Theoretical Population Biology 106 (2015) 22–31

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

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

Nicole E. Drakos, Lindi M. Wahl *

*Applied Mathematics, Western University, London, Ontario N6A 5B7, Canada*

ARTICLE INFO

*Article history:*
Received 29 October 2014
Available online 14 September 2015

*Keywords:*
Mobile genetic elements
Branching processes
Markov chain
Mobile promoters
Extinction probability
Extinction times

ABSTRACT

Theoretical approaches are essential to our understanding of the complex dynamics of mobile genetic elements (MGEs) within genomes. Recently, the birth–death-diversification model was developed to describe the dynamics of mobile promoters (MPs), a particular class of MGEs in prokaryotes. A unique feature of this model is that genetic diversification of elements was included. To explore the implications of diversification on the longterm fate of MGE lineages, in this contribution we analyze the extinction probabilities, extinction times and equilibrium solutions of the birth–death-diversification model. We find that diversification increases both the survival and growth rate of MGE families, but the strength of this effect depends on the rate of horizontal gene transfer (HGT). We also find that the distribution of MGE families per genome is not necessarily monotonically decreasing, as observed for MPs, but may have a peak in the distribution that is related to the HGT rate. For MPs specifically, we find that new families have a high extinction probability, and predict that the number of MPs is increasing, albeit at a very slow rate. Additionally, we develop an extension of the birth–death-diversification model which allows MGEs in different regions of the genome, for example coding and non-coding, to be described by different rates. This extension may offer a potential explanation as to why the majority of MPs are located in non-promoter regions of the genome.



## 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 (Orgel and Crick, 1980). For instance, transposable elements have been linked to hybrid dysgenesis in *Drosophila* (Rubin et al., 1982) and to deleterious mutations in bacteria and yeast (Kleckner, 1981). However, they can also be beneficial to an organism, as in the case of plasmids conferring antibiotic resistance (Frost et al., 2005). 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 (Charlesworth and Charlesworth, 1983; Langley et al., 1983), and the relatedness between copies in a family (Ohta, 1985; Slatkin, 1985; Brookfield, 1986; Hudson and Kaplan, 1986). The effects of selective pressures in limiting copy number have also been studied in some detail (Hickey, 1982; Charlesworth and Charlesworth, 1983; Golding et al., 1986; Edwards and Brookfield, 2003; Le Rouzic et al., 2007), as have the complex histories of transposable element lineages within genomes (Le Rouzic and Deceliere, 2005; Le Rouzic et al., 2007).

For MGEs in prokaryotes, both branching process and Markov chain approaches have been used to predict the distribution of copy number within genomes (Sawyer and Hartl, 1986, Moody, 1988, Basten and Moody, 1991, Hartl and Sawyer, 1988, Bichsel et al., 2013, van Passel et al., 2014, but also see Dolgin and Charlesworth, 2006, Wagner, 2006). 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 (Bichsel et al.,

* Corresponding author.
*E-mail addresses:* ndrakos@uwo.ca (N.E. Drakos), lwahl@uwo.ca (L.M. Wahl).

http://dx.doi.org/10.1016/j.tpb.2015.09.001
0040-5809/

2013; Hartl and Sawyer, 1988; van Passel et al., 2014). 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 (Hartl and Sawyer, 1988; Bichsel et al., 2013).

Mobile promoters (MPs) are a newly proposed class of MGEs (Matus-Garcia et al., 2012). 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 (Baños et al., 2009), and there is anecdotal evidence of rewiring of silent genes through the recruitment of promoters (Lee and Bernard, 2010; Stoebel and Dorman, 2010). Additionally, promoter sequences are highly conserved even between distantly related species (Nijveen et al., 2012; Matus-Garcia et al., 2012) indicating they may have the same origin. Furthermore, a recent study has shown that regulatory switching can occur through HGT of regulatory regions (Oren et al., 2014). This lends credence to the theory that transcriptional rewiring may be achieved through the recruitment of MPs (Nijveen et al., 2012).

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 (van Passel et al., 2014; Nijveen et al., 2012; Matus-Garcia et al., 2012). 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 (van Passel et al., 2014) 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 transferred 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 (Bichsel et al., 2013), 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 (Thomas and Nielsen, 2005).

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. Methods

### 2.1. The birth–death-diversification model

In Matus-Garcia et al. (2012) 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 van Passel et al. (2014). 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 within a single genome with very similar sequences; for example, 80% sequence identity over 50 nucleotides was used as a threshold in van Passel et al. (2014), 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 out of all the MGE families within this collection of genomes, that have $n$ copies. To calculate this, we first find all the families with $n$ copies within each genome, and then sum over all the genomes. 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 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 transferred 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, meaning that the probability of this occurrence is less than 0.001. 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 $\eta$. The birth–death-diversification model originally included *de novo* creation of elements in rate $\eta$ (van Passel et al., 2014). However, since this paper is mainly interested in lineages of MGEs, we will consider $\eta$ as only describing HGT rates. Additionally, HGT is likely much more common than *de novo* creation of elements, and thus *de novo* creation may make a negligible contribution to this rate.

The resulting copy-level model can be expressed as an infinite system of ordinary differential equations (ODEs) that describe the