August 2015

The Effect of Diversification on the Dynamics of Mobile Genetic Elements in Prokaryotes: The Birth-Death-Diversification Model

Nicole E. Drakos
The University of Western Ontario

Supervisor
Dr. Lindi M. Wahl
The University of Western Ontario

Graduate Program in Applied Mathematics

A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

The Effect of Diversification on the Dynamics of Mobile Genetic Elements in Prokaryotes: The Birth-Death-Diversification Model

(Thesis Format: Integrated-Article)

by

Nicole E. Drakos

Graduate Program in Applied Mathematics

A thesis submitted
in partial fulfillment of the requirements for
Master of Science

The School of Graduate and Postdoctoral Studies
Western University
London, Ontario, Canada

Abstract

Mobile genetic elements (MGEs) are ubiquitous among prokaryotes, and have important implications to many areas, such as the evolution of certain genes, bioengineering and the spread of antibiotic resistance. In order to understand the complex dynamics of MGEs, mathematical models are often used. One model that has been used to describe the dynamics of mobile promoters (a class of MGEs) is the birth-death-diversification model. This model is unique in that it allows MGEs to diversify to create new families. In this thesis, I analyze the dynamics of this model; in particular, I examine equilibrium distributions, extinction probabilities and mean time until extinction for MGE lineages. I find that diversification indirectly increases MGE propagation through increased horizontal gene transfer rates; therefore, diversification increases population growth rates and decreases extinction probability. Overall, this work indicates that diversification of elements should be considered in order to fully understand the dynamics of MGEs in prokaryotes.

Keywords: mobile genetic elements, markov chains, mobile promoters, extinction probability, extinction times

The Co-Authorship Statement

The work in Chapter 2 has been submitted for publication:

Nicole E. Drakos and Lindi M. Wahl, 2015. Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: the birth-death-diversification model. Under revision for *Theoretical Population Biology*.

Acknowledgments

First, I would like to thank my supervisor, Dr. Lindi Wahl, for her support during my degree. She is one of the most encouraging, inspiring people I have had the pleasure to know, and I am very grateful for her guidance. I would also like to thank the administrative staff in the Department of Applied Mathematics, Cinthia MacLean and Audrey Kager. They are always available to answer questions, and help create an excellent working environment. I am also indebted to the many professors who have taught me throughout my undergraduate and graduate degree, and to my fellow graduate students for their support and helpful discussions. Finally, I would like to thank the Ontario Graduate Scholarship program for their financial support.

Table of Contents

List of Figures

Chapter 1

n − number of members in a population

λ_n − rate a population moves from state n to $n+1$ in a classic birth-death process

μ_n − rate a population moves from state n to $n-1$ in a classic birth-death process

π_n − steady state proportion of populations of size n

X_n − extinction probability of a population of size n

$P_n(t)$ − probability a population that starts with n members at time zero will be extinct before time t

T_n − expected extinction time of a population of size n given that it will go extinct

Chapter 2

u − rate of duplication for the copy-level model

w − rate of deletion for the copy-level model

v − rate of diversification for the copy-level model

η − rate of horizontal gene transfer for the copy-level model

η_c − critical value of horizontal gene transfer at which the population will not change in size

$\hat{w}$, $\hat{v}$, $\hat{\eta}$ − rates normalized by the duplication rate, u (for example, $\hat{w} = w/u$)

$\hat{w}_p$, $\hat{v}_p$, $\hat{\eta}_p$ − best fit values found for mobile promoter data

X_n − extinction probability for a lineage of mobile genetic elements with n copies

$P_n(t)$ − probability a population that starts with n members at time zero will be extinct before time t

$G_n(t)$ − cumulative distribution function for extinction times of a family that has n copies at time zero

$g_n(t)$ − probability density function for extinction times of a family that has n copies at time zero

T_n − expected extinction time of a population of size n given that it will go extinct

$C_i(t)$ − the expected number of families with i copies at time t

$\bar{C}_i$ − the expected number of families with i copies at equilibrium

c_i − the expected proportion of families with i copies at equilibrium

$k-$ longterm exponential growth rate of MGE families

$\rho-$ the probability that a new mobile genetic element that arrives either through a duplication or HGT event will be inserted in the coding region

$u_1,\, w_1,\, v_1 -$ duplication, deletion and diversification rates in coding region for 2-D copy-level model

$u_2,\, w_2,\, v_2 -$ duplication, deletion and diversification rates in non-coding region for 2-D copy-level model

$\tilde{u}_1,\, \tilde{w}_1,\, \tilde{v}_1,\, \tilde{u}_2,\, \tilde{w}_2,\, \tilde{v}_2 -$ rates normalized by HGT rate, η (for example, $\tilde{u_1} = u_1/\eta$)

$X_{n_1,n_2} -$ extinction probability for a lineage of mobile genetic elements with n_1 copies in the coding region, and n_2 copies in the non-coding region

$P_{i,j}(t) -$ probability a population that starts with i copies in the coding region, and j copies in the non-coding region at time zero will be extinct before time t

$R_{i,j} -$ the expected number of families with i copies in the coding region,and j copies in the non-coding region

$r_{i,j} -$ the proportion of families with i copies in the coding region, and j copies in the non-coding region

$p -$ rate of diversification for the family-level model

$q -$ rate of loss for the family-level model

$\mu -$ rate of horizontal gene transfer for the family-level model

$\hat{q},\, \hat{\mu} -$ rates normalized by the growth rate, p (for example, $\hat{q} = q/p$)

$\hat{q}_p,\, \hat{\mu}_p -$ best fit values found for mobile promoter data

$F_m(t) -$ the expected number of genomes with m families at time t

$\bar{F}_m -$ the expected number of genomes with m copies at equilibrium

$f_m -$ the expected proportion of genomes with m families

$\bar{f} -$ average number of families per genome

$G -$ the total number of genomes

List of Acronyms

MGE – Mobile genetic element

HGT – Horizontal gene transfer

TE – Transposable element

IS – Insertion sequence

MP – Mobile promoter

CDF – Cumulative distribution function

PDF – Probability density function

Chapter 1

Introduction

1.1 Mobile Genetic Elements

Mobile genetic elements (MGEs) are segments of the genome that are involved with the movement of DNA, and are universally present in both eukaryotes and prokaryotes [21]. Movement of DNA can either be intracellular (within a cell), or intercellular (between cells). Prokaryotes have three main mechanisms for obtaining DNA intercellularly: transformation, transduction and conjugation. Collectively, these processes of exchanging genes between organisms are known as horizontal gene transfer (HGT). Transformation is when a cell acquires DNA from its surroundings, and is often a side-effect of nutrient uptake. The other two mechanisms involve bacteriophages and plasmids respectively, and will be discussed in greater detail in later sections.

The dynamics of MGEs have important implications for evolution and also have applications in other areas such as the spread of antibiotic resistance [18]. Additionally, MGEs known as transposable elements (TEs) can be used in a laboratory setting; for instance TEs can be used to help determine the function of DNA sequences [5] or in recombinant DNA techniques [33]. Furthermore, transposable elements may have been crucial in the evolution of specific genes, such as telomerase [39] and RAG genes [1]. Finally, MGEs have been linked to numerous diseases, including: hemophilia [56], colon cancer [42], breast cancer [37], and muscular dystrophy [28].

MGEs can be categorized into four classes: transposons, plasmids, bacteriophages and self-splicing molecular parasites [53].

1.1.1 Transposons

Intracellular movement in DNA is often caused by transposable elements (TEs). Referred to as "jumping genes", TEs can be divided into three classes: insertion sequence elements (ISs), Class 1 transposable elements and Class 2 transposable elements. Class 1 transposable elements are present only in eukaryotes, while Class 2 elements and ISs are present in both prokaryotes and eukaryotes.

ISs are the simplest of these three classes, as they only encode for genes which mediate their own movement. This is in contrast to Class 1 and 2 elements which also carry accessory genes. Generally, ISs consist of a transposase gene that is flanked by inverted repeat sequences. Although ISs do not contain accessory genes, they can cause the movement of other genes through a phenomenon known as composite transposition. In this process, two nearby ISs transpose together, along with the genetic material that lies between them.

As mentioned previously, Class 1 and 2 transposable elements also contain accessory genes. Class 1 transposable elements, retrotransposons, create an RNA copy of themselves, and then use this RNA intermediate to insert a DNA copy somewhere else in the genome. On the other hand, Class 2 transposable elements, DNA transposons, are excised from their current location and then inserted elsewhere using transposase. These two mechanisms are called "copy and paste" and "cut and paste" methods respectively [34]. Insertion sequences are more similar to DNA transposons, and generally use a "cut and paste" method as well.

It should be noted that although some TEs act by a "cut and paste" mechanism, transposition can still cause increased copy numbers of the transposable element. In diploid cells, this is because the excised site can be restored using homologous recombination [17, 20]. Duplication can also happen if transposition occurs during certain phases of the cell cycle; for instance, if a segment of the genome transposes from a replicated portion of the genome to a region that has not yet been duplicated it will result in an extra copy of the transposon [12].

Overall, transposons cause most of the intracellular genetic movement in prokaryotes. Additionally, these genetic elements are able to transfer to plasmids, or become incorporated in bacteriophages. Hence, they are also involved in intercellular movement.

1.1.2 Plasmids

Plasmids are self-replicating pieces of DNA, that are usually circular and double stranded. Plasmids can be transferred intercellularly using a process called conjugation. In this process, the donor cell forms a pillus that connects to the recipient cell, and then plasmids are passed from the donor to the recipient. Bacterial cells often use this method to transfer antibiotic resistance genes to each other [18].

1.1.3 Bacteriophages

Bacteriophages are viruses that infect bacterial cells. They inject their DNA into the host cell, and begin to make copies of themselves. Sometimes, the phage accidentally incorporates host DNA into one of its capsids, and this new phage will consequentially transfer bacterial DNA to a new host. This phenomenon is termed transduction.

1.1.4 Self-splicing Molecular Parasites

Finally, we will briefly mention self-splicing molecular parasites. This category of MGEs describes genetic elements that are able to splice themselves into the genome without the aid of another enzyme (e.g. transposase). This category includes such elements as Group I and II introns and inteins.

1.2 Regulatory Elements as Mobile Genetic Elements

A more recent topic in mobile genetics is the role of regulatory elements as MGEs. Traditionally, HGT is thought of as the movement of genes between organisms. However, there is evidence that non-coding regions of the genome can also be transferred through HGT [35, 49]. It is possible that cells can use this mechanism to transfer regulatory regions of DNA, and thus alter gene expression. This is known as transcriptional rewiring.

1.2.1 Transcriptional Rewiring

Gene expression involves reading DNA and using it to create a gene product, such as a protein. The first step in this process is gene transcription, in which a cell's DNA is converted into RNA. Next, RNA is translated into protein molecules, which are involved in cell functions. A region of DNA called the promoter is the signal for the initiation of transcription.

Promoters often have specific characteristics that allow the initiation of transcription; for instance, in essential "housekeeping" genes in *Escherichia coli* approximately 10 base pairs (bps) before the start of the gene is the sequence "TATAAT", and then around -35 bps upstream is the sequence "TTGACA". On average, these sequences are 17 bps apart. This is shown in Figure 1.1.

Since promoters are needed to initiate transcription, they are crucial in gene regulation. Cells respond to their external environment by changing gene expression, and there are many mechanisms by which cells achieve this. Transcription factors such as repressors or activators can alter the ability of RNA to bind to a promoter;

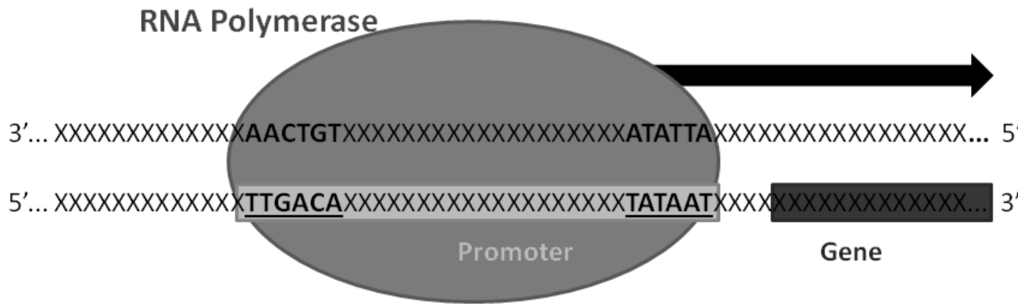

Figure 1.1: Diagram of a promoter. Polymerase binds to the promoter, and transcribes the gene. The top strand is the template strand, and the bottom strand is the coding strand. The arrow indicates the direction of transcription.

repressors bind to a region of DNA termed the operator and prevent transcription and activators bind to a region of DNA termed enhancers and increase transcription [36]. Transcriptional rewiring occurs when gene expression is altered through changes in transcription factors, promoters, operators or enhancers.

There has been plenty of speculation that regulatory regions of DNA are mobile, and this mobility is used to alter gene expression [50]. For instance, there is anecdotal evidence that recruitment of promoters can be used to activate silenced genes [32, 55]. Additionally, Oren et. al [49] recently found that HGT of regulatory regions of DNA is common to prokaryotes, and can indeed cause rapid transcriptional rewiring. Therefore it is reasonable to suspect that regulatory elements can act as a class of MGEs.

1.2.2 Mobile Promoters

One proposed class of mobile regulatory elements are mobile promoters (MPs). In a recent study of 1362 prokaryotic genomes, more than 4000 families of putative MPs were identified [35, 41]. To identify possible MPs, the authors looked for homologous sequences upstream of non-homologous coding regions, in the region where promoters should be located. Presumably, if these sequences are the same, one explanation is that the promoters have moved at some point in evolution. By widening the search criteria, this dataset was increased to 40,000 promoter sequences [57].

1.3 Mathematical Models of Mobile Genetic Elements

Many models have been developed to describe the dynamics of MGEs. Authors commonly compare the equilibrium distribution of their models to data to determine which processes are important in MGE dynamics; the processes considered include duplication, excision, selection, HGT, drift, mutation and recombination. Additionally, models have been developed to address specific questions regarding MGE dynamics, as discussed in the sections to follow.

1.3.1 Eukaryotic Models

The focus of this thesis is the dynamics of MGEs in prokaryotic genomes. Nonetheless, it is valuable to mention models that have been developed to describe eukaryotic populations. Much work has been done modeling MGEs in eukaryotes [6–10, 14, 19, 22, 24, 25, 27, 29, 30, 40, 43–48, 54] (for reviews see [11, 31]).

Most of these models explore how different factors such as mutation, selection, recombination, gene conversion and genetic drift affect the equilibrium states of MGEs [6, 8, 9, 22, 27, 29, 30, 43, 45, 48]. Other studies examine divergence of TE families, to gain insight on why TE families tend to be homogeneous [7, 25, 46, 47, 54]. Additionally, many studies consider more specific evolutionary questions [10, 14, 19, 24]; for instance, it is possible that TEs were involved in sexual differentiation [14, 24]. It is also speculated that TEs may be responsible for some speciation and extinction events [19].

1.3.2 Prokaryotic Models

Prokaryotic MGEs are often simpler to model since prokaryotes reproduce asexually and only possess one chromosone; therefore recombination and gene conversion between homologous chromosomes do not have to be considered. Most authors have taken a population biology approach, and modeled the dynamics of MGEs as branching processes [2–4, 13, 15, 23, 38, 51, 57], though game theory approaches have also been employed [58]. In their simplest form, these models allow the number of MGEs present in a genome to grow through transposition/duplication, while growth is limited due to deletion events and/or through selection. One important factor that is particularly important in prokaryotic models is the consideration of HGT, which has been included in [2–4, 23, 38, 57].

Sawyer and Hartl created multiple models to describe the distribution of TEs in prokaryotes [51]. They assumed that cells without TEs acquire them at a constant rate, and fitness of the host is reduced as a function of TE copy number. Deletion of TEs was ignored, since there is evidence that the rate of deletion is much lower than the rate of transposition [16]. In further work, some of these models were applied to data from 6 strains of ISs in *E. coli* [52]. Selection was incorporated by either assuming a slower growth rate or increased death rate in cells carrying TEs. The authors found that fitness reduces very slightly with increased copy number.

In the above work, authors assumed that MGEs either decreased the fitness of their host, or were selectively neutral in order to determine the factors involved in MGE dynamics. However, there is some debate as to whether MGEs are ubiquitous among species because they somehow increase fitness of their hosts, or if they are simply "selfish" elements, and are able to persist despite their harmful effects. This is addressed directly in [2, 13, 38], where they explore whether it is possible for MGEs to persist even if they are purely "selfish". These papers model the evolution of TEs using a branching process, and find that high *de novo* acquisition of TEs (either through mutation or HGT) can allow TEs to establish, even if they are detrimental to their hosts' fitness. This is also examined in [3, 4] where the authors found similar results when the approach was applied to ISs. Additionally, these authors suggested that ISs do not have a large effect on host fitness, and may in fact be selectively neutral.

Prokaryotic models have also been used to address specific questions regarding MGEs. For instance, Hartl and Sawyer examined why the presence of unrelated ISs is correlated in *E. coli*, and suggested that this is because they are often transferred together on plasmids [23]. Further, Edwards and Brookfield studied the effect of fluctuating environments on MGE populations [15]. Additionally, Wagner used a game theory approach to determine whether composite transposition provides an evolutionary advantage [58]. The author considered "selfish" ISs that only transpose themselves, and "cooperative" ISs that also transpose accessory genes. He found that "cooperation" is not an evolutionarily stable strategy, and thus predicted that composite transposition only exists now due to pressure from antibiotics, but will eventually disappear.

Finally, we will mention the birth-death-diversification model [57], which is the focus of this thesis. This model was originally developed to describe the evolution of families of MPs, however it can be applied to other classes of MGEs, such as TEs or ISs. A detailed description of this model can be found in Chapter 2, but its unique feature is that it considers genetic diversification of families of MGEs. The authors defined families to be sequences in the promoter regions that shared 80% identity over at least 50 nucleotides. Genetic diversification is defined to have occurred when the MP sequence is sufficiently different from the original to be considered a new family. In this thesis we demonstrate that diversification is important to consider, because of its indirect effect on HGT; diversification can influence the equilibrium distribution, extinction probability, and extinction times of MGE populations.

1.4 Markov Processes

A Markov process is a stochastic system that satisfies the Markov property. This property simply means that the system is "memoryless"; any predictions about the future state of the system can be made solely based on the current state. Some of these states can be absorbing states, which means the probability of leaving the state is zero. For biological models, absorption states often correspond to extinction. Some properties of Markov processes that are of interest are equilibrium distributions, absorption probabilities and mean time until absorption.

A Markov chain is a Markov process in which the object can move from one state to another with specific probabilities. Alternatively, in a continuous-time Markov chain, the object moves with a constant probability per unit time, which yields exponentially-distributed waiting times. In the next section, we will illustrate how to derive the equilibrium distribution, extinction probability and mean extinction time for a simple continuous-time model.

1.4.1 An Example: The Classic Birth-Death Model

Consider the classic birth-death model used to model a population of size n, illustrated in Figure 1.2. In this model, there are n members in the population, with $n \geq 0$. The system can move from state n to $n+1$ with rate λ_n, and from n to $n-1$ with rate μ_n. A state n is absorbing if $\lambda_n = \mu_n = 0$.

First, we will derive the equilibrium distribution for this general model, π_n, which is defined for $n \geq 0$. An equilibrium distribution gives a steady-state proportion of population sizes; when a set of populations is in this steady-state, the distribution will not change in time. Since π_n is defined as a proportion, it is required that $\sum\limits_{n=0} \pi_n = 1$.

The equilibrium solution can be calculated by using the fact that at equilibrium the rate at which individuals leave a state is equal to the rate they enter; i.e. $(\lambda_n+\mu_n)\pi_n =$

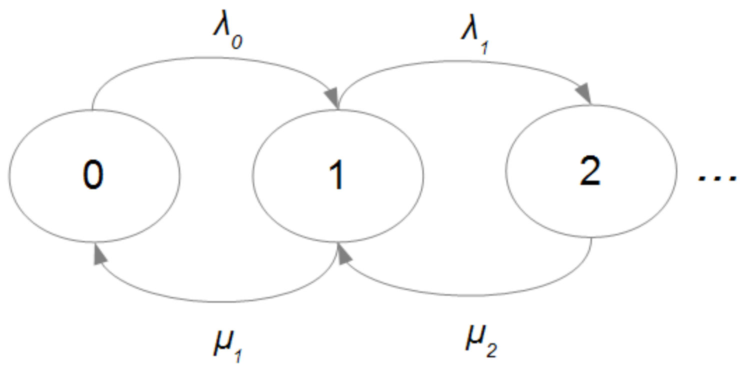

Figure 1.2: The birth-death process. The population size increases from n to $n+1$ members at rate λ_n, and decreases from n to $n-1$ members at rate μ_n.

$\lambda_{n-1}\pi_{n-1} + \mu_{n+1}\pi_{n+1}$. Using this, it can be shown (see [26]) that the equilibrium distribution of the birth-death process is given by:

$$\begin{aligned} \pi_0 &= \left(1 + \sum_{k=1}^{\infty} \frac{\lambda_{k-1}\lambda_{k-2}\cdots\lambda_0}{\mu_k\mu_{k-1}\cdots\mu_1}\right)^{-1} \\ \pi_n &= \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}\pi_0 \ . \end{aligned} \tag{1.1}$$

Next, consider the extinction probability of this system. In order for extinction to be possible, $n = 0$ needs to be an absorbing state. Therefore, for this example we will consider the linear birth-death process for which birth and death rates are given by $n\lambda$ and $n\mu$ respectively. Note that there is no non-zero equilibrium distribution in this case—the population will either become extinct or grow to an infinite size.

It is straightforward to calculate the extinction probability for the linear birth-death process. If there are n members in the population, a birth event will happen first with probability $\dfrac{\lambda}{\mu+\lambda}$, and similarly a death event will happen first with probability $\dfrac{\mu}{\mu+\lambda}$. This leads to a recurrence relation for the probability a population with n members will go extinct, X_n:

$$X_n = \frac{\lambda}{\mu+\lambda} X_{n+1} + \frac{\mu}{\mu+\lambda} X_{n-1} \ . \tag{1.2}$$

For extinction probabilities, $X_0 = 1$. In cases where there is a non-zero survival probability, we also have the constraint that $\lim_{n\to\infty} X_n = 0^+$. If no solutions exist that satisfy these boundary conditions, the solution is simply $X_n = 1$ for all n, which is clearly a solution to Equation (1.2). Therefore, the extinction probabilities are given by:

$$X_n = \begin{cases} \left(\frac{\mu}{\lambda}\right)^n, & \text{if } \mu < \lambda \\ 1 & \text{otherwise.} \end{cases} \tag{1.3}$$

Finally, we will calculate the expected extinction time for the linear birth-death process. For this, we will define $P_n(t)$ as being the probability that a family with n members at time 0 will be extinct before time t. In the first infinitesimal amount of time, Δt, there are three possibilities: there is a birth event with probability $n\lambda\Delta t$, a death event with probability $n\mu\Delta t$, or the population does not change size with probability $1 - n(\lambda+\mu)\Delta t$. Therefore, we can express the following:

$$P_n(t+\Delta t) = n\lambda\Delta t P_{n+1}(t) + n\mu\Delta t P_{n-1}(t) + [1 - n(\lambda+\mu)\Delta t] P_n(t) \ . \tag{1.4}$$

If we take the limit as $t \to 0$ this equation becomes an ordinary differential equation (ODE). Further, it is clear that $P_0(t) = 1$ and $P_n(0) = 0$ for $n \neq 0$. Therefore, the probability the population will be extinct before time t is summarized in the following equations:

$$\frac{dP_n(t)}{dt} = n\lambda P_{n+1}(t) + n\mu P_{n-1}(t) - n(\lambda+\mu)P_n(t) \quad \text{for } n > 0$$
$$P_0(t) = 1 \tag{1.5}$$
$$P_n(0) = 0 \quad \text{for } n > 0.$$

From here, the cumulative distribution function (CDF) for extinction time can be found by dividing the probability of going extinct before time t, $P(t)$, by the overall probability of going extinct, X_n. The probability distribution function (PDF) is the derivative of the CDF. Then, the mean extinction time for a population of size n, T_n, is simply the expectation value found by integrating over the product of the PDF and time, t. This is summarized in the following equation:

$$T_n = \frac{1}{X_n}\int_0^\infty t\frac{dP_n(t)}{dt}dt \ , \tag{1.6}$$

where X_n and $P_n(t)$ are given in Equations 1.3 and 1.5 respectively.

In conclusion, for this simple example we have illustrated techniques to solve for equilibrium distributions, extinction probabilities and extinction times. In this thesis we will apply the same techniques to a more complicated model that describes the dynamics of MGEs in prokaryotes.

Bibliography

[1] A. Agrawal, Q.M. Eastman, and D.G. Schatz. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, 394(6695):744–751, 1998.

[2] C.J. Basten and M.E. Moody. A branching-process model for the evolution of transposable elements incorporating selection. *J Math Biol*, 29(8):743–61, 1991.

[3] M. Bichsel, A.D Barbour, and A. Wagner. The early phase of a bacterial insertion sequence infection. *Theor Popul Bio*, 78(4):278–288, 2010.

[4] M. Bichsel, A.D. Barbour, and A. Wagner. Estimating the fitness effect of an insertion sequence. *J Math Biol*, 66(1):95–114, 2013.

[5] D. Botstein and D. Shortle. Strategies and applications of *in vitro* mutagenesis. *Science*, 229(4719):1193–1201, 1985.

[6] J.F. Brookfield. Interspersed repetitive DNA sequences are unlikely to be parasitic. *J Theor Biol*, 94(2):281–299, 1982.

[7] J.F. Brookfield. A model for DNA sequence evolution within transposable element families. *Genetics*, 112(2):393–407, 1986.

[8] J.F. Brookfield. The population biology of transposable elements. *Philos Trans R Soc Lond B Biol Sci*, 312(1154):217–26, 1986.

[9] B. Charlesworth and D. Charlesworth. The population dynamics of transposable elements. *Genet Res*, 42:1–27, 1983.

[10] B. Charlesworth and C.H. Langley. The evolution of self-regulated transposition of transposable elements. *Genetics*, 112(2):359–383, 1986.

[11] B. Charlesworth and C.H. Langley. The population genetics of *Drosophila* transposable elements. *Annu Rev Genet*, 23(1):251–287, 1989.

[12] J. Chen, I.M. Greenblatt, and S.L. Dellaporta. Molecular analysis of Ac transposition and DNA replication. *Genetics*, 130(3):665, 1992.

[13] R. Condit, F.M. Stewart, and B.R. Levin. The population biology of bacterial transposons: a priori conditions for maintenance as parasitic DNA. *Am Nat*, pages 129–147, 1988.

[14] E.S. Dolgin and B. Charlesworth. The fate of transposable elements in asexual populations. *Genetics*, 174(2):817–27, 2006.

[15] R.J. Edwards and J.F. Brookfield. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Mol Biol Evol*, 20(1):30–7, 2003.

[16] C. Egner and D.E. Berg. Excision of transposon Tn5 is dependent on the inverted repeats but not on the transposase function of Tn5. *Proc Natl Acad Sci U S A*, 78(1):459–463, 1981.

[17] W.R. Engels, D.M. Johnson-Schlitz, W.B. Eggleston, and J. Sved. High-frequency P element loss in *Drosophila* is homolog dependent. *Cell*, 62(3):515–525, 1990.

[18] T.J. Foster. Plasmid-determined resistance to antimicrobial drugs and toxic metal ions in bacteria. *Microbiol Rev*, 47(3):361, 1983.

[19] L.R. Ginzburg, P. M. Bingham, and S. Yoo. On the theory of speciation induced by transposable elements. *Genetics*, 107(2):331–341, 1984.

[20] G.B. Gloor, N.A. Nassif, D.M Johnson-Schlitz, C.R. Preston, and W.R Engels. Targeted gene replacement in *Drosophila* via P element-induced gap repair. *Science*, 253(5024):1110–1117, 1991.

[21] J.P. Gogarten and J.P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, 3(9):679–687, 2005.

[22] G.B. Golding, C.F. Aquadro, and C.H. Langley. Sequence evolution within populations under multiple types of mutation. *Proc Natl Acad Sci U S A*, 83(2):427–31, 1986.

[23] D.L. Hartl and S.A. Sawyer. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? *Genetics*, 118(3):537–41, 1988.

[24] D.A. Hickey. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101:519–31, 1982.

[25] R.R. Hudson and N.L. Kaplan. On the divergence of members of a transposable element family. *J Math Biol*, 24(2):207–15, 1986.

[26] D.L. Isaacson and R.W. Madsen. *Markov Chains: Theory and Applications*. John Wiley and Sons, 1976.

[27] N. Kaplan, T. Darden, and C.H. Langley. Evolution and extinction of transposable elements in Mendelian populations. *Genetics*, 109(2):459–480, 1985.

[28] K. Kobayashi, Y. Nakahori, M. Miyake, K. Matsumura, E. Kondo-Iida, Y. Nomura, M. Segawa, M. Yoshioka, K. Saito, M. Osawa, et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature*, 394(6691):388–392, 1998.

[29] C.H. Langley, J.F. Brookfield, and N. Kaplan. Transposable elements in mendelian populations. I. A theory. *Genetics*, 104(3):457–71, 1983.

[30] A. Le Rouzic, T.S. Boutin, and P. Capy. Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A*, 104(49):19375–80, 2007.

[31] A. Le Rouzic and G. Deceliere. Models of the population genetics of transposable elements. *Genet Res*, 85(3):171–81, 2005.

[32] D. Lee and O. Bernard. Adaptive evolution of *Escherichia coli* K-12 mg1655 during growth on a nonnative carbon source, l-1,2 propanediol. *J Appl Environ Microbiol*, 76(13):4158–4168, 2010.

[33] V.A. Luckow, S.C. Lee, G.F. Barry, and P.O. Olins. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. *J Virol*, 67(8):4566–4579, 1993.

[34] M. Lynch. *The Origins of Genome Architecture*. Sinauer Associates, Inc., 2007.

[35] M. Matus-Garcia, H. Nijveen, and M.W. van Passel. Promoter propagation in prokaryotes. *Nucleic Acids Res*, 40(20):10032–40, 2012.

[36] W.R. McClure. Mechanism and control of transcription initiation in prokaryotes. *Annu Rev Biochem*, 54(1):171–204, 1985.

[37] Y. Miki, T. Katagiri, F. Kasumi, T. Yoshimoto, and Y. Nakamura. Mutation analysis in the BRCA2 gene in primary breast cancers. *Nature Gen*, 13(2):245–247, 1996.

[38] M.E. Moody. A branching process model for the evolution of transposable elements. *J Math Biol*, 26(3):347–57, 1988.

[39] T.M. Nakamura, G.B. Morin, K.B. Chapman, S.L. Weinrich, W.H. Andrews, J. Lingner, C.B. Harley, and T.R. Cech. Telomerase catalytic subunit homologs from fission yeast and human. *Science*, 277(5328):955–959, 1997.

[40] V. Nanjundiah. Transposable element copy number and stable polymorphisms. *J Genet*, 64(2-3):127–134, 1985.

[41] H. Nijveen, M. Matus-Garcia, and M.W. van Passel. Promoter reuse in prokaryotes. *Mob Genet Elements*, 2(6):279–281, 2012.

[42] M. Nyström-Lahti, P. Kristo, N.C. Nicolaides, S. Chang, L.A. Aaltonen, A. Moisio, H.J. Järvinen, J. Mecklin, K.W. Kinzler, B. Vogelstein, et al. Founding mutations

and Alu-mediated recombination in hereditary colon cancer. *Nat Med*, 1(11):1203–1206, 1995.

[43] T. Ohta. Population genetics of selfish DNA. *Nature*, 292, 1981.

[44] T. Ohta. Theoretical study on the accumulation of selfish DNA. *Genet Res*, 41(01):1–15, 1983.

[45] T. Ohta. Population genetics of transposable elements. *Math Med Biol*, 1(1):17–29, 1984.

[46] T. Ohta. A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics*, 110(3):513–24, 1985.

[47] T. Ohta. Population genetics of an expanding family of mobile genetic elements. *Genetics*, 113(1):145–159, 1986.

[48] T. Ohta and M. Kimura. Some calculations on the amount of selfish DNA. *Proc Natl Acad Sci U S A*, 78(2):1129–1132, 1981.

[49] Y. Oren, M.B. Smith, N.I Johns, M.K. Zeevi, D. Biran, E.Z. Ron, J. Corander, H.H. Wang, E.J. Alm, and T. Pupkol. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc Natl Acad Sci USA*, 111(45):16112–16117, 2014.

[50] J.C. Perez and E.A. Groisman. Evolution of transcriptional regulatory circuits in bacteria. *Cell*, 138(2):233–44, 2009.

[51] S. Sawyer and D. Hartl. Distribution of transposable elements in prokaryotes. *Theor Popul Biol*, 30(1):1–16, 1986.

[52] S.A. Sawyer, D. E. Dykhuizen, R.F. DuBose, L. Green, T. Mutangadura-Mhlanga, D.F. Wolczyk, and D.L. Hartl. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics*, 115(1):51–63, 1987.

[53] J.L. Siefert. Defining the mobilome. In *Horizontal Gene Transfer*, pages 13–27. Springer, 2009.

[54] M. Slatkin. Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics*, 110(1):145–58, 1985.

[55] D. Stoebel and C. Dorman. The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. *Mol Biol and Evol*, 27(9):2105–2112, 2010.

[56] E. Sukarova, A.J. Dimovski, P. Tchacarova, G.H. Petkov, and G.D. Efremov. An Alu insert as the cause of a severe form of hemophilia A. *Acta Haematologica*, 106(3):126–129, 2000.

[57] M.W. van Passel, H. Nijveen, and L.M. Wahl. Birth, death, and diversification of mobile promoters in prokaryotes. *Genetics*, 197(1):291–299, 2014.

[58] A. Wagner. Cooperation is fleeting in the world of transposable elements. *PLoS Comput Biol*, 2(12):e162, 2006.

Chapter 2

Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: the birth-death-diversification model

2.1 Introduction

Mobile genetic elements (MGEs) are regions of DNA that are involved with the movement of genetic material within and between genomes, typically containing the genetic code for a protein that mediates their own movement. These elements are nearly universally present throughout the domains of life, but are particularly active in prokaryotes. Consistent with the "selfish DNA" hypothesis, MGEs often reduce the fitness of their hosts [81]. For instance, transposable elements have been linked to hybrid dysgenesis in *Drosophila* [82] and to deleterious mutations in bacteria and yeast [71]. However, they can also be beneficial to an organism, as in the case of plasmids conferring antibiotic resistance [66]. Due to their ubiquitousness and impact on cellular function, MGEs are of immense importance in genetics.

The dynamics of MGEs within genomes have been previously studied using a range of theoretical approaches. For transposable elements in eukaryotes, models that consider factors such as mutation, recombination and drift have successfully predicted the number of transposable element copies within a genome [63, 72], and the relatedness between copies in a family [62, 70, 79, 84]. The effects of selective pressures in limiting copy number have also been studied in some detail [63, 65, 67, 69, 73], as have the complex histories of transposable element lineages within genomes [73, 74].

For MGEs in prokaryotes, both branching process and Markov chain approaches have been used to predict the distribution of copy number within genomes ([60, 61, 68, 77, 83, 88], but also see [64, 89]). These models explicitly include a "birth" process, duplication or transposition, which increases the number of MGE copies, as well as a "death" term, excision or deletion, which reduces copy number. For prokaryote lineages, horizontal gene transfer (HGT) is clearly an important process and this is reflected in several approaches [61, 68, 88]. These techniques have allowed us to infer, for example, the relative importance of HGT and selection in maintaining and limiting insertion sequences in bacterial genomes [61, 68].

Mobile promoters (MPs) are a newly proposed class of MGEs [76]. The extreme plasticity of prokaryotic genomes implies that transcriptional rewiring is of major importance in prokaryotic evolution. New genes acquired through horizontal gene transfer (HGT) can be silenced by the recipient cell [59], and there is anecdotal evidence of rewiring of silent genes through the recruitment of promoters [75, 86]. Additionally, promoter sequences are highly conserved even between distantly related species [76, 78] indicating they may have the same origin. Furthermore, a recent study has shown that regulatory switching can occur through HGT of regulatory regions

[80]. This lends credence to the theory that transcriptional rewiring may be achieved through the recruitment of MPs [78].

While there is no direct evidence that a class of promoters act as MGEs, nearly 40,000 potential MPs have been identified by sequence analysis [76, 78, 88]. To describe the distribution of these MPs both within and among genomes, a mathematical model of the dynamics of MGEs was developed. A dataset collected from all available prokaryotic genomes [88] and statistical model selection were used to reduce the model and determine which terms and processes were necessary to describe the distribution of MPs in prokaryotes.

The resulting birth-death-diversification model is similar to a classical birth-death Markov chain, but has two key differences. First, it was necessary to include the process of genetic diversification of MGEs in order to obtain a satisfactory description of the MP data. Diversification occurs when the sequence of an element changes so that it is substantially different from the original sequence; if we consider an evolutionary lineage of MGEs, with diversification a new lineage of related MGEs branches from the original family. Since genome sequencing is continually improving our ability to identify multiple related families of MGEs within genomes, accounting for diversification may become increasingly important in describing MGE dynamics.

Second, model selection concluded that all rate terms were best described by linear processes, except HGT, which was best fit at a constant rate. In other words, the probability that a MGE is transfered to a new genome by HGT does not increase linearly with the number of copies of the MGE in the donor genome. A constant HGT term was likewise suggested in a rigorous model selection exercise describing the dynamics of the insertion sequence IS5 [61], and is reasonable considering the large number of external factors influencing HGT. For example, a phenomenon termed surface exclusion prevents the transfer of genes to recipient cells that already carry similar genes [87].

We thus expect that both diversification, and HGT at a constant rate, may be critical to modeling the distributions of MPs, insertion sequences, and other MGEs in prokaryotic genomes. However, the influence of these processes on the longterm fate of MGE lineages has not yet been elucidated. In particular, it is unknown how diversification and HGT affect the extinction probability of an extant lineage, nor how they affect the expected distribution of copy number within MGE families, or the distribution of MGE families among genomes.

In the sections to follow, we derive extinction probabilities, extinction times and stationary distributions for the birth-death-diversification model, and illustrate how these measures of the longterm fate of MGEs depend on both diversification and HGT. We find that the interplay of these two processes is subtle; while diversification does not increase the number of MGEs in the lineage, it can nonetheless increase both survival probability and longterm growth rates, but only in the presence of HGT. We also derive similar results for an extension of the model which allows MGEs in different regions in the genome, for example coding and non-coding regions, to be described by different rates.

2.2 Methods

2.2.1 The Birth-Death-Diversification Model

In [76] the promoter regions of all available prokaryote genomes were compared, and sequence similarities were used to identity families of closely related promoters within each genome. A model to describe these data was developed in [88]. Statistical model selection techniques were used to determine which processes should be included in the model, and whether the rates for these processes were constant or varied linearly with the number of MGEs in the genome. The resulting model and rates are described below.

We model a collection of prokaryote genomes, each of which may contain a number of MGE families. A family is defined as elements with very similar sequences; for example, 80% sequence identity over 50 nucleotides was used as a threshold in [88], where MP families were found to have on average over 95% sequence identity. These families may be of different sizes, that is, each family contains some integer number of (nearly identical) copies of the MGE.

The copy-level model describes the number of MGE families, C_n, out of all the MGE families within this collection of genomes, that have n copies. In the copy-level model, an MGE family can gain a copy by a duplication (birth) event, which occurs at rate nu for a family with n copies. Similarly, a copy may be lost lost due to a deletion (death) event, which occurs at rate nw. Additionally, new families are created if a copy within the family diversifies. Diversification includes mutational processes that would make this copy sufficiently different from the other copies in the family, for example if one copy of the MGE sequence obtains an insertion. In this case the original family loses a copy and a new family of one copy, a singleton family, is created. Thus, we make the reasonable assumption that the newly diversified sequence is not similar to any pre-existing MGE family.

The final process included in the model is HGT. We assume that HGT occurs through replicative transfer, that is, the donor cell does not lose a copy through this event. We further assume that the probability that the recipient genome already contains a copy of the transfered MGE is negligible. This assumption is justified for MPs, since each genome in this dataset contains on average three out of over 4000 distinct families. Since the model describes the overall number of MGE families with n copies, the net effect of HGT is thus to add singleton families. Each family, irrespective of the number of copies in the family, contributes a HGT event at rate η.

The genome-level model describes the number of genomes, F_m, that carry m MGE families. Recall that diversification, as described above, changes one family of n copies to a family of $n-1$ copies and a new singleton family. Thus each diversification event adds a family to the genome. At the genome level, diversification is therefore analogous to a birth event. For a genome with m families, the rate of diversification is mp. Similarly, deletion as described above could result in the loss of a family; this occurs at rate mq. Finally, a new family is created if a new element enters the genome through HGT, which occurs at rate μ, independent of the number of families already carried by the genome.

The resulting model is shown in Figure 2.1. Both the copy-level and the genome-level models can be expressed as an infinite system of ordinary differential equations (ODEs) as shown in Equations (2.13) and (2.19).

We should also note that the two levels are linked. As derived in detail in [88], we can calculate the genome-level rates given the copy-level rates and distribution. New families are created when any families with more than one copy have a diversification event. Families are lost when a one copy family has a deletion event. Finally, new families enter a genome through HGT whenever any families have an HGT event. Since μ is the HGT rate per genome, this is dependent on the average number of families per genome, $\bar{f}$. Therefore, the genome-level rate parameters can be expressed as follows, where c_n is the proportion of families with n copies:

$$\begin{aligned} p &= \sum_{n=2}^{\infty} nvc_n \\ q &= c_1 w \\ \mu &= \bar{f} \sum_{n=1}^{\infty} \eta c_n = \bar{f}\eta \end{aligned} \tag{2.1}$$

.

On the copy-level this process is similar to a classic birth-death model, but has the important difference that the process is non-conservative. A family can start in one state, and then diversify into two different states; formally, we have a multitype

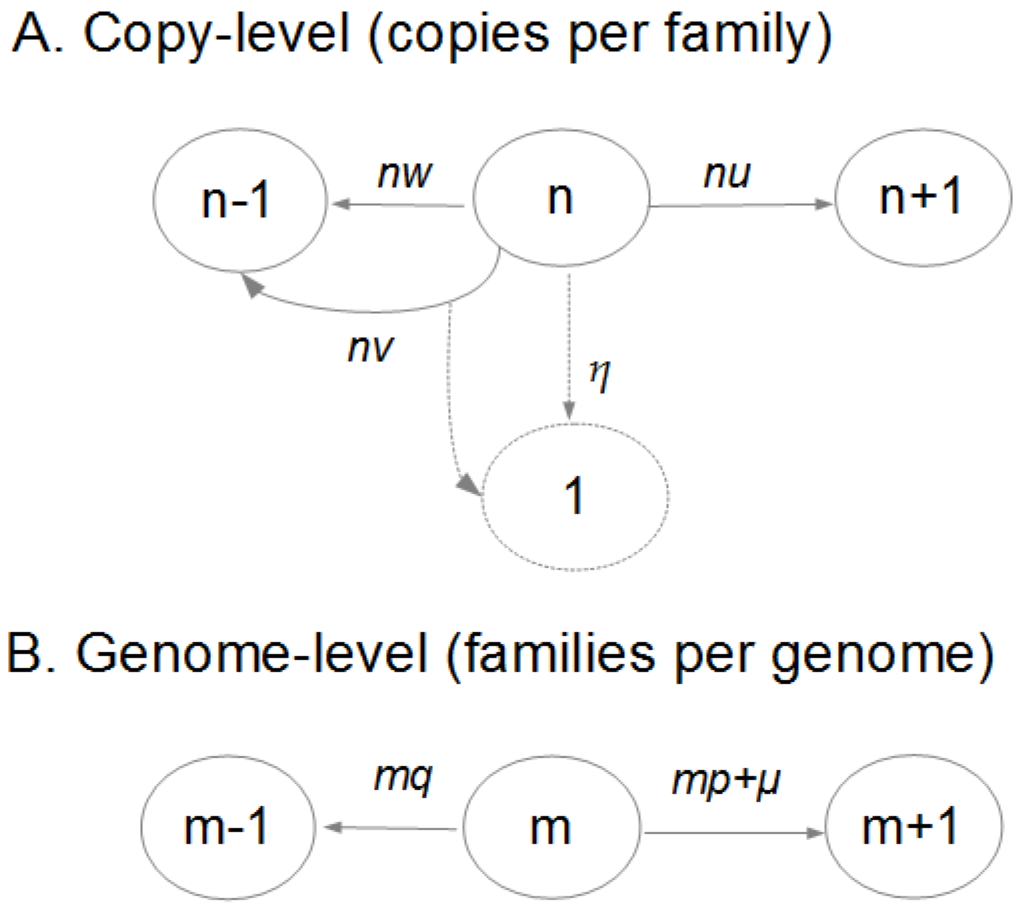

Figure 2.1: Birth-death-diversification model. **A** The copy-level model describes the number of copies of MGEs per family. Parameters are duplication (u), deletion (w), diversification (v) and horizontal gene transfer (η). Note that irrespective of copy number, each family creates new families through horizontal gene transfer at constant rate η. **B** The genome-level model describes the number of MGE families per genome. Parameters are diversification (p), loss (q) and horizontal gene transfer (μ).

branching process rather than a birth-death process. We also note that $n = 0$ is an absorbing state, and that η is nonlinear. In contrast, while the genome-level model also has a nonlinear term, μ, the process is conservative, has no absorbing state, and can be described by a nonlinear birth-death model.

We note that these models do not include selection. Although it seems unreasonable to expect that MGEs have no effect on fitness, analyses of the mobile promoter data [88] and of the insertion sequence IS5 [61] both considered selection, and found that selection parameters were not statistically justified in describing the available data. This issue is discussed in greater detail in [88].

Despite these findings, we still expect that the location of MGEs should affect their dynamics. Therefore, we have extended the copy-level model into a 2-D model, that describes the number of copies in two different classes, such as regions of the genome. In this paper we will illustrate the use of the 2-D model by describing MGEs in coding and non-coding areas, but the model can be used for any classes of interest. For instance, MP sequences have been identified in both promoter and non-promoter regions of prokaryote genomes [76, 78]. The 2-D model is shown in Figure 2.2.

In the 2-D model, n_1 is the number of copies per family in the coding region, and n_2 is the number of copies in the non-coding region. The rates of duplication, deletion and diversification can differ between regions: in the coding regions these rates are u_1, w_1 and v_1, while in the non-coding region they are u_2, w_2 and v_2. As before, η is the HGT rate. Finally, ρ is the probability that a new MGE that arrives either through a duplication or HGT event will be inserted in the coding region. This model allows us to explore, for example, whether the dynamics of MGEs are more rapid in the coding or non-coding regions of the genome. We can also test for selective effects that were not justified in the 1-D model. For example, in the coding region of the genome, the deletion rate may be higher or the insertion rate may be reduced, possibly due to

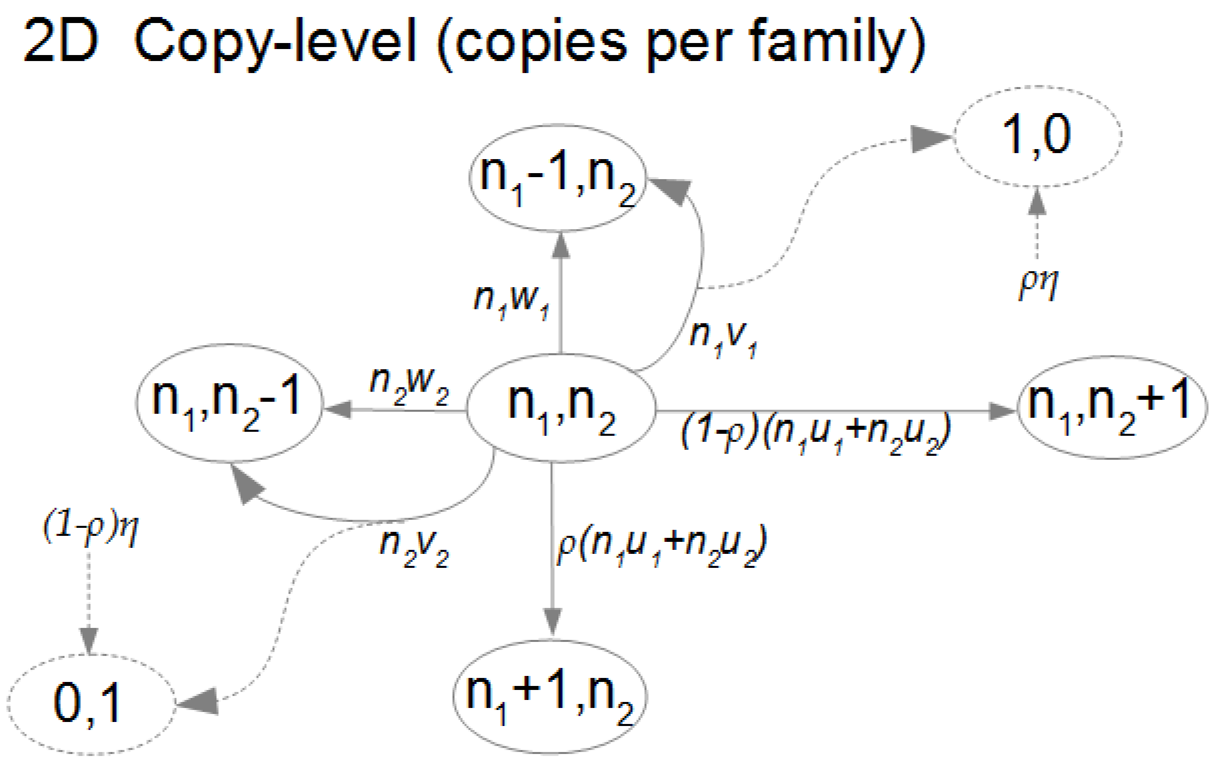

Figure 2.2: 2-D model for families with n_1 copies in the coding region, and n_2 copies in the non-coding region. The rates for duplication, deletion, and diversification are u_1, w_1 and v_1 in the coding region, and u_2, w_2 and v_2 in the non-coding region. Each family creates new families through horizontal gene transfer at constant rate η. New copies are inserted in the coding region with probability ρ.

lethal insertions. System (2.18) gives the formal expression of the 2-D model as an infinite system of ODEs.

We wish to examine extinction probabilities, extinction times and equilibrium states for MGEs. For the copy-level model we are interested in the probability that a new family will go extinct, and how this probability changes with diversification and HGT. For both levels, we wish to determine an equilibrium solution, and then make predictions regarding the longterm behavior of the distributions currently found in prokaryotic genomes. Finally, we will specifically consider the implications of these model predictions for mobile promoters.

2.2.2 Extinction Probability

We would like to compute the probability that a novel MGE lineage is ultimately lost. When a novel MGE first appears, by some *de novo* mutation process, it appears as a single copy in a single genome in the population, forming a new singleton family. In this section, we compute the probability that this family, including all branches of the family that arise through diversification and HGT, ultimately goes extinct. Since extinction involves losing every copy of the novel MGE, we address this question at the copy level.

1-D Model

Let X_n denote the extinction probability of a lineage of MGEs that starts with n copies. The probabilities that a duplication, deletion, diversification or HGT event is the first to occur are given by $\frac{nu}{\delta}$, $\frac{nw}{\delta}$, $\frac{nv}{\delta}$ and $\frac{\eta}{\delta}$ respectively, where $\delta = n(u+w+v)+\eta$. Therefore, an expression for X_n is given by:

$$X_n = \frac{nu}{\delta}X_{n+1} + \frac{nw}{\delta}X_{n-1} + \frac{nv}{\delta}X_{n-1}X_1 + \frac{\eta}{\delta}X_nX_1, \quad n \geq 1 \ . \tag{2.2}$$

Boundary conditions and (2.2) lead to the the following recurrence relation:

$$
\begin{gathered}
nuX_{n+1} - (n(u+w+v) + \eta(1-X_1))X_n + n(w+vX_1)X_{n-1} = 0 \quad n \geq 1 \\
X_0 = 1 \\
\lim_{n\to\infty} X_n = 0^+ \ .
\end{gathered} \tag{2.3}
$$

It should be noted that solutions satisfying the boundary conditions of Equation (2.3) do not always exist; then, the extinction probabilities are simply $X_n = 1$ for all n. Closed form solutions of (2.3) can be found for two cases: (i) $\eta = 0$ and (ii) a linear η term (i.e. replace η with $n\eta$). Interestingly, the second case is also asymptotic in the sense that as v approaches infinity, any family with more than 1 copy will instantaneously split into different families. Therefore all families will be linear with respect to η. These asymptotic solutions were found by using an ansatz of the form $X_n = a^n$.

Case i: In the $\eta = 0$ case, the solution is the same as the linear birth-death model:

$$
X_n = \begin{cases} \left(\dfrac{w}{u}\right)^n, & \text{if } w < u \\ 1 & \text{otherwise.} \end{cases} \tag{2.4}
$$

This case is an upper limit, since η acts as a birth term and hence reduces extinction probability. Additionally, as n approaches infinity, terms including η in equation (2.3) become negligible, so this is also the approximate solution for large n, even when η is non-zero.

Case ii: In the second case, η acts as a linear birth term, and the solution becomes:

$$
X_n = \begin{cases} \left(\dfrac{w}{u+\eta}\right)^n, & \text{if } w < u+\eta \\ 1 & \text{otherwise.} \end{cases} \tag{2.5}
$$

Since η reduces extinction probability through the creation of new copies, this case gives a lower bound on solutions to equation (2.3).

The extinction probability for non-asymptotic cases was calculated numerically from equation (2.3) and verified using Monte Carlo simulation. Note that the recurrence relation (2.3) has a single degree of freedom, such that a value of X_1 determines all the X_i. To obtain numerical solutions to this infinite system of equations, a straightforward approach is to impose a maximum family size, N, and determine X_1 such that $X_N = 0$; N is then increased until the value of X_1 converges. In the Monte Carlo simulation, a MGE family is initiated with n copies of MGEs, and duplication, deletion, diversification or HGT events are simulated with probabilities nu/δ, nw/δ, nv/δ and η/δ, until the family becomes extinct, or reaches a threshold size at which the probability of further extinction is negligible. In the figures to follow, we use an initial system size of $N = 100$, a convergence tolerance of 10^{-4}, and a conservative threshold size of 500 copies. Simulation data are shown for 10^5 replicates for each data point.

2-D Model

A similar analysis was performed for the 2-D model. We let X_{n_1,n_2} be the probability that a family will go extinct if it has n_1 copies in the coding region, and n_2 copies in the non-coding region. Then:

$$\begin{gathered}
(n_1u_1 + n_2u_2)\rho X_{n_1+1,n_2} + (n_1u_1 + n_2u_2)(1-\rho)X_{n_1,n_2+1} \\
+(\eta\rho X_{1,0} + \eta(1-\rho)X_{0,1} - (n_1(u_1+w_1+v_1) + n_2(u_2+w_2+v_2) + \eta))X_{n_1,n_2} \\
+n_1(w_1 + v_1X_{1,0})X_{n_1-1,n_2} + n_2(w_2 + v_2X_{0,1})X_{n_1,n_2-1} = 0 \qquad n_1, n_2 \geq 0,\ n_1+n_2 \neq 0 \\
X_{0,0} = 1 \\
\lim_{n_1\to\infty} X_{n_1,n_2} = \lim_{n_2\to\infty} X_{n_1,n_2} = 0^+ \ .
\end{gathered} \tag{2.6}$$

An important insight is that if the rates in the two regions are the same ($u_1 = u_2$,

$w_1 = w_2$ and $v_1 = v_2$), these equations are equivalent to the system derived for the 1-D model, (2.3), with $X_{n_1,n_2} = X_{n_1+n_2}$.

If the rates in the two regions differ, the solution to (2.6) can be found numerically and verified using Monte Carlo simulation. In this case, the numerical solution was found iteratively. First, we created a matrix of size $N = 50$ where each entry represented an extinction probability. Using equation (2.6), we expressed X_{n_1,n_2} as a function of the other extinction probabilities, and recalculated each X_{n_1,n_2} until the solution converged with tolerance 10^{-4}. The matrix size N was then increased, until the final set of solutions converged within the same tolerance. The Monte Carlo simulation was performed as described for the 1-D model.

2.2.3 Extinction Times

1-D Model

To determine the extinction times, we will first define $P_i(t)$ as the probability a family that has i copies at time 0 will be extinct before time t. Consider the probability that a family is extinct by time $t + \Delta t$. If Δt is small, the probabilities that a birth, death, diversification or HGT event happened within time Δt are $iu\Delta t$, $iw\Delta t$, $iv\Delta t$ and $\eta\Delta t$ respectively. The probability that no event happened in this time is $1 - (i(u + w + v) + \eta)\Delta t$. Therefore, we obtain the following expression:

$$\begin{aligned} P_i(t + \Delta t) = {} & iu\Delta t P_{i+1}(t) + iw\Delta t P_{i-1}(t) + iv\Delta t P_{i-1}(t)P_1(t) + \eta\Delta t P_i(t)P_1(t) \\ & + [1 - (i(u + w + v) + \eta)\Delta t]P_i(t) \qquad i \geq 1 \\ & P_0(t) = 1 \ . \end{aligned} \tag{2.7}$$

If we take the limit as Δt goes to 0, this is simply a system of differential equations:

$$\frac{dP_i}{dt} = i(w + vP_1)P_{i-1} - \left[i(u + w + v) + \eta(1 - P_1)\right] P_i + iuP_{i+1} \qquad i \geq 1$$
$$P_0(t) = 1 \ . \tag{2.8}$$

Additionally, we can use the initial condition that $P_i(0) = 0$ for $i \geq 1$.

Note that if we set $\frac{dP_i}{dt} = 0$ we recover the expression for the extinction probability. That is, as time increases, the probability that the family will go extinct by time t approaches the total extinction probability.

First, we can obtain the cumulative distribution function (CDF), $G_i(t)$:

$$G_i(t) = \frac{P_i(t)}{X_i} \ , \tag{2.9}$$

where X_i is the extinction probability, described in (2.3).

By definition, the probability density function (PDF) , $g_i(t)$, can be found from the derivative of the CDF.

$$g_i = \frac{dG_i}{dt} = \frac{1}{X_i}\frac{dP_i}{dt} \tag{2.10}$$

Finally, we can calculate the expected extinction time of a family that starts with i copies, T_i as follows:

$$T_i = \int_0^\infty t g_i(t) dt \tag{2.11}$$

Extinction times were solved both numerically and through Monte Carlo simulation. Equation (2.8) was integrated with a Runge Kutta routine (ODE45 in Matlab), and the integral from (2.11) was approximated with a trapezoidal rule. The Monte Carlo simulation was performed as described for extinction probabilities; the time step, Δt, was chosen such that the probability of an event happening at each step was approximately 0.01.

2-D Model

The extinction times for the 2-D copy level can be derived in the same way as for the 1-D level. The probability that a family that starts with i copies in the coding region, and j copies in the non-coding region will be extinct by time t is represented by $P_{i,j}$, which satisfies the following equation:

$$\begin{gathered}\frac{dP_{i,j}}{dt} = (iu_1 + ju_2)\rho P_{i+1,j} + (iu_1 + ju_2)(1-\rho)P_{i,j+1} \\ +[\eta\rho P_{1,0} + \eta(1-\rho)P_{0,1} - i(u_1 + w_1 + v_1) - j(u_2 + w_2 + v_2) - \eta]P_{i,j} \\ +i(w_1 + v_1 P_{1,0})P_{i-1,j} + j(w_2 + v_2 P_{0,1})P_{i,j-1} \quad i,j \geq 0,\ i+j \neq 0 \\ P_{0,0}(t) = 1 \\ P_{i,j}(0) = 0 \quad i,j \geq 0,\ i+j \neq 0 \\ . \end{gathered} \tag{2.12}$$

This system can be solved numerically, as described for the 1-D model.

2.2.4 Equilibrium Solution

1-D Copy-Level Model

The birth-death-diversification model can be expressed as an infinite system of ordinary differential equations (ODEs), where C_i is the expected number of families with i copies at time t:

$$\begin{gathered}\frac{dC_1}{dt} = -(u+w+v)C_1 + 2(v+w)C_2 + \sum_{i=1}^{\infty}(iv+\eta)C_i \\ \frac{dC_i}{dt} = (i-1)uC_{i-1} - i(u+w+v)C_i + (i+1)(v+w)C_{i+1}, \quad i \geq 2\ . \end{gathered} \tag{2.13}$$

We seek an equilibrium solution, $\bar{\mathbf{C}}$, in which the number of families with i elements does not change in time, $\dfrac{d\bar{\mathbf{C}}}{dt} = 0$. Using an ansatz of the form $\bar{C}_i \propto \frac{1}{i}a^i$ we recover solutions given by:

$$\bar{C}_i = A\frac{1}{i}\left(\frac{u}{w+v}\right)^i \quad , \tag{2.14}$$

which exists under the conditions:

$$\begin{gathered} \eta = \eta_c = \left[\ln\left(\frac{w+v}{w+v-u}\right)\right]^{-1} \frac{u(w-u)}{(w+v-u)} \\ u < w+v \quad , \end{gathered} \tag{2.15}$$

where A is a constant that is determined from the initial condition. We can normalize this solution to give the proportions of families with i copies, c_i.

$$c_i = \left[\ln\left(\frac{w+v}{w+v-u}\right)\right]^{-1} \frac{1}{i}\left(\frac{u}{w+v}\right)^i \quad , \tag{2.16}$$

As this model requires that η be nonnegative, this restricts our solution further to $u \leq w$.

Note that the equilibrium solution (2.16) only exists when the rate constants are precisely balanced as described in (2.15). As this situation is unlikely in reality, we can also look for stationary distributions, that is, solutions in which the total number of copies are changing in time, but their proportions remain the same. These solutions will be given by:

$$\frac{d\mathbf{C}}{dt} = k\mathbf{C} \quad , \tag{2.17}$$

where k is a constant. Clearly k is an eigenvalue of the coefficient matrix of the ODE system (2.13), and gives the longterm exponential growth rate; its corresponding eigenvector gives the longterm distribution. While η_c corresponds to $k = 0$, the condition $\eta > \eta_c$ implies an increasing population and $\eta < \eta_c$ implies a decreasing population, where η_c is defined in (2.15).

We obtained the longterm growth rate, k, when $k \neq 0$ by finding the eigenvalue with the largest real part of the coefficient matrix of system (2.13). Using the convergence criteria outlined above, the eigenvalue was determined numerically with a finite matrix size, N, and N was increased until the eigenvalue converged.

2-D Copy-Level Model

For the two dimensional model, $R_{i,j}$ is the number of families with i copies in the coding region, and j copies in the non-coding region. This can also be expressed as a system of ODEs:

$$\frac{dR_{1,0}}{dt} = -(u_1 + w_1 + v_1)R_{1,0} + 2(v_1 + w_1)R_{2,0} + (v_2 + w_2)R_{1,1} + \sum_{\substack{i=0 \\ i+j\neq 0}}^{\infty}\sum_{j=0}^{\infty}(iv + \rho\eta)R_{i,j}$$

$$\frac{dR_{0,1}}{dt} = -(u_2 + w_2 + v_2)R_{0,1} + 2(v_2 + w_2)R_{0,2} + (v_1 + w_1)R_{1,1} + \sum_{\substack{i=0 \\ i+j\neq 0}}^{\infty}\sum_{j=0}^{\infty}(jv + (1-\rho)\eta)R_{i,j}$$

$$\frac{dR_{i,0}}{dt} = \rho(i-1)u_1R_{i-1,0} - i(u_1 + w_1 + v_1)R_{i,0} + (i+1)(w_1 + v_1)R_{i+1,0} + (w_2 + v_2)R_{i,1}, \quad i \geq 2$$

$$\frac{dR_{0,j}}{dt} = (1-\rho)(j-1)u_2R_{0,j-1} - j(u_2 + w_2 + v_2)R_{0,j} + (j+1)(w_2 + v_2)R_{0,j+1} + (w_1 + v_1)R_{1,j}, \quad j \geq 2$$

$$\begin{aligned}\frac{dR_{i,j}}{dt} = {} & \rho((i-1)u_1 + ju_2)R_{i-1,j} + (1-\rho)(iu_1 + (j-1)u_2)R_{i,j-1} \\ & -(i(u_1 + w_1 + v_1) + j(u_2 + w_2 + v_2))R_{i,j} \\ & +(i+1)(w_1 + v_1)R_{i+1,j} + (j+1)(w_2 + v_2)R_{i,j+1}, \quad i,j \geq 1 \ .\end{aligned} \tag{2.18}$$

To estimate the longterm growth rate of the system, we integrate (2.18) numerically.

Genome-Level Model

We can similarly solve for an equilibrium solution to the genome-level model. When condition (2.15) is met at the copy level, both levels will be at equilibrium. When condition (2.15) is not met, however, one of the parameters of the genome-level model, μ, will not be constant. This HGT rate scales with the average number of families per genome, $\bar{f}$. Intuitively, if the number of families with n copies is growing, for all n, in a finite population of genomes, the mean number of families per genome is likewise growing. Thus the equilibrium distribution derived for the genome-level model is valid only when the copy-level model is at equilibrium.

Similar to our previous approach, the genome-level can be expressed as a system as ODEs:

$$\frac{dF_m}{dt} = ((m-1)p + \mu)F_{m-1} - (m(p+q) + \mu)F_m + (m+1)qF_{m+1}, \; m \geq 0 \quad (2.19)$$

where F_m is the number of genomes containing m families of MGEs (and we define $F_{-1} = 0$). Using the software package Maple, we are again able to find an analytic solution in the case of $\frac{d\bar{\mathbf{F}}}{dt} = 0$, under the constraint $p < q$:

$$\bar{F}_m = \frac{G}{m!}\left(\frac{p}{q}\right)^m \frac{\Gamma(m + \frac{\mu}{p})}{\Gamma(\frac{\mu}{p})}\left(1 - \frac{p}{q}\right)^{\mu/p} \quad , \qquad (2.20)$$

where G is the number of genomes and $\Gamma(x)$ is the gamma function.

Then, the proportion of genomes with m families, f_m, is given by:

$$f_m = \frac{1}{m!}\left(\frac{p}{q}\right)^m \frac{\Gamma(m + \frac{\mu}{p})}{\Gamma(\frac{\mu}{p})}\left(1 - \frac{p}{q}\right)^{\mu/p} \quad . \qquad (2.21)$$

Unlike the copy-level model, we do not have to consider stationary solutions. In this model there are no absorbing states, and the system will approach this equilibrium as long as the condition $p < q$ holds, and the parameters p, q and μ are constant. If $p > q$, the number of MGE families will increase without bound.

2.3 Results

It should be noted that it is the ratios of the parameters, rather than their actual values, that affect the extinction probabilities and equilibrium state of the system. We chose to normalize the 1-D copy-level and genome-level parameters by u and p respectively, using notation $\hat{w} = w/u$ for example, or $\hat{q} = q/p$. For the 2-D copy-level model, because the parameter η is shared by both regions of the genome, we normalize by η, for example $\tilde{w_1} = w_1/\eta$. In previous work, best fit values for MPs were found to be $w/u = 0.9810$, $v/u = 0.0424$, $\eta/u = 0.0965$, $q/p = 1.24$ and $\mu/p = 1.09$ [88]. We will refer to these values from the promoter dataset as $\hat{w}_p$, $\hat{v}_p$, $\hat{\eta}_p$, $\hat{q}_p$ and $\hat{\mu}_p$, respectively.

An important interaction occurs in this model between diversification and HGT. Since HGT is not dependent on copy number, there is more HGT when there are many families with few copies, as compared to few families with the same number of total copies. Therefore, diversification acts indirectly to increase the amount of HGT.

2.3.1 Extinction Probability

For the 1-D copy-level model, the extinction probability for a family of size 1, X_1, is illustrated in Figure 2.3. We find that increasing HGT, η, promotes survival, which is reasonable since each HGT event adds a MGE copy. We also find, however, that increasing diversification, $\hat{v}$, increases survival, but only in the presence of HGT. Since a diversification event has no net effect on copy number, this is an interesting result (see Section 2.4). Finally, we note that the parameters estimated for MPs, $\hat{w}_p$, $\hat{v}_p$ and $\hat{\eta}_p$, predict an extinction probability of $X_1 = 0.96$.

In Figure 2.4, we examine the extinction probability for larger families, finding that increasing family size, n, reduces the extinction probability as expected. In both asymptotic cases, note that extinction probability obeys an exponential law with

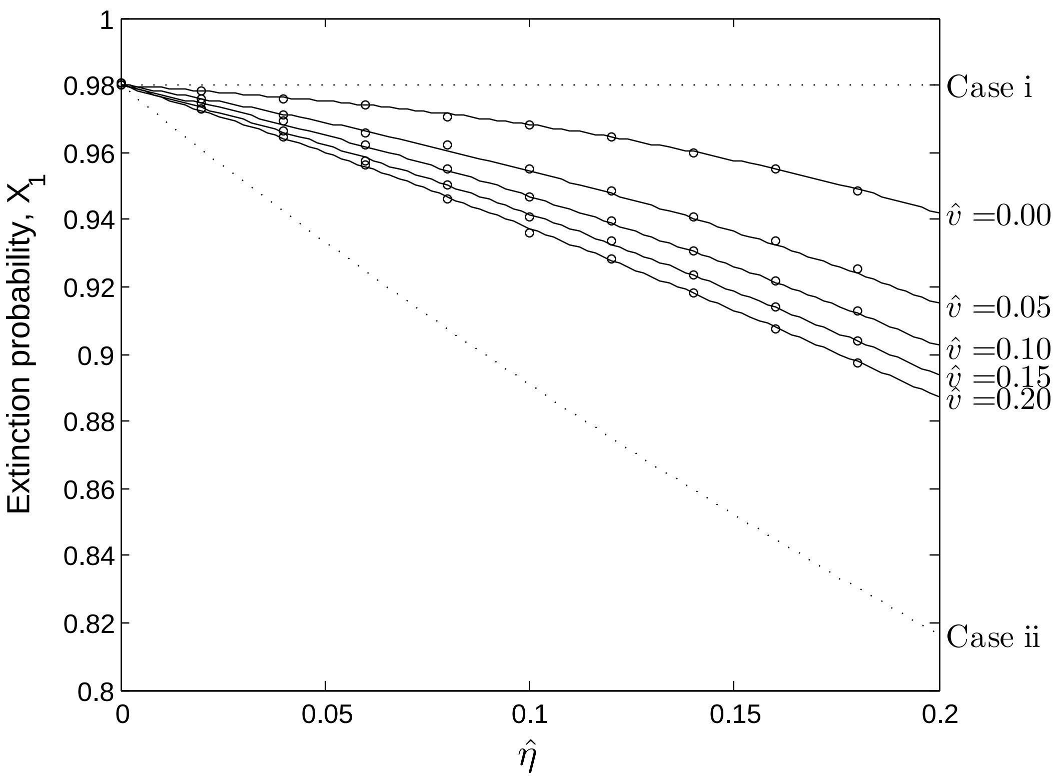

Figure 2.3: Extinction probabilities of one member families, X_1, for varying diversification ($\hat{v}$) and horizontal gene transfer ($\hat{\eta}$) rates. The value of the deletion rate parameter is $\hat{w} = 0.98$. Results are shown for both the numerical solution (solid line), and Monte Carlo simulation (circles). The asymptotic cases are described in equations (2.4) and (2.5), and depicted as dotted lines.

respect to n. More generally, however, the solution to equation (2.3) does not obey an exponential law as seen in the figure inset. In particular, for small families, the extinction probability is lower than would be predicted by an exponential law.

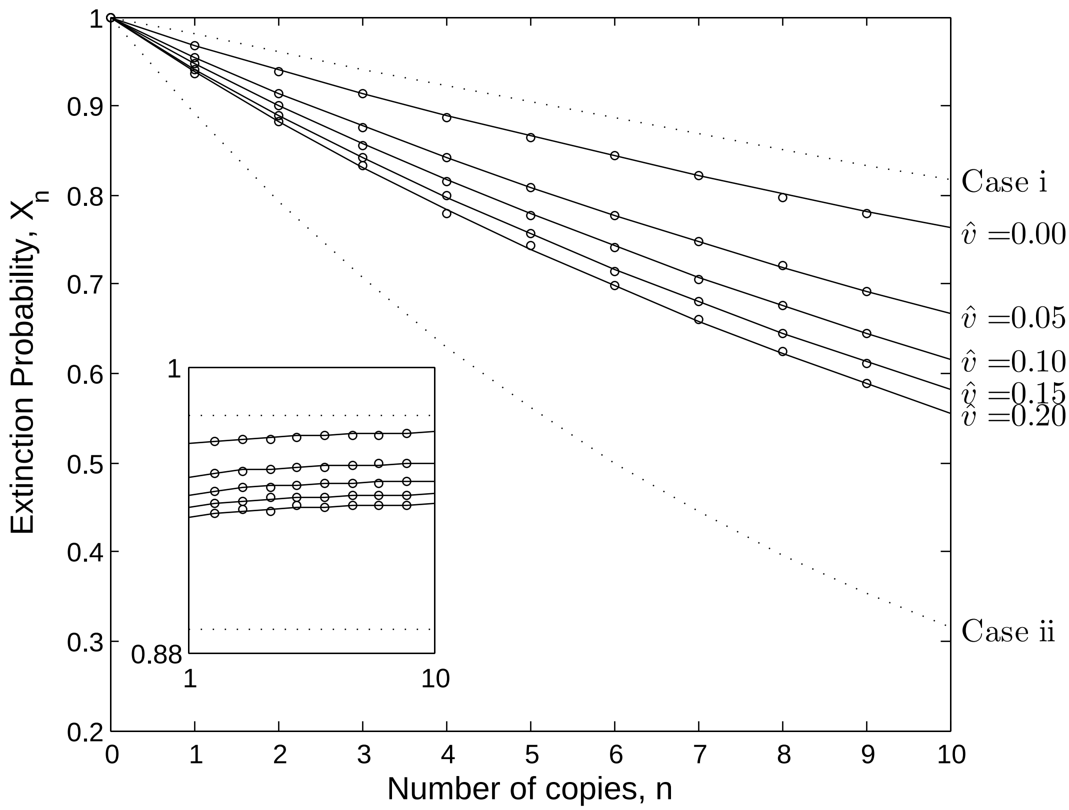

Figure 2.4: Extinction probabilities of n member families, X_n, for varying diversification rates ($\hat{\upsilon}$). Values used for HGT and deletion rates are $\hat{\eta} = 0.1$ and $\hat{w} = 0.98$ respectively. Results are shown for both the numerical solution (solid line), and Monte Carlo simulation (circles). Asymptotic cases are described in equations (2.4) and (2.5), and depicted as dotted lines. The inset shows the same data but with $\sqrt[n]{X_n}$ on the y-axis. With the exception of the asymptotic cases, the solutions do not obey an exponential law. All solutions are for n integer; lines are plotted continuously to guide the eye.

For the 2-D model, we compare the extinction probability of one member families if they are initially present in the coding region to those in the non-coding region in Figure 2.5. When $\rho = 0$, the extinction probability for one member families in the non-coding region, $X_{0,1}$, is the same as in the 1-D case using the rate parameters for the non-coding region (u_2, w_2, v_2 and η). As ρ increases from 0 to 1 this value approaches the extinction probability for the coding region (u_1, w_1, v_1 and η). The opposite is true when considering $X_{1,0}$ from $\rho = 1$ to $\rho = 0$. A striking feature of these predictions is the substantial influence of the initial condition: even when all further copies have equal chances of being mobilized to coding or non-coding regions ($\rho = 0.5$), the extinction probability sensitively depends on where the MGE lineage begins. This is because a critical factor in extinction is whether the initial copy itself is deleted before it duplicates.

2.3.2 Extinction Times

The distribution of extinction times is shown in Figure 2.6, and expected extinction times for one member families, T_1, are shown in Figure 2.7. Increasing diversification and HGT rates decrease expected extinction times even though the families are less likely to go extinct. Parameters estimated for MPs, $\hat{w}_p$, $\hat{v}_p$ and $\hat{\eta}_p$ yield an expected extinction time for one member families to be $T_1 = 3.7u^{-1}$.

2.3.3 Equilibrium Solution

Examples of numerical solutions to the 1-D copy-level ODEs, equation (2.13), are shown in Figure 2.8 for three cases; $\eta < \eta_c$, $\eta = \eta_c$ and $\eta > \eta_c$, where η_c is defined in (2.15). The behavior is as expected: after a transient period, the system reaches a stationary distribution in which total copy number either decreases, stays the same, or increases with time depending on the value of η.

The effects of parameters on the longterm growth rate, $\hat{k}$, are illustrated in Figure 2.9. Increases in HGT, $\hat{\eta}$, increase the growth rate, while increases in deletion, $\hat{w}$, decrease growth. Note however that diversification, $\hat{v}$, causes increased growth, but only in the presence of HGT. Similar to our predictions for extinction probability, this appears counter-intuitive since diversification events do not change total copy

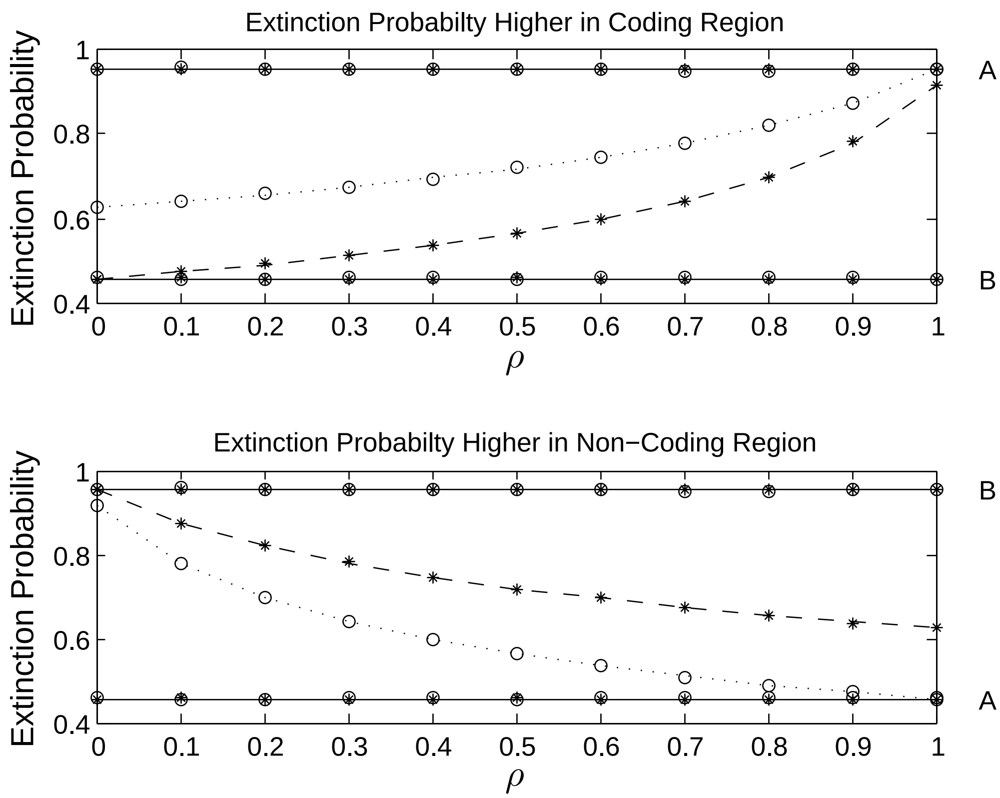

Figure 2.5: Extinction probability versus the probability that new MGEs are inserted into coding regions, ρ, for the 2-D model (2.6). Dotted lines and dashed lines represent the numerical solution of the extinction probability for a single promoter copy in a coding region, $X_{1,0}$, or non-coding regions, $X_{0,1}$, respectively. Solid lines illustrate cases in which there is no difference in the rates between coding and non-coding regions, that is, all parameters are set to either the coding region (A) or non-coding region (B) parameters. In these cases $X_{1,0} = X_{0,1}$ for all ρ. The results of Monte Carlo simulation are shown with circles for $X_{1,0}$, and stars for $X_{0,1}$. In the top panel $\tilde{w}_1 = 9.8$ and $\tilde{w}_2 = 4.9$, and in the bottom $\tilde{w}_1 = 4.9$ and $\tilde{w}_2 = 9.8$. Other parameter values are $\tilde{u}_1 = \tilde{u}_2 = 10$ and $\tilde{v}_1 = \tilde{v}_2 = 0.4$.

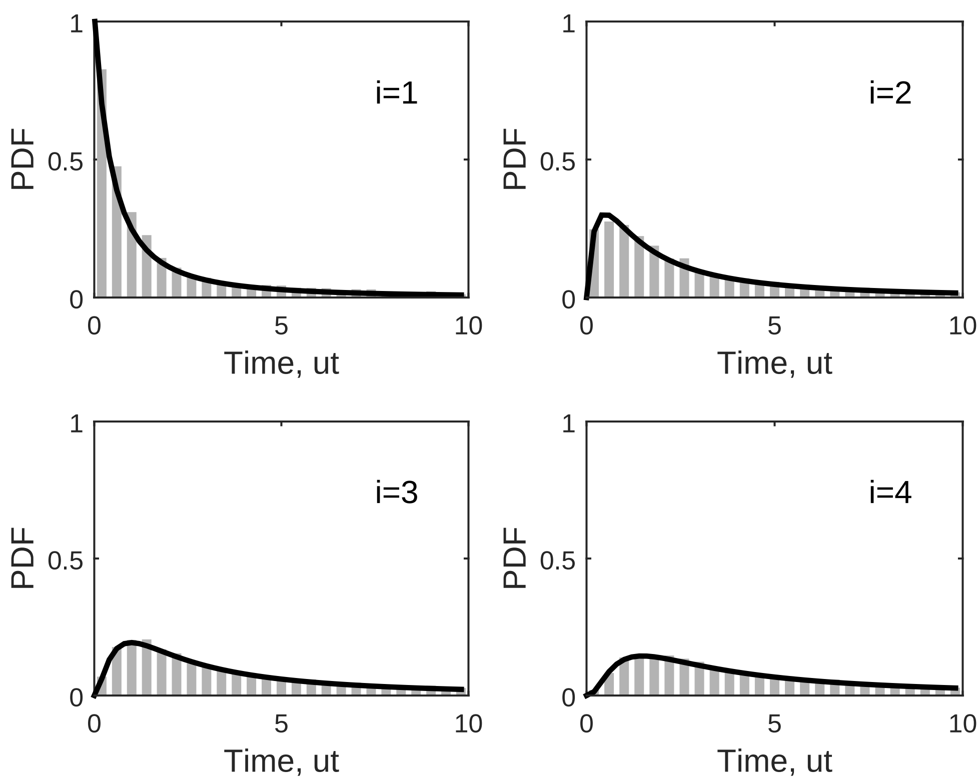

Figure 2.6: Histograms of extinction times for families that have i copies at time zero. The stochastic simulation was performed 1×10^4 times for each plot. The numerical solution to (2.10) is plotted with a solid line. Parameter values used were $\hat{w} = 0.98$, $\hat{v} = 0.04$, and $\hat{\eta} = 0.1$.

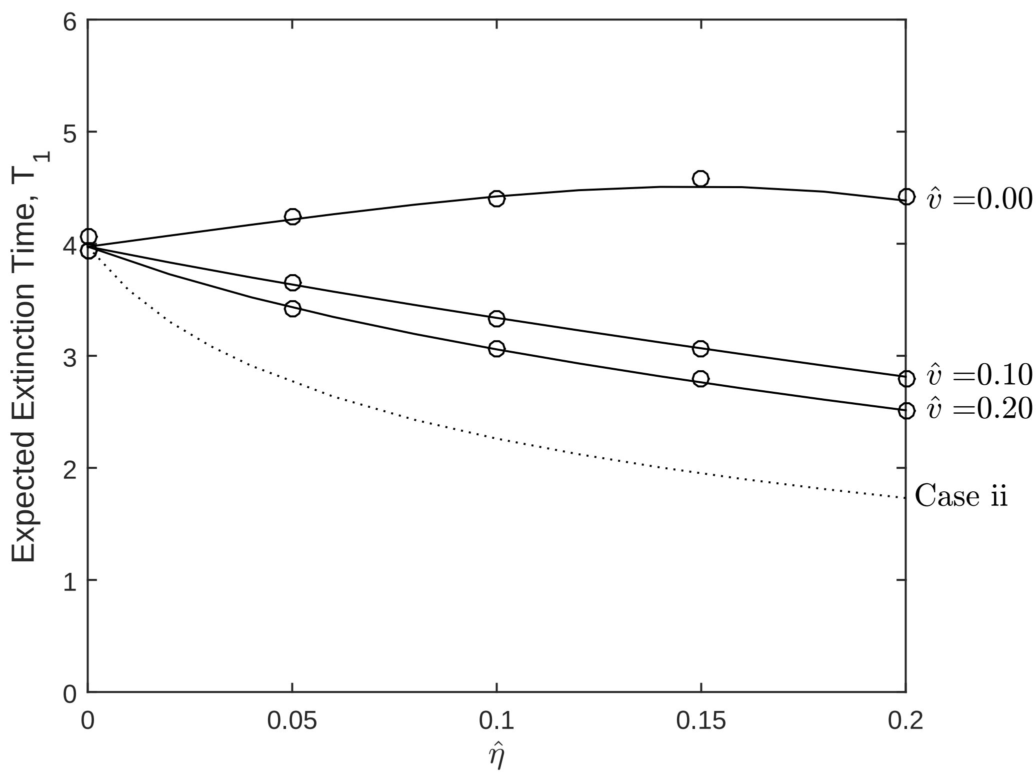

Figure 2.7: Expected extinction times for families that have 1 copy at time 0, T_1. Both numerical (solid line) and Monte Carlo simulation (circles) are shown. Numerical solutions were found using Equation (2.11). Time is non-dimensionalized with respect to u.

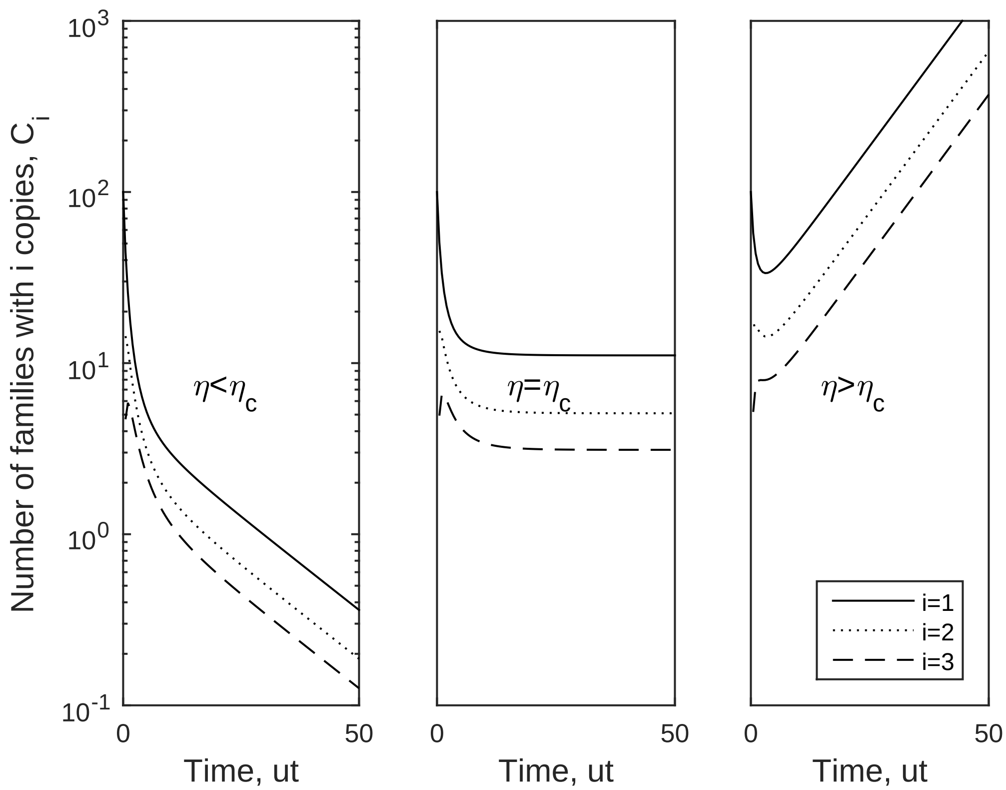

Figure 2.8: Numerical solution to copy-level model (2.13). The number of families with i copies is plotted versus time, where time is non-dimensionalized with respect to the duplication rate, u. Parameters values are $\hat{v} = 0.04$ and $\hat{w} = 1.05$. The values of $\hat{\eta}$ were set to 0, $\hat{\eta}_c$ and $2\hat{\eta}_c$, where η_c is defined in (2.15). In all three cases, the population appears to reach a stationary distribution after a transient. System (2.13) was numerically integrated using a Matlab routine (ODE15s) with an initial condition of one family with 100 copies.

number; we will return to this observation in the Section 2.4. We also note that the parameters estimated for mobile promoters predict a growth rate of $k = 0.0336u$.

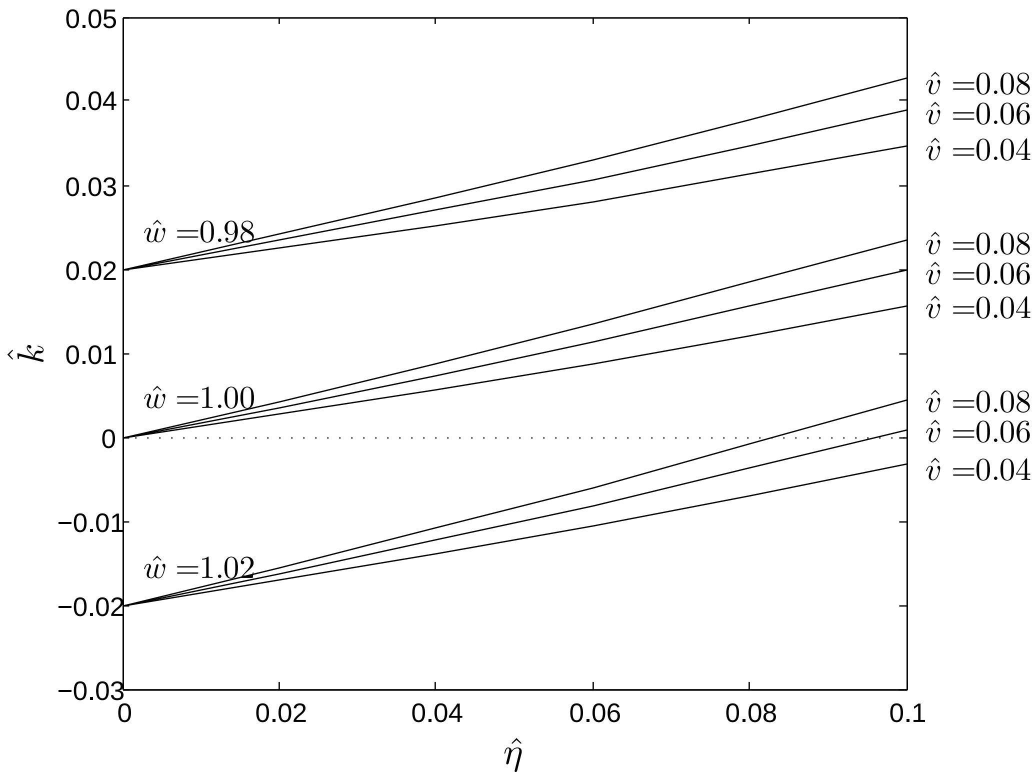

Figure 2.9: Growth rate of the number of families. The longterm growth rate $\hat{k}$, is plotted for varying deletion, $\hat{w}$, diversification, $\hat{v}$, and horizontal gene transfer, $\hat{\eta}$, rates. The dotted line at $\hat{k} = 0$ corresponds to the true equilibrium solution (2.16).

The longterm growth rate for the 2-D model is shown in Figure 2.10. For $\rho = 0$ the growth rate agrees with the rate from the 1-D model using parameters in the non-coding region, and similarly for $\rho = 1$ we recover the rate from the 1-D model using parameters in the coding region. The growth rate varies linearly with ρ. Therefore, if the rates in the two regions differ, the overall growth rate will be a weighted average of values in the two regions.

Finally, we examined how HGT, μ, affects the distribution of families per genome.

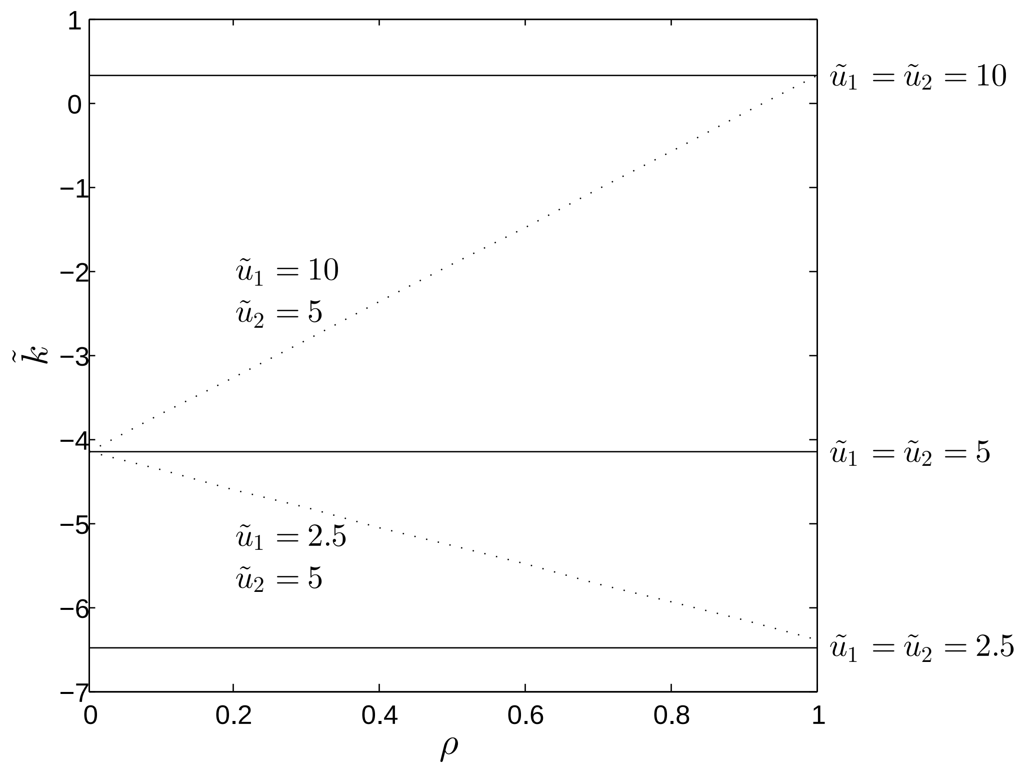

Figure 2.10: The longterm growth rate, $\tilde{k}$ for the 2-D copy-level model versus the probability a new element will be inserted into the coding region (2.18). Parameter values are $\tilde{w}_1 = \tilde{w}_2 = 9.8$ and $\tilde{v}_1 = \tilde{v}_2 = 0.4$. Solid lines lines are used when the parameters are the same in both regions. To calculate $\tilde{k}$, the system was integrated numerically using a fourth order Runge-Kutta method until the slope converged (within a tolerance of 10^{-8}). The system was originally of size $i + j < N$, where $N = 100$. N was increased in steps of 50 until the slope converged (within a tolerance of 10^{-4}).

These results were plotted using the analytic solution (2.21), and are shown in Figure 2.11. As μ is increased, the peak in the distribution shifts to higher m values. While most genomes have no MGE families when HGT rates are low, higher rates of HGT yield distributions in which the vast majority of genomes carry many families of MGEs.

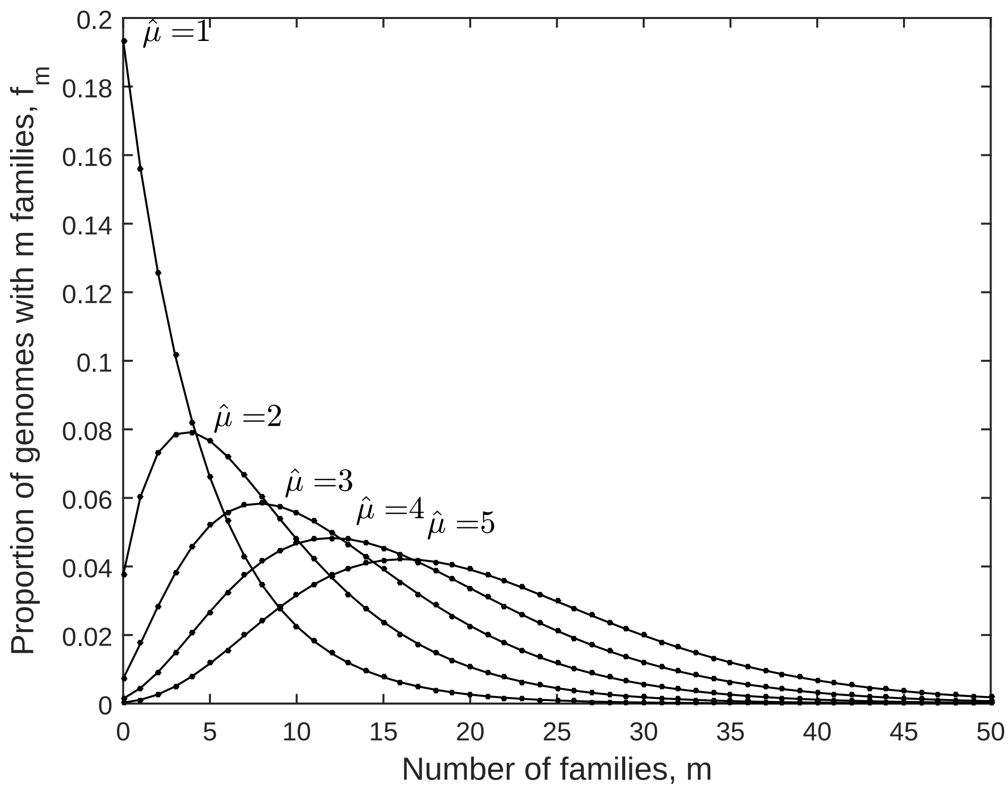

Figure 2.11: Analytic solution (circles) to the genome-level model (2.21), for varying $\hat{\mu}$, the rate at which new families enter a genome through horizontal gene transfer. The parameter value used for the loss term was $\hat{q} = 1.24$. Lines have been added to guide the eye.

The MP data set for all sequenced prokaryote genomes yields a monotonically decreasing distribution of families per genome [88]. However, in examining individual species from this data set, examples of distributions with non-zero peaks are evident; Figure 2.12 shows four such anecdotal cases. Although the copy-level growth rate,

k, was estimated to be non-zero for the MP dataset, and thus the distribution of families per genome cannot strictly be at equilibrium, a comparison of Figure 2.12 with Figure 2.11 suggests that HGT rates for MPs within species are higher than the average rate estimated from the pooled prokaryote data.

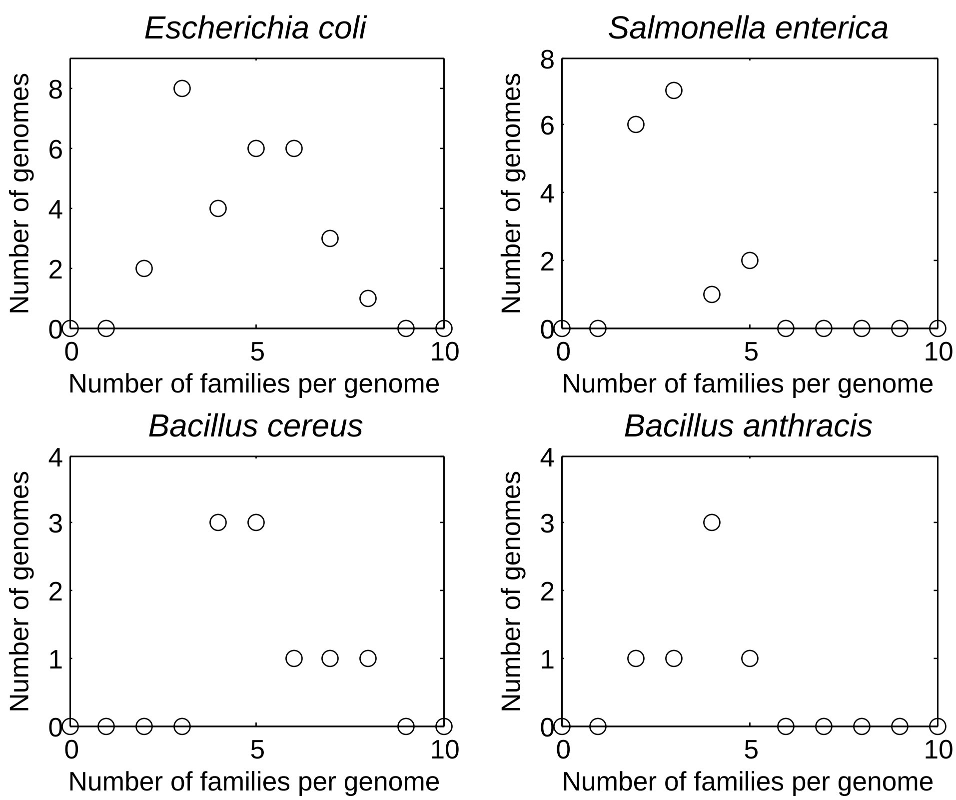

Figure 2.12: Anecdotal data from four species included in the mobile promoter study [88]. Histograms show the number of sequenced genomes in the species with varying numbers of families per genome. The equivalent histogram for data pooled from all prokaryote species is monotonically decreasing [88], consistent with the expectation that within species, HGT rates are higher.

2.4 Discussion

The key novel feature of the model we examine is diversification. Diversification changes the dynamics in part because the process is non-conservative, and cannot be described as a standard Markov process. One of the most interesting results to emerge is that diversification increases the survival probability of a lineage of MGEs, but only in the presence of HGT. Increased diversification does not change the total number of element copies, but rather changes the distribution; it creates more families with fewer members. In contrast, HGT does create another copy and therefore increases survival. Because HGT occurs at a constant rate per family, independent of the number of copies in the family, HGT is more effective if the elements are distributed into many small families. Therefore, high diversification rates will increase the rate of HGT, and thus reduce the extinction probability.

We also found that for the 1-D copy-level model to have an equilibrium distribution, the parameter values must obey a strict relationship (2.15). Otherwise, there can exist a stationary distribution, in which the total number of families is changing, but the ratio of family sizes remains the same. Since in reality it is not likely that the parameters will perfectly satisfy (2.15), we expect that the number of families with n copies is changing in time for most MGEs. A consequence of this result is that the genome-level equilibrium distribution is unlikely to be satisfied. This is because the HGT rate at the genome level, μ, depends on the average number of families per genome, and so changes as the overall number of families increases or decreases.

A further prediction of interest is that HGT changes the shape of the family distribution, that is, the number of families per genome. In previous work [88], the probability mass function was shown to be monotonically decreasing with increasing family number. However, we demonstrate here that if HGT rates are large enough, there is a non-zero peak in the distribution, and as HGT increases, the location of this peak occurs at a higher family number. Therefore, the shape of the distribution of families per genome for a specific MGE should reflect HGT rates for that element. In future work we could apply this approach to compare HGT rates of different MGEs in a single species, or related MGEs across different species, simply by comparing their family-number distributions.

The predictions above address the birth-death-diversification model as a general approach for describing the longterm fate of MGE lineages. Since numerical parameter values for this model have previously been estimated for MPs, we can make further quantitative predictions specific to these elements. For example, these parameters predict that new families of promoters become extinct 96% of the time. Furthermore, we used these parameters to estimate that one member families en route to extinction have an expected extinction time of $T_1 \approx 4/u$. If we assume the birth rate, u, is on the order 10^{-5} per element per generation [85, 88], this corresponds to an extinction time of 4×10^5 generations. This is a relatively short time, and suggests that extant MP families represent only a small fraction of those that have arisen throughout genomic evolution.

In Section 2.4, we found that the deletion rate must exceed the duplication rate for the copy-level model to be in equilibrium. Interestingly, this condition was not met in the promoter data set, which had a higher duplication rate. Similarly, a study of insertion sequences in *Escherichia coli* [85] estimated that duplication slightly exceeded deletion. We thus predict that for MPs, the system will not be in equilibrium but will reach a stationary distribution that grows at a rate $k = 0.0336u$. If we again estimate u is on the order 10^{-5} per element per generation, this result predicts that the number of MP families is increasing at 3×10^{-7} families per generation. This rate is sufficiently low that for MPs it is reasonable to assume that the rate parameters for the genome-level model are approximately constant. Equilibrium at the genome level further requires that the birth rate of families be less than the death rate, a condition which holds in the estimated MP parameters. Nonetheless, we reiterate that in the MP dataset, the genome level parameters are not strictly constant, and while the model offers several important insights, its predictive ability in the longterm is clearly limited.

In Section 2.3, we illustrated the application of the 2-D model to coding and non-coding regions of the genome, however the model could be used to divide MGEs into other regions or categories. For example, the mobile promoter data was initially collected in "promoter" regions of the prokaryote genome, immediately upstream from coding sequences [76]. Copies of promoter sequences thus identified were subsequently detected in non-promoter regions of the genome [88]. A comparison of these two data sets revealed that only $13,000$ out of $40,000$ potential MPs were located in the promoter region [88]. We suggest that estimating the parameter rates of the 2-D model for MPs in promoter and non-promoter regions could offer an explanation as to this surprising prevalence of MPs in non-promoter regions.

Finally, for MPs, we demonstrated anecdotally that for several species, the distribution of families per genome is not monotonically decreasing. This suggests that HGT rates can vary considerably among species, and is consistent with the expectation that HGT rates are higher within, rather than between species. Unfortunately, the data from individual species are not yet sufficient to support model fitting. We also note that a non-zero peak in the family distribution could simply be an artifact of the highly non-uniform sample of genomes for which full sequence data are available.

Extensions to the approach we outline here are readily apparent. For instance, we assume that waiting times for diversification are exponentially distributed. Although diversification could occur through the deletion or insertion of a length of sequence inside a MGE, diversification could also occur through successive point mutations. For the latter case, waiting times based on a gamma distribution would clearly be more accurate.

2.5 Conclusion

As applied to mobile promoters, this paper gives the first estimate of the extinction probability of novel MP lineages, and predicts that the number of copies per family is not in equilibrium, but is growing very slowly. We also provide an avenue for exploring the intriguing observation that a substantial fraction of MPs persist in non-promoter regions of the genome. More generally, we predict that as long as HGT occurs at a roughly constant rate per family, both the survival and growth rate of mobile genetic elements will be increased by genetic diversification; this effect will be greater where HGT is higher. We propose a new approach for estimating HGT rates by examining the distribution of MGE families per genome, and analyze a unique stochastic process that describes the dynamics of a wide range of mobile genetic elements in prokaryotes.

Bibliography

[59] R.C. Baños, A. Vivero, S. Aznar, J. García, M. Pons, C. Madrid, and A. Juárez. Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genet*, 5(6):e1000513, 2009.

[60] C.J. Basten and M.E. Moody. A branching-process model for the evolution of transposable elements incorporating selection. *J Math Biol*, 29(8):743–61, 1991.

[61] M. Bichsel, A.D. Barbour, and A. Wagner. Estimating the fitness effect of an insertion sequence. *J Math Biol*, 66(1):95–114, 2013.

[62] J.F. Brookfield. A model for DNA sequence evolution within transposable element families. *Genetics*, 112(2):393–407, 1986.

[63] B. Charlesworth and D. Charlesworth. The population dynamics of transposable elements. *Genet Res*, 42:1–27, 1983.

[64] E.S. Dolgin and B. Charlesworth. The fate of transposable elements in asexual populations. *Genetics*, 174(2):817–27, 2006.

[65] R.J. Edwards and J.F. Brookfield. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments. *Mol Biol Evol*, 20(1):30–7, 2003.

[66] L.S. Frost, R.Leplae, A.O. Summers, and A.Tussaint. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*, 3:722–32, 2005.

[67] G.B. Golding, C.F. Aquadro, and C.H. Langley. Sequence evolution within populations under multiple types of mutation. *Proc Natl Acad Sci U S A*, 83(2):427–31, 1986.

[68] D.L. Hartl and S.A. Sawyer. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? *Genetics*, 118(3):537–41, 1988.

[69] D.A. Hickey. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101:519–31, 1982.

[70] R.R. Hudson and N.L. Kaplan. On the divergence of members of a transposable element family. *J Math Biol*, 24(2):207–15, 1986.

[71] N. Kleckner. Transposable elements in prokaryotes. *Annu Rev Genet*, 15:341–404, 1981.

[72] C.H. Langley, J.F. Brookfield, and N. Kaplan. Transposable elements in mendelian populations. I. A theory. *Genetics*, 104(3):457–71, 1983.

[73] A. Le Rouzic, T.S. Boutin, and P. Capy. Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A*, 104(49):19375–80, 2007.

[74] A. Le Rouzic and G. Deceliere. Models of the population genetics of transposable elements. *Genet Res*, 85(3):171–81, 2005.

[75] D. Lee and O. Bernard. Adaptive evolution of *Escherichia coli* K-12 mg1655 during growth on a nonnative carbon source, l-1, 2-propanediol. *Appl Environ Microb*, 76(13):4158–68, 2010.

[76] M. Matus-Garcia, H. Nijveen, and M.W. van Passel. Promoter propagation in prokaryotes. *Nucleic Acids Res*, 40(20):10032–40, 2012.

[77] M.E. Moody. A branching process model for the evolution of transposable elements. *J Math Biol*, 26(3):347–57, 1988.

[78] H. Nijveen, M. Matus-Garcia, and M.W. van Passel. Promoter reuse in prokaryotes. *Mob Genet Elements*, 2(6):279–81, 2012.

[79] T. Ohta. A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics*, 110(3):513–24, 1985.

[80] Y. Oren, M.B. Smith, N.I Johns, M.K. Zeevi, D. Biran, E.Z. Ron, J. Corander, H.H. Wang, E.J. Alm, and T. Pupkol. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc Natl Acad Sci U S A*, 111(45):16112–16117, 2014.

[81] L.E. Orgel and F.H.C. Crick. Selfish DNA: The ultimate parasite. *Nature*, 284:604–607, 1980.

[82] G.M. Rubin, M. Kidwell, and P.M. Bingham. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Annu Rev Genet*, 29:987–94, 1982.

[83] S. Sawyer and D. Hartl. Distribution of transposable elements in prokaryotes. *Theor Popul Biol*, 30(1):1–16, 1986.

[84] M. Slatkin. Genetic differentiation of transposable elements under mutation and unbiased gene conversion. *Genetics*, 110(1):145–58, 1985.

[85] A. Sousa, C. Bourgard, L.M. Wahl, and I. Gordo. Rates of transposition in Escherichia coli. *Biol Lett*, 9(6):20130838, 2013.

[86] D. Stoebel and C. Dorman. The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. *Mol Biol Evol*, 27(9):2105–12, 2010.

[87] C.M. Thomas and K.M. Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol*, 3:711–21, 2005.

[88] M.W. van Passel, H. Nijveen, and L.M. Wahl. Birth, death, and diversification of mobile promoters in prokaryotes. *Genetics*, 197(1):291–9, 2014.

[89] A. Wagner. Cooperation is fleeting in the world of transposable elements. *PLoS Comput Biol*, 2(12):e162, 2006.

Chapter 3

Conclusion

In summary, this thesis examines a mathematical model used to describe the dynamics of MGEs in prokaryotes. Specifically, we analyze equilibrium distributions, extinction probabilities and extinction times. The work presented here emphasizes the importance of considering diversification of MGE families. Though diversification does not directly act to increase the number of copies of an element, it indirectly increases HGT rates and therefore copy number. Additionally, this work develops an extension of the birth-death-diversification model that separates the genome into two different sections. For example, the dynamics in coding and non-coding regions can be described by different rates. Alternatively, the genome could be split into regulatory and non-regulatory regions, and this could offer insight at to why the majority of MPs appear to be located in non-regulatory regions [92, 93].

Though this model has offered many interesting insights into the dynamics of prokaryotic MGEs, mainly through the relationship between diversification and HGT, there are many simplifications to the model which could be addressed in future work.

The main weakness in the model is that it predicts that if an MGE population does not go extinct, it will grow without bound. This biologically impossible result is likely due to the fact that selection is ignored. While many studies have shown that MGEs are likely selectively neutral [90, 91, 93], there is clearly an upper limit to the number of MGEs one cell can possess. For the model explored in this thesis, van Passel et al. [93] considered three possible forms of selection for the MP data: (1) fitness decreases exponentially with increased numbers of MGE families (2) fitness decreases linearly with increased families, and (3) fitness decreases uniformly, dependent only on the presence or absence of MGEs. The authors found that none of these models were justified statistically. However, the authors did not consider fitness that depended on the copy number of MGEs, and this might be a more accurate description. Therefore, it could be useful to examine selection models in which fitness decreases as a function of copy number. If one of these selection models were justified using the MP data, perhaps the birth-death-diversification model could be extended to incorporate selection.

Another simplification to the birth-death-diversification model is that it describes diversification waiting times as being exponentially distributed. Presumably, diversification would happen through a series of exponentially-distributed mutation events, until the sequence was altered enough to be considered a new family. Therefore, a more accurate model might require that an MGE go through numerous mutation events before it is considered a new family. However, since each element can still go through duplication, deletion and HGT events, this complicates the model immensely.

Additionally, the main result of this thesis—that diversification is important in MGE dynamics—hinges on the idea that HGT rates are not dependent on copy number. This is supported by results for ISs [91] and MPs [93]. However, more in-depth work on the relationship between HGT rates and copy number in MGEs could be interesting.

Finally, this model suggests that diversification is beneficial to a MGE because the element is more likely to be transferred to other cells through HGT. However, there is a cost to diversification which this model does not consider. If a mutation happens in the region of the MGE that involves transposition, the MGE could be less likely to transpose. Additionally, if the MGE carries a beneficial accessory gene, a diversification event could change the function of this gene, so it was no longer beneficial, or perhaps even detrimental to a cell's fitness. Therefore, there should be some penalty to having a diversification event. This question could also be addressed using game theory, in an approach similar to Wagner's [94]. For instance, if we consider a plasmid with one transposition site, the original MGE and a diversified version would compete for this same site. The diversified gene might have a reduced ability to transpose, but would be more likely to be transferred to another host. In future work this could be used to determine whether the benefit of diversification is worth the possible loss of functionality.

In conclusion, MGEs are ubiquitous across species and have applications to a wide range of topics including antibiotic resistance, evolution and biological engineering. Therefore, understanding the dynamics of these elements is of utmost importance. Many models have been developed to describe the dynamics of MGEs, but little attention has been given to the diversification of elements, likely because the relationship between diversification and HGT is subtle. Here we have mentioned numerous avenues that could be pursued to further explore this finding. Overall, our work indicates that diversification of elements should be considered in order to fully understand the dynamics of MGEs in prokaryotes.

Bibliography

[90] M. Bichsel, A.D Barbour, and A. Wagner. The early phase of a bacterial insertion sequence infection. *Theor Popul Bio*, 78(4):278–288, 2010.

[91] M. Bichsel, A.D. Barbour, and A. Wagner. Estimating the fitness effect of an insertion sequence. *J Math Biol*, 66(1):95–114, 2013.

[92] M. Matus-Garcia, H. Nijveen, and M.W. van Passel. Promoter propagation in prokaryotes. *Nucleic Acids Res*, 40(20):10032–40, 2012.

[93] M.W. van Passel, H. Nijveen, and L.M. Wahl. Birth, death, and diversification of mobile promoters in prokaryotes. *Genetics*, 197(1):291–299, 2014.

[94] A. Wagner. Cooperation is fleeting in the world of transposable elements. *PLoS Comput Biol*, 2(12):e162, 2006.

Curriculum Vitae

Name:	Nicole Drakos
Post-Secondary Education and Degrees:	Master of Science in Applied Mathematics, candidate Mathematical Biology with Scientific Computing The University of Western Ontario, 2013-
	Bachelor of Science Honors Double Major in Applied Mathematics and Astrophysics The University of Western Ontario, 2011-2013
	Bachelor of Medical Science Honors Specialization in Medical Science The University of Western Ontario, 2007-2011
Honors and Awards:	Ontario Graduate Scholarships ($15,000) The University of Western Ontario, 2014-2015
	The Dillon Gold Medal Highest graduating average in Applied Mathematics module The University of Western Ontario, Awarded June 2013
	The Western Gold Medal in Astrophysics Highest graduating average in Astrophysics Major module The University of Western Ontario, Awarded June 2013
Related Work Experience:	Teaching Assistant The University of Western Ontario, 2013-2015
	Research Assistant The University of Western Ontario, 2013-2015
Publications:	Drakos, N.E. and Wahl, L.M. Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: the birth-death-diversification model. *Under Revision for Theoretical Population Biology*