THE RATE OF LATERAL GENE TRANSFER IN *BACILLACEAE* EVOLUTION

MODELING THE RATE OF LATERAL GENE TRANSFER IN *BACILLACEAE* GENOMIC EVOLUTION

By
DANYA KONRAD, B. Sc. (HONS.)

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Masters of Science

McMaster University

MASTERS OF SCIENCE (2008)
(Biology)

McMaster University
Hamilton, Ontario

TITLE: Modeling the rate of lateral gene transfer in *Bacillaceae* genomic evolution

AUTHOR: Danya Konrad, B. Sc. Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: xii, 107

ABSTRACT

Genome evolution is not always shaped by a Darwinian-fashion of vertical inheritance from ancestral lineages. The historical gene content of a species contains many atypical gene sequences showing high similarity to those of distantly related taxa. This evolutionary phenomenon is referred to as lateral gene transfer (LGT). Lateral gene transfer permits the exchange of genetic material across lineages, completely ignoring any concept of taxonomic boundary. The rapid acquisition of foreign genes into bacterial genomes has greatly obscured the historical phylogeny of prokaryotes. In this thesis we calculate the rate of LGT on a *Bacillaceae* phylogeny, to determine the extent to which it controls species evolution. First, we examined the evolution of the phylogeny according to a simple model of maximum likelihood. We assume equal rates of gene insertion and deletion on the phylogeny and show high rates of evolution in the genomes of *B. anthracis*, *B. cereus*, and *B. thuringiensis* (Bc group), representative of adaptive evolution. We then improved the model to account for differential rates of gene insertion and deletion, thus offering a more realistic model of gene evolution. Again, we demonstrate that members of the Bc group are rapidly evolving, with the rate of gene insertion being significantly higher than the rated of gene deletion. Finally, we evaluate the sole effect of LGT on the phylogeny in a simple birth-death analysis with immigration. We show that LGT is the main vehicle of gene acquisition when the number of gene families substantially increases from external taxa to members of the Bc group. Collectively, our findings suggest that the *Bacillaceae* genome is rapidly expanding, and that laterally transferred genes may facilitate adaptive evolution and subsistence in a new niche.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my husband, Andrew. He has held my hand throughout the trials and tribulations of university life and has always shown me the light, even at my darkest times. In his continued encouragement and support I was given the necessary strength to achieve all my academic goals. His diligent work ethics and good nature have always been an inspiration to my interpersonal and intellectual endeavors. He is my best fan and number one supporter.

Secondly, I would like to thank my family. Their endearing compassion and support has helped me become the person I am today. My parents' persistent acknowledgment and praise of my academic success convinced me to pursue my dreams as a Masters student. The academic interest of my older brother, Brandon, in the sciences helped me realize my love of mathematics and motivated me to also focus my studies in science. I would also like to thank my younger brother, Nathan, for being my 'student-guinea pig' and letting me practice my role as an educator with him, from which I have gained valuable skills that are essential to my future career as a science teacher.

The enrolment and completion of my thesis could not have been achieved without support and leadership of my supervisor, Dr. Brian Golding. His continued push for excellence made me persevere at my studies, and helped me realize a newfound appreciation for computational biology. I am in constant admiration of his eminent knowledge and am truly grateful to have experienced his scholarly coaching during my time in the lab. Brian is a formidable mentor, who always seeks what is best for his students.

Finally, I would like to thank my fellow lab mates and friends who have helped make my stay at McMaster University so enjoyable. It was through the novel research of Dr. Weilong Hao that I was introduced to my thesis project. His genuine enthusiasm over microbial evolution in our many discussions on the topic made his research interests contagious, and sparked continued interest in my own thesis. I am also indebted to the rest of my lab mates, Dr. Melanie Huntley, Dr. Wilfried Haerty, Melanie Lou, Wilson Sung, and Stephanie Sun, who helped alleviate the stress of graduate work by providing optimal working conditions in a familial environment. Lastly, I could not have completed my ‘rollercoaster ride’ as a graduate student without the unconditional support of my lovely ladies, Melanie Huntley, Ally Maclean, Maria Abou Chakra, Laura Smallbone, Abha Ahuja, Alexandra Pontefract, and Andrea Morash (a.k.a The LSB Mofia), and our coffee breaks to guide me through. Their friendship is truly one I will carry with me to future endeavors.

Contents

List of Figures

List of Tables

Part I

INTRODUCTION

It has been a central focus of evolutionary inquests to reconstruct a universal 'tree of life' detailing the historical relationship among all species. With the rise of whole genome sequencing, substantial evolutionary classification could be made on the basis of gene content. As more sequences became available, sound conclusions on the hierarchal decent of a species could be drawn from larger and more reliable data sets (Gu and Zhang 2004; Snel *et al.* 2005). It soon became apparent that analyses of different gene content data yielded different tree topologies for the same phylogeny (Woese 1987; Mirkin *et al.* 2003; Gu and Zhang 2004). For example, when informational genes, those involved in transcription, translation, and replication, and operational genes, those involved in metabolism, structure, etc., are used to infer the universal phylogeny of life, both produce different trees (Rivera *et al.* 1998; Jain *et al.* 1999; Mirkin *et al.* 2003). The reason for the incongruity is attributed to the lateral transfer of genetic material across taxa, otherwise known as lateral gene transfer (LGT) (Doolittle 1999; Bushman 2002).

Lateral gene transfer, or horizontal gene transfer (HGT), is in direct apposition to Medelian inheritance where descendants inherit genes vertically from parental taxa. It introduces atypical genes into the genetic sequence (Daublin *et al.* 2003) and is characterized by genes owing high levels of similarity to genes found in distantly related taxa and by phylogentic relationships that are inconsistent with most other genes (Gogarten *et al.* 2002). It has greatly obscured the evolutionary relationship among the Archaea, Bacteria, and Eukarya domains, as well as, the identification of a sole last common ancestor (Woese 1987; Doolittle 1999). The exchange of genetic material across the universal phylogeny is evident in many taxonomic genomes, including the acquisition of bacterial mitochondria and chloroplast into the Archaea and Eukarya domains (Doolittle 1999). As such, the last common ancestor for the evolutionary tree of life is usually depicted as a community of interacting species that consistently trade genetic information, rather than a single organism.

The exchange of genetic information from LGT can occur from: viral transduction of a gene into host DNA, transformation of a gene from the surrounding environment, or direct physical transfer of a gene from another cell (Bushman 2002; Gogarten *et al.* 2002). In order for a newly transferred gene to survive, it must undergo cell division and be inherited with ancestral gene content (Berg and Kurland 2002; Bushman 2002). Those genes that successfully persist in the phylogeny may provide a selective advantage to the organism, helping it readily adapt to adverse environmental conditions (Lan and Reeves 1996; Rivera *et al.* 1998; Gogarten *et al.* 2002; Daubin *et al.* 2003a; Jain *et al.* 2003; McLysaght *et al.* 2003; Hao and Golding 2004, 2006; Lake and Rivera 2004; Novozhilov *et al.* 2005; Marri *et al.* 2006, 2007). The impact of LGT on genome innovation is most prominent in prokaryotic evolution (Lan and Reeves 1996; Gogarten *et al.* 2002; Snel *et al.* 2002; Jain *et al.* 2003; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Galtier 2007; Linz *et al.* 2007; Marri *et al.* 2006, 2007). By introducing novel genes into the genome, LGT induces genetic divergence and can assist the evasion of antibiotics (Berg and Kurland 2002; Gogarten *et al.* 2002; McLysaght *et al.* 2003), invasion of new hosts (Doolittle 1999; Daubin *et al.* 2003a; Mirkin *et al.* 2003; Marri *et al.* 2006), and growth in the presence of pollutants (Bushman 2002). Because LGT is so common in prokaryotes, bacteria are emerging as fundamental tools in studying the influence of LGT on phylogenetic evolution.

The incidence of LGT in prokaryotes has been explored by a variety of approaches, including: distance-based methods (Snel *et al.* 1999), parsimony (Snel *et al.* 2002; Daubin et al. 2003a, b; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Hao and Golding 2004), birth-and-death models (Berg and Kurland 2002; Gu and Zhang 2004; Huson and Steel 2004; Novozhilov *et al.* 2005), genome signatures (Karlin *et al.* 1997; Karlin 1998; Karlin *et al.* 1999), and maximum likelihood analyses (Gu 2001; Kunin and Ouzounis 2003; Huson and Steel 2004; Lake and Rivera 2004; Hao and Golding 2006; Marri *et al.* 2006, 2007; Linz *et al.* 2007). Most methods use gene content data to infer incidences of gene gain and gene loss on a phylogeny. In distance models, the degree of similarity between two genomes is used to infer the extent of LGT in shaping gene content. Evolutionary distance is quantified as the relative proportion of shared genes between the taxa and construes the phyletic pattern of gene gain and gene loss (Snel *et al.* 1999). Measures of taxonomic similarity can also be inferred from genome signatures. An organism's genomic signature defines the relative abundance of dinucleotides in a sequence and is used in comparison with other genome signatures to determine the degree of divergence between taxa (Karlin *et al.* 1997; Karlin and Mrazek 1997; Karlin 1998; Karlin *et al.* 1999). Here, an evolutionary relationship is constructed on the basis that closely related taxa have more similar genome signatures than distantly related taxa (Karlin *et al.* 1997; Karlin and Mrazek 1997; Karlin 1998; Karlin *et al.* 1999).

The significance of LGT in driving prokaryotic evolution has also been noted in current models of maximum parsimony (Snel *et al.* 2002; Daubin *et al.* 2003a, b; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Hao and Golding 2004). Maximum parsimony aims to reconstruct phylogenies based on genomic arrangements representative of minimal evolutionary change (Felsenstein 1988, 2004; Mirkin *et al.* 2003). Some models may apply a gain penalty with LGT to obtain rates more reflective of a natural evolutionary course (Snel *et al.* 2002; McLysaght *et al.* 2003). Once the most parsimonious scenario of gene insertion and deletion is assumed on the phylogeny, the extent to which LGT effects genomic context can be inferred.

Gene insertions and deletions have also been modeled according to a stochastic birth-and-death process of evolutionary growth (Berg and Kurland 2002; Kurland *et al.* 2003; Gu and Zhang 2004; Huson and Steel 2004; Novozhilov *et al.* 2005). In this approach, mathematical equations describing gene gain (birth) and gene loss (death) are used to simulate probable scenarios of genome evolution. Some methods employ sophisticated algorithms to monitor changes in genomic populations when the additional forces of mutational inactivation, selection, intraspecific horizontal gene transfer, population size (Berg and Kurland 2002) and interspecific horizontal gene transfer (Novozhilov *et al.* 2005) are assumed to affect the rate of gene insertion and deletion. Other models utilize simple gene content data to infer the basic evolutionary rates of gene gain and loss (Gu and Zhang 2004).

Recently, models describing the role of gene insertion and gene deletion on bacteria evolution have adopted the powerful approach of maximum likelihood (Gu 2001;

Kunin and Ouzounis 2003; Huson and Steel 2004; Lake and Rivera 2004; Hao and Golding 2006; Marri *et al.* 2006, 2007; Linz *et al.* 2007). Likelihood methods try to approximate a phylogenetic tree that maximizes the probability of the observed data according to a specified model (Felsenstein 1988, 2004; Brocchieri 2000). Often phyletic patterns of gene presence and absence are cataloged to identify the rates of gene insertion and deletion required to maximize the outcome of the proposed model (Gu 2001; Kunin and Ouzounis 2003; Huson and Steel 2004; Hao and Golding 2006; Marri *et al.* 2006, 2007). Statistical methods, like the Markov model (Lake and Rivera 2004; Galtier 2007) and the Poisson distribution (Linz *et al.* 2007), have also been successfully incorporated in the likelihood framework to infer the rate of LGT on prokaryotic evolution. The intent of our research is to assess and improve the modeling of LGT with the goal to better understand the dynamics of this important process. Using a group of Gram-positive Bacillus bacteria, we successfully reconstruct the evolutionary history of the phylogeny and provide a rigorous portrayal of LGT in shaping the genome via our constant-rate maximum likelihood model, differential-rate maximum likelihood model, and birth-death model.

First we model the rate of LGT according to a constant-rate maximum likelihood scheme. Equal rates of gene insertion and deletion were assumed on each branch and phyletic patterns of gene presence and absence were used to infer the rates required to maximize the likelihood. The optimal rates of evolution were determined for four different rate cases: a single constant rate across all branches ($\alpha = \beta = \gamma$), a rate distinguishing members of *B. anthracis*, *B. cereus*, and *B. thuringiensis* (the Bc group) from the rest of the phylogeny (α, $\beta = \gamma$), an additional rate along the branch leading to the Bc group (α, β, γ), and branch-specific insertion/deletion (indel) rates ($\alpha_1,\alpha_2,\ldots,\alpha_{23}$). An algorithm that continually tests the boundary of a subset of three rate estimates was used to infer the optimal rates. Statistical testing via bootstrap sampling and standard deviation measures from the likelihood curve confirm the robustness of the constant-rate likelihood model. Results of the likelihood ratio test and Akaike Information Criterion (AIC) both suggest that the evolutionary history of the *Bacillaceae* bacteria is best portrayed when branch specific indel rates are assumed across the phylogeny.

The likelihood model was then modified to calculate differential rates of gene insertion and gene deletion for the *Bacillaceae* phylogeny. By acknowledging that the rates at which a gene is inserted or deleted need not be equal, the enhanced model provides a more accurate reflection of bacterial evolution. Powell's (1964) maximum convergence algorithm was applied to the above rate cases and gene presence/absence patterns, to infer the optimal rate(s) of gene insertion and gene deletion. Higher rates of gene insertion were observed in most rate cases, signifying genome growth and possible adaptive evolution. Strong statistical support for the differential rate conditions is granted in the small estimates of standard deviation in the bootstrap samples and likelihood curve measurements. Once again, the likelihood ratio test and AIC confirm that genome evolution is most accurately modeled when different rates of gene insertion and gene deletion are assigned to each branch.

After the role of LGT on *Bacillaceae* evolution was extensively explored in the maximum likelihood analyses, a birth-death model of genomic growth was applied to infer the optimal rates of gene insertion and deletion. Interesting conclusions can be drawn when different models of evolution are assumed on the same phylogeny, as congruencies between the models can assist in identifying the true nature of genome evolution. In the birth-death model, a separate 'immigration' parameter was used to determine the sole rate of LGT on the phylogeny. Therefore, gene 'birth' results from either a duplication or LGT event and gene 'death' results from a deletion. Phyletic patterns of gene families were used to infer the optimal rate of gene insertion and gene deletion according to a maximum likelihood scheme. Only the simple evolutionary scenario of an equal duplication, LGT, and deletion rate across all branches was assumed on the phylogeny. Further improvement of our current birth-death model to infer the evolutionary rates of the other rate models using a larger set of gene patterns, will help provide a more accurate depiction of *Bacillaceae* evolution.

Chapter 1

Maximum likelihood model of *Bacillaceae* evolution under equal rates of gene acquisition and loss

1.1 ABSTRACT

Maximum likelihood models are emerging as important tools in identifying the evolutionary forces that drive bacterial genomic growth. The statistical framework of the method allows for rapid and efficient calculation of those parameters required to maximize the outcome of a proposed model. In phylogentic studies, likelihood models can predict the influential forces of gene insertion, gene deletion, and/or lateral gene transfer (LGT) that shape the bacterial genome, thus revealing the unique evolutionary relationships that exists among the species. We applied a likelihood-based approach to a group of thirteen closely related *Bacillaceae* species, to reveal the genetic history of the group. Four evolutionary rate scenarios were assumed on the phylogeny and the

insertion/deletion rate(s) (indel rate) of each were calculated. The optimal rate parameters produced by the model were then statistically verified via estimates of standard deviation inferred from bootstrap sampling and variance in the likelihood curve. Both statistical tests indicate little deviation in the predicted optimal indel rates, thus supporting the assumptions of our maximum likelihood model. Through the innate consistency and robustness of such models, confidence in identifying the true evolutionary mechanism governing bacterial genome growth is granted.

1.2 INTRODUCTION

Models of maximum likelihood often serve as the general means of statistical inference in phylogenetic reconstruction studies. In fact, with the expansion in computational capacity, maximum likelihood has been deemed superior to other methods of statistical inference, like parsimony and distance matrix methods (Huelsenbeck 1995). The method of maximum likelihood was founded by Fisher (1912) (Felsenstein 2004) and its statistical implications for phylogenetic reconstruction were justified by Felsenstein (1981) (Schrago 2006). In phylogenies, likelihood analyses provide estimates of branch length or infer the hypothetical gene states of ancestor taxa (Felsenstein 1988). The innate sufficiency of the model allows it to predict parameter estimates from which no other statistics, based on the same data set, can provide additional information (Fisher 1922). Furthermore, accuracy in predicting the true parameter values is observed to increase with the number of tested sample sets (Fisher 1922; Felsenstein 1988). For these reasons, the statistics of maximum likelihood have proved extremely beneficial in the reconstruction of phylogenies from genomic data.

Maximum likelihood is a simple statistical test that is utilized in a vast array of computational algorithms. Essentially, likelihood methods try to approximate an unknown parameter that maximizes the outcome of a known model. As defined by Fisher (1922), the frequency of this event is a likelihood rather than a probability because it depends on the occurrence of past events and, as such, the predicted outcomes may not sum to one. Maximum likelihood is an ideal statistic because it tests for the parameters best suited to fit the model when varying restrictions of the model are imposed on the same data set. In phylogenetic analysis, models of maximum likelihood attempt to identify the phylogenetic arrangement that best represents the evolutionary succession of a given genome. This is done by estimating the phylogenetic tree (*T*) that maximizes the probability of the observed data (*D*) under a specific model (*M*) (Felsenstein 1988, 2004; Brocchieri 2000). The statistical relationship is formally represented in the conditional probability:

$$P(D|\, T, M).$$

Many phylogenetic reconstruction studies (Gu 2001; Huson and Steel 2004; Lake and Rivera 2004; Hao and Golding 2006) have utilized the aforementioned relationship to describe the role of gene insertions, gene deletions, and lateral gene transfer (LGT) on bacterial evolution. In their model, Hao and Golding (2006) were able to determine the rate of evolution for a group of Gram-positive *Bacillaceae* bacteria under the assumption that genes are inserted and deleted at equal rates. The group consisted of thirteen completely sequenced *Bacillaceae* genomes of high similarity: *B. anthracis* Ames, *B. anthracis* "Ames ancestor," *B. anthracis* Sterne, *B. thuringiensis*, *B. cereus* ZK, *B. cereus* ATCC 10,987, *B. cereus* ATCC 14,579, *Geobacillus kaustophilus*, *B. licheniformis*, *B. subtilis*, *B. clausii*, *B. halodurans*, and *Oceanobacillus iheyensis*. Because the gene sequences of *B. anthracis*, *B. cereus*, and *B. thuringiensis* are so similar, they were grouped as the Bc group. The predicted phylogeny (see Fig. 1.1) was assumed to evolve according to three separate rate cases: a single constant rate, α (Case 1 in Fig. 1.2); two rates, α and β, separating the Bc group from the rest of the phylogeny (Case 2 in Fig. 1.2); and three rates, α, β, and γ, where γ defines the rate along the branch leading to the Bc group (Case 3 in Fig. 1.2). The rates for each case were estimated from the observed gene presence/absence patterns inferred for the phylogeny (see Table 1.5, for a list and frequency of the most common phyletic gene patterns). Those rate values that maximized the likelihood (see Table 1.1) were denoted optimal for genomic evolution. Although their model of maximum likelihood was successful in predicting the indel rate for the *Bacillaceae* group, the robustness of the model must be verified through statistical scrutiny. Only after the statistics of the model are examined, can the fittingness of the inferred parameter values to the algorithm be confirmed.

Statistical tests are often employed to verify the robustness of a theoretical model under varying statistical environments. With current improvements in computational technology, the solution of many intricate statistical algorithms can be achieved at the aid of a computer (Efron and Tibshirani 1991). One such method of computational efficiency is the bootstrap. Bootstrap estimates are fairly unbiased (Efron and Tibshirani

1986) and may be applied to almost any statistic (Efron 1979a; Efron and Gong 1983; Efron and Tibshirani 1986). The error of a sample set is assessed in terms of bias, standard error (Efron and Gong 1983), variance of the sample mean (Efron 1979a), and/or prediction error (Efron and Tibshirani 1986). These estimates help determine the accuracy of the original data. The bootstrap algorithm operates on the premise of a random number generator that draws random points, $x_1, x_2,..., x_\mathrm{n}$, from the original data set, Z. Each point is drawn independently and with replacement, thus generating an independent and random bootstrap sample, $X = (x_1, x_2,..., x_n)$, of the original data set (Efron 1979b). The procedure is repeated for a large number of trials, m, to establish a sufficient set of randomly independent bootstrap samples, $Y(1)$, $Y(2),\ldots,Y(m)$, required for statistical testing. A sample size of 1000 bootstraps estimates is usually adequate for error analysis (Efron and Tibshirani 1986). This form of bootstrapping is typically referred to as the Monte Carlo approximation of the bootstrap distribution (Efron 1979a, Efron and Gong 1983; Efron and Tibshirani 1986). Often, estimates of standard deviation are used to infer the statistical accuracy of a sample. The standard deviation of a bootstrap reflects the theoretical error that occurs when an arbitrary data sample is identical to the observed data distribution (Efron and Tibshirani 1991). As such, the standard error of the original sample can be inferred directly from the standard deviation of the bootstrap data. Standard deviation can also be linked to the estimated parameter mean of the bootstrap sample. This correlation creates an interval of confidence from which many statistically accurate measures of error may be inferred (Efron and

Tibshirani 1986). Both statistical attributes, standard deviation and mean, were considered in testing the certainty of the original data.

Variance assessment of the likelihood curve is another means of statistical validation. Altering the parameter values in the proposed model generates a curve representative of the likelihood function. From the graphical surface, the statistical relevance of the observed results is affirmed. Geometrical analysis begins by evaluating the negative second derivative of the likelihood function at the maximum estimate, otherwise known as curvature (Edwards 1972; Felsenstein 1988; Schrago 2006):

$$\text{Curvature} = -\frac{d^2L}{d\theta^2}\ .$$

The reciprocal of this value:

$$1/\text{Curvature} = -1/\frac{d^2L}{d\theta^2}$$

taken at the maximum estimate is the observed formation, and corresponds to the radius of curvature of the support function (Edwards 1972). The square root of the observed formation approximates the standard deviation, or span, of the maximum likelihood estimate (Edwards 1972). It approximates an interval of confidence, reflective of the width of the likelihood curve at the maximum value (Edwards 1972), from which the statistical significance of the proposed model is inferred. Confidence intervals symmetrical about the estimate (Edward 1972) and of limited range usually confirm the robustness of the model. Geometrically, this narrow span is representative of functions of extreme curvature with rapidly decreasing slopes on either side of the maximum. Such centralization of the function provides a precise estimate of the true maximum likelihood value. In addition, support for the estimate is anticipated to increase with increasing sample size (Edwards 1972). The statistical implications of these confidence intervals were used to reveal the inherent variation of the original test sample.

The intent of the proposed research is to assess the reliability, via bootstrap testing and calculation of standard deviation from the likelihood curve, of the gene insertion/deletion rates inferred by Hao and Golding (2006) for the *Bacillaceae* group. For each rate model they assumed on the phylogeny (Case 1 – Case 3 in Fig. 1.2), the bootstrap data and standard deviation calculations will confirm the accuracy of the predicted maximum likelihood estimates. An additional rate model of branch specific insertion/deletion rates, α_1, $\alpha_2,\ldots,\alpha_{23}$, will be assessed in like manner (Case 4 in Fig. 1.3). Divergence between the original and bootstrapped data will be evaluated according to the mean and standard deviation obtained for the parameters of each model. Deviation of the bootstrap mean from the original insertion/deletion rates might indicate possible bias in the original data set. Further uncertainty in the primary data set will be measured by estimating the standard deviation of the maximum likelihood curve. This deviation calculation is intuitive of the relative distance between the estimated rate value and the expected rate value. Incidences of low deviation indicate little departure from the inferred rates, with certainty in the estimated values granted in narrow error margins. Such localization will help establish the true maximum likelihood, and thus, validate estimates of the original data. Taken together, the statistical implications attained by both the bootstrap and variance calculations will provide valuable insight regarding the ideal rates of bacterial gene insertion/deletion.

1.3 THE MODEL

The maximum likelihood model of Hao and Golding (2006) offers a simple algorithm that quickly and efficiently solves the likelihood of a given phyletic pattern. The model assumes independent evolution of phylogenetic site and lineage so that the likelihood at each point may be determined separately. The probability equations used in the likelihood analysis are formed on the basis that a gene present at a descendant site may be determined from the known ancestral gene state. Thus, assuming equal rates of gene insertion and deletion, the conditional probabilities:

$$Prob(P_d|P_a,t) = v/(u+v) + e^{-(u+v)t}[1\text{-}v/(u+v)],$$

$$Prob(A_d|P_a,t) = u/(u+v) - e^{-(u+v)t}[1\text{-}v/(u+v)],$$

$$Prob(P_d|A_a,t) = v/(u+v) - e^{-(u+v)t}[1\text{-}u/(u+v)],$$

$$Prob(A_d|A_a,t) = u/(u+v) + e^{-(u+v)t}[1\text{-}u/(u+v)],$$

represent all possible gene states of descendant taxon, where P is gene presence, A is gene absence, d is descendant node, a is ancestral node, v is the rate of gene insertion and u is the rate of gene deletion (Hao and Golding 2006). These probability relationships may be simplified to the form:

$$Prob(P_{d,a}|P_{d,a},t) = 1/2 + (1 + e^{-2t}).$$

The likelihood of a given pattern is obtained by calculating the probability that a gene (x) is present at an ancestor node (G), given the genetic states of the descendants (E and F), see Figure 1.4. This practical reconstruction method makes use of the recursion principle, and thus, was appropriately coined pruning by Felsenstein (2004). Here, the likelihood of the ancestral genetic state is represented by the product of the observed gene present/absent patterns in the descendants (Hao and Golding 2006):

$$L^x_{\mathrm{G}}(P) = (Prob(P_d| P_a, t_1) * L^x_{\mathrm{E}}(P) + Prob(A_d| P_a, t_1) * L^x_{\mathrm{E}}(A)) \\ x \ (Prob(P_d| P_a, t_2) * L^x_{\mathrm{F}}(P) + Prob(A_d| P_a, t_2) * L^x_{\mathrm{F}}(A)).$$

Therefore, starting likelihood calculations at the tip of the tree requires the likelihood of a gene being present in either descendant (E or F) to be one, $L^x(P) = 1$, and the likelihood of it being absent to be zero, $L^x(A) = 0$. Another simplification of the model is that the likelihood of the ancestral genetic state (x) is determined equally by gene presence and absent patterns at the root (Hao and Golding 2006):

$$Q^x = (L^x_{\mathrm{G}}(P) + L^x_{\mathrm{G}}(A))/2.$$

Gene sites absent from the data must also be considered in the model. Hao and Golding (2006) accounted for this by adopting the approach devised by Felsenstein (1992) in his restriction site analysis of phylogenies. For such data, likelihood calculations of the genetic state are based on the probability of the gene being present in at least one species:

$$Q^x_+ = \frac{Q^x}{1 - Q^x_-}$$

where Q^x_- is the likelihood of the gene being absent in all species. Once an estimate of likelihood is established for a single point, the value is used to infer the transition probability of the successive ancestor node. This pattern continues upwards the phylogenetic tree. At the root of the tree, the overall likelihood of the phylogeny is the sum of the conditional likelihoods (Felsenstein 1992, 2004). Therefore, the sum likelihood of a tree with n phylogentic sites is (Hao and Golding 2006):

$$Q = \prod_{x=1}^{n} Q^{x}{}_{+}.$$

Often, the likelihood function is reverted to its log counterpart:

$$\log(Q) = \sum_{x=1}^{n} \log(Q^{x}{}_{+}),$$

so that the mathematical relations of extremely small likelihood values are easier to access. This simplified algorithm was employed to calculate likelihood, under the assumption of an equal gene insertion and deletion rate, in the statistical testing of the model.

1.4 METHODS

Genomic rate data for the *Bacillaceae* group was statistically tested to verify the results obtained by Hao and Golding (2006). In their model, genomic evolution was inferred under the assumption that genes are inserted and deleted at equal rates. Thirteen completely sequenced *Bacillaceae* genomes of high similarity were examined to reconstruct the evolutionary history of the bacterial group. The frequency of gene presence and absence patterns across the phylogeny were assessed in a maximum likelihood analysis, used to estimate the optimal indel rate of evolution.

A bootstrap algorithm described by Efron (1979b) was first implemented to test the significance of the estimated insertion/deletion rates. For computational efficiency, a script was written to perform the bootstrap analysis. Bootstrap samples where generated using a random number generator to randomly select a phylogenetic pattern, from the original 7228 gene presence/absence patterns tallied for the *Bacillaceae* group (Hao and Golding 2006). Each pattern was drawn independently and with replacement from the original data set. A total of 7228 gene patterns were selected from the original presence/absence patterns to establish a new bootstrap sample. In total, 1000 bootstrap samples were taken for each rate parameter, thus establishing 1000 independent estimates of each variable in a rate case. All of the rates cases defined in Figure 1.2 *(*$\alpha = \beta = \gamma$; α, $\beta = \gamma$; and α, β, γ), and the additional case of branch specific insertion/deletion rates, $\alpha_1,\alpha_2,\ldots,\alpha_{23}$ (Fig. 1.3), were considered in the bootstrap analysis. A script borrowed from a fellow colleague, Dr. Weilong Hao, was used to calculate the optimal insertion/deletion rates using an algorithm called "golden". The algorithm golden uses a bracket interval to pinpoint the best estimate(s) of gene insertion/deletion required to maximize the likelihood. Once 1000 independent estimates of each rate variable in a rate case was obtained, the data was analyzed using the statistical package STATA 7.0 (Stata Corporation, College Station, TX). The mean and standard deviation of the inferred insertion/deletion rates was calculated for each of the parameters. From these statistical estimates, a confidence interval for each rate variable was established and used to assess the accuracy of the original model of gene evolution (Hao and Golding 2006). The confidence interval for each variable is scored in Table 1.2.

The statistical variance of the actual gene model (Hao and Golding 2006) was evaluated from the observed curvature at the maximum likelihood estimate. This graphical interpretation of variance follows the maximum support method devised by

Edwards (1972). For each of the defined rate cases, the rate parameters were altered to establish values reflective of the true likelihood function. Each rate parameter was altered separately, thus generating a likelihood curve representative of only that estimate. The likelihood curve for each rate was obtained by increasing and decreasing the optimal rate values by 1%, respectively. The degree of increase and decrease for consecutive rate estimates was kept consistent to establish values reflective of a symmetrical distribution. A total of nine rates: four rates greater than the optimal rate, four rates less than the optimal rate, and the optimal rate, were plotted against their associated likelihoods in the statistical software system STATA 7.0. In order to get single variable estimates, a two dimensional approach was adapted to graph changes in likelihood when the rate of insertion/deletion varied for only one parameter while the others remained constant. A likelihood function of the form:

$$L(\theta) = a\hat{\theta}^2 + b\hat{\theta} + c$$

was estimated, via regression analysis, for each individual rate parameter. From the second partial derivative of this quadratic function:

$$d^2L/d\theta^2 = 2a$$

a measure of curvature at the maximum estimate was inferred. The negative inverse of this value:

$$- [d\theta^2/d^2L] = -1/2a$$

is an estimate of the statistical variance of the suggested model. From the variance estimate, a measure of standard deviation can easily be calculated. Values for standard deviation were used to assess the validity of the data produced by the model. These values are tabulated in Table 1.2 in the corresponding insertion/deletion rate cases.

To evaluate the parameter restrictions that best fit the maximum likelihood model, the likelihoods estimated for each rate case were compared using the likelihood ratio test. When multiple hypotheses (H_1 and H_2) are developed for the same model, it is important to identify that distribution with parameters best suited for the model. The superior model often consists of those parameters that generate the greatest maximum likelihood under the given assumptions (Fisher 1922). As defined by Fisher (1922), the likelihood ratio test is simply a comparison ratio of the likelihood of one hypothesis to another:

$$\Lambda(x) = L_1/L_2,$$

where x defines the data set. The more reliable hypothesis, based on the assumed parameter distributions, is the one that returns the observed results more frequently (Fisher 1922). For example, if the likelihood ratio of H_1:H_2 is 4:1, then we would expect H_1 to return more reliable results, more often. Although Fisher's (1922) comparison ratio tells us the relative expectancy of a given hypothesis, it does not explicitly assess the quality of the decision. In order to verify which model best describes the observations, the significance of the predicted likelihoods need to be statistically evaluated. One such test statistic is Wilks (1938) chi-square approximation of the likelihood ratio test. In his theorem, Wilks (1938) states that the likelihood ratio test asymptotically follows a chi-square distribution. The relationship is formally represented by:

$$-2 \ln \lambda = \chi^2$$

where λ is the likelihood ratio and the degree of freedom is the difference in the number of parameters between the two hypothesis. This chi-square approximation was used to evaluate the statistical significance of the maximum likelihood values obtained for the different rate models (see Table 1.3).

The statistics of a model often depend upon the restrictions imposed on the parameters. Therefore, to ensure the number of parameters did not falsely identify the superiority of a model, the estimated likelihoods were also evaluated according to the Akaike Information Criterion (AIC) (Akaike 1972). The AIC is founded on principles of maximum likelihood estimation and information theory criterion. It uses the Kullback-Leibler (1951) definition of information:

$$I(H_1,H_2) = \int H_1(x) \log[H_1(x)/H_2(x)]dx,$$

where H_1 and H_2 are defined as before, to identify the optimal estimate based on the informational divergence (separation/distance) between two models (Akaike 1972). In order to determine the number of parameters that best fit the suggested model, the AIC of each rate case was calculated. The AIC of a model is defined by:

$$\mathrm{AIC} = -2 \ln(L) + 2k,$$

where L is the maximum likelihood and k is the number of parameters of the model (Akaike 1972). The parameter restrictions that result in the minimal information theoretic criterion estimate (MAICE) are the best fit for the model. If two hypotheses estimate the same MAICE then the principle of parsimony is considered, where the model with fewer parameters denotes the superior choice (Akaike 1974). Support for a given model was evaluated by taking the difference between the MAICE and the alternate hypothesis. In general, the greater the distance between the two hypotheses, the more likely the tested hypothesis is not the best model, given the data. Burnham and Anderson (2002) developed a general support scheme based on the difference in AIC values, to assess the statistical support for an alternative model: substantial support (0-2), considerably less support (4-7), and essentially no support (>10). The statistics of the AIC are listed in Table 1.4.

1.5 RESULTS

The maximum likelihood analysis performed by Hao and Golding (2006) was statistically tested using the bootstrap and curvature methods of statistical inference. Support for an observed insertion/deletion rate was evaluated in terms of closeness to the statistically inferred estimate. Overall, little deviation from the proposed rate was observed in both the bootstrap and curvature approximations (see Table 1.2). Even the gene presence/absence patterns predicted by the bootstrap testing (Table 1.6) were very similar to the observed phylogenetic patterns (Table 1.5). These results are indicative of robust rate estimates for the *Bacillaceae* group.

Under the assumption of a sole constant rate of gene evolution (Case 1 in Fig. 1.2), the statistics of the bootstrap sample and likelihood curve were consistent with the proposed insertion/deletion rate. The confidence interval of the bootstrap approximation is very narrow, 0. 5187 ± 0.0252, with a sample mean almost identical to the predicted optimal rate of 0.5175. Additionally, from the acute curvature of the maximum likelihood curve, little deviation, 0.0069, is observed in the data set.

Following the grouping of the Bc group by Hao and Golding (2006), statistical verification was performed on the case of two separate rates (Case 2 in Fig. 1.2), α and β, influencing the rate of *Bacillaceae* gene evolution. Branches in the Bc group were estimated to evolve under the assumed rate of α, while the rest of the phylogeny evolved at the rate of β. The parameter means of the bootstrap sample for α and β, 4.572 and 0.3487 respectively, are very similar to the purposed indel rates of 4.564 and 0.3487, respectively. Estimates for standard deviation also indicate little divergence in the suggested rate values, with lower deviation observed in rate β. Standard deviation estimates from the likelihood curve are relatively analogous to the bootstrap results, predicting little variation in the rate data with less variation in β, 0.0051, than in α, 0.1098.

Based on clear differences in the rate of genome evolution between the Bc group and the rest of the phylogeny, a third rate scenario was developed to model the rate of *Bacillaceae* evolution. A new rate γ was used to describe gene evolution on the branch leading to the Bc group (Case 3 in Fig. 1.2). In the bootstrap analysis, the estimated sample means for α, β, and γ, 4.015, 0.2836, and 1.274 respectively, are comparable to their observed insertion/deletion rates, 4.011, 0.2837, and 1.273 respectively. Once again, estimates of standard deviation from the bootstrap sample and the likelihood curve are in close approximation of the observed rate values. In the bootstrap data, the measure of deviation for β is less than that of γ, followed by α. A similar pattern is observed in the curvature estimates of the likelihood curve, where the values for α, β and γ, are 0.0899, 0.0045, and 0.0430 respectively.

An additional case of *Bacillaceae* rate evolution was also considered in the extended study. Individual insertion/deletion rates were assigned to each branch in the phylogeny (α_1, α_2,..., α_{23}), to establish rate parameters unique to the evolutionary history of each member (Case 4 in Fig. 1.3). Rates were ordered from left to right on the phylogeny, beginning at the most recent members, the BC group. Overall, the statistics of the bootstrap samples and likelihood curve show little deviation in the predicted optimal rates and are in agreement with each other. In the bootstrap data, the estimated insertion/deletion rates for members outside of the Bc group are nearer the optimal rate values. The confidence intervals of these members are also more limited in their range. Only small variation in the standard deviation values of rates α_{13}, α_{17}, and α_{23}, is noted between the two test statistics (see Table 1.2). Rate α_{20}, describing the branch leading to Bk and Bh, is observed to have the highest deviation in both data sets, with an estimate of zero curvature.

When the likelihood ratio test was applied to the estimated maximum likelihood of each rate case, greater support for a model is observed as the number of parameters increase (see Table 1.3). The parameter distribution supporting the best model is that of rate Case 4 (Fig. 1.3), with a maximum likelihood of -34864.39 ($\chi^2 = \Delta 2$ LnL > 31.41 with d.f. = 20). The results of the AIC test also agree with those of the likelihood ratio test (see Table 1.4). A sequential increase in the goodness of fit of parameters to a given model is noted down Table 1.4, as the distance between the AIC values increase. Again, rate Case 4 appears to best represent the genomic data, owing to a MAICE of 69774.78.

1.6 DISCUSSION

In order to justify the assumptions of a given model, the statistics of alternative hypotheses need to be analyzed and evaluated to determine the criterion that best fits the data. The preferred hypothesis is usually the one with the greatest consistency and agreement among its test statistics. The alternate rate hypotheses (Case 1 – Case 3 in Fig. 1.2) suggested by Hao and Golding (2006) and the additional case of independent branch evolution (Case 4 in Fig. 1.3), were evaluated in terms of the maximum likelihood model. The statistics of each rate model were compared to determine the number of parameters that best fit the genome data. Together, the statistical results reveal that rate Case 4 most accurately models the rate of evolution governing the *Bacillaceae* phylogeny.

The bootstrap and maximum likelihood curve results are quite consistent for the three rate cases (Case 1 – Case 3) assumed by Hao and Golding (2006). The mean rate values predicted by bootstrap sampling are extremely close to the estimated optimal insertion/deletion rates. Measures of standard deviation also support the inferred indel rates and, together with the bootstrap mean, create narrow confidence intervals indicating little divergence between the data sets (see Table 1.2). Similar results are observed in the statistical analysis of the likelihood curve. The standard deviation values obtained for the three rate cases are comparable to those obtained in the bootstrap samples. Only the estimate for rate α in Case 3, 0.0899, varies slightly between the two test statistics.

Like the statistical results of the first three rate cases, little deviation is noted between the bootstrap samples and curvature data in the fourth rate case (Case 4 in Fig. 1.3). Most of the estimated standard deviation values agree between both test statistics. Only slight variation is noted in some of the values obtained in the curvature analysis: α_{13}, 0.0089, α_{17}, 0.0981, and α_{23}, 0.0696. Rate α_{20} is observed to have infinite variation in the curvature data. Altering the optimal indel rate of α_{20} resulted in no difference in the maximum likelihood estimate. As such, when the rate values were plotted against their associative likelihoods, the graph resembled a horizontal line, indicating infinite deviation in the rate estimate. The predicted bootstrap means are very close to the optimal insertion/deletion rates estimated by the maximum likelihood model. Again, the bootstrap mean and standard deviation of each rate establish narrow confidence intervals, thus providing continued support for the likelihood model.

The results of the likelihood ratio test identify rate Case 4 as the most statistically significant hypothesis. As the number of parameters increase from Case 1 – Case 4, there is a clear succession of increase in the estimate for maximum likelihood, with Case 4 having the highest estimate of -34864.39. Doing a chi-square comparison of the maximum likelihoods obtained from the different rate cases, confirms that the indel rate of *Bacillaceae* evolution is best described by the parameter assumptions of Case 4 ($\chi^2 = \Delta 2$ LnL > 31.41 with d.f. = 20). To ensure that the choice of Case 4 as the best likelihood method is not simply a product of the number of parameters defined in the model, the different rate hypothesis were also evaluated according to the AIC. This statistical identification method tests for the superior hypothesis based on the maximum likelihood estimate and number of parameters, rather than on levels of significance (Akaike 1974). It offers a practical way to mathematically identify the best likelihood model among a series of hypothesis. The AIC measures certainty in the different parameter restrictions assumed by various models, all of which must be derived from the same number of observations (Akaike 1974). The strength of a result is ranked as the degree of informational difference between competing hypotheses. The greater the distance between two hypotheses, the more likely the challenging hypothesis is not the best model. One of the most prominent features of the test is that the order of computing the AIC for alternate hypothesis is not important (Akaike 1974). Therefore, multiple hypotheses can be arbitrarily compared, regardless of the sequential increase in parameter restrictions. Applying AIC to the four defined rate cases, reveals that the number of parameters in rate Case 4 give the MAICE, 69774.78, and are best fit for the likelihood model. These results confirm those obtained by the likelihood ratio test, and provide further support for the parameter assumptions of rate Case 4. Although both test statistics identify Case 4 as the 'best-fit' model, it is important to note that it is only the best model out of the hypotheses offered. Also, the results of the tests are only as good as the data and observations they originate from. Therefore, based on the hypotheses offered (Case 1 – Case 4), the rate of *Bacillaceae* evolution is most accurately portrayed when branches are assumed to change independently, according to branch specific indel rates.

1.7 ACKNOWLEDGMENS

This research was supported by a National Science and Engineering Research Council of Canada (NSERC) grant to G.B.G.

Table 1.1. The optimal indel rates and the associated maximum likelihood, as inferred by Hao and Golding (2006) for the three rate cases defined in Figure 2.

Rate Case	MLE	LnL
$\alpha = \beta = \gamma$	0.51	-40277
α	4.42	-36902
$\beta = \gamma$	0.35	
α	3.92	
β	0.28	-36128
γ	1.23	

Table 1.2. Optimal rates of gene insertion/deletion as predicted by the maximum likelihood analysis, bootstrap testing, and curvature method for rate Cases 1 – 3 in Figure 2 and rate Case 4 in Figure 3.

		Bootstrap	**Curvature**
Rate	**MLE**	**MLE ± St. Dev**	**St. Dev**
$\alpha = \beta = \gamma$	0.5175	0. 5187 ± 0.0252	0.0069
α	4.564	4.572 ± 0.1558	0.1098
$\beta = \gamma$	0.3487	0.3487 ± 0.0055	0.0051
α	4.011	4.015 ± 0.1156	0.0899
β	0.2837	0.2836 ± 0.0052	0.0045
γ	1.273	1.274 ± 0.0413	0.0430
α_1	2.351	2.705 ± 1.394	1.175
α_2	1.117	0.6902 ± 0.8404	0.8175
α_3	14.62	15.15 ± 3.241	2.988
α_4	9.211	9.218 ± 0.7634	0.6934
α_5	5.519	5.483 ± 0.5245	0.4586
α_6	11.67	11.69 ± 0.6135	0.6124
α_7	1.562	1.560 ± 0.1959	0.1306
α_8	0.4434	0.4429 ± 0.0183	0.0176
α_9	0.2760	0.2757 ± 0.0164	0.0155
α_{10}	0.2531	0. 2528 ± 0.0149	0.0147
α_{11}	0.2473	0. 2465 ± 0.0122	0.0121
α_{12}	0.3096	0.3102 ± 0.0145	0.0137
α_{13}	0.0172	0.0554 ± 0.0291	0.0089
α_{14}	13.74	13.73 ± 0.6344	0.6104
α_{15}	1.184	1.193 ± 0.1675	0.1578
α_{16}	7.206	7.226 ± 0.5856	0.5303

		Bootstrap	Curvature
Rate	MLE	MLE ± St. Dev	St. Dev
α_{17}	1.030	1.038 ± 0.1481	0.0981
α_{18}	1.321	1.323 ± 0.0418	0.0427
α_{19}	0.5345	0.5373 ± 0.0257	0.0818
α_{20}	3.674×10^{-11}	1.188×10^{-4} ± 0.0017	infinite
α_{21}	0.4991	0.4988 ± 0.0517	0.0493
α_{22}	0.1701	0.1704 ± 0.0205	0.0188
α_{23}	1.456	1.155 ± 0.2158	0.0696

Table 1.3. Results of the likelihood ratio test for the four rate cases (Case 1 – 3 in Fig. 1.2 and Case 4 in Fig. 1.3) assumed on the *Bacillaceae* phylogeny.

Rate Case	LnL	- Δ2LnL	df	P-value
Case 1	-40276.543557	-	-	-
Case 2	-36901.410594	-6750.26592	1	3.84
Case 3	-36126.560876	-1549.69944	1	3.84
Case 4	-34864.390508	-2524.34074	20	31.41

Table 1.4. AIC statistics calculated from the maximum likelihood estimated by each rate case (Case 1 – 3 in Fig. 1.2 and Case 4 in Fig. 1.3). The AIC value of rate Case 4 gives the MAICE.

Rate Case	LnL	k	-2 Ln(L) + 2k	Difference from MAICE
Case 1	-40276.543557	1	80555.0871	10780.3061
Case 2	-36901.410594	2	73806.82118	4032.04018
Case 3	-36126.560876	3	72253.12174	2478.34074
Case 4	-34864.390508	23	69774.781	-

Table 1.5. The number of genes with the most common phyletic patterns in the *Bacillaceae* group as observed by Hao and Golding (2006).

Number of genes	**Ba$_1$**	**Ba$_2$**	**Ba$_3$**	**Bt**	**Bc$_1$**	**Bc$_2$**	**Bc$_3$**	**Gk**	**Bl**	**Bs**	**Bk**	**Bh**	**Oi**
1139	1	1	1	1	1	1	1	1	1	1	1	1	1
1024	1	1	1	1	1	1	1	0	0	0	0	0	0
285	0	0	0	0	0	0	0	0	1	1	0	0	0
194	1	1	1	0	0	0	0	0	0	0	0	0	0
156	0	0	0	0	0	0	0	1	0	0	0	0	0
148	0	0	0	0	0	0	0	0	0	0	0	1	0
132	1	1	1	1	1	1	0	0	0	0	0	0	0
128	0	0	0	0	0	0	0	0	0	0	1	0	0
119	0	0	0	0	0	0	0	0	0	1	0	0	0
118	0	0	0	0	0	0	0	0	0	0	0	0	1
109	1	1	1	1	1	0	1	0	0	0	0	0	0
103	1	1	1	0	1	0	0	0	0	0	0	0	0
99	1	1	1	1	1	0	0	0	0	0	0	0	0
96	1	1	1	1	1	1	1	1	1	1	1	1	0
90	1	1	1	1	1	1	1	0	1	1	1	1	1
85	0	0	0	0	0	0	0	0	1	0	0	0	0
3203	Other patterns												

Table 1.6. An example of one bootstrap result of the number of genes with the most common phyletic patterns in the *Bacillaceae* group.

Number of genes	**Ba$_1$**	**Ba$_2$**	**Ba$_3$**	**Bt**	**Bc$_1$**	**Bc$_2$**	**Bc$_3$**	**Gk**	**Bl**	**Bs**	**Bk**	**Bh**	**Oi**
1111	1	1	1	1	1	1	1	1	1	1	1	1	1
1034	1	1	1	1	1	1	1	0	0	0	0	0	0
253	0	0	0	0	0	0	0	0	1	1	0	0	0
174	1	1	1	0	0	0	0	0	0	0	0	0	0
163	0	0	0	0	0	0	0	1	0	0	0	0	0
151	1	1	1	1	1	1	0	0	0	0	0	0	0
149	0	0	0	0	0	0	0	0	0	0	0	1	0
133	0	0	0	0	0	0	0	0	0	0	1	0	0
119	0	0	0	0	0	0	0	0	0	0	0	0	1
115	0	0	0	0	0	0	0	0	0	1	0	0	0
114	1	1	1	1	1	0	0	0	0	0	0	0	0
104	1	1	1	0	1	0	0	0	0	0	0	0	0
95	1	1	1	1	1	1	1	1	1	1	1	1	0
88	0	0	0	0	0	0	0	0	1	0	0	0	0
87	1	1	1	1	1	1	1	0	1	1	1	1	1
85	1	1	1	1	1	0	1	0	0	0	0	0	0
3253	Other patterns												

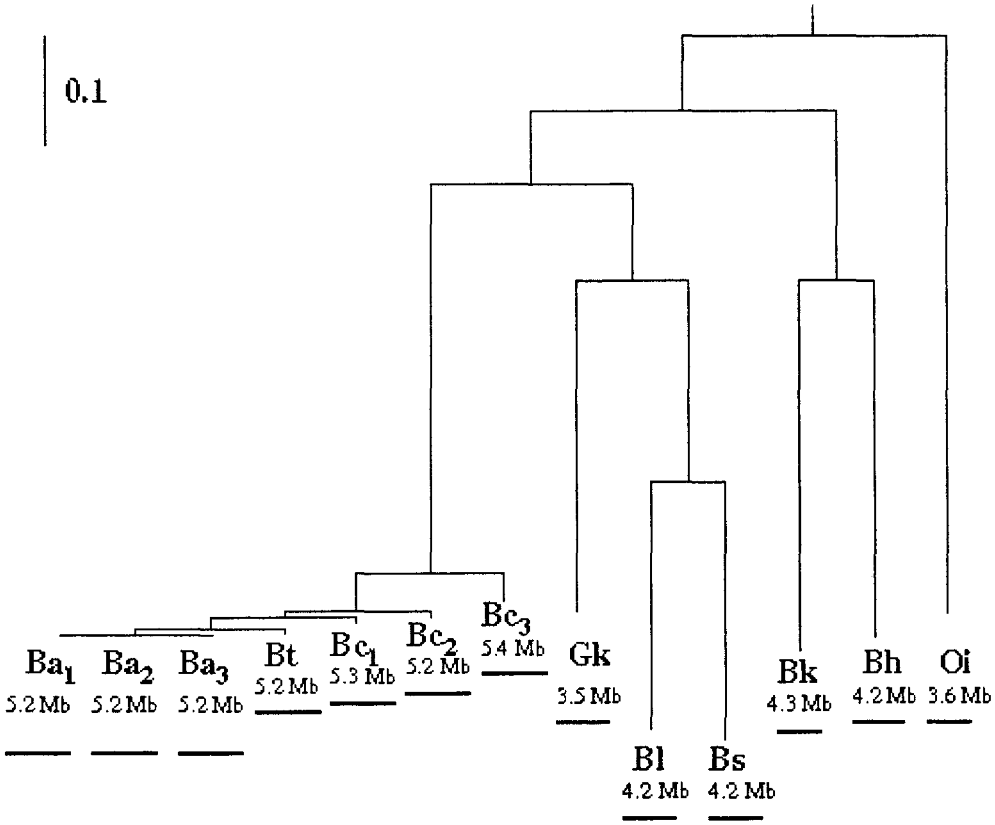

Figure 1.1. The evolutionary branching order of the *Bacillaceae* phylogeny inferred by Hao and Golding (2006), using the concatenated gene sequences of *gmk*, *glpF*, and *pycA*. Members used to construct the phylogenetic history of the group include: *Bacillus anthracis Ames* (Ba_1), *Bacillus anthracis* "Ames Ancestor" (Ba_2), *Bacillus anthracis* Sterne (Ba_3), *Bacillus thuringiensis* (Bt), *Bacillus cereus* ZK (Bc_1), *Bacillus cereus* ATCC 10,987 (Bc_2), *Bacillus cereus* ATCC 14,579 (Bc_3), *Geobacillus kaustophilus* (GK), *Bacillus licheniformis* (Bl), *Bacillus subtilis* (Bs), *Bacillus clausii* (Bk), *Bacillus halodurans* (Bh), and *Oceanobacillus iheyensis*. Note, time is scaled as the expected number of nucleotide substitutions per site.

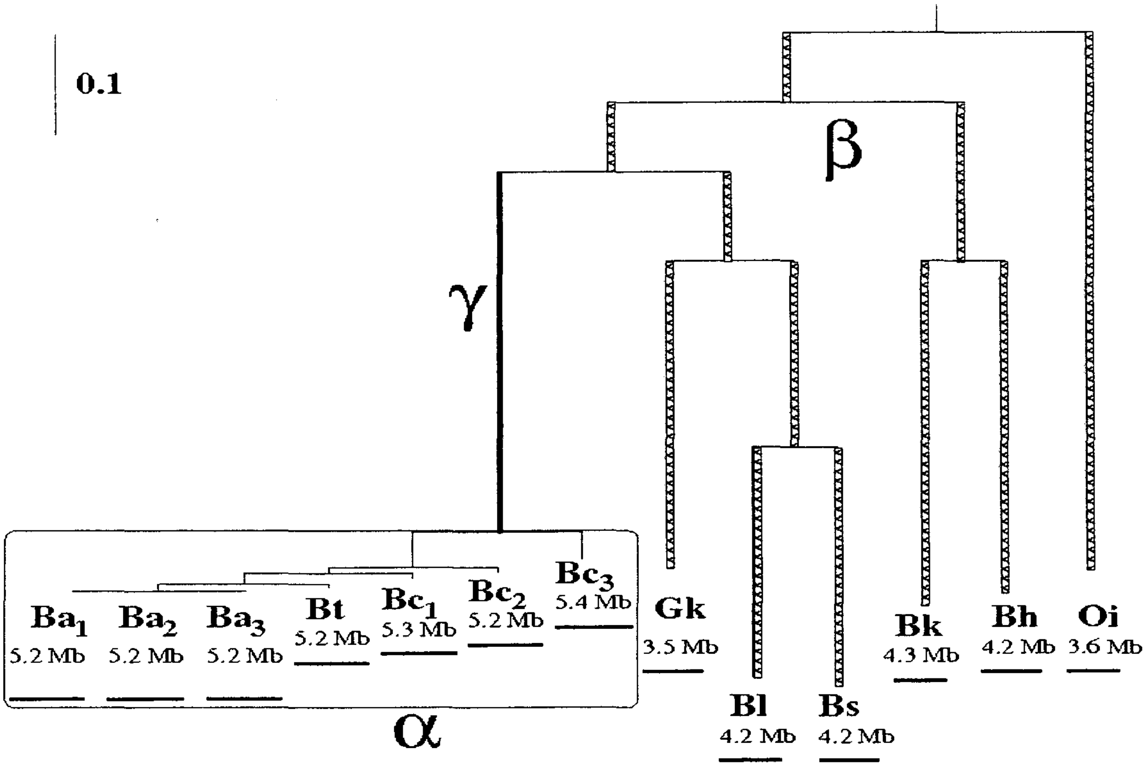

Figure 1.2. The different evolutionary scenarios assumed on the *Bacillaceae* phylogeny by Hao and Golding (2006). The Bc group is boxed off and evolves at rate α, while the remaining hatched section of the phylogeny evolves at rate β. Rate γ denotes the rate of divergence between the two groups. Case 1: a single constant rate throughout the phylogeny ($\alpha = \beta = \gamma$). Case 2: two rates differentiating the Bc group (α, $\beta = \gamma$). Case 3: three rates differentiating the Bc group and the branch leading to the group (α, β, γ).

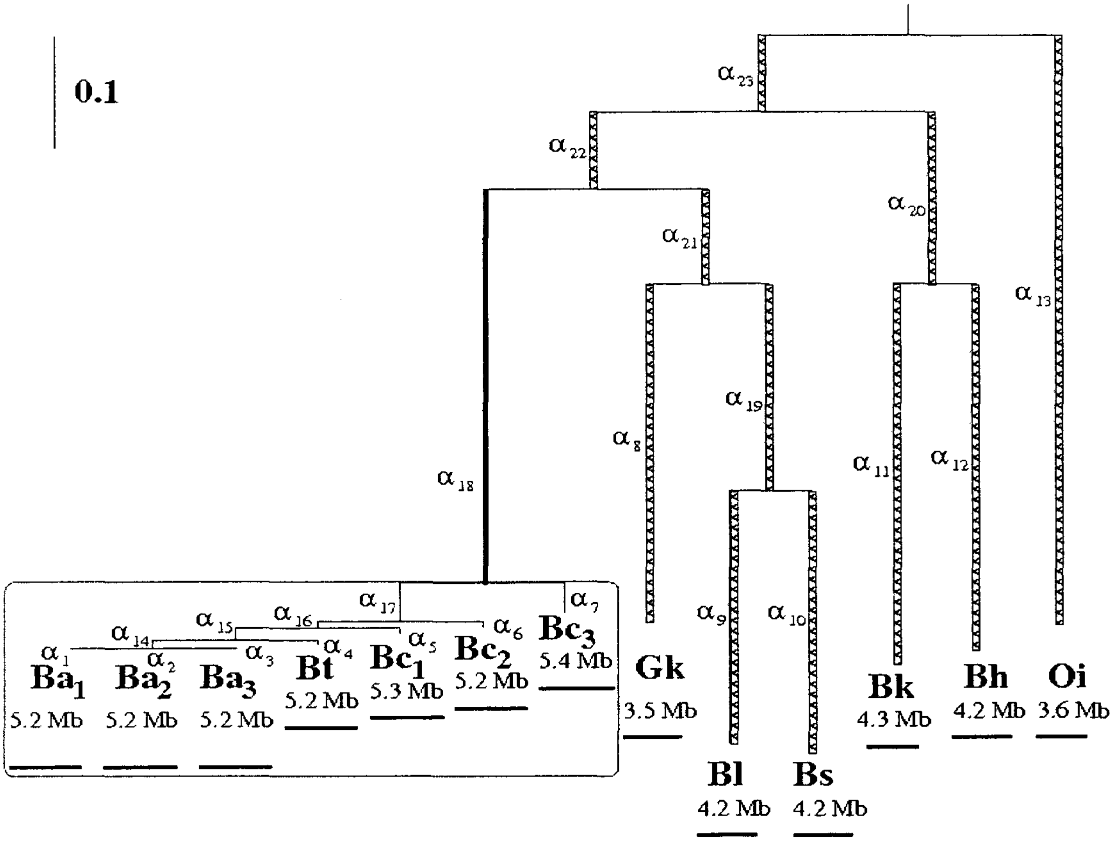

Figure 1.3. Additional rate model of *Bacillaceae* evolution assumed on the phylogeny. Individual rates of gene evolution were used to differentiate each branch of the phylogeny. Case 4: branch specific indel rates in Chapter 1 and branch specific insertion and deletion rates in Chapter 2 (α_1, α_2,..., α_{23}).

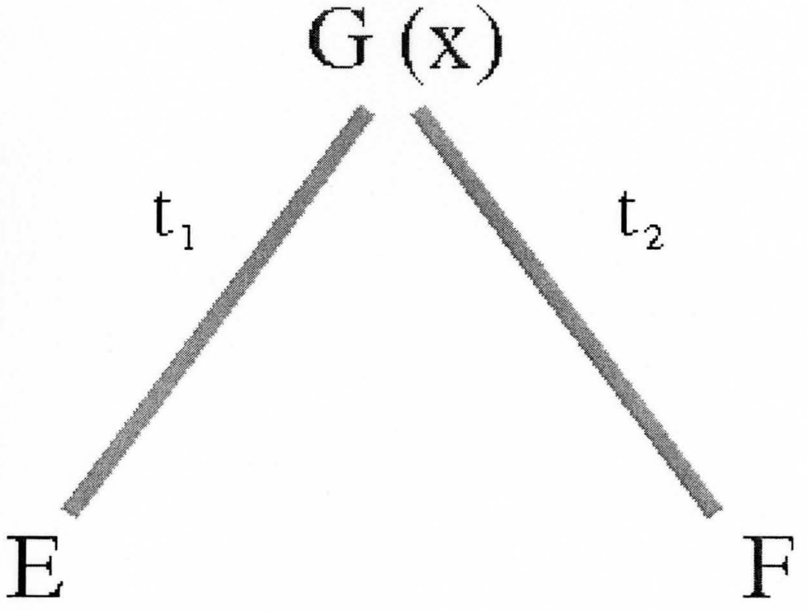

Figure 1.4. The branching relationship representing the simplest case for determining the likelihood of the ancestral state, using the recursion principle. The likelihood that gene *x* is present at the ancestor node G depends on the genetic states of the descendent taxa E and F, separated by t_1 and t_2 generations.

Chapter 2

Differential-rate maximum likelihood model of *Bacillaceae* evolution

2.1 ABSTRACT

The genomic history of bacteria is largely influenced by the dynamic interactions of gene insertions, gene deletions, and incidences of lateral gene transfer (LGT) that occur during sequence evolution. Increased rates of LGT observed during prokaryotic evolution have prompted the conclusion that it is the most prominent force controlling the topology of the taxa. Through the constant introduction of alien genes into the phylogeny, LGT can dramatically alter gene content, rapidly driving species divergence. To determine the role of each evolutionary factor in shaping the history of a phylogeny, the extent to which each impacts genomic evolution needs to be quantified. Using a multidimensional maximum likelihood model, we calculated the optimal rates of genome evolution for a group of *Bacillaceae* bacteria. Differential rates of gene insertion and

gene deletion were assumed on the phylogeny, and their evolutionary patterns were inferred under four different rate models. Overall, the phylogeny evolves according to a higher rate of gene insertion than gene deletion, with increased rates of gene acquisition observed in members belonging to *B. anthracis*, *B. cereus*, and *B. thuringiensis* (the Bc group). Statistical testing of the inferred optimal rates by bootstrap sampling and curvature measurements revealed little deviation in the observed data set. In fact, only slight deviation is noted when independent branch evolution is assumed on the phylogeny (Case 4), with the *Bacillus anthracis* group showing the greatest variation in the predicted rate values. Both the likelihood ratio test and Akaike Information Criterion (AIC) reveal that the history of the *Bacillaceae* phylogeny is best modeled when independent rates of evolution are assigned for each branch (Case 4). The strong statistical support of the model, along with the inferred optimal rates of gene insertion and gene deletion, confirm the robustness of the maximum likelihood algorithm and help provide valuable insight to the true nature of bacterial evolution.

2.2 INTRODUCTION

As databases of whole genome sequences continually expand and become readily available to research, increased interest has been directed at understanding the evolutionary history of the genes. Of particular interest are the microbial genomes. With the accumulation of fully sequenced genomes, bacteria are emerging as the ideal subject in modeling genome evolution. The genome of a bacterium is shaped by many processes, including: gene duplication, gene loss, lateral gene transfer (LGT), and mutation (Snel *et al.* 2002; Gu and Zhang 2004; Lake and Rivera 2004; Hao and Golding 2004, 2006; Novozhilov *et al.* 2005; Marri *et al.* 2007). As reviewed in Doolittle (1999), lateral gene transfer is the exchange of genetic material across taxa and has been denoted the most prominent factor regulating prokaryotic evolution (Lan and Reeves 1996; Gogarten *et al.* 2002; Jain *et al.* 2003; Mirkin *et al.* 2003). Therefore, microbial research often aims to identify incidences of LGT and analyze the resulting effect on genomes of closely related species. Most models of bacterial evolution (Snel *et al.* 2002; Lake and Rivera 2004; Hao and Golding 2004, 2006; Linz *et al.* 2007; Marri *et al.* 2007) only consider the rate of gene insertion and gene deletion, assuming that both processes occur at equal rates. It is not always the case, however, that the genomes are shaped in such a balanced manner. For example, in the poxvirus study by McLysaght e*t al.* (2003) the rate of gene loss was found to vary across the poxvirus genome and higher rates of gene insertion were observed in the orthopox group. Therefore, the evolution of a genome may be influenced by a variety of forces, including unequal rates of gene insertion and gene deletion (Berg and Kurland 2002; Huson and Steel 2004; Novozhilov *et al.* 2005). As such, it is important to devise a model that provides an accurate depiction of microbial evolution by considering the effect of varying rates on genome evolution.

In order to explore the differential evolution of a genome, the maximum likelihood method of Hao and Golding (2006) was adapted to estimate separate rates of gene insertion and gene deletion. The new model was applied to the same group of Gram-positive *Bacillaceae* bacteria (Hao and Golding 2006) so that the results could be compared to the previous findings. Gene presence/absence patterns used to infer the maximum likelihood estimate (MLE) of insertion and deletion were also obtained from their likelihood study (see Table 1.5 for the most commonly noted patterns and their frequency). Calculation of the maximum likelihood and the corresponding optimal rates of gene insertion and gene deletion were achieved using Powell's (1964) method of quadratic convergence. For a detailed description of the maximization algorithm applied, please see section 2.2 of this chapter.

The proposed convergence model examines the role of varying gene insertion and gene deletion rates on a bacterium genome. The multidimensional maximum likelihood analysis was applied to the four rate scenarios defined in Chapter 1: a single constant insertion rate and deletion rate throughout the phylogeny, $\alpha = \beta = \gamma$, (Case 1 in Fig. 1.2); two insertion and deletion rates differentiating the Bc group, α, $\beta = \gamma$, (Case 2 in Fig. 1.2); three insertion and deletion rates differentiating the Bc group and the branch leading to the group, α, β, γ, (Case 3 in Fig. 1.2); and branch specific insertion and deletion rates ($\alpha_1, \alpha_2, \ldots, \alpha_{23}$) (Case 4 in Fig. 2.1). For most of the rate parameters, the insertion rate was found to be considerably greater than the deletion rate, indicating growth in the genome. Collectively, the model provides an efficient means of estimating evolutionary rates and establishes supplementary support for the optimization algorithm detailed by Hao and Golding (2006).

2.3 POWELL'S CONVERGENCE ALGORITHM

Powell's (1964) convergence algorithm is an efficient method of optimization when the derivative of the function being maximized is unknown, and provides quick convergence to the global optimum, even when initial estimates are bad. Starting at some initial point x_0, the maximum of a function is obtained by moving along some direction z until the function is maximized in that direction. Upon reaching a maximum for the first vector, the algorithm then proceeds to the next directional vector, moving along it to reach a new maximum. This cycle repeats itself for the entire set of directional vectors. The goal of the algorithm is to find a set of vectors that are orthogonal to one another in order to find the function maximum most efficiently. When all the vectors of the directional set are orthogonal to one another, the function is said to be maximized. Therefore, the maximum of an n-dimensional function is achieved by moving along each directional vector, one vector at a time, to the maximum until a set of n mutually conjugate directions is obtained.

The computation of Powell's algorithm formally begins by saving the initial approximations as the starting point, P_0, and setting the initial set of directions u_i equal to the unit vectors, for $i = 1, 2, \ldots, n$. The program then applies the following steps to search for a maximum:

(i) Choose an initial point P_0 and set u_i equal to unit vectors, for $i = 1, 2, \ldots, n$

(ii) Optimize the likelihood starting from P_0 in the direction u_1 and label the resulting vector P_1.

(iii) Optimize the likelihood starting from P_1 in the direction u_2 and label the resulting vector P_2, Optimize the likelihood starting from P_2 in the direction u_3 and label the resulting vector P_3,

. . .

. . .

. . .

Optimize the likelihood starting from P_{n-1} in the direction u_n and label the resulting vector P_n

(iv) Set $u_i = u_{i+1}$ for $i = 1,\ldots, n - 1$.
Set $u_n = P_n - P_0$.

(v) Optimize the likelihood starting from P_n in the direction u_n and label the resulting vector P_0.

(vi) Return to step (ii) until the maximum of the function is reached or until some specified stopping criterion is met (for example, when the required accuracy is reached; Powell 1964).

In order to improve the efficiency of the algorithm along narrow valleys, only the current directions of largest increase can be used *in lieu* of the previous best estimates. This minor adjustment was considered in the multidimensional analysis because calculating the global maximum is complicated by the increase in the number of rate parameters from rate Case 1 – Case 4. Therefore, in the more complex rate scenarios, the modification allows for quick convergence in the new direction of a complicated landscape and minimizes the incidence of linear dependence in the direction set (Flannery *et al.* 1992).

2.4 METHODS

In the previous maximum likelihood analysis (Chapter 1), a likelihood algorithm adapted from Hao and Golding (2006) was used to infer the optimal rate of gene insertion/deletion (indel) for the *Bacillaceae* phylogeny. In total, thirteen fully sequenced Gram-positive *Bacillaceae* genomes were analyzed, including: *B. anthracis* Ames, *B.*

anthracis "Ames ancestor," *B. anthracis* Sterne, *B. thuringiensis*, *B. cereus* ZK, *B. cereus* ATCC 10,987, *B. cereus* ATCC 14,579, *Geobacillus kaustophilus*, *B. licheniformis*, *B. subtilis*, *B. clausii*, *B. halodurans*, and *Oceanobacillus iheyensis*. The genomes of *B. anthracis*, *B. cereus*, and *B. thuringiensis* were further considered as a single group, the Bc group, because studies previous to Hao and Golding (Ash et al. 1991; Priest et al. 2004) have revealed their sequences to be very similar (Hao and Golding 2006). Indel rates for the four defined rate scenarios (Case 1 - Case 3 in Fig. 1.2 and Case 4 in Fig. 1.3) were determined by using bracket intervals to pinpoint the rate required to maximize the likelihood. The robustness of the predicted rate values was confirmed using bootstrap results and by measuring deviation in the likelihood curve. Further statistical analysis by the likelihood ratio test and AIC, reveals the parameter assumptions of rate Case 4 to support the best model of phylogenetic evolution.

To provide a more accurate depiction of gene evolution in the *Bacillaceae* group, the maximum likelihood model offered by Hao and Golding (2006) was improved. In their model, genes are assumed to be inserted and deleted at an equal rate. Genomic evolution, however, is not necessarily shaped by a balance between gene insertion and gene deletion, but rather by unequal rates of acquisition and loss (Berg and Kurland 2002). In fact, gene insertions have been predicted to occur more often than gene deletions in some viruses (Daubin et al. 2003; McLysaght et al. 2003). Hao and Golding (2006) assumed equal rates of gene insertion and deletion to prevent the genome size from diverging to zero or infinity, especially for longer term gene content analysis. By assuming a single rate, they were able to efficiently and accurately model the indel rate of

Bacillaceae evolution under a simplified likelihood scheme. As such, the model offers a reliable base algorithm from which more complicated likelihood models can be derived. Following the assumptions of the aforementioned likelihood model, a multidimensional approach was adapted to determine the optimal rate(s) of insertion and deletion for the bacteria. The conditional probabilities:

$$Prob(P_d|P_a,t) = v/(u + v) + e^{-(u+v)t}[1-v/(u + v)],$$

$$Prob(A_d|P_a,t) = u/(u + v) - e^{-(u+v)t}[1-v/(u + v)],$$

$$Prob(P_d|A_a,t) = v/(u + v) - e^{-(u+v)t}[1-u/(u + v)],$$

$$Prob(A_d|A_a,t) = u/(u + v) + e^{-(u+v)t}[1-u/(u + v)],$$

where *P* is present, *A* is absent, *d* is descendant, *a* is ancestor, *v* is the rate of gene insertion, and *u* is the rate of gene deletion, were used to calculate the likelihood that a given gene was present at an ancestor node. The overall likelihood of the phylogeny was determined in similar manner as outlined in Chapter 1, using the equation:

$$\begin{aligned} L^x{}_\mathrm{G}(P) = (&Prob(P_d|\, P_a,\, t_1) * L^x{}_\mathrm{E}(P) + Prob(A_d|\, P_a,\, t_1) * L^x{}_\mathrm{E}(A)) \\ &x \quad (Prob(P_d|\, P_a,\, t_2) * L^x{}_\mathrm{F}(P) + Prob(A_d|\, P_a,\, t_2) * L^x{}_\mathrm{F}(A)). \end{aligned}$$

where G is the ancestor taxon, E is a descendant, and F is a descendant. The rate of gene insertion and gene deletion was inferred from the same gene presence and absence patterns used by Hao and Golding (2006).

Multidimensional maximization algorithms usually require the calculation of a conjugate gradient to determine the successive directions of increase of some function. When the derivative of a function cannot be determined, however, maximization of a quadratic function can be achieved by calculating conjugate directions. Powell (1964) describes such an efficient method of minimization/maximization, where conjugate directions are used to determine the minimum/maximum of a function by changing one variable at a time. Because derivatives cannot be determined for the likelihood function, Powell's optimization algorithm was applied to infer the rate of gene evolution.

The scripted formula outlined in Numerical Recipes for C (Flannery et al. 1992) for Powell's method of quadratic convergence, was altered to perform the multidimensional maximization of the modified likelihood function. The modified algorithm was applied to the three rate cases defined by Hao and Golding (2006): single constant insertion rate and deletion rate throughout the phylogeny ($\alpha = \beta = \gamma$), two insertion and deletion rates differentiating the Bc group (α, $\beta = \gamma$), three insertion and deletion rates differentiating the Bc group and the branch leading to the group (α, β, γ) (Fig. 1.2), and the additional rate case of branch specific insertion and deletion rates (α_1, α_2,..., α_{23}) (Fig. 2.1). For each rate case, the initial conjugate directions were set equal to the unit vectors. To estimate the maximum likelihood, the independent insertion and deletion rates were changed in accordance to the conjugate directions, until the optimal rates required to maximize the likelihood function were achieved. Upon completion of the n^{th} dimensional conjugate direction, the average direction moved, x_n-x_0, was used to replace the previous directions of increase, where x_0 is the initial rate and x_n is the final rate after maximizing in n directions. The entire process is then repeated. In the successive iterates, former direction estimates representing the largest increase in the function were ignored to minimize the occurrence of linearly dependent directions (Flannery et al. 1992). The entire procedure was repeated until the fractional tolerance of the function value was less than 0.1×10^{-22}.

For the evolutionary scenarios of higher dimension (Case 3 and Case 4), Powell's routine was run multiple times in order to establish the exact, global maximum estimate of the likelihood function. After each iterate the newly calculated insertion and deletion rate estimates were used to replace the previous initial estimates. This substitution was done to aid in the efficiency of the algorithm, as the new rate estimates should constitute a more reliable initial data set. The number of required iterates for the algorithm varied, depending on the initial rate values, with the more ambiguous values requiring more repetition.

Bootstrap sampling and curvature measurements were applied to the maximization algorithm to test the statistical significance of the differential gene insertion and gene deletion rate model. Methods for both statistical processes follow from the procedures as detailed in Chapter 1. To test the accuracy of the predicted insertion and deletion rates, bootstrap samples were generated for each of the defined rate cases (Case 1 - Case 4). Bootstrap sampling of each scenario was repeated for a total of 1000 iterates, producing 1000 independent parameter estimates of the different rates assumed on the phylogeny. The bootstrapped estimates were then evaluated in the statistical package STATA 7.0 (Stata Corporation, College Station, TX). The accuracy of an inferred insertion or deletion rate was determined from the confidence intervals, consisting of the mean and standard deviation, predicted for each parameter (Table 2.1).

Deviance in the actual convergence model was analyzed based on variance measures from the maximum likelihood curve. Independent likelihood curves were constructed for each parameter in the different rate cases, separately. This was achieved by varying only one parameter at a time while the remaining parameters were kept constant. For each likelihood curve, nine rates: four rates lower than the maximum estimate, four rates greater than the maximum estimate, and the maximum estimate, were used to estimate the curvature of the likelihood function. Estimates for rates greater than or less than the maximum estimate were calculated by increasing and decreasing, respectively, the optimal rate estimate by a specified value. For each rate predicted by altering the maximum estimate, the corresponding likelihood was calculated and plotted with the rate as a coordinate in the likelihood curve. Variance in the model was inferred from the negative inverse of the second derivative, or curvature, of the associated likelihood function, evaluated at the maximum rate estimate. This value was then used to measure the standard deviation of the likelihood model (Table 2.1).

To evaluate whether increasing the number of parameters of the likelihood model actually produces a better estimate of maximum likelihood, the likelihood ratio test (Fisher 1924) was applied to the four defined rate cases. The statistical significance of an estimate was determined using Wilks (1938) chi-square approximation of the likelihood ratio (see Chapter 1 for a more detailed description of Wilks' chi-square approximation). For each rate case, the predicted maximum likelihood was compared to the maximum likelihood value of the next consecutive rate case, following an order of increasing complexity from Case 1 to Case 4. The resulting ratios were then compared in a chi-square distribution, with the degree of freedom determined from the difference in the number of independent variables between two cases (Table 2.3).

It is often the case that the statistics of a model depend largely on the parameters being considered. Therefore, the statistical choice for the best-fit model needs to be validated. By penalizing a model based on the number of parameters it uses, the Akaike Information Criterion (AIC) identifies the best hypothesis void of significance levels (Akaike 1974). An AIC value can be calculated from the maximum likelihood of each rate case, using the equation:

$$\mathrm{AIC} = -2\ \ln(L) + 2k,$$

where k is the number of parameters defined in the model. The rate hypothesis that gives the minimum AIC estimate (MAICE) is considered the best model for the likelihood analysis. The AIC of each rate case was compared to the MAICE, to determine the statistical support of the competing hypothesis (Table 2.4).

2.5 RESULTS

Using Powell's (1964) algorithm, the multidimensional maximum likelihood analysis modeled the influence of unequal gene insertion and gene deletion rates on the evolution of the *Bacillaceae* phylogeny. The model was applied to a group of thirteen highly similar bacteria sequences, in order to explore the patterns of lateral gene transfer (LGT) in closely related species. Optimal rates of acquisition and loss were determined for each of the rate cases defined in Figure 1.2, and the additional rate case of branch specific insertion and deletion rates (Fig. 2.1). Once the rates required to maximize the likelihood function were achieved, the accuracy of the estimates were statistically tested by bootstrap sampling and variance measures from the maximum likelihood curve. Together, the observed results will aid in understanding the nature of LGT in shaping the *Bacillaceae* genome.

Initially, the phylogeny was assumed to evolve according to a constant rate of gene insertion and gene deletion across all branches of the phylogeny (Case 1 in Fig. 1.2). The likelihood function was maximized according to an insertion rate of 0.7326 and a deletion rate of 0.6132. Bootstrap testing and maximum likelihood curve approximations reveal very little deviation in the statistical estimates from the observed rate values (Table 2.1). In fact, the mean insertion rate of the bootstrap sample is identical to the observed rate. Additionally, the standard deviation estimates for the bootstrapped rates are very small, with the insertion rate having a slightly smaller deviation interval, 0.7346 ± 0.0057, than the deletion rate, 0.6136 ± 0.0275. Standard deviation in the likelihood curve is also minor for the insertion rate and the deletion rate, 0.0081 and 0.0151 respectively, and the predicted curves appeared to fit the data points well.

Two differential rates of evolution were used to describe the *Bacillaceae* phylogeny in the second rate model (Case 2 in Fig. 1.2). The high sequence similarity between *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* resulted in the grouping of these species into the *B. cereus* group (Bc group) (Hao and Golding 2006). Rate α was used to describe gene acquisition and loss within the Bc group and rate β is the evolutionary gene rate assumed on the remaining phylogeny. In the differential rate model, the inferred rate of gene insertion and gene deletion for α is 14.08 and 1.809, and for β 0.4049 and 0.3524, respectively. Statistical sampling of the two rates indicates little deviation from the proposed estimates, with the rates predicted for α showing slightly more deviation than the rates for β (Table 2.1). The confidence intervals predicted for the insertion rate, 14.09 ± 0.3920, and deletion rate, 1.812 ± 0.1082, of α, and the insertion rate, 0.4051 ± 0.00089, and deletion rate, 0.3516 ± 0.0169, of β, are relatively small and the predicted mean rates are extremely close to the optimal insertion and deletion rates. Additionally, only minor disparities in the rate values are observed when testing the variance via the curvature method. Measures of standard deviation inferred from the likelihood curve are small for both rate α, 0.2457 for gene insertion and 0.1022 for gene deletion, and rate β, 0.0067 for gene insertion and 0.0118 for gene deletion, thus providing continued support for the proposed model.

In the third model of *Bacillaceae* evolution, a third rate was added into the phylogeny (Case 3 in Fig. 1.2). It is evident from the phylogenetic tree that members of the Bc group have a lager genome size than the rest of the *Bacillaceae* phylogeny. Therefore, following the reasoning of Hao and Golding (2006), the incidence of gene insertion and gene deletion between these two groups might be expected to occur at different rates. Rate γ was used to describe gene acquisition and loss along the branch separating the two parts of the phylogeny. The optimal rates of insertion and deletion estimated by this model are 9.909 and 2.289 for rate α, 0.3633 and 0.3376 for rate β, and 1.592 and 0.6351 for rate γ. Again, both statistical tests provide strong support for all rate estimates. The inferred gene insertion and deletion rates are predicted to have narrow confidence intervals from the bootstrap samples. Rate β has the most consistent values, 0.3638 ± 0.0074 for gene insertion and 0.3378 ± 0.0159 for gene deletion, followed by rate γ, and then rate α. Estimates of standard deviation from the curvature are reflective of the bootstrap results. Only modest dispersion in the data is predicted, again with rate β showing slightly less deviation, 0.0065 for gene insertion and 0.0116 for gene deletion, than rate γ and rate α. All values used to construct the likelihood curves appear to fit the model well.

For the first three rate scenarios, it is interesting to note that the rate of gene insertion to gene deletion is always greater. In fact, in the second and third rate cases, the relative number of gene insertions to gene deletions increases towards the Bc group, with the Bc group having the highest insertion rates. The phylogeny outside the Bc group appears to evolve more steadily, owing to an insertion rate that is almost par with the deletion rate. In rate Case 3, even rate γ is predicted to have a higher rate of insertion to deletion in comparison to the branches outside the Bc group.

The final rate model assumed on the phylogeny was branch specific insertion and deletion rates (Case 4 in Fig. 2.1). By allowing each branch to evolve independently, the unique evolutionary rates influencing a single *Bacillaceae* genome can be observed and compared to the rest of the group. The optimal rates of gene insertion and gene deletion for rates α_1 - α_{23}, starting from left to right on the phylogenetic tree (Fig. 1.1), are listed in Table 2.1. Although, the bootstrap results produced reliable confidence intervals for most members of the phylogeny, slightly more deviation is observed in the rates estimated for the Bc group. Rates α_1, α_2, and α_3, all belonging to the *Bacillus anthracis* group, show some departure from the optimal values, particularly in the insertion rate estimated for α_1. The rate of gene insertion, 1.684, and gene deletion, 0.4339, for α_{18}, representing the branch leading to the Bc group, are quite similar to those predicted for rate γ in Case 3. Extreme deviation is observed in rates α_1, α_{13}, α_{17}, α_{20}, and α_{23}. Rates α_1 and α_{17} both belong to the Bc group, where α_1 describes the branch leading to *Bacillus anthracis* Ames and α_{17} represents the branch leading to *Bacillus cereus* ATCC 14,579. The insertion rate of α_1 is extremely small, 1.227×10^{-13}, and it has a wide confidence interval of $3.44 \times 10^{-11} \pm 4.38 \times 10^{-11}$. Likewise, the deletion rate of α_{17} is also very small, 1.467×10^{-11}, and a reliable estimate of deviation could not be achieved in the bootstrap sampling. Rates α_{13}, α_{20}, and α_{23} all occur outside the Bc group, closer to the root of the phylogeny. Rates α_{13}, representing the branch leading to *Oceanobacillus iheyensis*, and α_{23}, both have extremely low optimal rates of gene insertion, and thus, statistical departure from the inferred values could not be determined in the confidence intervals. Rate α_{20} has a low deletion rate of 3.042×10^{-11}, with a confidence interval of $4.83 \times 10^{-11} \pm 2.54 \times 10^{-11}$. Deviation estimates from the likelihood curve produced similar statistical results as the bootstrap testing. Once again rate estimates for members of the Bc group show slightly more deviation. Even the curvature analysis for α_1, α_2, and α_3, resulted in poorer fitting of the likelihood curve. The minute nature of rates α_1, α_{13}, α_{17}, α_{20}, and α_{23} made it difficult to determine additional plot values that deviated from the maximum likelihood, and thus, a reliable likelihood curve could not be constructed. As a result, the data points reflected a horizontal line rather than a curve, and this reflects a standard deviation/variance approaching infinity.

Most of the branches are observed to evolve according to a higher rate of insertion than deletion in rate Case 4. Only rates α_1, α_{13}, and α_{21} – α_{23} have MLE deletion rates greater than the insertion rate. An increase in the relative number of insertions to deletions is also observed in the branches nearer the Bc group. The Bc group has the highest rates of gene insertion, with the exception of α_1. Outside the Bc group, the rest of the phylogeny evolves at a more steady pace, with insertion rates more similar to deletion rates. The incidence of greater gene deletion along the branch leading to *Oceanobacillus iheyensis* (α_{13}) is most engaging, as this species has a genome size of only 3.6 Mb.

A likelihood ratio test was done to assess the statistical significance of the proposed rate cases (Case 1 - Case 4). The results indicate that altering the model to include more rate parameters provides a larger estimate of the maximum likelihood (Table 2.3). The insertion and deletion parameters considered in rate Case 4, give the best estimate of maximum likelihood, -34823.650, indicated by the Δ2LnL value of 3837.776 with 40 degrees of freedom.

To ensure that the choice of Case 4 as the best model of *Bacillaceae* evolution is not skewed by the number of parameters considered in the model, the rate cases were also evaluated according to AIC. In support of the likelihood ratio test results, the MAICE, 69739.3, is defined by the parametric assumptions of rate Case 4. When the AIC values of the alternate rate hypothesis were compared to the MAICE, little statistical support is observed for the parameter values of the opposing hypothesis of rate Case 1 – Case 3 (Table 2.4).

2.6 DISCUSSION

When attempting to reconstruct the phylogenetic history of a set of taxa, the predicted gene tree does not always agree with the species tree. Inconsistencies between the branching patterns reveal the inherent diversity of the genome resulting from gene insertions, gene deletions, and/or lateral gene transfer (LGT) (Snel *et al.* 2002, 2005; Mirkin *et al.* 2003; Kunin and Ouzounis 2003; Gu and Zhang 2004; Hao and Golding 2004, 2006; Novozhilov *et al.* 2005; Linz *et al.* 2007; Marri *et al.* 2007). Recent models of prokaryotic evolution (Jain *et al.* 2003; Mirkin *et al.* 2003; Hao and Golding 2004, 2006; Galtier 2007; Linz *et al.* 2007; Marri 2007) have revealed the rampant and pronounced influence of LGT in shaping bacterial evolution. Through the exchange of genetic material across species, LGT incorporates new genes into the genome and initiates divergence within the taxonomic group. Often, the presence of atypical genes, with closer resemblance to genes found in distantly related species, can only be explained by LGT (Gogarten *et al.* 2002; Daubin *et al.* 2003).

Cataloging the patterns of gene presence and absence may identify lateral gene transfer in closely related taxa (McLysaght *et al.* 2003; Hao and Golding 2004, 2006; Charlesworth and Eyre-Walker 2006; Marri *et al.* 2007). It is important to consider genomes of high similarity because any irregularities in the sequences can be attributed to LGT (Hao and Golding 2006; Marri *et al.* 2007). In our model, a group of thirteen fully sequenced *Bacillaceae* genomes of high similarity were used to infer the extent of LGT on the evolution of the phylogeny. The method of maximum likelihood offers an approach to estimate the rate of LGT when the rates of gene insertion and gene deletion are assumed unequal. Criterion for optimization and patterns of gene presence and absence used to estimate maximum likelihood were adopted from Hao and Golding (2006). Applying the model to the group of *Bacillaceae* bacteria reveals the extensive and important role of LGT in shaping the genome.

By allowing for differential rates of gene insertion and gene deletion, values for the evolutionary rates assumed on the phylogeny (Case 1 - Case 4) were estimated using the proposed likelihood framework. The likelihood model reveals that the estimates are consistent and reliable. The efficiency of the model and its ability to be easily manipulated, make it a good algorithm to infer the rapid rate of LGT within a phylogeny. This is clear in the robust rate estimates predicted for the individual parameters defined in the different rate cases (see Table 2.1). Almost all of the values calculated for the maximum likelihood estimates are well supported, indicated by the narrow error margin of the bootstrap samples and curvature variance. In fact, only the increasing complex case of branch independent rate parameters shows some deviation in the estimated dataset. The optimal rates predicted for α_1, α_{13}, α_{17}, α_{20}, and α_{23} have extremely wide confidence intervals and exhibit infinite variation. Rates α_1, α_{13}, and α_{23} have a relatively large rate of gene deletion to gene insertion, creating high variation within the individual datasets. As a result, the instability of the estimates made it difficult to construct reliable confidence intervals and maximum likelihood curves. Likewise, rates α_{17} and α_{20} have an extremely high rate of gene insertion to gene deletion, contributing to high deviation in the predicted rates. Because these rates are so low in magnitude relative to the maximum likelihood, the resulting likelihood curves resemble a horizontal line rather than an arc. Hence, the curvature of the line is zero, resulting in infinite variation, and similarly, infinite deviation. The relative degree of gene insertion to gene deletion, and vice versa, observed among these rates will be further considered in our discussion on the evolutionary patterns observed to influence evolution in the *Bacillaceae* genome.

When estimating the optimal rates of gene insertion and gene deletion, multiple trials were run to ensure that the exact values required to maximize the likelihood were obtained. This was done because the maximization algorithm would often return alternate optimal rates for the same maximum likelihood. In order to evaluate the inconsistency in the predicted optimal rates further, a 3-dimensional surface plot of the simplest rate Case (Case 1) was evaluated (Fig. 2.2). Plotting various insertion and deletion rates against their associative likelihoods reveals that the resulting surface resembles a saddle. Therefore, because the algorithm maximizes one rate at a time, it is very easy for other rates to fall off the edge of the saddle and skew the results. Looking at the contour map of these results (Fig. 2.3), it is clear that points can fall on the border of the saddle and are unable to reach the maximum value because they continually fall down the edge of the structure. To limit such inconsistencies in the data set, all rate parameters need to be jointly maximized at the same time.

When the insertion/deletion rates of the equal rate model (Table 2.2) were compared to the differential rate model, little deviation is observed between the two data sets, as predicted by Powell's (1964) algorithm. Overall, the rates estimated by the single parameter model are more consistent than those estimated under varying rates of gene insertion and gene deletion. Rates assumed on the phylogeny for the first three rate cases (Case 1 – 3 in Fig. 1.2) are quite consistent between both models. The only discrepancy between the two sets occurs in the rates estimated for the forth case of branch specific evolution (Case 4 in Fig. 2.1). Greater deviation is observed in more of the assumed rates of the differential model than in comparable rates of the single rate model. In fact, only rate α_{20} of the single rate model exhibits high deviation, likely owing to the very small, predicted estimate of the parameter, 2.719×10^{-11}.

The predicted optimal rates suggest the *Bacillaceae* phylogeny is evolving according to a higher rate of gene insertion than gene deletion, especially for the first three rate cases (Case 1 – Case 3 in Fig. 1.2). Such elevated rates of gene insertion are representative of a growing genome and may induce adaptive evolution in new habitats (Lan and Reeves 1996). When branches are assumed to evolve separately, however, the dominance of insertion over gene deletion changes in branches closer to the root of the tree (see Case 4 in Table 2.1). Rates belonging to the Bc group have the highest rate of gene insertion to gene deletion. In particular, rates α_3-α_7, α_{14}, α_{16} and α_{17} show extremely high levels of gene acquisition to loss and are likely the result of larger genome sizes and smaller branch lengths. Only rate α_1 of the Bc group has a higher deletion than insertion rate. This peculiar incident is likely the product of the small branch length originally predicted for the branch leading to *Bacillus anthracis* Ames. Because the branch is so small, the estimated number of gene insertions and/or deletions along the branch can fluctuate greatly, causing high deviation in the predicted rates. Outside the Bc group, most of the phylogeny evolves in a more gradual manner.

Although the relative number of gene insertions to deletions is still greater, the values are closer in magnitude. As the branches approach the root of the phylogeny, a greater rate of gene deletion to insertion is observed for α_{13} and $\alpha_{21} - \alpha_{23}$. This result is very interesting and may indicate genome shrinkage. For example, a higher deletion rate for α_{13}, representing the branch leading to *Oceanobacillus iheyensis*, is indicative of a smaller genome size of 3.5 Mb.

A hierarchy of increasing maximum likelihood estimates, as the number of parameters increase from Case 1 to Case 4, provides strong support that the rates of change are variable. Applying the likelihood ratio test to the obtained likelihoods reveals greater accuracy in the estimated maximum likelihood when more rate parameters are considered (Table 2.3). Rate Case 4 returned the highest maximum likelihood estimate of -34823.650, reflecting the superior precision of the more complicated rate model ($\chi^2 = \Delta 2$ LnL > 55.76 with d.f. = 40). The results of the AIC test (Table 2.4) also identify rate Case 4 as the superior hypothesis. The number of parameters and the maximum likelihood of Case 4 produced the MAICE of 69739.3, while none of the other cases had comparable support. Therefore, the rate of *Bacillaceae* evolution is best modeled when branches evolve according to independent rates of gene insertion and gene deletion.

In the maximum likelihood analysis, optimal rates of gene insertion and gene deletion were assumed to be unequal. Differential rates of acquisition and loss were considered in order to provide a more realistic model of bacterial evolution. The genomes of prokaryotes are dynamically shaped by gene gains, gene losses, and LGT, but these factors do not necessarily occur in such a balanced manner. For example, gene insertions were found to dominant the lineage of the enterobacteria group, alpha-proteobacteria group, and the *Streptococcus* group studied by Daubin *et al.* (2003), and assist in the adaptive evolution of poxviruses (McLysaght *et al.* 2003). Many models of bacterial evolution (Snel *et al.* 2002; Mirkin *et al.* 2003; Novozhilov *et al.* 2005) have attempted to infer the varying degree to which each evolutionary factor controls the innate gene content. These models often assume conditional penalties on certain rate parameters, thereby limiting the biological procession of inheritance. Although the likelihood model operates in a fixed genome, no restrictions are applied in the calculation of insertion and deletion rates. The likelihood model accounts for the differential rates of insertion, deletion, and LGT governing the phylogeny and identifies the dominating factor(s) influencing genome evolution. From these predicted rates, the true nature of *Bacillaceae* evolution may be better inferred.

2.7 ACKNOWLEDMENTS

This research was supported by a National Science and Engineering Research Council of Canada (NSERC) grant to G.B.G.

Table 2.1. Optimal insertion and deletion rates as predicted by the multidimensional maximum likelihood analysis using Powell's algorithm (1964), bootstrap testing, and curvature method for the different rate cases (Case 1 – Case 3 in Fig. 1.2 and Case 4 in Fig. 2.1). The rate of gene insertion is *v* and the rate of gene deletion is *u*.

			Bootstrap	**Curvature**
Rate		**MLE**	**MLE ± St. Dev**	**St. Dev**
$\alpha = \beta = \gamma$	v	0.7346	0.7346 ± .0057	0.0081
	u	0.6132	0.6136± 0.0275	0.0151
α	v	14.08	14.09 ± 0.3920	0.2457
	u	1.809	1.812 ± 0.1082	0.1022
$\beta = \gamma$	v	0.4049	0.4051 ± 0.0089	0.0067
	u	0.3524	0.3516 ± 0.0169	0.0118
α	v	9.909	9.906 ± 0.2852	0.2191
	u	2.289	2.285 ± 0.1167	0.1090
β	v	0.3633	0.3638 ± 0.0074	0.0065
	u	0.3376	0.3378 ± 0.0159	0.0116
γ	v	1.592	1.591 ± 0.0603	0.0448
	u	0.6351	0.6359 ± 0.0513	0.0403
α_1	v	1.227×10^{-13}	$3.44\times 10^{-11} \pm 4.38 \times 10^{-11}$	infinite
	u	4.109	4.158 ± 2.055	2.017
α_2	v	2.162	2.128 ± 2.168	2.022
	u	1.030	0.9949 ± 1.007	0.9577
α_3	v	20.52	20.60 ± 6.902	6.778
	u	16.22	16.26 ± 4.107	4.163
α_4	v	24.43	24.34 ± 2.221	1.898

			Bootstrap	Curvature
Rate		MLE	MLE ± St. Dev	St. Dev
	u	3.027	3.010 ± 0.6923	0.6240
α_5	v	21.42	21.37 ± 1.551	1.392
	u	2.130	2.136 ± 0.3992	0.3827
α_6	v	24.70	24.71 ± 1.743	1.459
	u	9.297	9.295 ± 0.8556	0.7823
α_7	v	5.845	5.844 ± 0.3864	0.3151
	u	0.4116	0.4068 ± 0.1805	0.1324
α_8	v	0.5608	0.5581 ± 0.0254	0.0218
	u	0.4714	0.4724 ± 0.0433	0.0352
α_9	v	0.5025	0.5009 ± 0.0284	0.026
	u	0.1475	0.1483 ± 0.0240	0.0213
α_{10}	v	0.4247	0.4245 ± 0.0265	0.0238
	u	0.1349	0.1352 ± 0.0223	0.0202
α_{11}	v	0.3953	0.3958 ± 0.0206	0.0177
	u	0.1564	0.1564 ± 0.0228	0.0203
α_{12}	v	0.4595	0.4587 ± 0.0207	0.0193
	u	0.2153	0.2145 ± 0.0273	0.0233
α_{13}	v	1.082×10^{-10}	1.08×10^{-10} ± 0	infinite
	u	0.3345	0.3353 ± 0.0245	0.0205
α_{14}	v	38.87	38.94 ± 1.907	1.699
	u	7.074	7.052 ± 0.6440	0.6092
α_{15}	v	2.367	2.359 ± 0.4755	0.4254

Rate		MLE	Bootstrap MLE ± St. Dev	Curvature St. Dev
	u	0.7089	0.7037 ± 0.1888	0.1802
α_{16}	v	13.57	13.45 ± 1.429	1.250
	u	2.033	2.017 ± 0.7029	0.5655
α_{17}	v	2.584	2.593 ± 0.2754	0.2044
	u	1.467×10^{-11}	$1.47 \times 10^{-11} \pm 0$	infinite
α_{18}	v	1.684	1.685 ± 0.0524	0.0394
	u	0.4339	0.4319 ± 0.0453	0.0336
α_{19}	v	0.9476	0.9469 ± 0.0449	0.0383
	u	0.0131	0.0173 ± 0.0181	0.0211
α_{20}	v	0.0182	0.0353 ± 0.0435	0.0431
	u	3.042×10^{-11}	$4.83 \times 10^{-11} \pm 2.54 \times 10^{-11}$	infinite
α_{21}	v	0.2275	0.2152 ± 0.0651	0.0509
	u	1.626	1.624 ± 0.1951	0.1560
α_{22}	v	0.1385	0.1384 ± 0.0307	0.0255
	u	0.2604	0.2633 ± 0.0548	0.0464
α_{23}	v	8.304×10^{-11}	$8.30 \times 10^{-11} \pm 0$	infinite
	u	4.480	4.497 ± 0.2429	0.1986

Table 2.2. Optimal insertion/deletion rates predicted by Powell's (1964) maximization algorithm, bootstrap testing, and curvature method for the different rate cases (Case 1-Case 3 in Figure 1.2 and Case 4 in Figure 1.3).

		Bootstrap	**Curvature**
Rate	**MLE**	**MLE ± St. Dev**	**St. Dev**
$\alpha = \beta = \gamma$	0.7184	0.7186 ± 0.0077	0.0081
α	6.454	6.464 ± 0.2279	0.1493
$\beta = \gamma$	0.4632	0.4629 ± 0.0076	0.0061
α	5.326	5.320 ± 0.1552	0.1163
β	0.3719	0.3716 ± 0.0062	0.0055
γ	2.244	2.247 ± 0.0991	0.1003
α_1	2.937	2.876 ± 1.498	1.465
α_2	1.396	1.398 ± 1.018	1.019
α_3	18.28	18.20 ± 3.678	3.735
α_4	11.73	11.77 ± 1.019	0.8964
α_5	7.194	7.167 ± 0.6849	0.593
α_6	15.39	15.40 ± 0.8264	0.8110
α_7	1.701	1.749 ± 0.4232	0.1774
α_8	0.6022	0.6036 ± 0.0269	0.0246
α_9	0.3513	0.3530 ± 0.0241	0.0208
α_{10}	0.3329	0.3321 ± 0.0220	0.0199
α_{11}	0.3257	0.3254 ± 0.0168	0.0162
α_{12}	0.4105	0.4109 ± 0.0190	0.0184
α_{13}	0.0108	0.0139 ± 0.0108	0.0120

Rate	MLE	Bootstrap MLE $\pm$ St. Dev	Curvature St. Dev
α_{14}	17.83	17.81 $\pm$ 0.8182	0.7945
α_{15}	1.380	1.390 $\pm$ 0.2170	0.2054
α_{16}	9.128	9.175 $\pm$ 0.8201	0.6958
α_{17}	1.551	1.516 $\pm$ 0.3279	0.1373
α_{18}	2.365	2.366 $\pm$ 0.1074	0.1023
α_{19}	0.7471	0.7465 $\pm$ 0.0394	0.0372
α_{20}	2.719×10^{-11}	$5.65 \times 10^{-11} \pm 3.19 \times 10^{-11}$	infinite
α_{21}	0.6029	0.5994 $\pm$ 0.0755	0.0710
α_{22}	0.1949	0.1949 $\pm$ 0.0281	0.0263
α_{23}	2.055	2.032 $\pm$ 0.10907	0.0943

Table 2.3. Results of the likelihood ratio test for the four rate cases (Case 1 – Case 3 in Fig. 1.2 and Case 4 in Fig. 2.1) assumed on the *Bacillaceae* phylogeny.

Rate Case	LnL	- Δ2LnL	df	P-value
Case 1	-41672.978	-	-	-
Case 2	-37167.489	9010.978	2	5.99
Case 3	-36742.538	849.902	2	5.99
Case 4	-34823.650	3837.776	40	55.76

Table 2.4. AIC statistics calculated from the maximum likelihood estimated by each rate case (Case 1 – Case 3 in Fig. 1.2 and Case 4 in Fig. 2.1). The AIC value of rate Case 4 gives the MAICE.

Rate Case	LnL	k	-2 Ln(L) + 2k	Difference from MAICE
Case 1	-41672.978	2	83349.956	13610.656
Case 2	-37167.489	4	74342.978	4603.678
Case 3	-36742.538	6	73497.076	3757.776
Case 4	-34823.650	46	69739.3	-

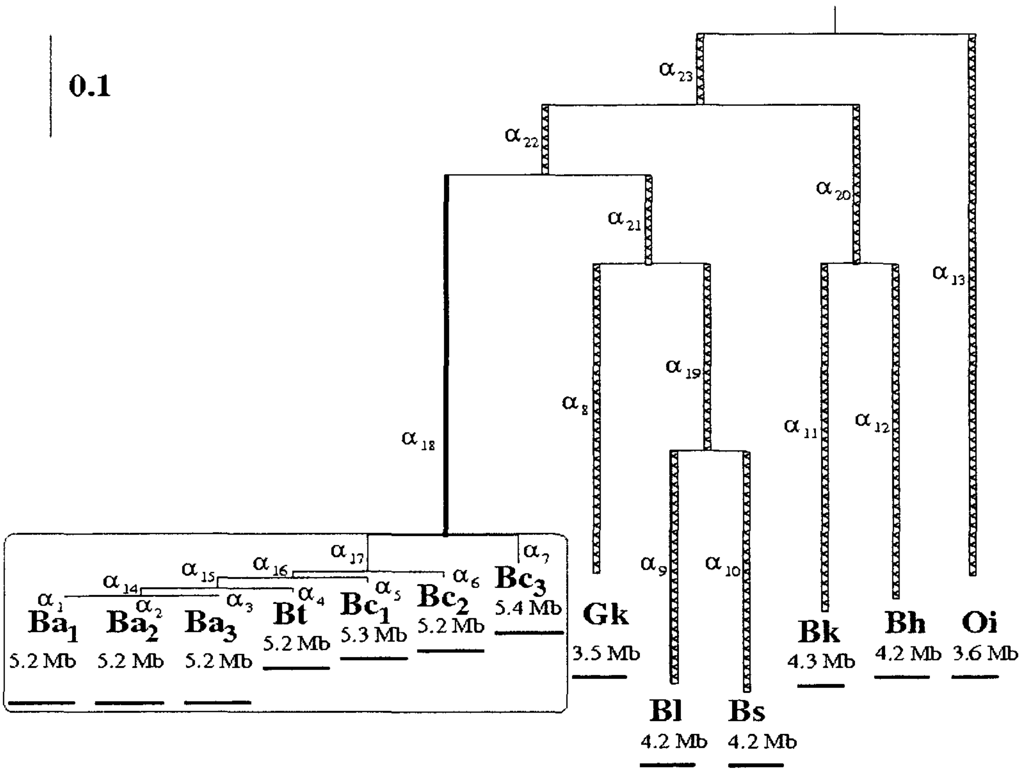

Figure 2.1. Individual rates of branch evolution assumed on the *Bacillaceae* phylogeny. Case 4: branch specific insertion and deletion rates ($\alpha_1,\alpha_2,\ldots,\alpha_{23}$).

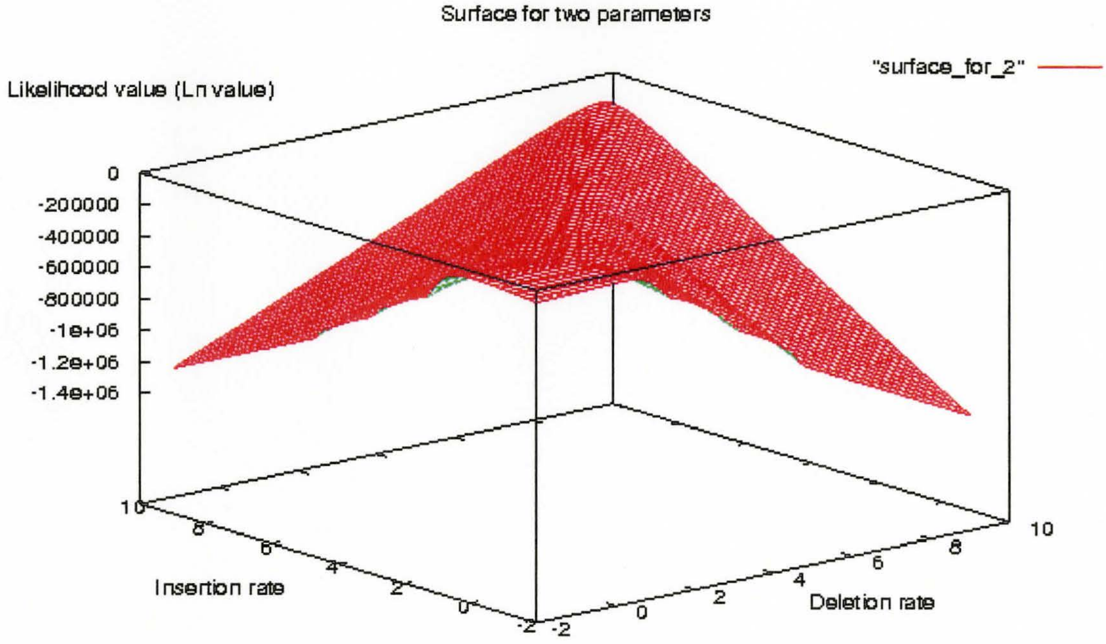

Figure 2.2. The likelihood surface of various insertion and deletion rates modeled under the simple case of a single constant rate α assumed on the phylogeny (Case 1 in Fig. 1.2). The plotted points resemble a saddle with steep sloping edges bordering the maximum likelihood estimate located at the top.

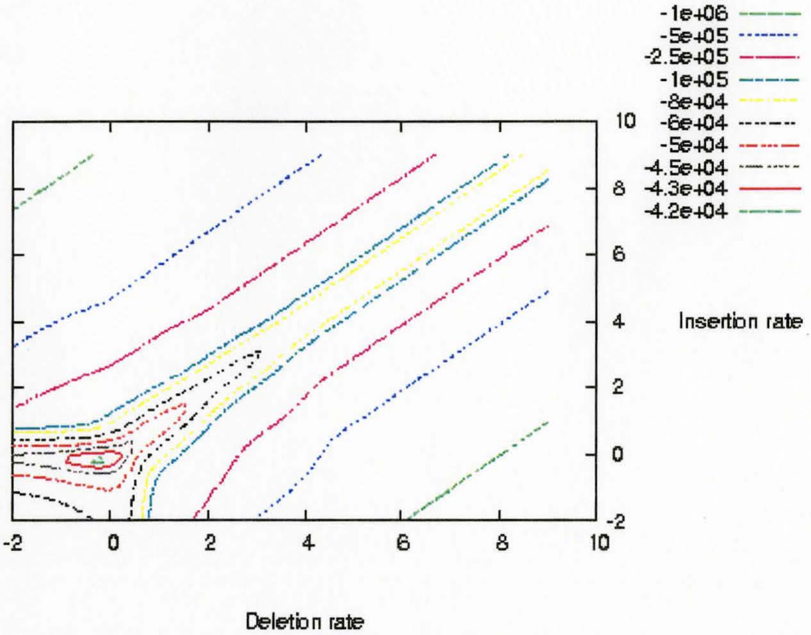

Figure 2.3. Contour map of the likelihood surface produced using various rates of gene insertion and deletion for the simple rate model of a single rate α assumed on the phylogeny (Case 1 in Fig. 1.2). The contour lines are concentrated around the maximum likelihood estimate and rapidly drop off around the surrounding edges.

Chapter 3

A birth-death model of lateral gene transfer in *Bacillaceae*

3.1 ABSTRACT

Lateral gene transfer (LGT) is an important source of evolution in prokaryotic genomes. Acquisition of novel genes via LGT can promote adaptive evolution and help a species survive a new niche. As such, many stochastic models have been developed to infer the role of LGT in prokaryotic evolution. From gene presence and absence data, the evolutionary history of a phylogeny can be reconstructed and the rate of LGT determined. Here, we employ a simple birth-death model with immigration to calculate the optimal rate of gene duplication, LGT, and gene deletion for a group of thirteen fully sequenced *Bacillaceae* genomes. The rate of duplication, LGT, and deletion was assumed constant across the entire phylogeny and only six phyletic patterns were considered in the study. Based on the evolutionary model analyzed and the data set used, elevated accounts of

LGT are only noted when there is a substantial influx of genes from the outer phyletic branches to members belonging to *B. anthracis*, *B. cereus*, and *B. thuringiensis* (the Bc group). In future studies the entire set of phylogenetic patterns should be used in the analysis and the assumptions of the birth-death model should be statistically verified.

3.2 INTRODUCTION

Laterally transferred genes have been extensively noted in microbial genomes (Lan and Reeves 1996; Gogarten *et al.* 2002; Snel *et al.* 2002; Jain *et al.* 2003; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Linz *et al.* 2007; Marri *et al.* 2006, 2007). Together with gene insertions and gene deletions, lateral gene transfer (LGT) alters the current and ancestral gene content and promotes evolution (Snel *et al.* 1999; Gogarten *et al.* 2002; Snel *et al.* 2002; Kunin and Ouzounis 2003; Mirkin *et al.* 2003; Lake and Rivera 2004; Novozhilov *et al.* 2005). By introducing foreign genes into the genome, it rapidly increases species diversity and can help the species adapt to adverse environmental conditions (Lan and Reeves 1996; Gogarten *et al.* 2002; Daubin *et al.* 2003a; McLysaght *et al.* 2003; Hao and Golding 2004, 2006; Lake and Rivera 2004; Marri *et al.* 2006, 2007). Although the importance of LGT in bacterial evolution is widely acknowledged, some advocate that its role in controlling genome progression is exaggerated (Kunin and Ouzounis 2003; Kurland *et al.* 2003; Kurland 2000; Kurland 2005). But, many studies of prokaryotic evolution have been offered that confirm the dominant impact of LGT in shaping genomic history

In bacteria, LGT can be detected by analyzing the gene content of closely related species. Because these genomes exhibit high sequence similarity, any abnormalities in gene composition or codon usage may be representative of LGT (Daubin *et al.* 2003a). The presence and absence of a gene is used to reconstruct the evolution of gene content and the phylogenetic relationships may be inferred according to the method of maximum parsimony (Snel *et al.* 2002; Daubin et al. 2003a, b; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Hao and Golding 2004) or calculation of the evolutionary distance between two genomes (Snel *et al.* 1999). Other studies (Karlin *et al.* 1997; Karlin 1998; Karlin *et al.* 1999) use similarities in genome signatures, defined as the relative abundance of dinucleotides in a genome, to infer phylogenetic evolution, and have even identified prokaryotic LGT into animal mitochondria (Mt) genomes (Karlin *et al.* 1999). The evolution of gene content has also been examined using the approach of maximum likelihood (Gu 2001; Kunin and Ouzounis 2003; Huson and Steel 2004; Lake and Rivera 2004; Hao and Golding 2006; Marri *et al.* 2006; Linz *et al.* 2007; Marri *et al.* 2007). In most likelihood models, phyletic patterns of gene insertion and gene deletion are used to reconstruct the evolutionary relationship of the phylogeny. Other models employ Markov processes (Galtier 2007) or the Poisson distribution (Linz *et* al. 2007) in their likelihood analysis to infer the evolutionary history of the genome.

Stochastic processes of birth and death have also been employed to model the rate of gene acquisition and loss governing bacterial genome growth (Berg and Kurland 2002; Gu and Zhang 2004; Huson and Steel 2004; Novozhilov *et al.* 2005). The statistical implications of the birth-death model were first introduced in the correlation study of death and diminishing surname frequency by Watson and Galton (1875) (Novozhilov *et al.* 2006). Early applications of the theory attempted to model the growth of a population based on incidences of birth, death, and irregular mutation events in a given time period (Yule 1925). It was not until the results of the generalized birth and death process were completely formulated (Kendal 1948a), that the biological importance of the method became evident (Novozhilov *et al.* 2006; Nee 2006). Since then, many derivatives of the general birth-death process have been successfully applied in models of phylogenetic reconstruction (Harvey 1994; Nee *et al.* 1994). In prokaryotes, the 'birth' of a gene results from duplication or LGT and the 'death' of a gene results from deletion. Birth and death models of evolutionary growth have been used in gene content studies to infer the rate of gene proliferation and loss in the genome (Gu and Zhang 2004; Huson and Steel 2004). In other phylogenomic studies, the external influences of: selection, drift, mutational inactivation, LGT between members of different species (Berg and Kurland 2002), and LGT between members of the same species (Novozhilov *et al.* 2005) are also included in the birth and death analysis. Because the rates of gene insertion and gene deletion depend upon many parameters, these models require the use of sophisticated algorithms to calculate the evolutionary rates controlling phylogenetic growth. In this study, we apply a simple birth-death model with immigration (Kendall 1948b; Karlin and McGregor 1958; Bailey 1964) to thirteen completely sequenced *Bacillaceae* genomes, to infer the rate of duplication, LGT, and deletion on the phylogeny. By focusing on only the rate of gene insertion and deletion void of external evolutionary forces, our model offers a clear and computationally feasible method of phylogentic reconstruction.

Differential rates of gene duplication, LGT, and gene deletion were assumed on the phylogeny and modeled according to a steady state of evolution on all branches. With the phylogentic patterns investigated, however, the influence of LGT on genome evolution was weak. Only the genetic sequences of the final two patterns show extensive accounts of LGT, as predicted with the high influx of genes from the out group *Oceanobacillus iheyensis* (Oi) and members of the Bc group. In fact, the first two phyletic patterns appear to produce optimal rate estimates that suggest no reasonable biological meaning. Clearly, the methodology of the current model needs to be improved. In future investigations, estimates for the optimal rates should be based on the entire set of observed gene family patterns, rather than only a single pattern, and the assumptions of the model should be subjected to further rigorous statistical testing to confirm the accuracy of the results.

3.3 THE MODEL

Simple birth and death models are commonly used to monitor the change in population size with respect to the per capita birth rate and death rate, at a given period in time. Changes in the state of the system can only occur between three possible transition states, as illustrated in Figure 3.1. Here, the incident of birth is denoted by the addition of an individual to the population, the incident of death is denoted by the subtraction of an individual from the population, and the incident of neither a birth nor a death occurring is simply denoted by a constant population size. All three transitional states and their associated probabilities may be summarized as follows:

$$\begin{array}{lll}
\text{birth} & n = n + 1, & vn\,dt + 0\,dt; \\
\text{neither} & n = n, & 1 - (v + u)n\,dt + 0\,dt; \\
\text{death} & n = n - 1, & un\,dt + 0\,dt;
\end{array}$$

where n is the size of the population, t is time, v is the birth rate, and u is the death rate (Kendall 1948; Bailey 1964). The incident of birth, and likewise death, is dependent upon the number of individuals in the population, and thus, its affect is measured as a factor of n. It is also important to note that only one of these events can occur in one instant of time and that the birth rate and death rate are assumed non-negative.

In order to model the impact of lateral gene transfer (LGT) on *Bacillaceae* evolution, the possibility of immigration was considered in the above simple birth and death process. Immigration is an external factor that contributes to an increase in population size (Karlin and McGregor 1958) and, together with the birth rate, is represented in the transitional growth of the system. In our model, the size of the population is measured as the number of gene in a family present in a given species. The rate of phylogenetic evolution can then be inferred from the probability of acquiring a new gene from duplication or LGT, or the probability of losing a gene from a deletion. The dynamics of the model are summarized in the following difference equations:

$$\begin{aligned}
p_0(t+1) &= -\lambda + up_1(t), \\
p_n(t+1) &= [v(n - 1) + \lambda]p_{n-1}(t) - [(v + u)n + \lambda]p_n(t) \\
&\quad + u(n + 1)p_{n+1}(t),
\end{aligned}$$

where t is time, v is gene duplication, λ is LGT, and u is gene deletion. Note, that t denotes the length of time separating a descendant from its ancestor and is measured as the expected number of nucleotide substitutions per site (Hao and Golding 2006). With the addition of immigration, the general from of the transitional probability matrix becomes:

$$
\begin{aligned}
M_{00} &= (1 - \lambda),\\
M_{01} &= u,\\
M_{n,n+1} &= u(n + 1),\\
M_{nn} &= 1 - [(v + u)n + \lambda],\\
M_{n,n-1} &= v(n - 1) + \lambda,\\
M_{N,\mathrm{N}-1} &= v(N - 1) + \lambda,\\
M_{NN} &= 1 - uN
\end{aligned}
$$

(Kendall 1948; Karlin and McGregor 1958; Bailey 1964),

where N is the maximum allowed number of genes in a family. The act of immigration occurs independently of population size and, hence, it is not scaled according to n. The birth rate and death rate still depend on the number of genes present in a family and, as before, all parameters are assumed positive. Note, that because the system calculates the transitional probability of a state, it requires all matrix entries to be less than one and the entries of each column to sum to one.

3.4 METHODS

Initial research on bacterial evolution aimed at identifying the optimal insertion and deletion rate shaping the phylogeny of a group of Gram-positive *Bacillaceae*. The thirteen fully sequenced bacterial species comprising the group include: *B. anthracis* Ames, *B. anthracis* "Ames ancestor," *B. anthracis* Sterne, *B. thuringiensis*, *B. cereus* ZK, *B. cereus* ATCC 10,987, *B. cereus* ATCC 14,579, *Geobacillus kaustophilus*, *B. licheniformis*, *B. subtilis*, *B. clausii*, *B. halodurans*, and *Oceanobacillus iheyensis*. High sequence similarity between the strains belonging to *B. anthracis*, *B. cereus*, and *B. thuringiensis* (Ash et al. 1991; Priest et al. 2004) lead to the supplementary grouping of these members into the Bc group. Equal rates of gene insertion and deletion (indel rate) were assumed on the phylogeny, and the maximum likelihood algorithm of Hao and Golding (2006) was modified (Chapter 1) to calculate the optimal rate of evolution under four different rate scenarios (Case 1 - Case 3 in Fig. 1.2 and Case 4 in Fig. 1.3). By continuously testing possible rate estimates encircling a subset of three points, the new algorithm was able to converge on those values representative of the optimal indel rate(s). Little deviation in the predicted estimates was observed when the algorithm was subjected to rigorous bootstrap sampling and variance measurements from the likelihood curve. The evolutionary assumptions of rate Case 4 achieve the highest estimate of likelihood and, as confirmed by the likelihood ratio test and Akaike Information Criterion (AIC), support the strongest model of *Bacillaceae* evolution.

The maximum likelihood model was then further modified to calculate varying rates of gene insertion and gene deletion for the phylogeny (Chapter 2). By acknowledging that the rate at which a gene is inserted or deleted need not be equal, the new model was able to provide a more accurate depiction of *Bacillaceae* evolution. The differential rate model was applied to the same group of closely related *Bacillaceae* species, and the rate of evolution was inferred according to four defined rate cases (Case 1 – Case 3 in Fig. 1.2 and Case 4 in Fig. 2.1). Powell's (1964) optimal convergence algorithm was employed to determine the insertion rate(s) and deletion rate(s) required to maximize the likelihood. The resulting optimal estimates were then statistically verified via bootstrap testing and curvature measurements from the likelihood curve. For both tests, almost no variation was observed in the rates predicted for the first three rate cases, and only modest deviation was observed in the estimates of Case 4. Once again, the likelihood ratio test and AIC rank rate Case 4 as the superior model, providing further support that the evolution of the phylogeny is most accurately reflected when independent rate parameters are assumed for each branch.

In order to satisfy a more concrete explanation of the factors regulating bacterial evolution, the same *Bacillaceae* phylogeny was investigated in our simple birth and death model (Section 3.3). When different modes of evolution are imposed on the same data set, it is interesting to note the degree of overlap between the results. Any congruencies between the models may validate the assumptions of the hypotheses and provide valuable insight on the evolutionary patterns governing bacterial growth. In the birth-death model, the rate of LGT was measured as a separate parameter, λ, alongside the rate of gene duplication. The rate of evolution was inferred from the number of genes present in a species. Therefore, an addition of a gene via gene duplication or LGT contributes to genomic growth, and the deletion of a gene results in genomic decay. This requires the state of having no members of a gene family to be the transitional probability of going from one gene to none, $M_{01} = u$, or the transitional probability of remaining at zero, $M_{00} = 1 - \lambda$. Note that the rate of gene duplication is not included in the transitional probabilities of having no members in a family, as duplication cannot occur when no genes are present. Also note that the state of the system can never be negative because a negative number of genes in a family is not biologically sensible. In similar manner, the probability of obtaining the maximum number of gene in a family, N, can only result from the transitional state of growth, $M_{N,N-1} = v(N-1) + \lambda$, or the transitional probability of remaining at N, $M_{NN} = 1 - uN$. Duplication of a gene cannot occur past N as it would overshoot the boundaries defined by the model and yield no biological meaning. Our model allows for a maximum of a hundred possible genes in one family, $N = 100$.

To determine the expected number of gene families for the phylogeny, the 7228 gene patterns cataloging the presence or absence of a gene (see Hao and Golding 2006 or Table 1.5) were altered to include the actual number of genes present in a species (see Table 3.1 for the most commonly noted patterns and their frequency). Calculating the likelihood of a particular gene family pattern follows the same approach as discussed in Chapter 1 and Chapter 2. The initial probabilities at the tips of the phylogeny are based on the observed number of genes present in a given species. Because this information is known for those species located at the tip of the phylogeny, the probability of obtaining the observed number of genes in the family is 1 and all other possibilities are 0. The likelihood of observing a particular gene pattern in an ancestral species is dependent upon the observed likelihoods of its descendant taxa, separated by t_1 and t_2 generations (see Fig 1.4). Therefore, the transitional probabilities of the given gene pattern must be determined for each descendant taxa. This is achieved by monitoring the change in the state of the system for each generation, or branch length, separating the descendant taxa from its ancestor. The likelihood of the ancestral gene pattern is then calculated as:

$$\begin{aligned} L^x_{\mathrm{G}}(P) = {} & (Prob(P_d|\, P_a, t_1) * L^x_{\mathrm{E}}(P) + Prob(A_d|\, P_a, t_1) * L^x_{\mathrm{E}}(A)) \\ & x \quad (Prob(P_d|\, P_a, t_2) * L^x_{\mathrm{F}}(P) + Prob(A_d|\, P_a, t_2) * L^x_{\mathrm{F}}(A)) \end{aligned}$$

where P is the presence of a gene family, A is the absence of a gene family, d is descendant node, a is ancestral node, G is the ancestor taxon, E is a descendant, and F is a descendant (Hao and Golding 2006). The process is repeated until the root of the phylogeny is reached. Note that, unlike the previous models of Chapter 1 and Chapter 2, the birth and death algorithm only reports the likelihood of a single gene pattern for each run of the program. In total, the evolution of six different gene family patterns (Table 3.2) was examined in the birth and death analysis.

At the root of the tree, the overall likelihood of the observed phyletic pattern is calculated by multiplying the predicted likelihood for the number of genes present in a family by the expected number of occurrences of a particular gene. Because the solution to the transitional matrix is too intricate and difficult to achieve, the Poisson distribution was used to approximate the expected probability for the number of genes in a family. For each phyletic pattern, the number of genes observed to occur across the thirteen species were averaged and summed with the corresponding averages for the rest of the phyletic patterns, to obtain the average number of genes present in a family, X, for a given time interval. Thus, the expected number of genes for the entire phylogeny is:

$$Prob(x|X) = \frac{X^x e^{-X}}{x!}$$

where x is the observed number of genes in a family. These probabilities were then multiplied by the root likelihood to obtain the overall likelihood of the tree. The resulting product can be an extremely small value and, therefore, the log of the overall likelihood is used.

Only the simple evolutionary scenario of a single rate for gene duplication, LGT, and gene deletion was assumed on the *Bacillaceae* phylogeny in the birth-death analysis (Fig. 3.2). Although Powell's (1964) maximization procedure was successful in determining the optimal rates of the differential likelihood model (Chapter 2), it proved too intricate and time consuming in calculating the rates for the birth and death model. Therefore, to decrease the time of the optimization process, a grid of rates within the natural logarithmic interval of 10^{-6} to10^{1} were tested to see which gave the best likelihood estimate. The interval was subdivided to test ten points equally spaced within the defined logarithmic boundary. For each consecutive trial, this required the value of the rate to be incremented by a factor of $10^{0.7}$, 5.011872336, up to the maximum allowed value of 10^{1}. Each rate estimate was then tried in combination with all possible estimates of the other two variables, to see which parameters produced the highest estimate of likelihood. Once the optimal values for duplication, LGT, and deletion were identified in the logarithmic interval, they were further increased or decreased by a factor of 1% to see if any slight deviation from the predicted optimal rate would yield a higher estimate of likelihood. Values that succeeded in raising the likelihood estimate were tested again in combination with the other rate estimates to see if a sequential increase or decrease in the rate parameters would again result in a higher estimate for the likelihood. This procedure was repeated until the altered rate parameters produced lower than maximal likelihood estimates, specifically, until the maximum likelihood was reached. The maximum likelihood estimate and optimal duplication, LGT, and deletion rates for each of the six gene patterns are listed in Table 3.3.

3.5 RESULTS

The evolution of gene family patterns was inferred for the *Bacillaceae* phylogeny using the birth and death model detailed in sections 3.3 and 3.4, under the assumption of a single constant rate of gene duplication, LGT, and gene deletion (Fig. 3.2). The optimal rates for each of the gene patterns considered, and the associated likelihoods, are listed in Table 3.3. Genomes of high sequence similarity were considered in the study, because the effect of LGT is easier to detect in gene abnormalities among closely related species. Previous genomic studies, using the method of maximum likelihood, were successful in estimating both the indel rate (Chapter 1) and rate of gene insertion and deletion (Chapter 2) for the phylogeny. In this study, differential rates of gene duplication, LGT, and gene deletion were imposed on the phylogeny, with the intent to identify the sole impact of LGT in genome evolution. To achieve this, specific patterns of gene families (Table 3.2) were chosen to reflect the role of LGT at different stages of evolution. Unfortunately, with the phylogenetic patterns selected, the extent to which LGT has shaped the *Bacillaceae* genome is poorly characterized. Although a rate of LGT is predicted for each pattern by the model, incorporating a greater number of tested gene patterns into the algorithm would constitute a more reliable data set and, thus, offer better insight on the importance of LGT in bacterial evolution.

The patterns listed in Table 3.3 are ordered down the table according to an increase in the observed number of genes in a family. Starting at the first pattern, only a single gene is detected in the genomes of Ba_2 and Ba_3. Applying the birth and death algorithm to this model reveals the rate of gene deletion to be the greatest, 0.010, followed by the rate of LGT, 9.789×10^{-4}. The effect of gene duplication does not appear to play a role in the evolutionary sequence. Similarly, the presence of a single gene across more members of the phylogeny in the next pattern produces comparable results, with a deletion rate of 1.160×10^{-4} and no rate estimates for both gene duplication and LGT. At the third gene pattern, however, the model predicts equal rates of LGT and gene deletion, 0.0087, for the phylogeny. This is also the first pattern where a rate estimate for gene duplication, although small, 0.0013, is reported. As the number of possible genes in a family increases, from pattern 3 to pattern 6, the optimal rate for LGT gradually surpasses that of gene deletion. The phylogenetic sequence of pattern 6 generates the highest estimate for LGT at 0.0551.

In all cases, the rate of gene duplication is less than the rate of LGT and gene deletion, with most gene patterns suggesting few to no duplication events. Only when a greater number of genes are considered in the phylogenetic pattern, can an estimate for the rate of gene duplication be inferred. In the fifth pattern, it is important to note that the optimal estimate for gene duplication approaches zero but is not zero. Here a duplication rate of 2.104×10^{-14}, along with the optimal LGT, 0.0132, and deletion rate, 1.970×10^{-4}, gives the maximum likelihood estimate of -36.95. The largest rate of gene duplication, 0.0026, estimated by the birth and death model occurs in the gene family pattern of sequence 6.

Together, the rates required to maximize the likelihood in pattern 6 constitute the largest group of values among the patterns tested. The highest rate estimates for both gene duplication and LGT are also predicted by this model. Although the rate of gene deletion is high in the model, a higher rate of deletion is observed for patterns 1 and 3, with the first pattern producing the largest deletion rate of 0.010. It is also interesting to note that the evolutionary assumptions of the second model (pattern 2 in Table 3.3) generate the highest estimate of likelihood.

3.6 DISCUSSION

The genomic history of bacteria is uniquely organized according to the sequence of gene duplications, gene deletions, and LGT events that occur during phylogenetic evolution (Snel *et al.* 1999; Snel *et al.* 2002; Kunin and Ouzounis 2003; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Lake and Rivera 2004; Novozhilov *et al.* 2005; Linz *et al.* 2007). Many computational models of phylogenetic reconstruction (Snel *et al.* 2002; Daubin *et al.* 2003a,b; McLysaght *et al.* 2003; Mirkin *et al.* 2003; Hao and Golding 2004, 2006; Marri *et al.* 2006, 2007; Galtier 2007; Linz *et al.* 2007) have been suggested to estimate the role of each genetic event in controlling bacterial evolution. In this study, a simple birth-death model with immigration was applied to a phylogeny of thirteen highly similar *Bacillaceae* genomes (Fig. 3.2). The rate of gene duplication, LGT, and gene deletion was inferred according to a constant rate of evolution on all branches, for the six phyletic patterns defined in Table 3.3. Although a model detailing the evolutionary history for a lager number of gene family patterns in the *Bacillaceae* group would offer a more accurate depiction of phylogenetic evolution, the patterns studied did provide some useful information on the historical dynamics influencing the phylogeny.

Previous models of *Bacillaceae* evolution were successful in determining both the constant (Chapter 1) and differential rate (Chapter 2) of gene insertion and gene deletion inherent to the phylogeny. A separate parameter for LGT was included in the simple birth and death analysis to infer the degree at which LGT solely affects gene content. The model was applied to the same *Bacillaceae* group studied in the maximum likelihood models of Chapter 1 and Chapter 2, to detect possible congruencies among the predicted results. Any agreement in the inferred rates will help support the assumptions of the methods, and assist in understanding the evolutionary patterns controlling phylogenetic progression. In the model, the current number of genes present in a family is obtained directly from the taxa at the tip of the phylogeny. This approach was taken because reconstruction analysis based on historical data can sometimes provide an overestimation or underestimation of the actual rates as a duplication followed by two deletions or a deletion of a gene before it duplicates in ancestral taxa (Harvey 1994) is not detected in extant taxa. Therefore, by working backwards towards the root of the tree, our model attempts to correct for any such uncertainties in the occurrence of a duplication, LGT, and/or deletion event in ancestral lineages.

Our birth and death model simply estimates the rate of gene duplication, LGT, and gene deletion void of any external evolutionary factors, like genetic drift or selection. Other models of gene acquisition and loss infer the rate of gene ‘birth’ and gene ‘death’ directly from gene content data (Gu and Zhang 2004) or from more sophisticated equations that include the possibility of mutational inactivation, selection, genetic drift, LGT among species of the same phylogeny (Berg and Kurland 2002), and the insertion of a gene from outside the phylogeny (Novozhilov *et al.* 2005). The role of gene insertion, via both duplication and LGT, and deletion in prokaryotic adaptation has been well documented (Heidelberg *et al.* 2000; Riehle *et al.* 2001; McLysaght et al. 2003; Gevers *et al.* 2004; Lolkema *et al.* 2008). Recent duplications in the cholera pathogen *Vibrio cholerae* have promoted its subsistence in adverse environments (Heidelberg *et al.* 2000) and together, the recurring succession of gene duplication and loss has helped *Escherichia coli* adapt to high temperatures (Riehle *et al.* 2001). The profound affect of LGT on bacterial evolution is also made evident in the poxvirus study of McLysaght *et al.* (2003) where the apoptosis inhibitor gene AMV-EPB_034, belonging to the *Amsacta moorei entomopoxvirus*, is noted most similar to the inhibitor gene in the insect *Bombyx morider* (Order: Lepidoptera). Thus, it is the intent of our birth and death model to simply infer the raw rate estimate of gene duplication, LGT, and gene deletion inherent to the *Bacillaceae* phylogeny.

When the most common phyletic patterns of gene families for the *Bacillaceae* group (Table 3.1) were tallied, the majority of the patterns were similar to those predicted in Table 1.5. Although the data set was altered to include the actual number of genes present, it is interesting to note that the most common phyletic patterns still reflect the presence and absence patterns of the original data set. For most of the phylogeny, only the presence of a single gene is noted. This implies that the presence of multiple genes among taxa does not occur as frequently as a single gene.

A general trend is observed in the optimal rate estimates for the phyletic patterns analyzed (Table 3.3). As the number of possible genes in a family increases from pattern 1 to pattern 6, the rate of LGT slowly increases and surpasses that of gene duplication at pattern 3. Additionally, there is almost no indication of gene duplication until the rate estimates of patterns 5 and 6. The rate of gene deletion appears to change intermittently depending on the phyletic pattern of gene families. A higher rate of LGT in the last two patterns is indicative of the large number of gene differences between the out group *Oceanobacillus iheyensis* (Oi) and members of the Bc group. In the fifth gene pattern, Oi is observed to have 15 fewer genes than, for example, *Bacillus anthracis Ames* (Ba_1) of the Bc group, as well as, 30 fewer genes in pattern 6. Furthermore, an increase in the number of available genes can prompt a greater chance of duplication, as evident by the appearance of the duplication rate in the last two patterns. Similar rates of gene deletion and LGT are observed for patterns 3 and 4, with the gene families of pattern 3 owing to equal rates. The equal contribution of both evolutionary factors in pattern 3 may reflect possible LGT events in *Bacillus clausii* (Bk), *Bacillus subtilis* (Bs), *Bacillus anthracis Sterne* (Ba_3), and *Bacillus anthracis* "Ames Ancestor" (Ba_2) and/or gene deletions in *Bacillus licheniformis* (Bl), *Geobacillus kaustophilus* (Gk), and the other members of the Bc group. Likewise, the rates predicted for pattern 4 may indicate possible LGT events in Gk, Ba_3, and Ba_2 and/or deletions in Bl, Bs, Bk, *Bacillus halodurans* (Bh), and the remainder of the Bc group.

The most puzzling results of the investigation occur in the rates estimated for the first two phyletic patterns. Both models evolve according to a higher deletion rate, with no indication of gene duplication and/or LGT in the second pattern. In fact, the optimal rate estimates of the second model give the highest maximum likelihood value among the six phyletic patterns tested. This result implies that the *Bacillaceae* genome is shrinking and, in the absence of gene duplication and LGT, will eventually be extinct. Although some prokaryotic growth models have predicted the rate of gene loss to be substantial greater than acquisition (Berg and Kurland 2002; Novozhilov *et al.* 2005), this is not expected with the observed pattern of gene families in these two sequences. In the first pattern, the presence of a single gene in Ba_2 and Ba_3 is expected to indicate a possible LGT event in these two species. While a rate of LGT is estimated, 9.789×10^{-4}, for this pattern, perhaps its affect is masked by a possible mass deletion event in the rest of the phylogeny. Unfortunately, no satisfactory conclusions can be offered for the results of the second pattern.

All things considered, our simple birth and death model provides a basic and straightforward algorithm from which the rate of *Bacillaceae* evolution can be inferred. In order to improve the preliminary results, the entire set of gene family patterns characteristic of the phylogeny should be included in the analysis. The consideration of a lager number of phyletic patterns will constitute a more reliable data set and establish a more accurate portrayal of the evolutionary factors shaping the genome. Furthermore, the present study only examines the very basic model of rate evolution, a constant rate of gene duplication, LGT, and gene deletion assumed on the phylogeny. Other evolutionary scenarios like: separate rates of evolution for the Bc group, determining a rate along the branch separating the Bc group from the rest of the phylogeny, and branch specific rates, should also be analyzed according to the birth and death process. By applying the assumptions of the model to various rate cases, the parameters that best model the rate of *Bacillaceae* evolution will be identified and will assist in understanding the true mechanism of prokaryotic development. Additionally, the assumptions of the birth and death model need to be statistically verified to confirm the accuracy of the proposed rates. Once statistical support is granted for the model, the algorithm can be applied to other studies of phylogenetic reconstruction to assist in understanding the role of LGT in microbial evolution.

3.7 ACKNOWLEGMENTS

This research was supported by a National Science and Engineering Research Council of Canada (NSERC) grant to G.B.G.

Table 3.1. Frequency of the most common gene family patterns in the *Bacillaceae* phylogeny.

Number of genes	**Ba$_1$**	**Ba$_2$**	**Ba$_3$**	**Bt**	**Bc$_1$**	**Bc$_2$**	**Bc$_3$**	**Gk**	**Bl**	**Bs**	**Bk**	**Bh**	**Oi**
948	1	1	1	1	1	1	1	0	0	0	0	0	0
734	1	1	1	1	1	1	1	1	1	1	1	1	1
251	0	0	0	0	0	0	0	0	1	1	0	0	0
191	1	1	1	0	0	0	0	0	0	0	0	0	0
156	0	0	0	0	0	0	0	1	0	0	0	0	0
148	0	0	0	0	0	0	0	0	0	0	0	1	0
129	1	1	1	1	1	1	0	0	0	0	0	0	0
127	0	0	0	0	0	0	0	0	0	0	1	0	0
119	0	0	0	0	0	0	0	0	0	1	0	0	0
118	0	0	0	0	0	0	0	0	0	0	0	0	1
105	1	1	1	1	1	0	1	0	0	0	0	0	0
95	1	1	1	1	1	0	0	0	0	0	0	0	0
89	1	1	1	0	1	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0	1	0	0	0	0
70	0	0	0	0	0	1	0	0	0	0	0	0	0
65	1	1	1	1	1	1	1	1	1	1	1	1	0
3798	Other patterns												

Table 3.2. The phyletic patterns of gene families tested for the *Bacillaceae* group.

Pattern													
Number	**Ba$_1$**	**Ba$_2$**	**Ba$_3$**	**Bt**	**Bc$_1$**	**Bc$_2$**	**Bc$_3$**	**Gk**	**Bl**	**Bs**	**Bk**	**Bh**	**Oi**
1.	0	1	1	0	0	0	0	0	0	0	0	0	0
2.	1	1	1	1	1	1	1	1	1	1	0	0	0
3.	1	4	2	1	1	0	0	1	1	3	2	0	0
4.	1	1	1	1	1	2	2	3	2	2	2	2	3
5.	55	53	53	50	50	50	50	47	46	46	45	45	40
6.	100	98	98	95	95	88	83	80	80	80	74	74	70

Table 3.3. Optimal rate of gene duplication, lateral gene transfer, and gene deletion, and the maximum likelihood, for each phyletic pattern listed in Table 3.2, as predicted by the birth and death model for the simple case of a single constant rate of evolution (Fig. 3.2).

Pattern Number		MLE	LnL
	v	0	
1.	λ	9.789×10^{-4}	-7.428
	u	0.010	
	v	0	
2.	λ	0	-5.068
	u	1.160×10^{-4}	
	v	0.0013	
3.	λ	0.0087	-25.21
	u	0.0087	
	v	0	
4.	λ	8.347×10^{-5}	-13.75
	u	2.410×10^{-5}	
	v	$\gtrsim 0$	
5.	λ	0.0132	-36.95
	u	1.970×10^{-4}	
	v	0.0026	
6.	λ	0.0551	-41.55
	u	0.0032	

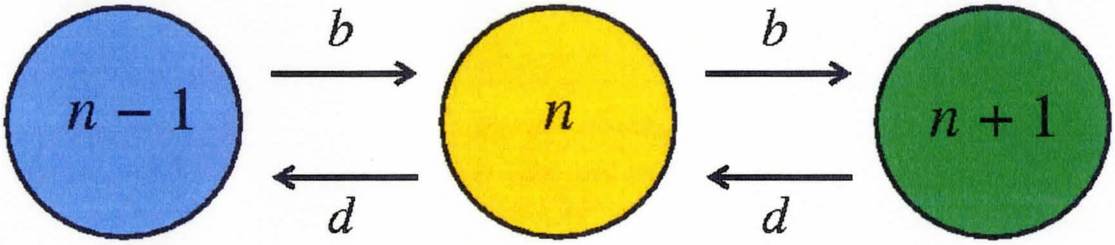

Figure 3.1. The basic birth and death model of population growth, where *n* is the number of individuals in the population. Births are represented by the state *n* + 1 and occur at rate *b*, while deaths are represented by the state *n* - 1 and occur at rate *d*. When there are no births and no deaths, the state of the population is *n*.

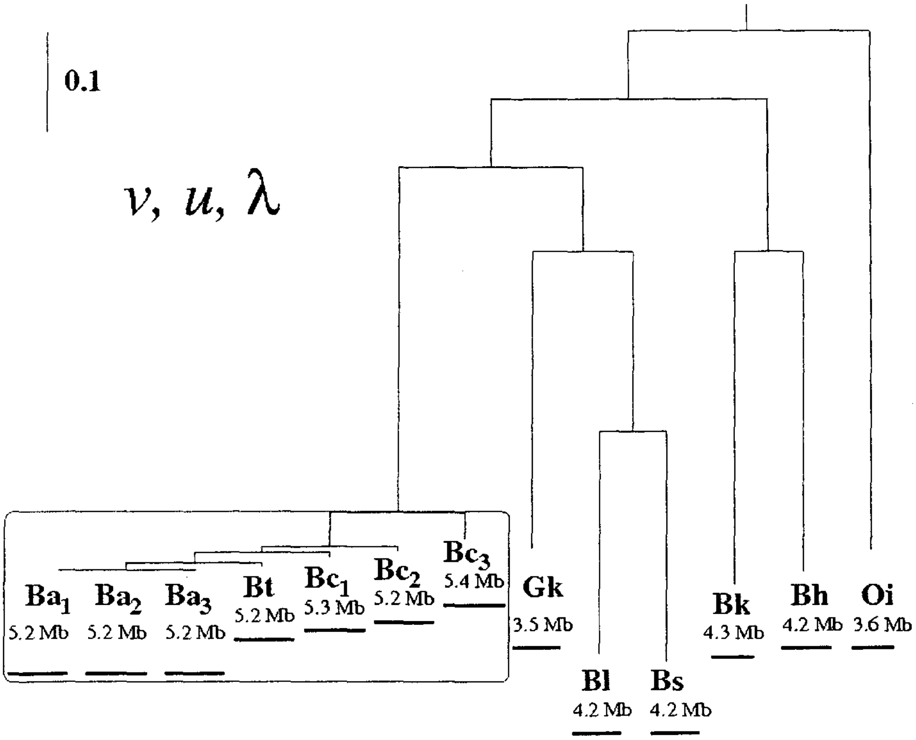

Figure 3.2. The evolutionary model of a constant duplication rate, v, lateral gene transfer rate, λ, and deletion rate, u, assumed on the *Bacillaceae* phylogeny in the birth and death analysis. Members belonging to the Bc group are defined within the boxed section.

Part II

CONCLUSION

The importance of LGT in shaping prokaryotic genomes is becoming increasingly clear, with much appreciation granted in its ability to promote high rates of evolution. Despite pervious claims (Kunin and Ouzounis 2003; Kurland *et al.* 2003; Kurland 2000; Kurland 2005), LGT is found to occur extensively during bacterial evolution and rapidly spreads genetic diversity across the phylogeny. By continuously incorporating novel genes into the taxonomic genome, species are provided the necessary gene pool that allows them to readily adapt to new niches (Hao and Golding 2006; Marri *et al.* 2007). Studies show that extensive accounts of LGT are likely responsible for the subsistence of bacteria in the presence of antibiotics (Berg and Kurland 2002; Gogarten *et al.* 2002; McLysaght *et al.* 2003) and invasion of new hosts (Doolittle 1999; Daubin *et al.* 2003a; Mirkin *et al.* 2003; Marri *et al.* 2006). Rapid rates of adaptive evolution have also been noted in the genomes of *Streptococcus* (Marri *et al.* 2006), *Bacillaceae* (Hao and Golding 2006), and *Corynebacterium* (Marri *et al.* 2007). Together, these findings present alarming evidence on the ever increasing ability of pathogens to resist medical defenses and induce disease onset.

Applying computational algorithms to the evolution of bacteria has greatly assisted in understanding the mechanisms of laterally transferred genes. By simulating possible courses of genome evolution, historical rates of change can be inferred on a phylogeny. When equal rates of gene insertion and gene deletion are assumed for the *Bacillaceae* phylogeny, the likelihood model predicts higher rates of evolution towards the Bc group. Similarly, the differential-rate likelihood model estimates higher rates of gene insertion for members of the Bc group and along branches leading to the group.

Such rapid exchange of gene content within this group may be indicative of adaptive evolution. Previous studies have revealed elevated accounts of nonsynonymous substitution in these *Bacillaceae* genomes, which are suggestive of selectively advantageous evolution (Hao and Golding 2006). Together, the high rates of gene acquisition and directional selection may help the *Bacillaceae* bacteria quickly adapt and facilitate growth in a new niche. Although both likelihood models provide robust rate estimates, the differential rate model is thought to better reflect the true nature of *Bacillaceae* evolution. Accordingly, the assumption of separate insertion and deletion rates along each branch of the phylogeny generates the most probable model of genomic growth.

The historical reconstruction of the *Bacillaceae* phylogeny was extended to a birth and death analysis in our final chapter. Due to time constraints, the optimal rate of evolution could only be inferred from a limited set of phyletic patterns. Growth of the phylogeny was also restricted to a constant rate of gene duplication, LGT, and deletion across all branches. High rates of LGT are noted when there is a substantial influx in the number of genes from the outer branches of the phylogeny to the inner branches of the Bc group. Indeed, this rapid expansion of the genome complements the elevated insertion rates of the likelihood analysis and could very well indicate adaptive evolution. From the patterns used and simple model studied, however, it is difficult to resolve the extent of LGT in shaping the *Bacillaceae* genome. Clearly, the assumptions of the current birth and death model need to be improved.

Future research should take two directions. First, the birth and death algorithm should be modified to calculate the rate of gene duplication, LGT, and deletion based on the total set of phyletic patterns and be applied to each rate scenario studied in the likelihood models. This will provide a more thorough investigation of genomic innovation in a more natural evolutionary context, like determining the branch-specific rates of evolution. Secondly, both likelihood algorithms and the birth-death model should be applied to the same phylogenies studied in other models of bacteria evolution. Thus, although the models may appear to produce satisfactory results for the rate *Bacillaceae* evolution, the findings may differ when the data set is applied to other models of phylogentic reconstruction. In order to further confirm the robustness of our algorithms, bacterial genomes investigated by other reconstruction studies, like *Streptococcus* (Marri *et al.* 2006) and those examined by Spencer *et al.* (2006), should be tested. Any concurrences between the rates estimated by the different methods may assist in identifying the true evolutionary mechanisms present in such bacterial genomes of interest.

Although LGT paints a hazy picture for the universal 'tree of life' some concept of a universal phylogeny can still be gained. It is important to recognize, however, that a species' genetic make-up is highly interdependent on the exchange of ancient gene content between the Archaea, Bacteria, and Eukarya domains (Doolittle 1999). Lateral gene transfer has succeeded in providing the necessary tools for innovative evolution and species proliferation from which modern-day taxa have emerged. Clearly, it is the essence of our being to which all species are united.

Part III

REFERENCES

Bibliography

Akaike, H. 1972. Information theory and an extension of the maximum likelihood principle. Available From: Parzen, E., Tanabe, K. and Kitagawa, G. 1998. *Selected Papers of Hirotugu Akaike (Springer Series in Statistics/Perspectives in Statistics)*. Springer, NY, pp. 199 – 213.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **6**: 716-723.

Ash, C., Farrow, J. A., Dorsch, M., Stackebrandt, E., and Collins, M. D. 1991. Comparative analysis of *Bacillus anthracis*, *Bacillus cereus*, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int. J. Syst. Bacteriol.* **41**: 343-346.

Bailey, N. T. J. 1964. *The elements of Statistical Processes.* John Wiley & Sons, Inc., New York, pp. 97-101.

Berg. O. G. and Kurland, C. G. 2002. Evolution of microbial genomes: Sequence acquisition and loss. *Mol. Biol. Evol.* **19**: 2265-2276.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multi-model inference: a practical information theoretic approach [online]. 2nd Ed. Springer. Available from http://www.myilibrary.com/Browse/open.asp?ID=948&loc=81 (Jan. 11, 2008).

Bushman, F. 2002. *Lateral DNA Transfer.* Cold Spring Harbor Library Press, Cold Spring Harbor, New York, pp. 1-67.

Charlesworth, J. and Eyre-Walker, A. 2006. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23**: 1348-1356.

Daubin, V., Lerat, E., and Perriere, G. 2003a. The source of laterally transferred g enes in bacterial genomes. *Genome Biol.* **4**: R57.

Daubin, V., Nancy, A. M. and Ochman, H. 2003b. Phylogenies and cohesion of bacterial genomes. *Science* **301**: 829-832.

Doolittle, W. F. 1999. Phylogenetic Classification and the Universal Tree. *Science* **284**: 2124-2128.

Edwards, A. W. F. 1972. *Likelihood.* Cambridge University Press, London, pp. 70-77.

Efron, B. 1979a. Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**: 1-26.

Efron, B. 1979b. Computers and the theory of statistics: Thinking the unthinkable. *SIAM J. Appl. Math.* **21**: 460-480.

Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jacknife, and cross-validation. *J. Amer. Statist. Assoc.* **37**: 36-38.

Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**: 54-75.

-------. 1991. Statistical data analysis in the computer age. *Science* **253**: 390-395.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* **22**: 521-565.

-------. 1992. Phylogenies from restriction sites: A maximum-likelihood approach. *Evolution Int. J. Org. Evolution* **46**: 159-170.

-------. 2004. *Inferring pylogenies.* Sinauer Associates, Inc., Sunderland, MA, pp. 248-260.

Fisher, R. A. 1924. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, Ser. A.* **222**: 309-368.

Flannery, B. P., Press, W. H., Teukolsky, S. A. and Vettering, W. T. 1992. *Numerical Recipes in C.* Cambridge University Press, London, pp. 412-430.

Galtier, N. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol.* **56**: 633-642.

Gevers, D., Vandepoele, K., Simillion, C. and Van de Peer, Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.* **12**: 148-154.

Gogarten, J. P., Doolittle, W. F. and Lawrence, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**: 2226-2238.

Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18**: 453-464.

Gu, X. and Zhang, H. 2004. Genome phylogentic analysis based on extended gene contents. *Mol. Biol. Evol.* **21**: 1401-1408.

Hao, W. and Golding, G. B. 2004. Patterns of bacterial gene movement. *Mol. Biol. Evol.* **21**: 1294-1307.

-------. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* **16**: 636-643.

Harvey, P. H., May, R. M. and Nee, S. 1994. Phylogenies without fossils. *Evolution* **48**: 523-529.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L. *et al.* 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae. Nature* **406**: 477-483.

Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. *Syst. Biol.* **44**: 17-48.

Huson, D. H. and Steel, M. 2004. Phylogenetic trees based on gene content. *Bioinformatics.* **20**: 2044-2049.

Jain, R., Rivera, M. C. and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexty hypothesis. *Proc. Natl. Acad. Sci.* **96**: 3801-3806.

Jain, R., Rivera, M. C., Moore, J. E. and Lake, J. A. 2003. Horizontal gene transfer accelerate genome innovation and evolution. *Mol. Biol. Evol.* **20**: 1598-1602.

Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. *Genomics* **1**: 598-610.

Karlin, S., Brocchieri, L., Mrazek, J., Campbell, A. M. and Spormann, A. M. 1999. A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. *Evolution* **96**: 9190-9195.

Karlin, S. and McGregor, J. 1958. Linear growth, birth and death processes. *J. Math. Mech.* **7**: 643-662.

Karlin, S., Mrazek. J. and Campbell, A. M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**: 3899-3913.

Kendall, D. G. 1948a. On the generalized "birth-and-death"process. *Ann. Math. Stat.* **19**: 1-15.

-------. 1948b. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometika* **35**: 6-15.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. *Ann. Math. Stat.* **22**: 79-86.

Kunin, V. and Ouzounis, C. A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**: 1589-1594.

Kurland, C. G. 2000. Something for everyone: Horizontal gene transfer in evolution. *EMBO Rep.* **2**: 92-95.

-------. 2005. What tangled wed: Barriers to rampant horizontal gene transfer. *Bioessays* **27**: 741-747.

Kurland, C. G., Canback, B. and Berg, O. G. 2003. Horizontal gene transfer: A critical view. *Proc. Natl. Acad. Sci.* **100**: 9658-9662.

Lake, J. A. and Rivera, M. C. 2004. Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol. Biol. Evol.* **21**: 681- 690.

Lan, R. and Reeves, P. R. 1996. Gene transfer is a major factor in bacterial evolution. *Mol. Bio. Evol.* **13**: 47-55.

Linz, S., Radtke, A. and von Haeseler, A. 2007. A likelihood framework to measure horizontal gene transfer. *Mol. Biol. Evol.* **24**: 1312-1319.

Lolkema, J. S., Dobrowolski, A. and Slotboom, D. 2008. Evolution of antiparallel two-domain membrane proteins: Tracing multiple gene duplication events in the DUF606 family. *J. Mol. Biol.* **378**: 596-606.

Marri, P. R., Hao, W. and Golding, G. B. 2006. Gene gain and gene loss in *Streptococcus*: Is it driven by habitat?. *Mol. Biol. Evol.* **23**: 2379-2391.

Marri, P., Hao, W. and Golding, G. B. 2007. The role of laterally transferred genes in adaptive evolution. *BMC Evol. Biol.* **7**: S8.

McLysaght, A., Baldi, P. F. and Gaut, B. S. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl. Acad. Sci.* **100**: 15655-15660.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y. and Koonin, E. V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**: 2.

Nee, S. 2006. Birth-death models in macroevolution. *Annu. Rev. Ecol. Evol. Syst.* **37**: 1-17.

Nee, S., May, R. M. and Harvey, P. H. 1994. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lon. B. Biol. Sci.* **344**: 305-311.

Novozhilov, A. S., Karev, G. P., and Koonin, E. V. 2005. Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.* **22**: 1721-1732.

Novozhilov, A. S., Karev, G. P., and Koonin, E. V. 2006. Biological applications of the theory of birth-and-death processes. *Brief. Bioinform.* **7**: 70-85.

Powell, M. J. D. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal.* **2**: 155-162.

Priest, F. G., Baker, M. Baillie, L. W., Holmes, E. C., and Maiden, M. C. 2004. Population structure and evolution of the *Bacillus cereus* group. *J. Bacteriol.* **186**: 7959-7970.

Riehle, M. M., Bennett, A. F. and Long, A. D. 2001. Genetic architecture of thermal adaptation *in Escherichia coli. Proc. Natl. Acad. Sci.* **98**: 525-530

Rivera, M. C., Rain, R., Moore, J. E. and Lake, J. A. 1998. Genomic evidence for two functionally distinct gene classes. *Natl. Acad. Sci.* **95**: 6239-6244.

Schrago, C. G. 2006. An empirical examination of the standard errors of maximum likelihood phylogenetic parameters under the molecular clock via bootstrapping. *Genet. Mol. Res.* **5**: 233-241.

Snel, B., Bork, P. and Huynen, M. A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21**: 108-110.

-------. 2002. Genomes in flux: The evolution of Archaeal and Proteobacterial gene content. *Genome Res.* **12**: 17-25.

Snel, B., Huyen, M. A., and Dutilh, B. E. 2005. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* **59**: 191-209.

Watson, H. W. and Galton, F. 1874. On the probability of extinction of families. *J. Anthrop. Inst. Of Great Britain and Ireland.* **4**: 138-144.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypothesis. *Ann. Math. Stat.* **9**: 60-62.

Woese, C. R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**: 221-271.

Yule, G. U. 1925. A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *J. R. Stat. Soc.* **88**: 433-436.