10-1-2015 12:00 AM

Evolution of Mobile Promoters in Prokaryotic Genomes.

Mahnaz Rabbani, *The University of Western ontario*

Supervisor: Dr. Lindi Wahl, *The University of Western Ontario*

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

Evolution of Mobile Promoters in Prokaryotic Genomes

(Thesis Format: Integrated-Article)

by

Mahnaz Rabbani

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 are important factors in evolution, and greatly influence the structure of genomes, facilitating the development of new adaptive characteristics. The dynamics of these mobile elements can be described using various mathematical and statistical models. In this thesis, we focus on a specific category of mobile genetic elements, *i.e.* mobile promoters, which are mobile regions of DNA that initiate the transcription of genes. We present a class of mathematical models for the evolution of mobile promoters in prokaryotic genomes, based on data obtained from available sequenced genomes. Our novel location-based model incorporates two biologically meaningful regions of the genome: promoter regions and other sites in the genome. We find the best model to describe the process using model selection techniques and reveal the most influential parameters in this dynamic process. We then compare the dynamics in these two regions of the genome with regards to the rates of four key processes: duplication, loss, diversification and horizontal gene transfer (HGT).

Keywords: mobile genetic elements, mobile promoters, promoter regions

Co-Authorship Statement

This thesis has been written by Mahnaz Rabbani under the supervision of Dr. Lindi Wahl. The work in chapter two is in preparation to be submitted to the Journal of Theoretical Biology.

Acknowledgments

I wish to express my sincere appreciation to my supervisor, Dr. Lindi Wahl, for giving me the opportunity to work under her invaluable supervision and be a member of her research team. I am extremely grateful to Dr. Wahl not only for her guidance, support and valuable advice but also for all her care, patience and encouragement. Sincere thanks to my colleagues and the examining committee.

I am indebted to M. Matus-Garcia, Mark W.J. van Passel and Harm Nijveen for providing the data used in this thesis, and Sharcnet by Compute Canada for providing us with the necessary computational resources.

I owe infinite thanks to my lovely parents, and my sisters for their continuous support and encouragement. I am in particular thankful beyond words to my older sister, Reihaneh, for all the love, attention, encouragement and support she gave me. Last but by no means the least, I would like to thank my dear friends, Aida and Amin, for giving me the home feeling since I came to Canada.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Prokaryotes and Their Genomes

All living organisms, except viruses, can be classified into three main domains: Bacteria, Archaea and Eukaryota. The first two branches are relatively simple organisms that are called prokaryotes. These are unicellular organisms with no nuclei or other membrane-bound organelles. From the evolutionary point of view, prokaryotes were the beginning stage of life, as two billion years ago eukaryotic cells evolved from the symbiosis of an archaeal host and a bacterium.

Prokaryotes' chromosomal DNA is located in the area of the cell's cytoplasm called the nucleoid, and has different DNA packaging from eukaryotes, which is known as supercoiling. Most prokaryotes have a single circular DNA molecule[1]. Moreover, unlike eukaryotes which reproduce sexually and typically carry two copies of each gene (diploid), most prokaryotes are asexual organisms and their genomes only have a single copy of each gene (haploid). Prokaryotic genomes carry much less noncoding DNA compared to eukaryotes, probably due to limited space in their single chromosomes. On average, 12% of prokaryote genomes consists of noncoding sequences as opposed to upwards of 98% in eukaryotes [Ahnert et al., 2008].

[1]However some exceptions have been discovered, such as *Vibrio cholerae* bacteria (causes cholera) that has two circular chromosomes [Trucksis et al., 1998], or the *Borrelia burgdorferi* bacteria (causes Lyme disease) which contains up to 11 copies of a single linear chromosome [Ferdows and Barbour, 1989].

Noncoding sequences of the genome perform various crucial functions such as initiating transcription, that is the first step of gene expression. Transcription refers to the process by which DNA is copied into RNA. One of the important transcriptional regulatory elements is a promoter, *i.e.* a DNA sequence that determines the DNA strand which must be transcribed and also the direction of transcription. The promoter indicates the transcription initiation site and launches transcription by providing a required binding site for RNA polymerase, *i.e.* the enzyme that is responsible for synthesizing RNA. The promoter region (PR) is located near (typically adjacent) to the transcription start site (TSS) and upstream in the genome from the coding sequence. Although eukaryotic promoters are relatively more complicated to recognize, in prokaryotes, the promoter region for many common genes is determined generally with two sequences, TATAAT and TTGACA, at roughly -10 bps and -35 bps upstream of the TSS, respectively.

This thesis studies and models mobile promoters, which are a sub-class of mobile genetic elements. Mobile promoter are of interest because they can affect the evolution of prokaryotic genomes, *i.e.* when the promoters activate silent genes or modify the expression of already present genes. The data used here includes strains of *E.coli* and other prokaryote genomes, 1362 genomes in total. *E.coli* is one of the most well-known species of prokaryote, and the most widely studied prokaryotic model organism. This is due in part to to its high growth rate (quick doubling time), which makes it a good candidate for laboratory culture. Hence it has been the primary model to study many biological phenomena e.g. bacterial conjugation, phage genetics, horizontal gene transfer, topography of gene structure, recombinant DNA, and the foundations of biotechnology and bioengineering discoveries resulting in more than ten Nobel prizes.

1.2 Mobile Genetic Elements (MGEs)

In 1950, when McClintock reported the existence of "controlling elements" in maize chromosomes [McClintock, 1950], nobody could guess that this discovery would be a new chapter in the old story of evolutionary research. The significance of her research wasn't understood initially and her work was ignored and rejected [Keirns, 2002]. However, more than thirty years later she was awarded the Nobel prize in Physiology or Medicine for the discovery of mobile genetic elements (MGEs). The term MGE, in general, refers to a wide range of DNA sequences with length from hundreds to a few thousand base pairs that have the ability to move within or between genomes, inserting themselves at other sites in the recipient genome [Craig et al., 2002]. Here, we first highlight the role of MGEs in evolution, and review some of the most important and prevalent types of MGEs. Then, we describe a specific type of MGE, *i.e.* mobile promoters, which are the focus of this thesis.

1.2.1 MGEs in Genome Evolution

MGEs can be considered genomic parasites, since they have no specific function in their host organisms (in the short term), and use host resources to copy themselves into the genome. Because of these characteristics, they are also referred to as "selfish DNA" or "junk DNA". In addition to natural selection, genome defense mechanisms by small RNA, RNA-mediated silencing, have evolved to protect genomes against these parasites [Blumenstiel, 2011]. However, these protective factors could not completely prevent the propagation of MGEs, and MGEs are in fact ubiquitous in nearly all organisms. For instance, MGEs constitute 85% of the maize genome [Schnable et al., 2009] and nearly half[2] of the human genome [Lynch and Walsh, 2007].

MGEs have a great influence on genome architecture [Kazazian, 2004], in partic-

[2]Up to 75% if we consider ancient mobilization events [Lynch and Walsh, 2007].

ular genome size [Touchon and Rocha, 2007]. These elements take advantage of their hosts to pass copies of themselves to future generations, and cause adverse mutations that have deleterious effects on their host [Pasyukova et al., 2004]. Nonetheless, evidence shows there is a mutual relation between MGEs and their host genomes and MGEs can have positive effects for species in the long term [Kazazian, 2004], often being the source of new adaptive characteristics in the organism e.g. antibiotic resistance.

There is overwhelming evidence demonstrating the crucial role of MGEs in the evolution of all organisms. For example in a recent study by Lynch et al. [2015] on the evolutionary origins of pregnancy in mammals, the authors revealed the surprising role of MGEs in the the mammalian transition from egg laying to live birth. They determined that ancient transposable elements (TEs), which are a sub-type of MGEs, are responsible for the emergence of the novel ability of pregnancy in early mammals. These TEs were the origin of the *cis*-regulatory elements that turned off the genes involved in the formation of the egg shell, and turned on other genes, which originally belonged to other organs and tissues, but in the uterus "were recruited to be expressed for new purposes". These functions include maternal-fetal communication, and the development of the maternal uterus immune system to protect the developing fetus [Lynch et al., 2015].

In comparison with other organisms, prokaryotic genomes contain a large fraction of foreign genes which are the result of Horizontal Gene Transfer (HGT) [Koonin et al., 2001]. HGT refers to movement of a DNA segment between organisms (intercellular). MGEs are agents of HGT, and exploring the dynamics of MGEs requires a good understanding of the process of horizontal/lateral gene transfer. This gene exchange between genomes plays a fundamental role in evolution, especially the evolution of bacteria [Ochman et al., 2000].

There are three main mechanisms for HGT: transformation, conjugation and

transduction (infection with bacteriophages). Transformation happens when the bacterium is in a particular state in which it is able to uptake DNA from the surrounding environment [Gyles and Boerlin, 2013]. Indeed, damaged and short DNA molecules exist in almost all environments and may persist for a very long time (more than half a million years) in ideal conditions. These DNA segments are created and quickly breakdown to very small pieces when organic matters are decomposed [Nielsen et al., 2007]. A recent experimental study shows that bacteria can uptake DNA molecules that belong to extinct species from thousands of years ago, through transformation, and insert them into their genomes [Overballe-Petersen et al., 2013]. The authors obtained DNA of a woolly mammoth from its 43,000 year old bone and mixed it with a contemporary bacteria. They suggest that transformation and therefore HGT, can take place with very ancient DNA sequences and may be one of the primary factors in early bacterial evolution.

We will explain the two other mechanisms of HGT, conjugation and transduction, in the next section as we describe the family of MGEs associated with each process.

1.2.2 Different Types of MGEs

MGEs are often categorized based on their features such as sequence characteristics or movement mechanisms [Siefert, 2009]. It is not, however, straightforward to categorize all of these elements disjointly and with no overlap, partly because our knowledge of MGEs is rapidly growing and new categories are introduced constantly. Here, we briefly review three main types of mobile genetics elements: transposable elements, plasmids and bacteriophages.

Transposable Elements (TEs)

Transposable elements are the "jumping genes" that go through the process of transposition described below. They are classified into two main groups, based on their

movement mechanism. The first group, Class I transposons, consists of retrotransposons. These elements transpose by a "copy and paste" mechanism, in which they are first transcribed to RNA and next reverse transcribed to DNA and then inserted into the target host genome. The second group, Class II transposons, consists of insertion sequences and DNA transposons. These elements move directly as a short sequence of DNA in a process called "cut and paste" [Lodish et al., 2000].

Insertion sequences (ISs) are the most prevalent mobile elements in prokaryotic genomes. They also have the simplest and smallest sequences compared to other types of TEs, including only the genes that are necessary for their mobility [Mahillon and Chandler, 1998]. Despite their simplicity, they play a crucial role in genome plasticity [Schneider and Lenski, 2004]. In contrast to ISs, retrotransposons and DNA transposons carry accessory genes beside the genes which are involved in their transposition. These elements are the most common forms of MGEs in eukaryotic genomes; notably numerous retrotransposons exist in plant and mammal genomes. It is worth mentioning that transposons in the maize genome were the first MGEs discovered [McClintock, 1950].

Plasmids

Plasmids are double-stranded DNA molecules that are separated from the chromosomal DNA of the cell and are smaller than it. Although there are some species with linear plasmids [Hinnebusch and Tilly, 1993], they typically have a circular structure. Plasmids are more common in bacteria but they are also found in eukaryotes. They have the ability to self-replicate and their core genes are those that encode replicative functions. They typically don't carry genes that are fundamental for the organism's survival, however they may have some beneficial genes for the organism. Antibiotic resistance is a well known example of a beneficial effect of plasmids for bacteria [Frost et al., 2005].

Plasmids can move to other cells through a process called *conjugation*, mentioned earlier as a HGT mechanism. This process occurs in three steps: creating a connection to the recipient cell through a mating-pair (pilus); signaling if the host environment is tolerable for transferring DNA such that transfer can happen; and finally transferring the plasmid to the host cell [Frost et al., 2005].

Bacteriophages

Bacteriophages or simply phages are defined as viruses that infect bacteria. Their genomes can be single or double stranded DNA or RNA, and either circular or linear. These viruses have played a major role in the development of molecular biology and genomics, in particular they are commonly used in genetic engineering (e.g. nanotechnology [Zhang, 2003]).

Phages are ubiquitous in bacterial populations and may be either virulent or temperate. The former replicates rapidly and results in lysis of the host cell, while the latter leads to lysogeny, in which the phage genome inserts into the host chromosome and then replicates with the host as a prophage [Frost et al., 2005].

1.2.3 Mobile Promoters: A Sub-Class of MGEs

As we mentioned before, a promoter is an essential regulatory element in the gene transcription process. In 2012, Matus-Garcia *et al.* published evidence of promoter mobilization in prokaryotic genomes, identifying "putative mobile promoters" or PMPs for short [Matus-Garcia et al., 2012]. These authors searched the promoter regions of the 1360 available sequenced genomes of prokaryotes and found more than 4000 families of mobile promoters in these regions. Two years later, van Passel et al. [2014] extended this work by searching whole genomes, rather than only the promoter regions, and found three times more copies of mobile promoters, overall.

These discoveries introduced a new aspect to the concept of HGT, *i.e.* evidence of

the transfer of entirely non-coding DNA sequences. As a result, regulatory elements in bacterial genomes can be considered a new class of MGEs, since recently it has been confirmed experimentally that these elements can transfer between genomes by a process called "horizontal regulatory transfer" (HRT) [Oren et al., 2014]. In HRT, the regulatory element is transferred alone, in contrast to HGT in which it is moved with adjacent regulated genes [Koonin, 2014]. These discoveries are remarkable because they are the first investigations to take into account the transfer of non-coding sequences.

1.3 Mathematical Models for MGEs

In the 1960s and 1970s, many mobile DNA sequences were discovered in bacterial genomes. These sequences had similar characteristics to the mobile elements in maize genomes discovered earlier by McClintock [1950]. Finally by 1980, the controversial theory of MGEs became well known and almost universally accepted [Dawkin, 1976, Orgel and Crick, 1980]. Since then, numerous experiments and theoretical works have explored the dynamics of these mobile elements from the evolutionary perspective. The growing number of genome sequencing projects, on the other hand, has resulted in more available genomic data and consequently more identified MGEs in sequenced genomes. This in turn has resulted in more intriguing questions about the origin, fate and impact of these mobile elements in evolution.

Mathematical and statistical models, despite the assumptions necessary for simplification, have contributed greatly to the interesting and helpful information deduced about the evolutionary dynamics of MGEs. While most of these models address MGEs generally, some consider a specific type of MGE and their particular interactions with their host genomes [Brookfield, 1991, Engels et al., 1990, Uyenoyama, 1985].

Transposable elements (TEs) have been the subject of research for a long time and are the most well-studied type of MGE in both eukaryotic and prokaryotic genomes [Rouzic and Deceliere, 2005]. Many mathematical models have been developed to describe their dynamics and explain their evolution by considering various key factors, including duplication, deletion, natural selection, self-regulation and genetic drift. Charlesworth and Charlesworth [1983] proposed one of the very first simulation models for TE dynamics in eukaryotes, in order to discover the factors limiting their spread. They discussed the dependence of fitness on TE copy numbers in *Drosophila* genomes. Many more models have been proposed to determine the distribution of TE copies and the factors controlling their abundance in genomes, and the effect of selection on populations of TEs; for examples see the works of Ohta [1985] and Hudson and Kaplan [1986] on eukaryotic genomes.

1.3.1 Modeling TEs in Prokaryotic Genomes

Modeling TE populations in prokaryotic genomes is different from the modeling process in eukaryotes. Horizontal gene transfer (HGT) is an important agent of TE dynamics in prokaryotes and should be considered in designing mathematical models; however these models are overall simpler since prokaryotes are haploid and asexual. Modeling progress in this field commonly consists of applying branching process and Markov chain approaches.

Markov processes are commonly used in modeling stochastic systems and are applied in a wide range of areas including population modeling and mathematical finance. A system is a Markov process if it has the "Markov property" *i.e.* being memoryless. In other words, in a Markov process, the outcome of the future state only depends on the present state and is independent of all past states. A Markov chain is a discrete time Markov process with a finite state space. A branching process is a particular Markov process mostly used for population modeling, in which each

individual of generation n produces a random number of offspring for the next generation, $n+1$, which is drawn from a probability distribution. This model is often used to study the basic reproductive rate, and the probability of ultimate extinction for the overall population.

A notable example of using branching processes is the work of Sawyer and Hartl [1986], where they proposed six models for TE dynamics in prokaryotes using data from three insertion sequences (*IS4, IS5* and *IS30*) in 71 strains of *E. coli*. They determined the equilibrium distributions of TE copy numbers and estimated the rate of different processes. They however did not consider deletion, because its rate is negligibly small in comparison with the transposition rate [Kleckner, 1981]. In most of the proposed models, Sawyer and Hartl [1986] assumed that TEs reduce the fitness of their hosts, however, they also examined a model in which TEs were assumed to have beneficial effects on their hosts. Similar studies were published later with analogous assumptions to demonstrate the advantages of TEs, and to argue against the theory that TEs are selfish DNA that exist as parasites in prokaryotes' genomes (for example see the study by Condit et al. [1988]). In later work, Hartl and Sawyer continued exploring the dynamics of six unrelated ISs (*IS1, IS2, IS3, IS4, IS5* and *IS30*) in the genomes of 71 strains of *E. coli*. They proposed various models with different assumptions regarding transposition and fitness, and estimated the positive correlation of HGT (mediated by plasmids) with the "presence or absence of different types of IS sequences" [Hartl and Sawyer, 1988, Sawyer et al., 1987].

Branching processes have been utilized in many other studies modeling TEs in prokaryotic populations. Moody [1988] proposed a probabilistic model for TEs in haploid populations in which he considered different factors affecting TE dynamics such as deletion, transposition and the probability of *de novo* acquisition. Moody discussed the relation between deletion and transposition rates and their effects on the stationary state. This work was later extended [Basten and Moody, 1991] by

investigating the impacts of selection on this system.

1.3.2 Modeling Mobile Promoters

Here we describe the standard birth-death model for mobile promoter evolution proposed by van Passel et al. [2014], which is also the basis of our model presented in the next chapter. Van Passel et al. [2014] searched 1360 available prokaryotic genomes and identified mobile promoters in these sequences. They then proposed a model which incorporates four main events that can occur for a copy of a mobile promoter in the genome; these are: duplication, loss, diversification and horizontal gene transfer. Figure 1.1 illustrates their model. Here we consider a family of PMPs with n copies. The parameter u denotes the rate at which each PMP is duplicated in the same genome, and w denotes the rate at which each PMP is lost from the genome. Therefore a family of n copies will create a family of $n+1$ and $n-1$ copies through duplication and loss events, respectively. The parameter v is assumed to be the rate of diversification in which the promoter sequence is changed. Hence due to diversification, a promoter family of n copies loses one promoter and will have $n-1$ copies in the next generation. The newly created promoter through diversification will be considered to be a new family with one PMP which is called a "singleton" family. Singletons are also created via HGT at rate η. Finally it should be noted that each copy of a PMP has an independent chance to undergo any of these four events, so that the overall rates would be nu, nw and nv. However van Passel et al. found that the diversification rate is independent of the number of existing MP copies in the genome, so the overall rate of diversification is η rather than $n\eta$.

In this thesis we propose an extended model which is a generalized location-based model for the evolution of mobile promoters in prokaryotic genomes. We use two datasets from previous work on mobile promoters. Unlike the standard model, our model is a two dimensional model which investigates the dynamics of mobile pro-

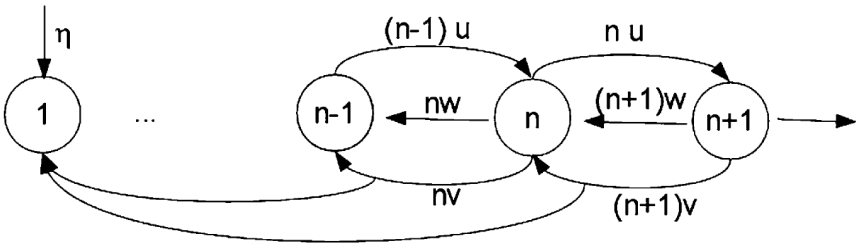

Figure 1.1: The standard 1D model for family of n mobile promoters. Four processes are involved in this dynamics: duplication (u), loss (w), diversification (v), and horizontal gene transfer (η). All processes, except HGT, occur in the rates related to number of the MP copies in genome. Figure reprinted from van Passel et al. [2014].

moters inside and outside promoter regions (PRs), separately. Following a standard birth-death process, in this 2-D model, each new copy of a mobile promoter, which is created by duplication (or received via HGT), inserts either inside the promoter region or at other sites of the genome. Moreover, we consider two different sets of acting rates for inside promoter regions and at other sites of the genome. Hence with this new model we can compare the rates of gene duplication, and loss of mobile promoters outside and inside promoter regions. We describe this model in more detail in the next chapter.

1.4 Computational Model Fitting

The next step after expressing a biological system in terms of parameters in mathematical expressions, *i.e.* the model formation process, is applying model selection techniques and statistical inferences to deduce the properties of the underlying biological system and to estimate the parameters of the model [Burnham and Anderson, 2002]. Both the standard birth-death model of van Passel et al. [2014], and also our location-based model, which we present in the next chapter, are defined based on a set of parameters, *i.e.* the rates of different processes; these rates should be derived from available genomic data. This is achieved by *fitting these models to the observed data*. More specifically, we search for the parameter values at which the observed

data is most likely to be generated by the model, or is most similar to simulated data produced by the model in its stationary state. This can be formulated as a classic optimization problem; where in the former we are using a log-likelihood loss function, and in the latter we minimize a Sum of Squared Errors (SSE).

Although optimization is a classic and well-studied computational problem, it is in most cases an NP-hard problem, and computationally expensive, especially when dealing with functions describing real world phenomena, which are non-convex and may have many local optima. This is also the case in our model. In these cases, common local minimization routines such as the Nelder-Mead simplex algorithm perform poorly. Instead, one should use a global optimization approach; these are typically Monte-Carlo-based stochastic methods, such as simulated annealing. Here, we applied the "Basin Hopping" global optimization approach which tries to find the best possible global minimum of a function. Rooted in physics, basin hopping is an iterative stochastic process built upon a local minimizer, which accepts or rejects the parameters in each iteration.

1.4.1 Model Selection and Accuracy

"all models are wrong, but some are useful."

Box 1976

It is far from realistic to expect to find the full and exact truth regarding a complicated biological system from a mathematical model with its many idealizations and simplifications. Although a model cannot reveal the complete reality of a system, it can provide useful insights into the underlying process, if it is well-defined and selected properly. Model selection is the process of choosing the best model from a set of candidate models, in an attempt to answer this simple question: what is the best model to use to describe a system?

In the past two decades, many techniques have been developed for model selection,

including stepwise procedures such as backward elimination and forward selection to find the right number of variables, cross-validation techniques to avoid overfitting, and criterion-based procedures which are defined based on a tradeoff between the simplicity and accuracy of the model [Burnham and Anderson, 2002]. Two well-known criteria used in the latter class are the Akaike information criterion (AIC) and the the Bayesian information criterion (BIC), which are used in this thesis.

More formally, let k denote the number of parameters of the model, and n the sample size, or the number of observations, then the *Akaike information criterion* (AIC) is defined as:

$$AIC = 2k - 2ln(L)$$

where L is the likelihood of the data. The *Bayes Factor and Bayesian information criterion* (BIC) is closely related to the AIC and is defined as:

$$BIC = kln(n) - 2ln(L)$$

In order to find the most qualified model by Akaike and Bayesian information criteria, one should search the set of candidate models to find the model with the lowest AIC or BIC. In fact, these criteria provide the possibility of comparing different models by making a trade off between the complexity of the model and the information lost, since to reach the minimum AIC value, the number of parameters, k, should reduced while the likelihood, L, should be increased. Finally it should be noted that the AIC and BIC cannot confirm the validity of the general model. In other words, by applying these criteria we do not figure out whether our general model is defined appropriately. The two following variations, AIC and BIC corrected, are also proposed for when the sample size is relatively small, where n denotes the sample size.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

$$BICc = BIC - kln(2\pi)$$

Applying AICc prevents overfitting which can happen when using the AIC if the sample size, n, is not larger than k^2.

1.4.2 Relative Probability

Although the model with the lowest AIC (or BIC) is considered to be the "best" model to fit the data, it is possible that several candidate models have relatively similar AIC values. In this situation, the relative probability can be used to determine the statistical significance of the selected model. The relative probability, R is computed as $R = \exp\left((AIC_{min} - AIC)/2\right)$.

The model with the lowest AIC always has relative probability 1.0, however there could be several candidate models that cannot be rejected, if their relative probabilities are also high (for example, higher than 0.05). This situation did not occur in the model selection process described in the following chapter. However techniques such as Bayesian model averaging [Hoeting et al., 1999] can be used when several candidate models cannot be rejected based on relative probability.

1.5 Thesis Statement and Contribution

Here, we extend the previous model of the birth, death and diversification of mobile promoters [van Passel et al., 2014], and propose a novel location-based extension. The new model incorporates two biologically meaningful parts of the genome: *i)* Inside promoter regions and *ii)* other sites of the genome. The model considers four key factors in genome alteration: *i)* duplication, *ii)* loss, *iii)* diversification, *iv)* horizontal

gene transfer (HGT). The research question of this thesis is to determine whether the rates of gene duplication and loss of PMPs have meaningful differences outside and inside promoter regions. If yes, we would like to shed light on the biological reasons underlying these differences.

In the next section, we present our general model in detail, and present and discuss the results and findings. Our main finding can be summarized as "Mobile promoters are much more stable in promoter regions of the genome than in other regions."

Bibliography

Ahnert, S. E., Fink, T. M., and Zinovyev, A. How much non-coding DNA do eukaryotes require? *Journal of Theoretical Biology*, 252(4):587–592, 2008.

Basten, C. J. and Moody, M. E. A branching-process model for the evolution of transposable elements incorporating selection. *Journal of mathematical biology*, 29(8):743–761, 1991.

Blumenstiel, J. P. Evolutionary dynamics of transposable elements in a small RNA world. *Trends in genetics : TIG*, 27(1):23–31, 1 2011. doi: 10.1016/j.tig.2010.10.003.

Box, G. E. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Brookfield, J. Models of repression of transposition in PM hybrid dysgenesis by P cytotype and by zygotically encoded repressor proteins. *Genetics*, 128(2):471–486, 1991.

Burnham, K. P. and Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

Charlesworth, B. and Charlesworth, D. The population dynamics of transposable elements. *Genetical Research*, 42(01):1–27, 1983.

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

Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M. *Mobile DNA ii*. 2002.

Dawkin, R. The selfish gene. *Oxford university Press*, 1:976, 1976.

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

Ferdows, M. S. and Barbour, A. G. Megabase-sized linear DNA in the bacterium Borrelia burgdorferi, the Lyme disease agent. *Proceedings of the National Academy of Sciences*, 86(15):5969–5973, 1989.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

Gyles, C. and Boerlin, P. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology Online*, page 0300985813511131, 2013.

Hartl, D. and Sawyer, S. A. Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*? *Genetics*, 118(3):537–541, 1988.

Hinnebusch, J. and Tilly, K. Linear plasmids and chromosomes in bacteria. *Molecular microbiology*, 10(5):917–922, 1993.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

Hudson, R. R. and Kaplan, N. L. On the divergence of members of a transposable element family. *Journal of mathematical biology*, 24(2):207–215, 1986.

Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science*, 303(5664): 1626–1632, 2004.

Keirns, C. Demythologizing McClintock, 2002.

Kleckner, N. Transposable elements in prokaryotes. *Annual review of genetics*, 15(1): 341–404, 1981.

Koonin, E. V. Horizontal transfer beyond genes. *Proceedings of the National Academy of Sciences*, 111(45):15865–15866, 2014.

Koonin, E. V., Makarova, K. S., and Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annual Reviews in Microbiology*, 55(1): 709–742, 2001.

Lodish, H. F., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., Darnell, J., et al. *Molecular cell biology*, volume 4. Citeseer, 2000.

Lynch, M. and Walsh, B. *The origins of genome architecture*, volume 98. Sinauer Associates Sunderland, 2007.

Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., Emera, D., Sheikh, S. Z., Grützner, F., Bauersachs, S., et al. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy, 2015.

Mahillon, J. and Chandler, M. Insertion sequences. *Microbiology and Molecular Biology Reviews*, 62(3):725–774, 1998.

Matus-Garcia, M., Nijveen, H., and van Passel, M. W. Promoter propagation in prokaryotes. *Nucleic acids research*, pages 40(20):10032–40, 2012.

McClintock, B. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, 1950.

Moody, M. E. A branching process model for the evolution of transposable elements. *Journal of mathematical biology*, 26(3):347–357, 1988.

Nielsen, K. M., Johnsen, P. J., Bensasson, D., and Daffonchio, D. Release and persistence of extracellular DNA in the environment. *Environmental biosafety research*, 6(1-2):37–53, 2007.

Ochman, H., Lawrence, J. G., and Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

Ohta, T. A model of duplicative transposition and gene conversion for repetitive DNA families. *Genetics*, 110(3):513–524, 1985.

Oren, Y., Smith, M. B., Johns, N. I., Zeevi, M. K., Biran, D., Ron, E. Z., Corander, J., Wang, H. H., Alm, E. J., and Pupko, T. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences*, 111(45):16112–16117, 2014.

Orgel, L. E. and Crick, F. H. Selfish DNA: the ultimate parasite. *Nature*, 284:604–607, 1980.

Overballe-Petersen, S., Harms, K., Orlando, L. A., Mayar, J. V. M., Rasmussen, S., Dahl, T. W., Rosing, M. T., Poole, A. M., Sicheritz-Ponten, T., Brunak, S., et al. Bacterial natural transformation by highly fragmented and damaged DNA. *Proceedings of the National Academy of Sciences*, 110(49):19860–19865, 2013.

Pasyukova, E., Nuzhdin, S., Morozova, T., and Mackay, T. Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. *Journal of Heredity*, 95(4):284–290, 2004.

Rouzic, A. L. and Deceliere, G. Models of the population genetics of transposable elements. *Genetical research*, 85(03):171–181, 2005.

Sawyer, S. and Hartl, D. Distribution of transposable elements in prokaryotes. *Theoretical population biology*, 30(1):1–16, 1986.

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

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. The B73 maize genome: complexity, diversity, and dynamics. *science*, 326(5956):1112–1115, 2009.

Schneider, D. and Lenski, R. E. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in microbiology*, 155(5):319–327, 2004.

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

Touchon, M. and Rocha, E. P. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular biology and evolution*, 24(4):969–981, 2007.

Trucksis, M., Michalski, J., Deng, Y. K., and Kaper, J. B. The Vibrio cholerae genome contains two unique circular chromosomes. *Proceedings of the National Academy of Sciences*, 95(24):14464–14469, 1998.

Uyenoyama, M. K. Quantitative models of hybrid dysgenesis: rapid evolution under transposition, extrachromosomal inheritance, and fertility selection. *Theoretical population biology*, 27(2):176–201, 1985.

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

Zhang, S. Fabrication of novel biomaterials through molecular self-assembly. *Nature biotechnology*, 21(10):1171–1178, 2003.

Chapter 2

A Location-Based Birth, Death and Diversification Model for Mobile Promoters in Prokaryotes

Mobile genetic elements (MGEs) are DNA sequences with the ability to move to new sites in genomes (within and between them). These genetic elements are ubiquitous in nearly all creatures despite the fact that natural selection and genome defense systems restrict their propagation to protect the genome [Blumenstiel, 2011]. Although MGEs are considered to be genomic parasites with detrimental effects on their hosts, evidence shows they may have long term benefits for the species [Kazazian, 2004]; antibiotic resistance in bacteria is one famous example [Boutoille et al., 2004]. MGEs indeed have a significant importance from an evolutionary point of view, particularly in prokaryotes [Ochman et al., 2000].

MGEs vary from relatively short sequences that only contain the genes required for their mobility, e.g. insertion sequences (ISs), to longer sequences that carry many accessory genes. These mobile elements are considered to be a key source of alteration in genome architecture [Kazazian, 2004], especially in genome plasticity. They

underlie the rapid dynamics of evolution in prokaryotic genomes [Frost et al., 2005], and are important features in transcriptional rewiring [Perez and Groisman, 2009], in which for example regulatory elements activate silent genes [Stoebel and Dorman, 2010]. Mobile promoters (MPs) are regulatory elements that are a sub-class of MGEs [Matus-Garcia et al., 2012, van Passel et al., 2014] and are considered to be one of the possible underlying factors in transcriptional rewiring [Nijveen et al., 2012].

A promoter is a region of DNA that marks the start of the transcription process. It has been recently discovered that these regulatory elements are mobile. Evidence of promoter mobility was first proposed by Matus-Garcia et al. [2012], following a search of 1360 sequenced prokaryote genomes. Matus-Garcia *et al.* searched promoter regions (PRs) of these genomes, as explained in greater detail in the following section, to identify putative mobile promoters (PMPs), which they defined to be "homologous promoter sequences with non-homologous coding sequences". Their dataset included 13,111 copies of PMPs overall. Van Passel et al. [2014] extended this work to search full genomes and expanded the previous data set to nearly 40,000 PMPs, providing strong evidence of promoter mobility. These two studies have two remarkable aspects. First, they introduce MPs as a new class of MGEs and produce datasets describing the distribution of MPs in prokaryotic genomes. Second, they are the first studies which present a new idea regarding horizontal gene transfer (HGT). More specifically, they provide evidence of the transfer of non-coding DNA sequences in isolation. Two years later, Oren et al. [2014] confirmed that regulatory elements can transfer without adjacent genes, a process that the authors called "horizontal regulatory transfer (HRT)".

As more MGEs have been discovered, more curiosity about the origins and fate of these elements has arisen. Questions regarding population dynamics and fitness effects, and the impact of factors such as horizontal transfer and drift have naturally emerged. Therefore much effort has been put into modeling these processes mathe-

matically. Approaches range from numerical models and simulations for the TE population in eukaryotic genomes [Rouzic and Deceliere, 2005] to more simple modeling approaches for ISs in prokaryotic genomes, such as branching processes and Markov chain models [Sawyer and Hartl, 1986]. In particular, van Passel et al. [2014] focused on mobile promoters and presented a birth-death-diversification model, in which they proposed a model for the distribution of MPs both within and among genomes. The extinction probability of MP lineages has also been determined [Drakos and Wahl, 2015].

In this article, we investigate the differences in the dynamics of MPs between promoter regions and other sites of the genome. In order to do that, we build a new two dimensional model based on the previously proposed model for the dynamics of MPs [van Passel et al., 2014]. The previous model is notable not only because it is the first proposed model for MPs, but it is also the first model for MGEs that considers the effect of genetic diversification. Our new location-dependent model is built to include four key parameters in these dynamics: duplication, deletion and diversification of MP copies, in addition to the acquisition of promoter copies via HGT. We then apply our model to data describing MPs in 1360 sequenced prokaryotic genomes. Our expectation was that MPs would exhibit greater stability in promoter regions, where their dynamics may have a crucial impact on the survival of the organism.

2.1 Methods

In this section, we introduce our location-based model for the dynamics of mobile promoters (MPs), which we call the promoter general model or *PGM* for short. This new model considers distinct rates for the main dynamic processes for two different parts of the genome, *i.e.* inside and outside of promoter regions. This model, when fitted to real data as presented in the next section, can be used to shed light on how the dynamics of MPs and the rates of these processes differ inside and outside promoter regions.

2.1.1 Promoter General Model (PGM)

The PGM is a two dimensional model which describes the distribution of mobile promoters in prokaryotic genomes. As stated previously the PGM is defined based on the main processes involved in the maintenance of mobile promoters, *i.e.* the rates of duplication, loss, diversification and HGT. The PGM considers different rates for these factors in two distinct regions of genome, and hence is able to reveal differences in the dynamics in these two regions. Since the PGM does not include any assumption specific to promoters, one may note that it can be applied to mobile genetic elements in prokaryotes in general, in order to compare any two distinct parts of the genome. Here, we consider each genome to consist of regions inside and outside of promoter regions; we then find the stationary state of the model and fit this simulated (predicted) data to the observed data for MPs, obtained from the previous studies on the available sequenced genomes, which is described later in detail.

Figure 2.1 illustrates the overall view of the PGM. It includes all the possible events that can happen for a family with n_1 promoters inside and n_2 promoters outside of promoter regions. This family of (n_1,n_2) copies can create a family of ($n_1 + i$,$n_2 + j$) in the next generation, where i and j could be 0 or 1.

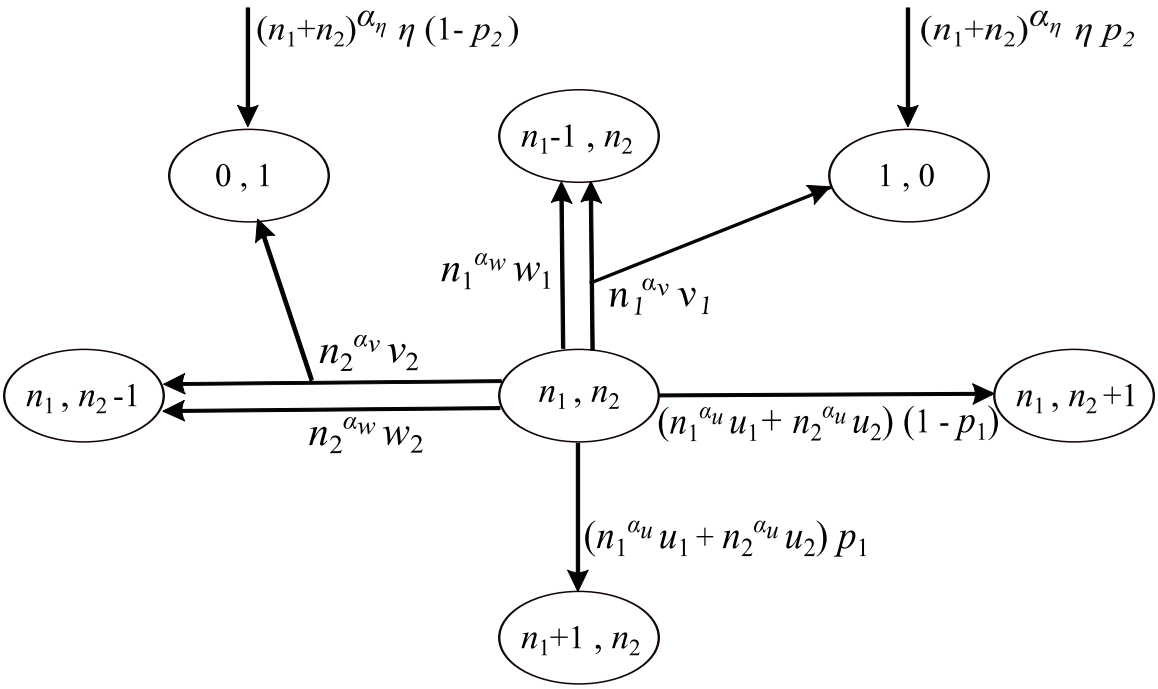

Figure 2.1: The PGM; 2D model for a family of (n_1, n_2) MPs, where n_1 and n_2 denote the number of MP copies inside and outside of promoter regions (PRs), respectively. Four processes are involved in these dynamics: duplication (u), loss (w), diversification (v), and horizontal gene transfer (η). These rates are differentiated by subscript 1 for inside and 2 for outside of PRs. A newly created promoter by duplication is inserted into PRs with the probability p_1, and to outside PRs with probability of $1 - p_1$. For instance, a duplication results in a transition from (n_1, n_2) to $(n_1 + 1, n_2)$ with the rate of $(n_1u_1 + n_2u_2)p_1$. Similarly, a promoter that is created via HGT is inserted in PRs with probability p_2 and to outside PRs with probability $1 - p_2$. The α-values are assigned to investigate the relations between the number of MP copies and the rate of each process, where if α is estimated to be 1, it confirms that the number of copies affects these rates, and otherwise $\alpha = 0$ and there is no effect.

We consider 4 main situations that can occur for a copy of a MP in the genome: duplication, loss, diversification and horizontal gene transfer (HGT). We take u to denote the rate at which each MP is duplicated in the same genome, noting that the mechanism of MP duplication is not yet completely understood. We take w as the rate at which each MP is lost from a genome, this could reflect either excision or loss of mobility through, for example a deletion. The parameter v represents the diversification rate. Diversification occurs when the sequence of the MP changes

sufficiently such that it would no longer be classified as the same family (less than 80% sequence identity). When diversification occurs, a family with (n_1,n_2) copies will become a family of ($n_1 - 1$,n_2) copies, for example, and create a new family with either (1,0) or (0,1) copies. The parameter η denotes the rate of HGT. We assume that the recipient genome does not already have a copy of the transferred MP, so that HGT always creates a new family with (1,0) or (0,1) copies. It should be noted that HGT preserves the original and makes a new MP copy.

Following a standard birth-death process, each MP copy has an independent chance for each process to occur, so the overall rates for a family of n copies would be nu, nw, nv and $n\eta$. We then go one step further with generalizing this by assuming n^α rather than n, where α is 0 or 1 [Bichsel et al., 2013]. Accordingly, the overall rates would be $n^{\alpha_u}u$, $n^{\alpha_w}w$, $n^{\alpha_v}v$, $n^{\alpha_\eta}\eta$. For instance, duplication would occur for a family of n MPs at rate u when α_u is 0 or at rate nu when α_u is 1.

The previous one dimensional model for MPs proposed in van Passel et al. [2014], also considers the four rates of u, v, w, and η for respectively duplication, loss, diversification, and horizontal gene transfer events. We divide each of these rates into 2 distinct parameters, *i.e.* first one for the rate inside promoter regions (PRs), distinguished by subscript 1, and the second one for the rate outside of the PRs, denoted by subscript 2. Thus the PGM includes u_1, w_1, v_1 for the rates of events in PRs and u_2, w_2, v_2 for the rates outside of PRs. Again as an example, duplication happens at rate n_1u_1 in PRs and at rate n_2u_2 outside of these regions. For simplicity, we did not assign two different rates to η.

Moreover, we include the probabilities p_1 and p_2, which consider the chance that a new copy of a MP, which is created through duplication or received via HGT, is inserted inside (or outside) the promoter region. In other words, a new MP created through duplication is inserted inside the PRs with probability p_1, and outside of the PRs with probability $1 - p_1$; the same follows for p_2 and a new MP created via HGT.

As mentioned above, single copies are created when genomes receive a new promoter via HGT, or when existing promoters diversify to a different sequence and are consequently considered to be a new family of promoter. Van Passel et al. named these "singletons", and reported a large number of MP families that only include a single copy of the MP in prokaryotic genomes. Here, as also shown in Figure 2.1, these singletons are denoted by either of these two states: (1,0) if the copy is located inside the PRs, or (0,1) if the copy is located outside of PRs.

2.1.2 Dataset Description

Each element of our dataset gives the observed number of MP families with (n_1,n_2) copies, where n_1 and n_2 denote the number of MPs inside and outside of promoter regions, respectively. To construct this dataset we benefit from two consecutive studies that have published datasets of MP families in promoter regions and the entire genome. In the first study Matus-Garcia et al. [2012] extracted 100 nucleotide segments, 150-50 nucleotides upstream from the translation start site, from all coding sequences in 1362 available sequenced prokaryotic genomes. They then searched these regions (which we will refer to as promoter regions, PR) for homologous sequences, in order to identify mobile regulatory elements. In their conservative search, they considered homologous sequences to be a MP family, if "they share 80% identity in at least 50 nucleotides and also have non-homologous up- and down-stream sequences". 13,111 MPs were found in 1043 genomes through intragenomic and intergenomic searching. The second study extended the previous work by searching entire genomes to identify sequences homologous to MPs that were identified previously. They found 3 times more MPs, 39,441 copies, of which 90% were located in noncoding regions of the genome [van Passel et al., 2014]. For each of the 4047 MP families identified in these studies, we use the dataset of Matus-Garcia's work for the number of promoters

in PRs, n_1, and the number of promoters outside of PRs, n_2, is calculated with a simple subtraction of n_1 from the observed counts per genome in the second study. It should be mentioned that the first study dataset does not include families with a single MP copy. As a result we assume $n_1 > 2$ in our model selection and data fitting process, however we do not make this assumption for n_2. We also note that the two studies used almost identical, but not identical, bioinformatics techniques. This resulted in small inconsistencies in the data, but affected less than 1.45 % of the data points in our study.

2.1.3 Model Selection and Data Fitting

In the model selection process, we fit the equilibrium distribution of the model to the observed data. Note that this distribution depends on the ratios of the rates, not each rate in isolation. Therefore data fitting was conducted based on u_2/u_1, w_1/u_1, w_2/u_1, v_1/u_1 and v_2/u_1 denoting different rates for duplication, deletion and diversification in the two distinct parts of the genome, and η/u_1 representing the rate of HGT. Together with the probabilities p_1 and p_2, that are constrained to be on the interval [0,1], the PGM has 8 parameters that can freely vary.

Furthermore the PGM also includes four exponents α_u, α_w, α_v and α_η that can be either 0 or 1 and make each process linear or constant. It is unclear at the outset whether each of these processes (parameters) is necessary to explain the observed data. For example, a model that includes duplication and deletion, but does not include diversification, may be sufficient to capture the data. We therefore consider possible subsets of the above parameters, creating nested models of the PGM, as our candidate models.

Given a set of parameter values, the PGM (and its nested/derived models) typically converge to a stationary state, which is the data generated by the model, denoted PGM_p. To fit this model to the observed data, then, we find the parameter values for

which the PGM_p is closest to the observed data, *i.e.* a classic optimization problem. To solve this minimization problem we use the basin-hopping algorithm (with the L-BFGS-B method for the local minimization function) from the SciPy [Jones et al., 2001] package in the Python programming language.

Our model selection technique is to simply choose the model with the maximum quality *i.e.* the simplest model with the lowest parameter numbers that could describe the observed data, whereas our full PGM is the most complex model. We measure the relative quality of a model compared to other models, using the Akaike information criterion (AIC) which is defined as: $AIC = 2k - 2ln(L)$ where k denotes the number of parameters of the model and L is the likelihood of the data. Our candidate models all originate from the general model, PGM, and each one has a subset of parameters of the general model. We also enumerate all possible combination of α-values and try each nested model with all of these combinations.

In generating candidate models, we assume duplication and deletion to be crucial processes in MP dynamics. The main reason for this assumption is the fact that models without these two parameters do not typically converge to a steady state distribution. Moreover, these two processes were previously found to be the most important parameters in the 1D MP model, and occur at a higher rate than any other parameters [van Passel et al., 2014]. Therefore we eliminate/prune all the nested models without at least one u and one w, which results in a reduction in the total number of candidate models. In total, we consider roughly 8700 nested models within the PGM.

2.1.4 Sensitivity

After having determined the best model to describe the MP distribution and consequently exploring the most influential parameters in these dynamics, we want to examine the robustness of our proposed model. In more detail, we evaluate the sensi-

tivity of the model's parameters, through a bootstrapping procedure involving 1000 random samples. Each sample contains 90% of the original data. In other words, for each sample we draw 90% of the available 4047 MP families, and repeat the data fitting step for that sample given our selected PGM model, to estimate the parameter values for that sample. The purpose of this experiment is to see how the parameters vary on different samples, to confirm that we have not over-fitted our model, and to see which process is more robust in the MP dynamics by estimating the parameters' pattern of variation. We present the results of our sensitivity analysis in two figures, showing variation in the actual values of the parameters and also variation in the values when normalized by their median.

2.2 Results

2.2.1 Model Selection and Data Fitting

As explained in the previous section, we assumed our candidate models to be all the nested models of the PGM with at least two to eight free parameters, *i.e.* only eliminating those models without at least one parameter for duplication and one for deletion. However, we did investigate the majority of the nested models, enumerating all the candidate models, and hence we can claim that our model identification process is deterministic [Burnham and Anderson, 2002]. It should however be noted that our models did not take natural selection into account, based on the earlier study by van Passel et al. [2014] which did not find any evidence for selection in MP dynamics.

Figure 2.2 illustrates the AIC values for all candidate models, grouped by the number of free parameters. As shown in the zoomed figure, the minimum AIC value belongs to a sub-model with 6 free parameters. We can take this 6-parameter model to be the single best model due to a large gap between its AIC value and the AIC values of competing models, *i.e.* 10.82 with the next best model (also a 6-parameter model)

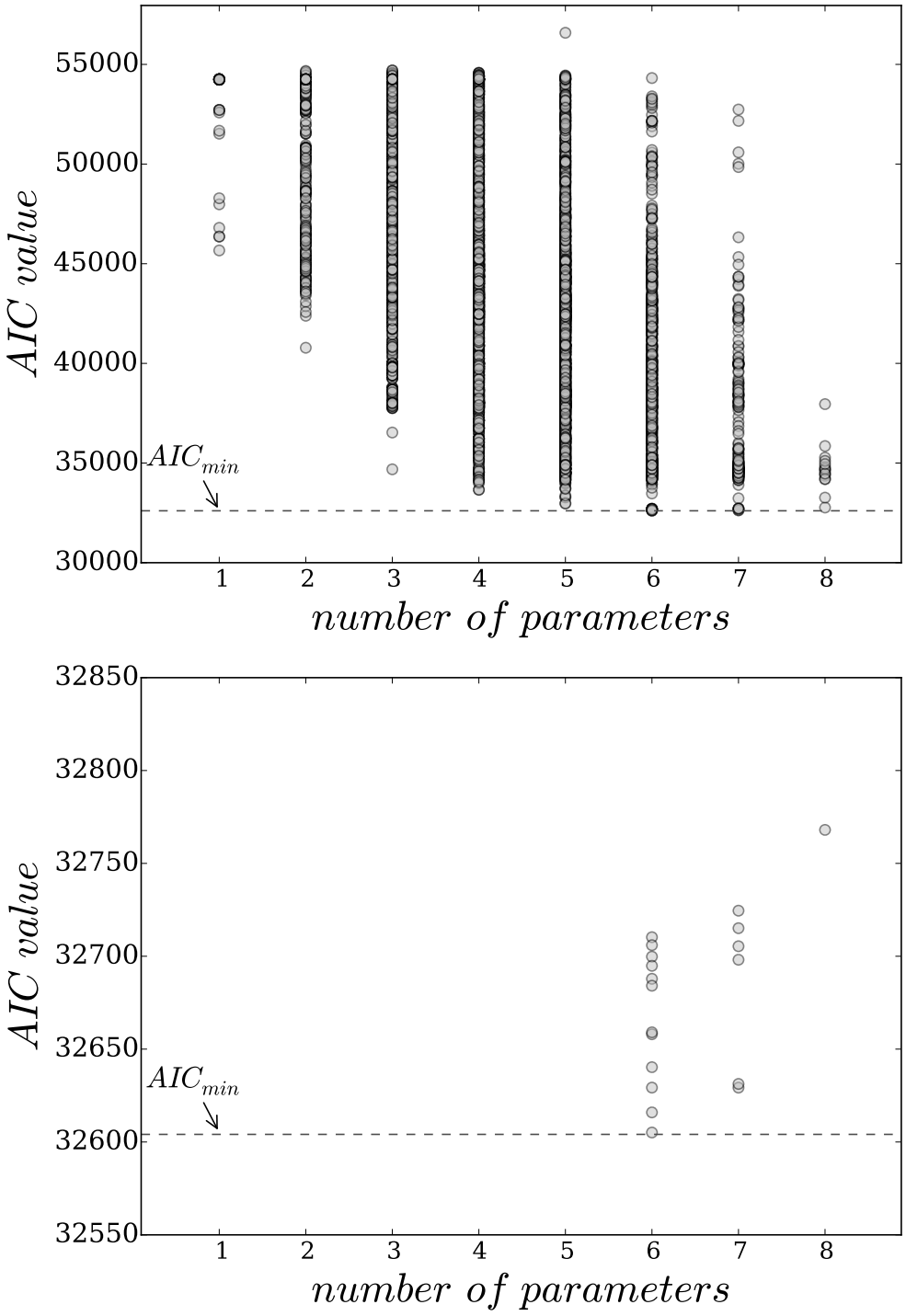

Figure 2.2: Quality of different models plotted as a function of the number of parameters in that model. The best model is the model with lowest AIC value which is a model with 6 free parameters, shown better in the zoomed figure at the bottom. This plot and the ones presented afterward are generated using the Matplotlib package [Hunter, 2007].

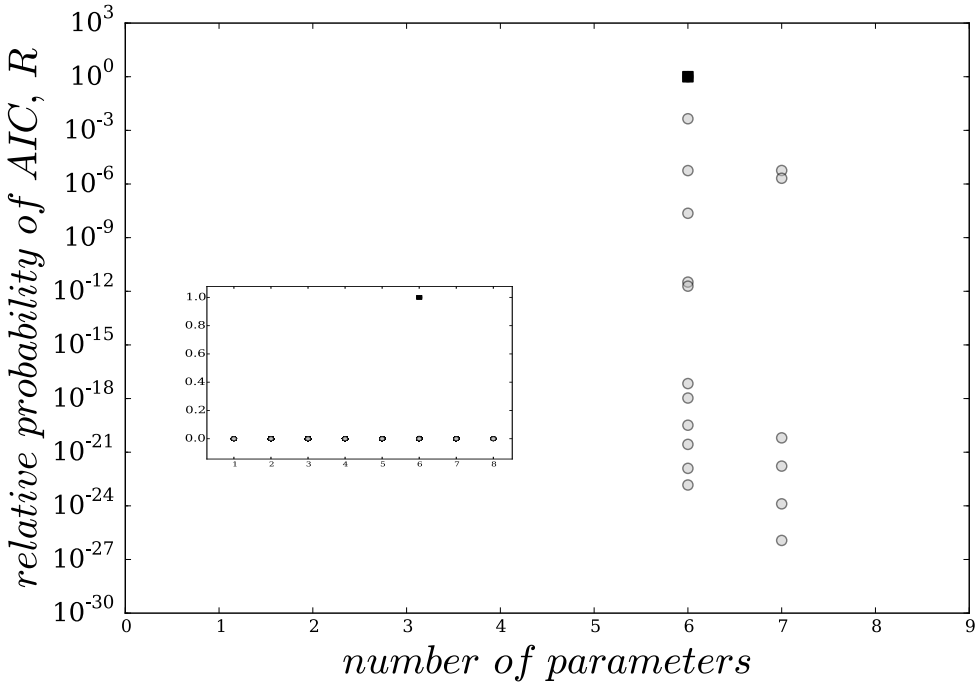

Figure 2.3: The relative probability of candidate models, plotted as a function of the number of parameters in that model. The relative probability is computed as $R = \exp((AIC_{min} - AIC)/2)$. Only models with $R > 10^{-30}$ are plotted. The square corresponds to the model with the minimum AIC. The inset represents the probabilities on a linear y-axis.

and 24.15 with the third best model (a 7-parameter model). We obtained the same result when using other criteria for model selection, namely Bayesian Information Criteria (BIC), and their corrected sample-size versions (*i.e.* AICc and BICc).

We can confirm that the selected model is significantly better than other candidate models also by comparing their relative probabilities, plotted in Figure 2.3.

Figure 2.4 shows the selected model and its parameters; the actual values of the parameters for this best model are summarized in the first row of Table 2.1.

During the model selection process, we considered all possible sets for α-values and examined each candidate model with all of these 16 options. The best data fitting was obtained when: $\alpha_u = 1$, $\alpha_w = 1$, $\alpha_v = 1$ and $\alpha_\eta = 0$. These exponential values reveal the dependency of duplication, loss and diversification rates on the number of

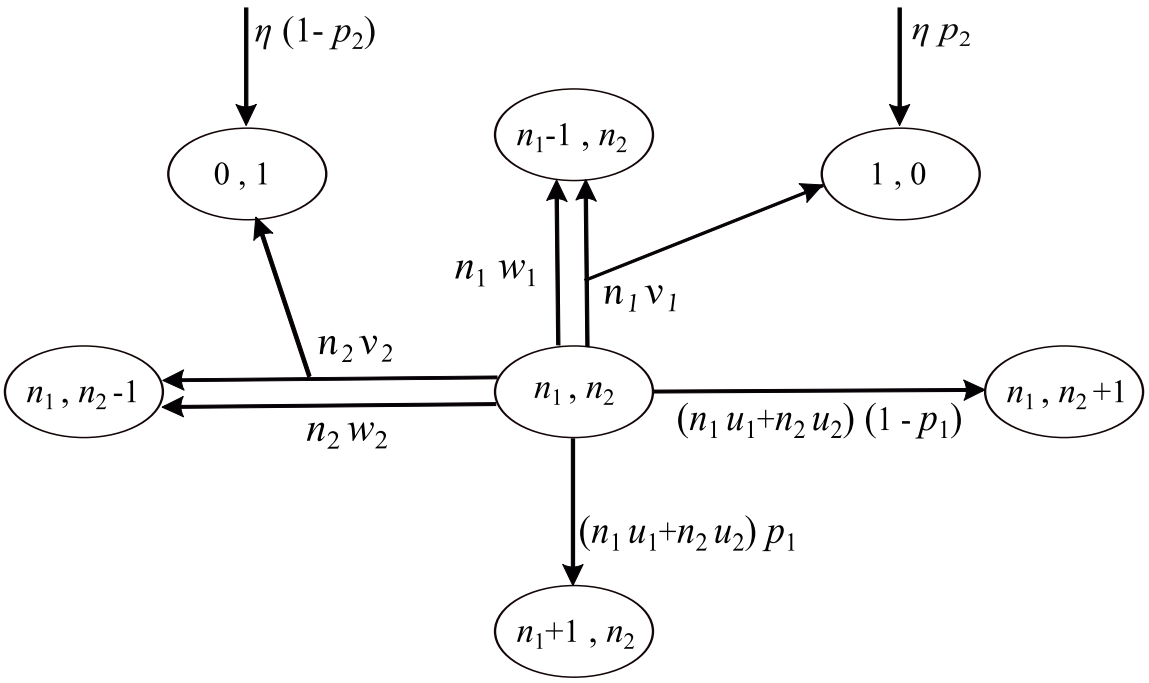

Figure 2.4: Here we can see the selected/simplified model with the actual values estimated for the α parameters, *i.e.* $\alpha_u = \alpha_v = \alpha_w = 1$ and $\alpha_\eta = 0$.

promoters in the MP family. In contrast, the exponent associated with HGT, α_η, implies that promoter families undergo HGT at a constant rate, independent of the number of copies in a family.

Examining the best model, we are able to determine the most important processes in MP dynamics. These parameters and their values are shown in Table 2.1. The values indicate a high rate for u_2/u_1 which implies that duplication occurs at a significantly higher rate outside of PRs. Moreover, the probability p_1 implies that these new promoters from duplication are mostly (90%) inserted outside of PRs. Deletion and diversification in PRs, w_1 and v_1 can be ignored as the best model does not include these parameters, implying their rates are negligible. In contrast, deletion and diversification events outside of PRs, w_2 and v_2, are important although their rates are lower than duplication and HGT. Finally HGT, η, happens at the highest rate, and roughly all the new promoters received via HGT are inserted in PRs. This may

explain the maintenance of MPs in PRs although the duplication rate is low in this regions.

parameters	u_2/u_1	w_2/u_1	v_2/u_1	η/u_1	p_1	p_2	AIC
best fit values	18.58	2.55	2.42	21.90	0.13	1.0	32605
sensitivity							
median	18.40	0.97	1.90	22.21	0.14	1.0	29373
mean	16.86	1.67	2.16	20.39	0.21	0.98	30828
min	1.09	0.00	0.68	1.79	0.04	0.83	29031
max	37.22	5.52	13.84	79.00	0.77	1.0	41489
s.d.	6.36	1.55	1.34	7.84	0.20	0.05	3889
cov	0.38	0.93	0.62	0.38	0.94	0.05	0.13

Table 2.1: Best fit values and sensitivity of the best model's parameters.

In Figure 2.5 we see the result of the best model fitted to the observed data, and in Figure 2.6 we show the comparison of fits for fixed values of either n_1 or n_2 (cross-sections). Thus we can confirm by inspection that the selected model has good agreement with the observed data.

2.2.2 Sensitivity Analysis

We performed the bootstrapping technique described above for the selected model to assess the sensitivity of its six parameters. Figure 2.7 shows the result of our sensitivity analysis for the parameters of the best model. These results are also presented in Table 2.1 where in most cases the mean and median of the bootstrapping result compares well with the best fit parameter value. The model seems to be robust to random sampling for all parameters, with the possible exception of w_2, the rate at which MP copies are lost from outside promoter regions. Note that the coefficient of variation (cov, *i.e.* s.d./mean) of w_2 is 0.93, indicating a wide variation in this parameter among bootstrap samples. We also note that, during the model selection process, we observed a high rate for u_2, the duplication rate outside PRs, for almost

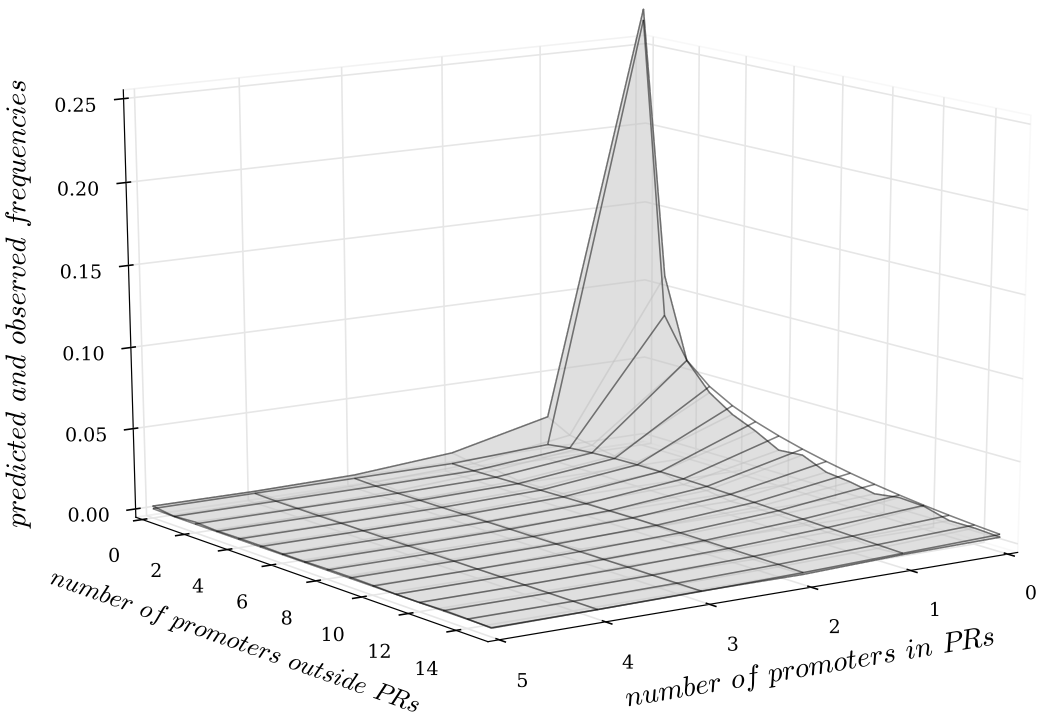

Figure 2.5: Result of best model fitted to the observed data. The shaded surface represents the distribution of observed/real data, while the predicted distribution is plotted on top as a wireframe. The exact parameter values resulting in this fit are reported in the first row of Table 2.1.

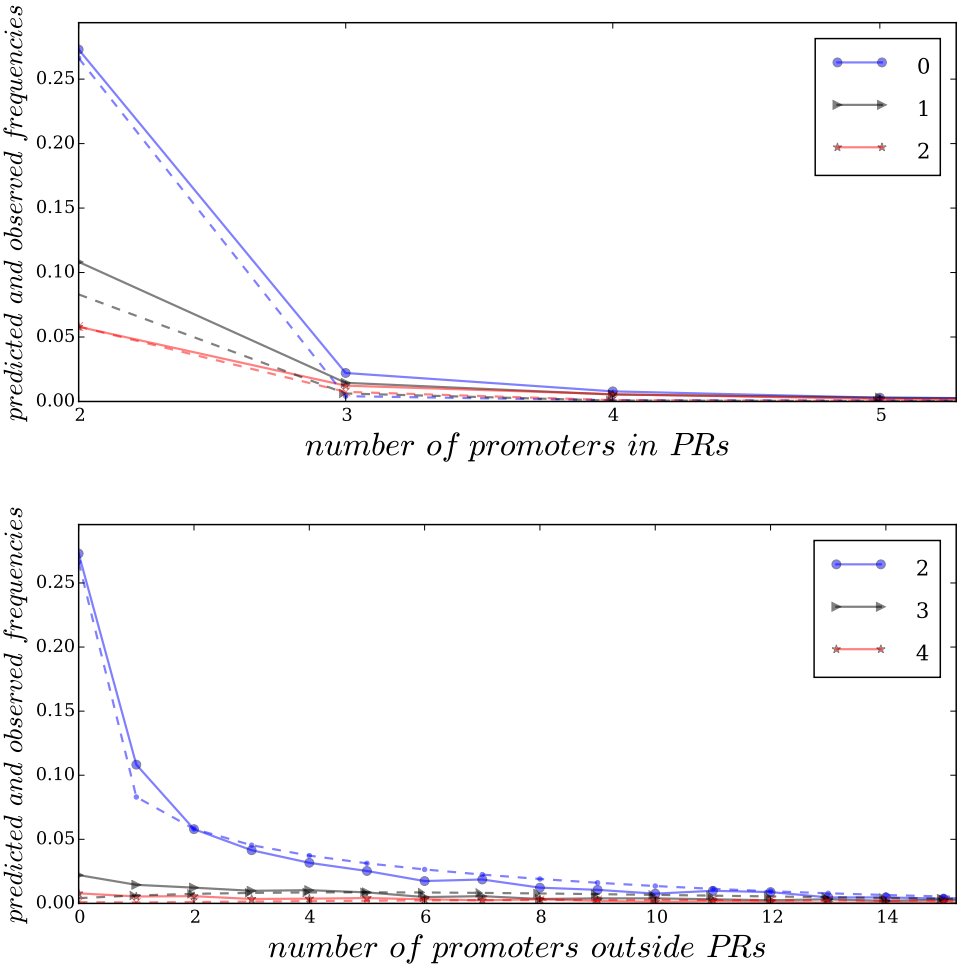

Figure 2.6: Cross-section of the 3D fit presented in Figure 2.5. The observed distribution is marked with solid line, while the dashed line represents the distribution predicted by the model.

all the nested models that had an acceptable fit. Thus our conclusion that u_2 is high relative to u_1 is robust to model selection and bootstrapping. On the other hand, although the cov of probability p_1 is high, this is largely due to the effect of outlier as illustrated in the boxplots of Figure 2.7.

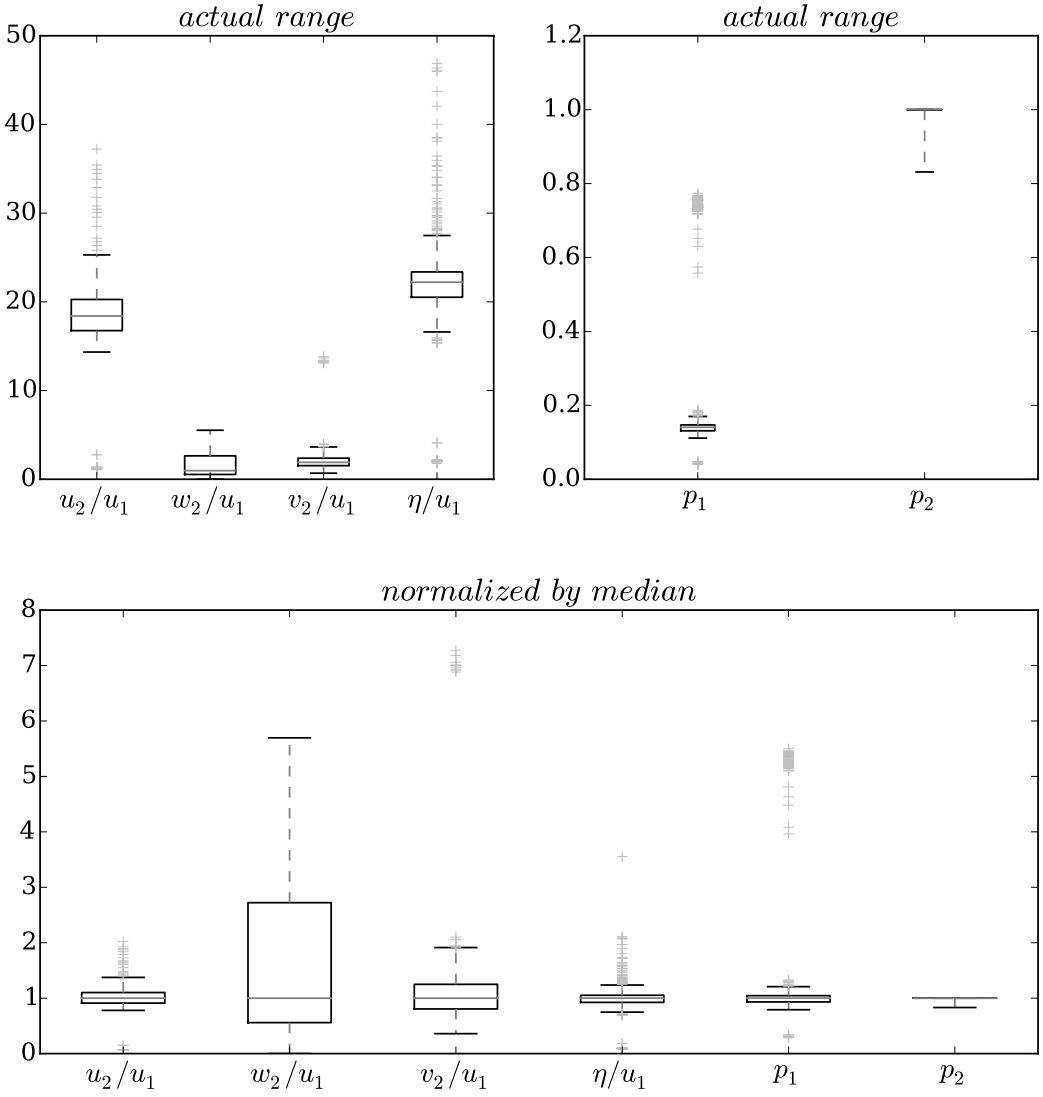

Figure 2.7: Descriptive statistics for values of the parameters for the best model, when fitted to 1000 samples from the data (each containing 90% of the whole dataset). Boxes show the first and third quartiles, Q_1 and Q_3, whereas the median (*i.e.* second quartiles) is marked by a horizontal line within each box. The variation outside the quartiles is illustrated by whiskers which are marked by dashed lines and denote the $1.5\,(Q_3 - Q_1)$. The points falling outside the whiskers' range are treated as outliers, and are plotted as individual points, denoted by a plus sign.

2.3 Discussion

One intriguing question addressed in this study is the relation between the number of MP copies and the rates of the underlying processes. We used α-exponents in our model in order to explore these relationships. Model selection and data fitting processes resulted in: $\alpha_u = \alpha_w = \alpha_v = 1$ and $\alpha_\eta = 0$. These values are in complete agreement with the α-values of the 1-D model published by van Passel et al. [2014] and also the model of IS5 insertion sequences proposed by Bichsel et al. [2013].

The rates of deletion and diversification in promoter regions, *i.e.* w_1 and v_1, contribute to the loss of promoter copies from PRs. These rates are predicted to be negligible in our model. One possible explanation could be very high selective pressure in these regions, such that promoter loss typically has a strong deleterious effect. If the organism does not survive, this results in a lack of associated data to study. In other words, the promoters in PRs possibly are constrained such that a change in their sequences may interfere with their functionality. Conversely, promoters outside of PRs are more capable of changing sequence since their diversification is less likely to be lethal for the genome.

Duplication in PRs, u_1, also occurs at a relatively low rate, while the probability p_2 indicates that almost all the promoters created via HGT are inserted in PRs. This high rate of promoter acquisition through HGT may explain the maintenance of MPs in PRs despite their low rate of duplication. These rates suggest that the main factor in the maintenance of MPs in PRs is HGT, since the duplication rate is low in these regions, and most promoters created via HGT are inserted in these regions.

Further, the reason that most of the promoters created via HGT are inserted in PRs may be answered by considering the HGT mechanism itself. One possible reason could be that promoters can find required homologous sequences for recombination in promoter regions more frequently than in other sites of the genome.

In contrast to the low rate of duplication in PRs, the duplication rate outside of

PRs, u_2, is high and as the probability p_2 implies, approximately 90% of the promoters newly created through duplication are inserted in these regions.

The mobile promoter data from entire genomes shows that over 27,000 copies of the total of 40,000 MPs exist outside promoter regions. This can be explained by the high rate of duplication in these regions. Although promoter loss through deletion and diversification outside of PRs occurs at a lower rate compared to duplication, these processes still play an important role in the MP dynamics in these regions. The parameters w_2 and v_2 represent the rate of deletion and diversification outside of PRs which together are estimated to be less than half of the duplication rate.

Overall, we observe that rates are lower or negligible in PRs compared with other regions, which implies a more unstable dynamics in non-promoter regions, and more stability in PRs.

There are many lines along which the work presented here could be extended. The first possible future direction is to apply and reconfirm the proposed PGM model with a more accurate dataset, one that uses the same scanning methods for searching PRs and other sites of genome. The second line to extend this work is to apply this location-based model to other types of mobile genetic elements, and investigate how and if their dynamics can be described with the proposed PGM model. The last but not least extension for our model is expand it to incorporate more factors, and cover more complex cases, for instance to explore the effects of selection.

Bibliography

Bichsel, M., Barbour, A., and Wagner, A. Estimating the fitness effect of an insertion sequence. *Journal of mathematical biology*, 66(1-2):95–114, 2013.

Blumenstiel, J. P. Evolutionary dynamics of transposable elements in a small RNA world. *Trends in genetics : TIG*, 27(1):23–31, 1 2011. doi: 10.1016/j.tig.2010.10.003.

Boutoille, D., Corvec, S., Caroff, N., Giraudeau, C., Espaze, E., Caillon, J., Plésiat, P., and Reynaud, A. Detection of an IS21 insertion sequence in the mexR gene of *Pseudomonas aeruginosa* increasing β-lactam resistance. *FEMS microbiology letters*, 230(1):143–146, 2004.

Burnham, K. P. and Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

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*, 2015.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

Jones, E., Oliphant, T., Peterson, P., et al. SciPy: Open source scientific tools for Python, 2001. URL `http://www.scipy.org/`. [Online; accessed 2015-08-28].

Kazazian, H. H. Mobile elements: drivers of genome evolution. *Science*, 303(5664): 1626–1632, 2004.

Matus-Garcia, M., Nijveen, H., and van Passel, M. W. Promoter propagation in prokaryotes. *Nucleic acids research*, pages 40(20):10032–40, 2012.

Nijveen, H., Matus-Garcia, M., and van Passel, M. W. J. Promoter reuse in prokaryotes. *Mobile genetic elements*, 2(6):279–281, 2012.

Ochman, H., Lawrence, J. G., and Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

Oren, Y., Smith, M. B., Johns, N. I., Zeevi, M. K., Biran, D., Ron, E. Z., Corander, J., Wang, H. H., Alm, E. J., and Pupko, T. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences*, 111(45):16112–16117, 2014.

Perez, J. C. and Groisman, E. A. Evolution of transcriptional regulatory circuits in bacteria. *Cell*, 138(2):233–244, 2009.

Rouzic, A. L. and Deceliere, G. Models of the population genetics of transposable elements. *Genetical research*, 85(03):171–181, 2005.

Sawyer, S. and Hartl, D. Distribution of transposable elements in prokaryotes. *Theoretical population biology*, 30(1):1–16, 1986.

Stoebel, D. M. and Dorman, C. J. The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. *Molecular biology and evolution*, 27(9): 2105–2112, 2010.

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

Chapter 3

Conclusion

In this work, we examined the dynamics of mobile genetic elements (MGEs), which play an important role in the evolution of all creatures. Prokaryotic genomes were our main focus, which we studied to answer intriguing questions about the distribution of MGEs. Although some types of MGEs, for example transposable elements (TEs), are well-studied, newly discovered classes of MGEs are still not completely understood and there are many questions about their dynamics to be answered. In particular, promoters are regulatory elements which recent evidence shows can be mobile within or between genomes, and therefore they are classified as a new type of MGE.

Following two recent studies on mobile promoters (MPs) by van Passel et al. [2014] and Matus-Garcia et al. [2012], we proposed a novel model for the dynamics of MPs which allowed us to explore MP dynamics with respect to their location in the genome. With this new model we were able to discover significant differences in the dynamics of mobile promoters in two different regions of the genome. In more detail, we considered regions inside of promoter regions and outside of promoter regions, and observed that mobile promoters duplicate, diversify, excise and transfer at considerably different rates in these regions. Our model is however more general, and one may apply this newly proposed location-based model to any two distinct

parts of the genome, for example coding and non-coding regions. Moreover, since in the model development process we did not make any specific assumption based on promoters, this model can also be applied to other types of MGEs.

Applying the model to the available data, we investigated the most important processes required to describe the dynamics of mobile promoters, and also the expected rates at which they occur. We discovered the simplest model that could describe the observed data, through enumerating a large number of candidate models and employing model selection techniques, resulting in a 6-parameter model. We further confirmed the statistical significance of the best model, and the robustness of its parameters through sampling and bootstrapping techniques. Hence we are confident that the biological interpretations provided in the thesis are well-supported by our computational results. One of the main biological findings of our work is that *mobile promoters are much more stable inside promoter regions, and most of their dynamic behaviour occurs outside of promoter regions.*

Finally, we should also point out that in our analysis we did not include natural selection. However, the fact that rates for deletion and diversification in promoter regions were predicted to be negligible by our results, may actually suggest the existence of strong selection in these regions. Thus, incorporating natural selection in the proposed model, and investigating its possible effects on the dynamics of mobile promoters is an interesting topic, and an important direction for future work. Another potential improvement in this model is to assume two different rates for the HGT process inside and outside of promoter regions. As mentioned in Chapter 2, we did not consider distinct rates for this process in order to avoid making the model too complicated. However, considering different rates for HGT could provide further insights about MP dynamics.

Bibliography

Matus-Garcia, M., Nijveen, H., and van Passel, M. W. Promoter propagation in prokaryotes. *Nucleic acids research*, pages 40(20):10032–40, 2012.

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

Curriculum Vitae

Name:	Mahnaz Rabbani
Post-Secondary Education:	Master of Science in Applied Mathematics, 2013-present The University of Western Ontario, London, Canada
	Bachelor of Science in Mathematics, 2007-2012 The KNT University of Technology, Tehran, Iran
Related Work Experience:	Teaching Assistant, 2013-2015 The University of Western Ontario
	Research Assistant, 2013-2015 The University of Western Ontario
	Primary School Instructor, Winter 2010 The Grassroots Support Organization, Tehran, Iran
Presentations and Posters:	The Birth-Death-Diversification Model of Mobile Genetic Elements in Prokaryotes: A Location-Based Model Using Mobile Promoter Data from Sequenced Prokaryotic Genomes; Presentation at the 2015 AMMCS-CAIMS Congress Waterloo, Ontario, Canada; June 7-12, 2015
	Evolution of Mobile Genetic Elements in Prokaryotes: Birth, Death, and Diversification Model using mobile promoters data in sequenced prokaryotic genomes; Poster Presented at the 2015 CRA Grad Cohort Workshop; San Francisco, California; April 10-11, 2015
	Rabbani, M. and Wahl, L.M. A Location-Based Birth, Death and Diversification Model for Mobile Promoters in Prokaryotes, working paper