estimation

For both $\hat{\lambda}$ and $\hat{\mu}$, the results are similar. A small part of the variance comes from within DNA sequences, which is to say between runs of the Quilg program for the same set of data. This suggests that variance would be reduced by the use of longer MCMC chains or by improving the rearrangement scheme to prevent becoming stuck with an inaccurate estimate of the parameters. However, the proportion of the variance between Quilg runs is relatively small, so the computer time needed to run longer chains might not be worthwhile.

Table 6.3: Sources of variation in $\hat{\mu}$

Source of variation in $\hat{\mu}$	df	SS	MS	F_s	Components
Among Ss	499	201.6930	0.4042	1.2457	8.85%
Among Gs within Ss	500	162.2318	0.3245	4.2788	55.18%
Among D within Gs	1000	75.8303	0.0758	14.5497	31.35%
Within D	2000	10.4236	0.0052		4.63%

Assuming normality of the estimates, a small but significant ($p < 0.01$ for the least significant example) amount of variance appears to be associated with differences between species trees. This kind of variance will tend to be unavoidable, because most researchers will likely be choosing their species to study for reasons other than accuracy of the estimates obtained for duplication and loss rates. There should be some room to examine exactly which shapes and branch lengths of species trees give the most precise estimates of λ and μ. The results in section 5.2 (page 74) show some specific examples of this.

There is a substantial amount of variance in the gene family trees for a particular species tree. This is in keeping with the analysis of known gene family trees in chapter 5. This is especially clear in subsection 5.2.1, where the differences in the likelihood surface from single gene family trees are shown. There, some gene family trees are much more informative for one of the parameters than the other.

A moderate amount of variance is attributable to different DNA sequences for a particular gene family tree. This suggests that methods which assume the sequences give a precise picture of the gene family tree will be greatly overestimating the certainty available with their data.

If we instead assume that variation in $\hat{\lambda}$ and $\hat{\mu}$ is multiplicative, transform-

Table 6.4: Sources of variation in $\log \hat{\lambda}$

Source of variation in $\log \hat{\lambda}$	*df*	SS	MS	F_s	Components
Among Ss	499	954.4687	1.9128	1.6024	17.85%
Among Gs within Ss	500	596.8448	1.1937	3.2375	40.97%
Among D within Gs	1000	368.7036	0.3687	8.0307	32.06%
Within D	2000	91.8239	0.0459		9.12%

Table 6.5: Sources of variation in $\log \hat{\mu}$

Source of variation in $\log \hat{\mu}$	*df*	SS	MS	F_s	Components
Among Ss	499	902.9287	1.8095	1.3662	11.82%
Among Gs within Ss	500	662.2518	1.3245	3.4049	45.60%
Among D within Gs	1000	389.0033	0.3890	8.1312	33.26%
Within D	2000	95.6816	0.0478		9.32%

ing the data by its log gives the results in tables 6.4 and 6.5. As the distribution of estimates appears by inspection to be closer to normality when the log of the estimates is taken, there is some justification for this transformation.

Chapter 7
SUMMARY

In this dissertation, I have presented a simple model of the evolution of members of a gene family. Using this model, I have shown how to calculate the likelihood of a particular gene family tree and thereby find the maximum likelihood estimates of λ and μ, the parameters of the model. By employing Markov chain Monte Carlo to search among different possible gene family trees, it was possible to show how to infer likelihoods for the parameters even when only DNA sequences are available, and the gene family tree was not known.

To examine the properties of inference of duplication and loss rates under a variety of circumstances, I implemented these methods in a pair of computer programs. The simpler of the programs, Blurp, calculates the probability of a particular gene family tree for a range of values of λ and μ. This allowed the examination of the ability to infer duplication and loss rates in a variety of circumstances. This in turn permitted the examination of inference with different gene family trees, the effect of the species tree on inference of rates, the ability to infer higher or lower rates of duplication and loss. In addition, robustness of the model was examined by simulating trees under different models and observing the effect on Blurp's inference of rates.

The more complex of the programs, Quilg, does not assume a single gene family tree, but rather sums over possible gene family trees. As there are a great many possible gene family trees, Quilg uses Markov chain Monte Carlo as an importance sampling method, concentrating the search on those gene family trees which provide a better explanation of the data. This program allowed the examination of the variability in estimation introduced by imprecise knowledge of the gene family tree.

The evolution of gene families is a fundamental question in evolutionary genetics, as it is potentially the source of many or most new genes. In addition, an understanding of gene family evolution is needed when using members of a gene family to address other evolutionary questions. I have attempted in this thesis to provide a useful tool in the analysis of gene family evolution. With the presented model and computer implementation, it is possible to infer parameters of duplication and loss rates in a gene family. With an improved knowledge of the rates of duplication and loss of genes, we will have a better picture of the rate and manner in which new genes have arisen. There are many more complicated models — of genome duplications, changing duplication and loss rates, or selection on the number of gene family members — which are not covered in this model. However, when each of these have been implemented, it will often be necessary to compare against a null model such as the simple birth-death process I have described.

Chapter 8
ADDITIONAL CONCERNS

In this chapter, I will examine some related evolutionary topics which might be analyzed with methods similar to those presented in this thesis.

8.1 Genomic Effects

8.1.1 Genomic duplication

It has been proposed (and also disputed) that the duplication of the entire genome has occurred in the history of a number of lineages (for example, see Nadeau and Sankoff 1997 on the mouse and human lineages). When such an event takes place, every locus in the genome simultaneously gives rise to another identical locus. This is not allowed in the model I have presented. However, it would be interesting to extend the model to include such events. One extreme model — similar to the parsimony model proposed by Guigó, Muchnik and Smith (1996) — would be to replace duplication events with genome duplication events. This would result in a model in which the rate of duplication events would not depend on the number of gene family members. Using the terminology from section 3.4 (page 31), letting the rate of genome duplications be λ_{genome}, and ignoring loss events, the probability of an internode not ending in a tip (including the point probability density of the event) is

$$L_{N'}(0) = e^{-t/\lambda_{\text{genome}}} L_N(0) \tag{8.1}$$

This is independent of s (the number of gene family members in the internode). Also note that without loss events, only $L(0)$ needs to be calculated, since all lineages will be observed.

With both genome duplication and simple loss events, the situation is somewhat more complicated. Unlike in the simple birth-death model, doomed lineages can arise from the same duplication event which leads to multiple surviving lineages. Because of this, it is easier to redefine the gene family tree, and explicitly include all genomic duplication events, even if they give rise only to doomed lineages. With this definition, in place of equation 8.1 we have

$$L_{N'}(i) = e^{-t/\lambda_{\text{genome}}} \sum_{k=0}^{\infty} p_{i\to k}(t) L_N(k+m) \tag{8.2}$$

Where m is the number of doomed lineages which arise from the genomic duplication event at the tip. The other terms are defined as shown before in section 3.4, though $p_{i\to k}$ is calculated with $\lambda = 0$, since we have not allowed simple duplication events.

It would be worthwhile to model both genomic duplication events and simple duplication events. With such a model, it would be possible to use likelihood ratio tests to compare a model with both simple duplications and genomic duplications with more restricted models with just simple duplications or just genomic duplications. This should allow the determination of whether simple duplications alone (or genomic duplications alone) can be rejected as an explanation of duplications in the gene families being studied.

8.1.2 Other birth-death models

Throughout this thesis, I have assumed a simple birth-death model for the duplication and loss of gene family members. However, as discussed in subsection 3.6.1 (page 37) and also subsection 3.7.2 (page 42), this is in some cases an unrealistic model. We might prefer to disallow the last member of the gene family from being lost in a lineage, or perhaps include other changes in the transition probabilities from the simple birth-death model.

My assumption of a simple birth-death process was made to simplify and speed the calculation of probabilities for gene family trees. Because of the use of the simple birth-death process, it was not necessary to keep track of loss events, and duplication events were only included in the tree when both resulting gene family members had surviving descendants. This is useful from a computational perspective, as loss events (and duplication events for which one of the resulting lineages is doomed) do not affect the probability of the data, since they have no effect on the branch lengths or topology between tips with DNA sequences.

Although this was a useful simplification to make, by relaxing this assumption, we would be able to include some potentially more realistic models. However, to do so, it would be necessary to keep track not only of surviving lineages, but also of doomed lineages. This is necessary because the introduction of a new doomed lineage can affect the probability of a new duplication or loss event.

A formulation of the gene family tree including all doomed lineages is certainly possible, and the calculation of probabilities on such a tree should not be difficult. However, when including doomed lineages, there are many more gene family trees among which to search. Coming up with a scheme to search efficiently among trees would be very important, especially in cases with large μ and λ, for which trees with many doomed lineages could have a relatively high likelihoods.

8.1.3 Rearrangements

Genome rearrangement through inversions has been modelled by Larget, Simon, and Kadane (2002). As described in section 2.7 (page 23), the authors modelled inversion events as occurring at a constant rate in each species, with each possible “reversal” having equal probability. They then used MCMC in a Bayesian framework to use mitochondrial gene order data to infer phylogenetic relationships. Rather than infer the rate of inversion events, though, the authors make a simple estimate of the inversion rate using a neighbor joining tree of the taxa using pairwise inversion distances. This rate is then used as a known value when evaluating the probability of proposed sets of inversions.

It would be possible, however, for the inversion rate to be inferred, instead. This could be done in using Bayesian updates to the inversion rate parameter, or the rate could be successively estimated as I have done for the estimation of duplication and loss parameters.

As suggested in Larget, et al. (2004), it would be worthwhile to add a model of transposition and duplication events to their framework. This would allow more accurate estimation of phylogenetic relationships from gene order data. It would also estimation of the relative frequencies of each kind of genomic rearrangement.

8.1.4 Gene conversion

The model presented in this thesis does not allow for gene conversion events, in which part or all of one locus is replaced by the corresponding part of another related locus. Such an extension would not, however, be impossible. Here I will briefly frame a few such models, and outline methods necessary for the implementation of probability calculations on such a model.

I will first examine the possibility of whole gene conversion. In this case, one locus is replaced in its entirety by another member of the gene family. Here the extension to the model is conceptually simple — This introduces a third kind of event, the whole gene conversion, along with an associated rate.

A gene conversion event results in a simultaneous deletion of one locus and the duplication of another locus in that same species.

The method by which these events are modelled affects the difficulty of their calculation. One plausible model would be that the rate of such events will depend on the number of pairs of existing members of the gene family. i.e., $\binom{\text{members}}{2}$ times an overall whole gene conversion rate. Though this appears to be a small change, the effect on the model is substantial. Since the rate of these duplication-loss events is not proportional to the number of gene family members, we can no longer treat events on each locus as independent.

Another plausible model is that the rate of these events will depend on the number of loci, not on the number of locus pairs. When such an event occurs, one gene is chosen at random as the replacing gene, and another (possible the same gene) is chosen as the replaced gene. In this case, the frequency of such events per locus does not depend on the total number of gene family members. However, even under this model, each event affects a pair of loci, and thus events at each locus are not independent.

Which of these models is a better choice? One can propose explanations which would support either. The locus pair model would correspond with the loci meeting at random in a meiosis, at which point the gene conversion would occur. The proportional model, on the other hand, would correspond with a cellular event (independent of the number of gene family members in the genome) which makes one particular member of the pair prone to gene conversion. In practice it may be possible to model both and test goodness of fit with the data. It may also be possible to perform laboratory tests for each of these events, and examine the frequency of gene conversion events in the presence of different numbers of gene family members.

In practice, however, it may well be best to use the per-locus model regardless of which model is more accurate. Even if the gene pair model is more accurate, the single gene model could still be very close if the number of gene family members is roughly similar throughout the tree.

If we allow gene conversion to replace part of a gene family member, rather then the entire locus, then the model and calculations are necessarily substantially more complex. First we must model the frequency of such events. This can be done as described above for whole gene conversion. In addition, there must be a model for the part of the gene affected by the conversion event. One such model would be to choose a random base from within the converted gene as the starting point. Going in a randomly chosen direction (3' or 5'), the converted DNA would continue for a geometrically distributed number of bases. If this goes beyond the end of the gene, we would not consider the bases which are not part of the gene. There is some justification for the use of geometric distribution; In Engels (1994), the author found a good fit of inferred gene conversion tracks to a geometric model.

This can result in a different history for different bases in each locus. The history of nearby bases would tend to be more closely related than that of more distant bases, but the degree of correlation between them is not obvious. The history of each individual base, however, could still be described by a tree.

This may be similar to the calculation of likelihoods on trees of related copies of a locus within a single population, as studied by Kuhner, Yamato, and Felsenstein (2000). The implementation of a method for probability calculations and rearrangements in such a "tree" space is difficult, but the basic methods have already been developed in that context.

8.2 Population Size Effects and Coalescent Times

My model of the birth-death process assumes that each individual in the population has the same number of members of the gene family — When there is a duplication, the new locus is found in all members of the population, and likewise all members of the population simultaneously lose the locus when there is a gene loss. There is assumed to be a single copy at each locus. These are equivalent to the assumption that there is a single haploid individual in each species.

When considering a larger population, base changes, gene losses, and gene duplications will take place in a single individual. Such changes will often be lost rather than fixed throughout the population. The process of fixation or loss of such a change could occur over a number of generations. It is possible that speciation might occur when a duplication or deletion is found in only part of the population. Each resulting species could then have the same or different frequencies of the mutation. In such a case, the common ancestor of the locus in the two species would exist prior to the time of speciation. It is also possible that a duplication event could take place in ancestral species and be fixed in one child species but not in the other.

Ideally, we would model duplications and losses as occurring in individual members of each species. This has been done for a number of studies of population parameters, such as in Kuhner, Yamato and Felsenstein (1995). However, combining the coalescent process used to describe the relationship between a sample of alleles and the birth-death process of gene duplications and losses could be complicated.

Following well known results of population genetics, the probability of fixation of a selectively neutral change will be $1/2N_e$, where N_e is the effective population size. As the frequency of these mutations in the population is proportional to N_e, this results in a roughly constant rate (though this assumes that the rate of neutral mutations does not differ with population size, which is not entirely correct). Because of this, the assumption of a constant rate for duplication and losses may be reasonable for neutral changes.

However, as was previously discussed in section 2.4, this does not hold if new duplications gain new, selectively advantageous functions. In this case, the rate of fixation of duplicate genes will depend on N_e. Thus, when considering models of selection on new gene duplicates, modelling population size (and changes in population size) would become more important. To properly model this, it may be necessary to attempt to integrate over possible ancestral population sizes, probably using MCMC. This has been done in certain simpler cases (without considering gene families) starting with Nielsen and Wakeley (2001) and Rannala and Yang (2003).

There are limits to the amount of information available about past population sizes and times of events. Once all sampled copies at a locus have coalesced, we will not have any further information about the size of the population. However, each speciation event will result in at least two copies at each locus, each representing the coalesced ancestor in each of the descendant species. Thus there will be some amount of information about population size even far from the tips of the tree.

8.3 Uncertainty in the Species Tree

Throughout this dissertation, the species tree has been assumed to be known both in topology and in branch lengths, based on analysis of other genes' sequences or other phylogenetically informative traits. However, the species tree is not necessarily known with certainty. Also, the sequences of the gene family under study themselves contain information about the species phylogeny. It should be possible to infer both the species phylogeny and the gene phylogeny using data from a gene family. Ideally it would be beneficial to use information from multiple gene families in order to determine the phylogeny.

Better yet, rather than assume a single phylogeny to be correct, one would prefer to sample over the possible phylogenies proportional to their probability, much as my method samples over possible gene phylogenies. This is a real possibility. In an MCMC method similar to that used for gene family rearrangements in this thesis, some of the proposed changes would need to be rearrangements of the species phylogeny, instead. Accomplishing this while keeping the gene phylogenies consistent with the DNA sequences could be difficult, though. The "rubber band" algorithm used in Rannala and Yang (2003) is a likely candidate for the role — Here a node in the species tree is reattached at a different point in the species trees, and any affected branches in the gene tree (the tree of related copies at a single locus, not to be mistaken for the gene family tree) are linearly rescaled to adjust to the altered branch lengths in the species tree. This would likely allow rearrangements of the species tree without too greatly reducing the probability of the sequences on the proposed new tree.

Though theoretically appealing, the usefulness of such a method may be limited. Where there are already large numbers of genes which have been sequenced in each of the species in the phylogeny, there will in general be little to gain by the addition of phylogenetic information from the studied gene family. The inclusion of uncertainty about the species phylogeny would most likely be of greater importance.

Another, simpler, way to allow for uncertainty in the phylogeny would be through resampling. This was suggested in a somewhat different context by Page and Cotton (2000) and originally in the context of phylogenies by Felsenstein (1985). By bootstrapping (or otherwise resampling) the sequence data used to infer the species phylogeny, one can obtain a phylogeny for each bootstrap replicate. By analyzing each of these bootstrap phylogenies as the assumed phylogeny, one could summarize over the results, allowing (somewhat imperfectly) for the lack of certainty about the species tree. However, currently the computational time for bootstrap replicates would be very high.

8.4 Related Fields

8.4.1 Speciation models

The literature of speciation and extinction processes has focused, for the most part, on tree imbalance. That is, the pattern of branching of trees is tested versus various null models of tree distributions. Very little of the research makes reference to explicit models of speciation and extinction, and so probabilities of trees are rarely calculated. Many of these tree imbalance methods were summarized and compared in Kirkpatrick and Slatkin (1993).

By modelling speciation and extinction explicitly, perhaps in a birth-death model, we would expect to be able to estimate those parameters much as has been done in this thesis for gene duplication and loss. The work of Paradis (1998) provides one probabilistic model of speciations. However, the model presented does not allow for extinction events. Another approach has been taken by Nicolas Salamin (personal communication), in which MCMC is used to sample many possible speciation-extinction trees.

Rannala and Yang (1996) model species phylogenies as coming from a birth-death process. Though this is presented as a nuisance parameter in the inference of species trees, their method could potentially be applied to the inference of speciation and extinction rates. A framework for doing so has been described in Felsenstein (2004), in which the author showed the use of a time transformation which should allow tests of speciation times versus a uniform distribution on that transformed time scale. This result is summarized in subsection 3.7.2 (page 42).

8.4.2 Host/parasite coevolution; Biogeography

In this thesis, I have examined the placement of a gene phylogeny within a species phylogeny. Though seemingly unrelated, there are similar topological issues involved in the fields of host/parasite coevolution and in certain formulations of biogeographic hypotheses.

In host/parasite coevolution, the researcher is interested in a group of related parasites and their hosts (also related to one another) and would like to find out how the parasites evolved on the hosts. Such a model involves a tree of the hosts, and can allow speciation, extinction, and host switching events for the parasite species. In a parsimony-based tree reconciliation model such at that used by Page and Charleston (1997), these speciation events are equivalent to duplication events in a gene family tree. Likewise extinction events correspond with gene losses. There is no simple equivalent to host switching events. However, if the gene duplication model allows for lateral transfers (see subsection 8.5.2), then these correspond with host switching. In the case of host/parasite coevolution, this has been studied in a probabilistic context by Huelsenbeck, Rannala and Larget (2000).

There is a very similar set of events in models of vicariance biogeography. In this case, the researcher is interested in a set of related species, and their host regions in an area which has been subdivided by historical geographic events. In this case, speciation is again equivalent to duplication on a gene family tree, extinction corresponds with gene losses, and moving from one geographic region to another (a dispersal event) would correspond with lateral transfer.

8.5 Genomic Context

8.5.1 Nearby sequences

Throughout this thesis, each locus has been assumed to consist of a certain set of bases specified by the researcher. However, it is unlikely that each duplication event affecting the loci in a gene family will duplicate precisely the same bases. If gene duplication events also contain some of the surrounding DNA, then there should be information about the relationship of gene family members in nearby DNA as well as in the locus, itself. This information would take the form of both additional bases to be considered in the analysis (resulting in more accurate inference of the relationship between gene family members) as well as from the boundaries and inferred order of such duplication events.

To consider genomic context, there needs to be a more detailed model of the gene duplication process. In addition to considering the frequency of such events, we must also model the extent of duplications. For example, one such model would be to assume the duplication event contains the entire gene family member, plus a geometrically distributed number of base pairs on either side of the gene, with an assumed or inferred value for the parameter of the distribution.

Thus, in this model, each duplication event is associated with the extent of the duplicated region. Ideally we would want to integrate over all possible extents for every duplication in the gene family tree. However, it may be more feasible to assign proposals for the length of the duplication as part of the MCMC sampling process. It is likely that some duplication events will contain DNA which is not ancestral to any of the sampled DNA sequences. This is possible even when we assume that very long sequences were sampled. This should speed the sampling of gene family trees somewhat, since the probability of the data from unsampled sequences is very easy to calculate.

8.5.2 Lateral transfer

My model for gene duplication and loss does not include the possibility that a locus may be introduced from another species. If we wish to model such events, another rate, the lateral transfer rate in each species, could be introduced. Unfortunately, the possibility of such an event will necessarily depend on the number of gene family members in the other species. This is because a species in which all members of the gene family have been lost cannot possibly be the source of a lateral transfer event. Thus some simplified methods of allowing these events would not be possible. Instead, it will likely be necessary to explicitly model each lateral transfer event.

An MCMC model of this kind has been described by Suchard (2005). In this paper, Suchard describes lateral transfer according to two different models, defined based on the topological changes they induce when there is a lateral transfer event. The method was then implemented and tested on a set of prokaryotic gene sequences.

8.5.3 A more general model

Although there are considerable difficulties which prevent its current implementation, it is worth considering a broader model of genomic rearrangement. On a genome we can imagine a set of operations which act to change, delete from, add to, and/or rearrange its sequence. Some examples would likely be

- Mutation at a single site
- Mutation at pairs or larger groups of sites (such as occurs due to UV radiation)

- Deletion of a stretch of bases
- Tandem duplication of regions
- Addition or loss of short repeats, or the initiation of a short repeat region
- Duplication to a region distant from the original sequence
- Replacement of a region by another region due to homologous recombination
- Transposition of a region to a distant location

In addition, each of these events will necessarily occur in a single individual genome, not instantaneously throughout the population.

The combination of these manipulations would be a very complicated graph which I will refer to as the genomic history H. However complicated the graph may become, the relationship of each individual site can always be represented by a tree.

This problem is a superset of the global multiple alignment question — Since H includes all rearrangements, duplications, and deletions, the (proposed) relative locations of all of the bases will be known. Thus, although each site in the resulting genomes may have a different tree than its neighbor, we will still be able to calculate the probability of the observed sequence at each base as $P(\text{data at site}|\text{tree of this site})$. If the site has undergone a duplication, this may be a tree of multiple related base pairs in each species. Since this model allows new sites to arise, it is possible that the tree of a particular site may not exist in all species being considered. In any case, the method of calculation will be the same. Conditioning on H and the parameters of the processes, multiplying these probabilities for all base pairs in all species (including the probability at each base just once), we could obtain $P(\text{genome sequences}|H, \theta)$.

Using this, we could imagine a similar approach to that used in this thesis to calculate

$$\begin{aligned} &P(\text{Genome sequences}|\text{Parameters of the processes}) \\ &\qquad = \sum_{H} P(H|\theta) P(\text{Genome sequences}|H, \theta)) \end{aligned} \tag{8.3}$$

Though it should be possible to calculate $P(\text{genome sequences}|H, \theta)$, $P(H|\theta)$ can be determined only if each of the processes is modelled in such a way as to allow the computation of probabilities. This should be possible, though choosing sensible models will not be easy.

Most difficult certainly would be the summation over possible Hs. The space of Hs is vast, and most will result in an unacceptably low likelihood. Therefore it will be necessary to use importance sampling to efficiently search H. For some kinds of rearrangements (for example, with transpositions), most such events will result in a very low likelihood. If the vast majority of proposed changes are rejected, it will likely take too many MCMC steps for the space to be satisfactorily sampled in a reasonable amount of time. (Methods such as simulated annealing would reduce but not fix this problem.) Therefore the schemes for proposed changes will need to chosen such that proposals will tend to have a reasonably high likelihood. This will probably be very difficult to achieve. An additional complication is that any proposed rearrangement of H will need to produce the correct number of bases in each species.

Though difficult, these problems may well be surmountable. Though effective methods for proposals of changes to H are not obvious, there has been work done (see, for example, the discussion of genomic rearrangement models in section 2.7, page 23) which suggests possible rearrangement schemes for certain kinds of events. Orthoparamap (Cannon and Young 2003) demonstrates another possible proposal method — A sufficiently accurate heuristic could be used to propose new rearrangements, which could then be altered with smaller rearrangements capable of reaching the entire space. There are, however, potential problems with this approach. The probability of making a particular proposal (needed for calculation of the Hastings ratio) could be very difficult to determine. And some heuristics – especially those which determine a single optimum set of arrangements – could systematically miss certain kinds of rearrangements, causing parts of the space to be completely unsearched.

Even if the difficulties are not solvable, there is nonetheless some merit in considering the formulation of such a complicated model, as it places these many seemingly disconnected evolutionary events into a single context. This could serve as a point of comparison with simpler models, and might provide a basis for simplifying approximations.

BIBLIOGRAPHY

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

Arvestad, L., A.-C. Berglund, J. Lagergren and B. Sennblad. Bayesian gene / species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19:7–15, 2003.

Arvestad, L., A.-C. Berglund, J. Lagergren and B. Sennblad. Gene tree reconstruction and orthology anaysis based on an integrated model for duplications and sequence evolution. In *RECOMB 04*, San Diego, California, USA, 2004.

Bailey, N. T. J. *The elements of stochastic processes*. John Wiley & Sons, Inc., New York, 1964.

Beerli, P. and J. Felsenstein. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152:776–773, 1999.

Blanchette, M., G. Bourque and D. Sankoff. Breakpoint phylogenies. In *Genome Informatics Series: Proceedings of the Workshop on Genome Informatics*, volume 15, pages 302–303. Universal Academy Press, 2004.

Blanchette, M., T. Kunisawa and D. Sankoff. Parametric genome rearrangement. *Gene*, 172:GC11–GC17, 1996.

Bourque, M. and P. A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research*, 12:26–36, 2002.

Cannings, C., E. A. Thompson and M. H. Skolnik. The recursive derivation of likelihoods on complex pedigrees. *Advances in Applied Probability*, 8:622–625, 1976.

Cannon, S. B. and N. D. Young. Orthoparamap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, 4(35), 2003.

Chen, K.-S., P. Manian, T. Koeuth, L. Potocki, Q. Zhao, A. C. Chinault, C. C. Lee and J. R. Lupski. Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nature Genetics*, 17:154–163, 1997.

Christiansen, F. B. and O. Frydenberg. Selection-mutation balance for two nonallelic recessives producing an inferior double homozygote. *American Journal of Human Genetics*, 29:195–207, 1977.

Clark, A. G. Invasion and maintenance of a gene duplication. *Proceedings of the National Academy of Science USA*, 91:2950–2954, 1994.

Clark, D. *Molecular Biology*. Elsevier Academic Press, San Diego, California, 2005.

Cotton, J. A. Analytical methods for detecting paralogy in molecular datasets. *Methods in Enzymology*, 395:700–724, 2005.

Dobzhansky, T. and A. H. Sturtevant. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.

E. Paradis. Testing for constant diversification rates using molecular phylogenies: A general approach based on statistical tests for goodnes of fit. *Molecular Biology and Evolution*, 15:476–479, 1998.

Edwards A. W. F. Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society B*, 32:155–174, 1970.

El-Mabrouk, N. Reconstructing an ancestral genome using minimum segments duplications and reversals. *Journal of Computer and System Sciences*, 65(3):442–464, 2002.

Engels, W. R. Analysis of conversion tract data with a geometric tract length distribution. Appendix to Hilliker, et al. (1994). *Genetics*, 137, 1994.

Fellows, M., M. Hallett and U. Stege. On the multiple gene duplication problem. In K.-Y. Chwa and O. H. Ibarra, editors, *Proceedings of the 9th International Symposium on Algorithms and Computation*, 1998.

Felsenstein, J. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16:183–196, 1981a.

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

Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.

Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.

Fisher, R. A. The sheltering of lethals. *American Naturalist*, 69:446–455, 1935.

Fitch, W. M. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19:99–113, 1970.

Fitch, W. M. Cautionary remarks on using gene expression events in parsimony procedures. *Systematic Zoology*, 28:375–379, 1979.

Fitch, W. M. Homology: A personal view of some problems. *Trends in Genetics*, 165(5):227–231, 2000.

Force, A., M. Lynch, B. Pickett, A. Amores and Y.-L. Yan. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545, 1999.

Garrigan, D. and S. V. Edwards. Polymorphism across an exon-intron boundary in an avian MHC class II B gene. *Molecular Biology and Evolution*, 16:1599–1606, 1999.

Geyer, C. J. Markov chain Monte Carlo maximum likelihood. In Keramidas, E.M., editor, *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface between Computer Science and Statistics*, pages 156–163. Interface Foundation, Fairfax Station, Virginia, 1991.

Geyer, C. J. and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90:909–920, 1995.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera and G. Matsuda. Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–168, 1979.

Gu, X. A simple evolutionary model for genome phylogeny based on gene content. In D. Sankoff and J. H. Nadeau, editor, *Comparative Genomics*, pages 515–523. Kluwer Academic Publishers, 2000.

Gu, X. and M. Nei. Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. *Molecular Biology and Evolution*, 16(2):147–156, 1999.

Guigó, R., I. Muchnik and T. F. Smith. Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213, 1996.

Haldane, J. B. S. The part played by recurrent mutation in evolution. *American Naturalist*, 67:5–9, 1933.

Hallett, M. T. and J. Lagergren. New algorithms for the duplication-loss model. In *RECOMB 00*, Tokyo, Japan, 2000.

Hannenhalli, S. and P. Pevzner. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of the Association for Computing Machinery*, 46(1):1–27, 1999.

Hannenhalli, S. and P. Pevzner. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of the Association for Computing Machinery*, 46(1):1–27, 1999.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark, et al. Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. *Genetics*, 137:1019–1026, 1994.

Horvath, J. E., J. A. Bailey, D. P. Locke and E. E. Eichler. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Human Molecular Genetics*, 10(20):2215–2223, 2001.

Huelsenbeck, J. P., B. Rannala and B. Larget. A Bayesian framework for the analysis of cospeciation. *Evolution*, 54(5):352–364, 2000.

J. B. Slowinski and R. D. M. Page. How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48(4):814–825, 1999.

Kendall, D. G. On the generalized "birth-and-death" process. *Annals of Mathematical Statistics*, 19:1–15, 1948.

Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

Kingman, J. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

Kirkpatrick, M. and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

Kubo, T. and Y. Iwasa. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution*, 49:694–704, 1995.

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

Kuhner, M. K., J. Yamato and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140:1421–1430, 1995.

Kuhner, M. K., J. Yamato and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–1401, 2000.

Larget, B., D. L. Simon and J. B. Kadane. Bayesian phylogenetic inference from animal mitochondrial genome arrangements (with discussion). *Journal of the Royal Statistical Society, Series B*, 64:681–693, 2002.

Larget, B., D. L. Simon, J. B. Kadane and D. Sweet. A Bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, 22(3):486–495, 2004.

Lynch, M. and A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154:459–473, 2000.

Lynch, M. and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155, 2000.

Lynch, M., M. OHely, B. Walsh and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159:1789–1804, 2001.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

Montesano, R., M. Hollstein and P. Hainaut. Genetic alterations in esophageal cancer and their relevance to etiology and pathogenesis: a review. *International Journal of Cancer*, 69(3):225–235, 1996.

Nadeau, J. H. and D. Sankoff. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147:1259–1266, 1997.

Nee, S., R. M. May and P. H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London, Series B*, 344:305–311, 1994.

Neyman, J. Molecular studies of evolution: a source of novel statistical problems. In S. S. Gupta and J. Yackel, editor, *Statistical decision theory and related topics*, pages 1–27. Academic Press, New York, 1971.

Nielsen, R. and J. Wakeley. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics*, 158:885–896, 2001.

Nowak, M. A., M. C. Boerlijst, J. Cooke and J. Maynard Smith. Evolution of genetic redundancy. *Nature*, 388:167–170, 1997.

Ohno, S. *Evolution by Gene Duplication*. Springer-Verlag, Berlin, 1970.

Ohta, T. On the evolution of multigene families. *Theoretical Population Biology*, 23(2):216–240, 1983.

Page, R. D. M. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogetics and Evolution*, 14:89–106, 2000.

Page, R. D. M. and J. A. Cotton. Genetree: A tool for exploring gene family evolution. In D. Sankoff and J. H. Nadeau, editor, *Comparative Genomics*, pages 525–536. Kluwer Academic Publishers, 2000.

Page, R. D. M. and J. A. Cotton. Vertebrate phylogenomics: reconciled trees and gene duplications. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale and T. E. Klein, editor, *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 536–547. World Scientific Publishing, 2002.

Page, R. D. M. and M. A. Charleston. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F. R. McMorris, F. S. Roberts and A. Rzhetsky, editor, *Mathematical Hierarchies in Biology*, pages 57–71. American Mathematical Society, Providence, Rhode Island, USA, 1997.

R. D. M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43(1):58–77, 1994.

Rannala B. and Z. Yang. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43:304–311, 1996.

Rannala B. and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656, 2003.

Ronquist, F. Parsimony analysis of coevolving species associations. In R. D. M. Page, editor, *Tangled Trees: Phylogeny, cospeciation and coevolution*, pages 22–64. University of Chicago Press, 2003.

Sankoff, D. Genome rearrangements with gene families. *Bioinformatics*, 15:909–917, 1999.

Simmons, M. P., C. D. Bailey and K. C. Nixon. Phylogeny reconstruction using duplicate genes. *Molecular Biology and Evolution*, 17:469–473, 2000.

Slowinski, J. B. and R. D. M. Page. How should species phylogenies be inferred from sequence data? *Systematic Biology*, 48(4):814–825, 1999.

Sokal, R. R. and F. J. Rohlf. Comparison of dendrograms using objective methods. *Taxon*, 11:33–40, 1962.

Sokal, R. R. and F. J. Rohlf. *Biometry*. W. H. Freeman and Company, San Francisco, 1969.

Suchard, M. A. Stochastic models for horizontal gene transfer: Taking a random walk through tree space. *Genetics*, 170(1):419–431, 2005.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. Phylogeny inference. In D. M. Hillis, C. Moritz and B. K. Mable, editor, *Molecular Systematics*, pages 411–501. Sinauer Associates, Sunderland, MA, USA, 1996.

Takahata, N., Y. Satta, and J. Klein. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics*, 130:925–938, 1992.

Thompson, E. A. *Human Evolutionary Trees*. Cambridge University Press, Cambridge, 1975.

Walsh, J. B. How often do duplicated genes evolve new functions? *Genetics*, 139:421–428, 1995.

Watson, J. D., N. H. Hopkins, J. W. Roberts, J. A. Steitz and A. M. Weiner. *Molecular Biology of the Gene*. Benjamin / Cummings Publishing Company, Inc., Menlo Park, California, 1987.

Yang, Z. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Systematic Biology*, 43:329–342, 1994.

Zmasek, C. and S. R. Eddy. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(14), 2002.

VITA

Lindsey Dubb was born in Burlingame, California and grew up in the San Francisco Bay Area. Currently he resides in Seattle, Washington. At the California Institute of Technology he earned a Bachelor of Science degree in Biology. In 2005 he earned a Doctor of Philosophy at the University of Washington in Genetics.