DIVERSITY AND EVOLUTION OF ECOSYSTEMS: FROM GENOMES TO THE BIOSPHERE

BY

CHI XUE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Associate Professor Taylor L. Hughes, Chair
Professor Nigel D. Goldenfeld, Director of Research
Assistant Professor Thomas E. Kuhlman
Professor Sergei Maslov

Abstract

In this thesis, I present research on biological systems at three different scales of space and time: biodiversity of ecological systems, the dynamics of repetitive elements and their diversity in the genome, and the development of phylogenetic trees in evolution. The unifying theme is the interplay between ecology and evolution, expressed within an ecosystem, within genomes, and over the evolutionary history of life.

Part I concerns biodiversity on the ecological scale. I study the "Kill the Winner" (KtW) hypothesis, a proposed solution to the biodiversity paradox questioning why many competitors can coexist in a single niche. The original KtW model is deterministic and expressed in terms of continuous biomass concentrations, and appears to predict the coexistence of species. Here I present a stochastic individual-level model for the KtW paradigm, representing populations as finite integers. We find an extinction cascade and a monotonic loss of diversity in time due to the stochasticity, thus failing to explain diversity in the presence of stochasticity. To solve this problem, we couple the coevolution of predators and prey with the KtW model, and show that the diversity of the stochastic system can arise from the constant population flux induced by the emergence of new mutants, although there are undoubtedly contributions from the spatial variations in populations too. Our results suggest that diversity reflects the dynamical interplay between ecological and evolutionary processes, and is driven by how far the system is from an equilibrium state.

Part II consists of three projects on the dynamics and diversity of repetitive DNA elements on the genomic scale. The first project is to develop a statistical mechanical model for the interaction between two types of DNA transposons, known as LINE and SINE. These mobile genetic elements are respectively autonomous and non-autonomous: SINE steals the machinery of LINE to complete its migration, and thus acts as a parasite. We have found that the demographic noise due to the discreteness of element copy numbers leads to noisy oscillations on the evolutionary time scale, in a similar way to that resulting in the predator-prey quasi-cycles in ecology. By viewing these DNA elements as predators and prey, we have shown that the dynamics in the genome can fruitfully be analyzed using the analogy to ecological models. In the second project, we look for the predicted quasi-cycles of LINE and SINE in the genomic history of the ancient fish coelacanth. We analyze the periodicity of the age distribution recorded in the genome by the molecular clock, and also develop a theoretical model to examine under what conditions can the cycles be recorded. Our analyses provide a procedure for future research work, but the conclusion is that the rapid deterioration of DNA transposons due to mutations means that the observational window is restricted to the last 50 million years, which is not long enough to conclude that the predicted oscillations are present. In the third project, we further explore the analogy of a genome to an ecosystem and DNA elements to organisms. We use the metric known as the rank-abundance distribution (RAD) from ecology to study the diversity of junk DNA "species". We have found that the RADs for all the 46 examined species can be reasonably fit by a power law, with very similar exponents. This universal RAD can help identify the underlying microscopic evolutionary processes of these DNA species. Our work demonstrates that applying ecological methods to study genomic elements may provide novel insights for genome functions and evolution.

Part III focuses on the development of phylogenetic trees on the evolutionary scale. The topology of phylogenetic trees has been found to obey a universal scaling law. The exponent lies in between the two extreme cases of completely balanced binary trees and completely imbalanced ones. We seek evolutionary processes that can generate the observed topology, and here study in particular the effect of niche construction on the large-scale structure of phylogenetic trees. In contrast to the conventional natural selection framework, which treats the environment independently of the organisms under selection, the niche construction theory views the feedback of organisms on their environment as a crucial and explicit process in evolution. We present a coarse-grained statistical model of niche construction coupled to simple models of speciation, and show that the resultant phylogenetic tree topology can exhibit a scale-invariant structure, through a singularity arising from large niche construction fluctuations. These results show in principle how the scaling laws of phylogenetic tree topology can emerge from rather general assumptions about the interplay between ecological and evolutionary processes.

To my parents.

Acknowledgments

I dedicate my sincere gratitude to my advisor, Professor Nigel Goldenfeld. As a coworker, Nigel is exemplary of outstanding scientists. He seems to know everything about everything, and is always eager to learn more. He is good at articulating his ideas and reasonings. And he always remembers our previous discussions. It has been very rewarding and pleasant to work with Nigel. Thanks to him, I've been introduced to some magnificent iceberg tips in the ocean of science, both within and outside physics. As an advisor, Nigel has offered tremendous guidance, insights and feedback to my research projects. I wouldn't be able to accomplish the work in this thesis without his valuable input. Through numerous instructions on matters ranging from chores as small as how to make an impressive talk opening to plans as big as how to choose a good research topic, Nigel has helped me build both a scientific skill set and mindset. I also greatly appreciate the efforts he has spent in keeping me on the right track toward the degree, from my very first day in the group till the current moment when I am finishing the thesis. As a mentor, Nigel cares about my general well-being and is always willing to share his insights on aspects of life. He is very generous with encouragements. I will always remember how he revitalized me at difficult moments along the stressful path through the graduate school, as well as his positive way of commenting on an ugly first draft of powerpoint slides. Among the many inspiring conversations, I constantly remind myself of the one we had at a lunch break during a Gordon Research Conference, over which he explained to me how different people handled different types of information, and how to accept and make use of the way in which our own brains were innately wired. I shall benefit from his enlightenment for many years to come.

I have had the pleasure of working with other brilliant scientists during the past years. Assistant Professor Oleg Simakov at the University of Vienna is our collaborator on the repetitive DNA element project. He is also the one who introduced me to the wonderful world of transposons. I am very grateful for working with him and greatly appreciate his feedback on the models. I thank Assistant Professor Thomas Kuhlman at the University of Illinois at Urbana-Champaign, for his helpful feedback on the LINE-SINE model and for involving me in the bacterial retroelement project. I also enjoy the joint Kuhlman-Goldenfeld group meeting, in which I've learned about some fantastic experiments and amazing organisms. I'm indebted to members of Nigel's group, for their sharp critical insights and warm discussions, which always help greatly improve my work. They are K. Michael Martini, Vikyath Rao, Hong-Yan Shih, Purba Chatterjee, Sang Hyun Choi, Minhui Zhu, Zhiru Liu, Farshid Jafarpour (former graduate student), Tommaso Biancalani (former postdoc) and Kit Sang Chu (former undergraduate student). I'm also grateful to students in Kuhlman's group for many interesting discussions. They are Nicholas Sherer, Gloria Lee, Davneet Kaur, Niko Urriola and Neil Kim (former graduate student).

I owe countless thanks to my family. Since I entered the middle school, I've been going farther and farther away, geographically, from my parents. Yet, they are always there, watching me growing, comforting my upsets, celebrating my accomplishments, and embracing every single bit of my life. They don't understand the research I do, nor can they read this acknowledgment. Still, they unconditionally trust me and support me. My gratitude to them is beyond words. I thank my husband, Xiongjie Yu, for the past nine years together. It's suddenly hard to think of specific things to thank him for, but he deserves endless gratitude for accompanying me through many memorable moments of my life till now. I'm thankful to my younger brother, Peng, for being a very supportive sibling and a very reliable friend. I'm also grateful to my two cats, for their generously sharing the habitat and mercifully sparing my precious coding computer and research notes.

I was supported by the National Aeronautics and Space Administration Astrobiology Institute (NAI) under Cooperative Agreement No. NNA13AA91A issued through the Science Mission Directorate, the National Science Foundation through Grant No. PHY-1430124 to the NSF Center for the Physics of Living Cells, and the L. S. Edelheit Family Biological Physics Fellowship from the Department of Physics, University of Illinois at Urbana-Champaign.

I owe special thanks to Professor S. Lance Cooper, who suggested me to explore the possibility of transition to theoretical physics, which led to years of spectacular journey.

Table of Contents

Chapter 1

Introduction

The science of biology branches into many subfields, studying objects of various scales. It ranges from as small as single molecules to as large as the entire biosphere. At the molecular level, DNA sequences record all the details of the blueprint to design a functional organism. At the ecosystem level, organisms interact with each other as well as with the environment, fighting for survival. Dynamics at these distinct scales are coupled via the process of evolution: genomic information determines how organisms behave, and the selection over behaviors promotes beneficial genotypes. The multiscale nature of biological systems, the important role of noise and fluctuations, and the strong interactions between their components, make them an interesting subject to study from the perspective of physical sciences, especially statistical physics. In this dissertation, I present projects applying techniques from statistical physics to study the dynamics in biological systems on multiple scales. As will be seen, intrinsically distinct systems can sometimes be described by the same type of minimal model, and applying methods in one field to another can generate insightful results.

From a mean-field point of view, the state of a system is represented by the continuous concentration of each component. This picture ignores both intrinsic and extrinsic stochasticity. In a stochastic system, the mean-field model often only describes the ensemble-averaged dynamics. The historical trajectory followed by a specific system, however, is like a single trial out of the ensemble and usually is very different from the average. Previous work has illustrated how demographic noise induced by the fact that species populations or molecule copy numbers are discrete integers can change the dynamics from the mean-field expectation, both in terms of time series and spatial distributions [1, 2]. The integration of stochasticity in these problems provides minimal models to described the observed phenomena without adding in detailed ingredients of the system. In this thesis, I also concentrate on the role of stochasticity in dynamics and incorporate intrinsic demographic noise in the modeling when applicable.

I report studies on questions of three different scales: the biodiversity in the ecological system, the dynamics of jumping genes and their diversity in the genome, and the development of phylogenetic trees in evolution. Part I is on the biodiversity in an ecosystem. I show that stochasticity destroys the predicted diversity in the deterministic "Kill the Winner" (KtW) hypothesis, and that incorporation of predator-prey coevolution maintains the diversity in the stochastic KtW paradigm.

Part II involves repetitive elements in the genome. I develop a minimal model to describe the interaction between two types of "jumping genes" or transposons, called LINE and SINE, and show that stochasticity leads to persistent noisy oscillations in their copy numbers, similar to prey-predator quasi-cycles in ecology. We further analyze the genome history of coelacanth to look for these predicted cycles, but the time series are not long enough to verify the predictions. We also use the rank-abundance distribution to investigate the diversity of repetitive elements in genomes of 46 species, showing that ecological concepts can be used to characterize the entities that reside in genomes.

Part III is on the phylogenetic tree in evolution. I address the question of the origin of the apparent scale invariance of phylogenetic trees, and show that this can arise from the interplay between niche construction and evolution. I present a theoretical model that integrate the niche construction of organisms as an explicit evolutionary process to study its impact on the resultant tree topology.

In the following sections, I briefly introduce the contents of remaining chapters, explain my contribution in each project, and list the publications.

1.1 Coevolution Maintains Diversity in the "Kill the Winner" Framework in Ecology

Diversity is one of the major topics of ecological study. It refers to the fact that multiple species coexist in the same ecosystem. Researches focus on how the species population is distributed and what is the mechanism that drives the diversity.

In Chapter 2, I introduce the biodiversity paradox and several resolutions to it. The biodiversity paradox [3–5] questions why many species competing for the same limited resource coexist, while the the competitive exclusion principle [6] predicts that only one species could survive eventually. I also review several methods of quantifying the diversity, including the diversity index [7], the species-abundance distribution (SAD) and the rank-abundance distribution (RAD). I discuss in detail the relation between the SAD and the RAD, and review models in the literature on these distributions.

In Chapter 3, I focus on one of the proposed resolutions to the biodiversity paradox, known as the "Kill the Winner" hypothesis [8, 9]. It argues that, in a system where species compete with each other for limited resources, there exist host-specific predators corresponding to these competitors. The predators control the population of each prey, preventing a winner from emerging and thus maintaining the coexistence of all species in the system. The original model assumes that the system is spatially homogeneous and uses deterministic ordinary differential equations of continuous biomasses to describe the population dynamics. It predicts coexistence of species in the sense that the equilibrium steady state has positive biomasses.

We develop an individual-level model for the KtW hypothesis to account for intrinsic demographic noise. We demonstrate that stochasticity causes the coexistence steady state in the deterministic KtW model to break down through a cascade of extinctions, leading to a loss of diversity. The reason of the breakdown is that the finite populations in the stochastic model always have a nonzero probability of reaching zero due to random fluctuation.

In Chapter 4, I develop a stochastic coevolving KtW model to revalidate the KtW theoretical framework, by introducing the coevolution of predators and prey. The coevolution arises when the prey mutate in phenotypic traits to escape from the predation and the predators mutate to catch up. It constantly introduces fit mutants into the system and thus prevents the elimination of species. We find that coevolution maintains the diversity of the stochastic KtW system, and we compute the diversity dependence on the mutation rate. The coevolving KtW model applies to systems that mutate frequently so that ecological interactions happen on the same time scale as the evolution. Our results suggest that diversity reflects the dynamical interplay between ecological and evolutionary processes, and is driven by how far the system is from an equilibrium state.

1.2 Dynamics and Diversity of Repetitive Elements in Genomes

The DNA in a genome is not 100% coding sequences. Instead, there are a large number of non-coding repetitive elements, or repeats, in genomes across all three domains of life. Their existence resolves the C-value paradox [10], which states that the complexity of an organism is not reflected in the genome size, because genomes can maintain enormous redundancy and also junk or non-coding regions. The human genome is about 45% junk, mostly composed of transposons [11]. Repetitive elements generally include two categories [12]: interspersed repeats, usually resulted from transposon activities, scattered all over the genome, and tandem repeats located adjacently to each other. Repetitive elements are a major driver of evolution, as their activities, such as expansion, contraction and migration, can interrupt the coding and regulatory sequences as well as cause misaligned pairing and unequal chromosome crossovers, resulting in both deleterious and advantageous mutations. Studying the dynamics of repetitive elements is thus crucial to understanding the evolution of species.

In Chapter 5, I review background knowledge on transposable elements (TEs), or transposons. Transposons are classified into autonomous elements that code all necessary enzymes for their activities, and non-autonomous elements that do not and so must rely on others' machinery [13]. They can also be categorized as DNA transposons following a “cut-and-paste” rule and retrotransposons obeying a “copy-and-paste” rule [13]. We focus on the autonomous LINE-1 and non-autonomous Alu elements, which are both abundant retrotransposons in the human genome. I review the molecular interactions between the two types of elements and discuss the parasitic dependence of Alu elements on LINE-1 elements.

In Chapter 6, I develop and solve a minimal individual-level model for the population dynamics of a pair of autonomous and non-autonomous transposons. I use LINE-1 and Alu elements as a model system. Our model predicts that demographic stochasticity generates persistent and noisy oscillations in the copy numbers of the transposons, similar to the predator-prey quasi-cycles in ecology. The characteristic time scale of the cycles is much longer than the cell replication time, and the state of the predator-prey oscillator is stored in the genome and transmitted to successive generations. By viewing these DNA elements as predators and prey, we have shown that the dynamics in the genome can fruitfully be analyzed using a mathematical analogy to an ecological model.

In Chapter 7, I report a search in the genomic history for the predicted quasi-cycles of a LINE-SINE pair. The history of the genome is annotated by the molecular clock, with the element age being reported as the number of point mutations in the DNA sequence. Researchers have found in several species so-called periodic expansion of transposons, visible as oscillations of transposon copy numbers along the age axis. We are interested in whether these cycles are induced by the intrinsic LINE-SINE interaction. Specially, we examine the periodic expansion in the transposon age distribution of coelacanth [14]. This “living fossil” species has remained in its current form for about 400 million years [15], which we interpret as evidence of a lack of external selection pressure. Thus we expect historical changes in its transposon composition to originate from intrinsic element interactions, rather than from external factors. We analyze the coelacanth transposon age distribution data to look for quasi-cycles, and also develop a theoretical model to investigate under what conditions can the quasi-cycles be recorded by the molecular clock. These analyses are the first attempt to explore the potential connection between the observed periodic expansion and the quasi-cycles of an autonomous/non-autonomous TE pair.

In Chapter 8, I study the diversity of all families of repetitive elements in the genome. We further explore the analogy of a genome to an ecosystem, and treat genomic elements as organisms that live in the system [16]. Element families are then like genomic species. Under this analogy, we use the rank-abundance distribution (RAD) in ecology as a characterizing metric to study the repeat diversity in the genomes of 46 species. We observe RADs of simple repeats that can be well-fitted as power laws with very similar exponents. This surprising universal distribution can help reveal the underlying microscopic evolutionary process, which will further provide insight to the question of whether the abundant repetitive elements in eukaryotic genomes are functional or just selfish "junk" DNA [17]. Our work also shows that applying ecological methods to study the genomic elements may provide insights from a novel perspective.

I also participated in a project that measured the bacterial growth defect induced by retroelements, in collaboration with Professor Thomas Kuhlman's group. We transferred two retroelements, human LINE-1 and the bacterial group II intron Ll.LtrB, into *E. coli* and *B. subtilis* cells, and observed that the invasion led to significant reduction in the population growth rate and eventually cell death. In addition, we found that retroelement lethality and proliferation was enhanced by the ability to perform eukaryotic-like nonhomologous end-joining (NHEJ) DNA repair. We showed by theoretical modeling that the only stable evolutionary consequence in simple cells was maintenance of retroelements in low numbers. We hypothesized that eukaryotes must have evolved methods to get around the growth defect associated with the transposons, and thus our work connected with the evolutionary history of the spliceosome. Our results suggested that NHEJ might have played a fundamental and previously unappreciated role in enabling the evolutionary transitions from simple to complex genomes and circular to linear chromosomes. The main results were obtained by K. Michael Martini and so are not reported in this thesis.

1.3 Effect of Niche Construction on the Evolution of Phylogenetic Trees

The conventional account of the evolutionary process is based upon natural selection [18]. Phenotypic variations first arise due to mutation and gene migration; then environmental selection and genetic drift determine how each phenotype frequency changes with time. The process results in adaptation to the environment with the survival of the fittest organisms. The environment, however, is treated as a boundary condition of the evolutionary process and especially does not depend on the status of the organisms under selection. A new theoretical framework called niche construction theory [19], in contrast, advocates for the impact of organisms' feedback on the environment as an evolutionary process. The result is a feedback between ecology and evolution, one example of which is so-called "Rapid Evolution" [20–22].

In Chapter 9, I briefly review main ideas of the niche construction theory and compare it with natural selection. In the niche construction framework, organisms are capable of modifying the environment, termed as the niche, and thus altering the selection pressure. Consequently, the evolutionary path of the organism is a result of the interplay of organisms and their environment. Natural selection, on the other hand, views niche construction behaviors as special phenotypic traits of certain organisms and does not distinguish their role in the evolution.

In Chapter 10, I introduce a metric to characterize the topology of a binary tree and review the observed universal topological scaling of phylogenetic trees. Phylogenetic trees represent the hypothetical evolutionary process derived from the relatedness of extant species, with internal nodes standing for inferred ancestor species. For an arbitrary node on the tree, we define A and C as the size and cumulative size, respectively, of its subtree. Technical details are provided in Chapter 10. The important point is that it appears that $C(A)$ of actual phylogenetic trees are found to be universal and obey $C(A) \sim A^{1.4}$ [23–25], which lies in between the two extreme cases of complete balanced binary trees and completely imbalanced binary trees.

In Chapter 11, I report attempts to formulate evolutionary process models to explain the observed power law scaling of phylogenetic trees. Especially, I include niche construction explicitly as a process in the evolution and study how it influences the resultant tree topology. We develop a Niche Inheritance Model in which the parent's niche passes on to the child species with some fluctuation due to the construction. We show that a large niche construction effect generates an apparent power-law regime in the topological metric of the tree. We show that the power-law regime emerges asymptotically as a model parameter tends to zero, using crossover scaling theory. These results show that over a wide range of tree size, niche construction effects can give rise in principle to power-law scaling in topological measure of phylogenetic trees. In short, niche construction can leave an indelible footprint on the evolutionary process.

1.4 My Contribution

The work elaborated in this dissertation is a result of close collaboration with other scientists. My advisor Professor Nigel Goldenfeld has provided intense motivation, input and feedback for all projects. Assistant Professor Oleg Simakov at University of Vienna has performed data analyses in the repetitive element project. I list my contribution in these projects below.

In the project on the biodiversity in ecology, I developed the stochastic individual-level model of the original "Kill the Winner" hypothesis and conducted the numerical simulations. I extended it to the generalized KtW model, performed numerical integrations of both its stochastic and deterministic versions. I also analyzed the linear stability of the mean-field version. These are documented in Chapter 3. I further proposed the coevolving KtW model in Chapter 4 and performed numerical simulations.

In the project on the dynamics of LINE-SINE transposon pair, I proposed the stochastic individual-level model for the L1-Alu pair in Chapter 6 and conducted the numerical simulations. I also derived and solved the stochastic differential equations for the power spectra and phase difference, and the mean-field equations for linear stability.

In the project on searching for quasi-cycles in the coelacanth genome in Chapter 7, Professor Oleg Simakov developed the phylogeny method of analyzing the sequence data, and generated the transposon age distribution of coelacanth. I examined the periodicity of the data and calculated the cross correlations of any two transposon families. I also developed the theoretical model to describe the age distribution, and analyzed the low-pass filter effect of the molecular clock .

In the project on the diversity of repetitive elements in genomes, reported in Chapter 8, the age distribution data of repeats in all species were generated and provided by Professor Oleg Simakov. I calculated and analyzed the rank-abundance distribution of repeats in each genome.

In the project on the bacterial growth defect induced by retroelements, I collaborated on developing the model for retroelement dynamics over evolutionary time.

In the project on the topology of the phylogenetic trees, I developed the Niche Inheritance Model in Chapter 11, performed numerical simulations and derived the mean-field $C(A)$ relation as a function of the tuning parameter. I also analyzed the critical scaling of $C(A)$ and performed the data collapse.

1.5 List of Publications

The work in Chapters 3 and 4 is published as Ref. [26]. The work in Chapter 6 is published as Ref. [27]. The work on the bacterial growth defect induced by retroelements is in review. The publications are listed below.

- Chi Xue, Nigel Goldenfeld, *Stochastic Predator-prey Dynamics of Transposons in the Human Genome*, Phys. Rev. Lett. **117**, 208101, (2016) [27]

- Chi Xue, Nigel Goldenfeld, *Coevolution Maintains Diversity in the Stochastic "Kill the Winner" Model*, Phys. Rev. Lett. **119**, 268101, (2017) [26]

- Gloria Lee, Nicholas A. Sherer, Neil H. Kim, Ema Rajic, Davneet Kaur, Niko Urriola, K. Michael Martini, Chi Xue, Nigel Goldenfeld, Thomas E. Kuhlman, *Testing the Retroelement Invasion Hypothesis for the Emergence of the Ancestral Eukaryotic Cell*, in review, Proc. Natl. Acad. Sci. U.S.A. [28]

Part I

Diversity at the Ecological Scale

Chapter 2

Introduction to the Diversity of Ecosystems

2.1 Biodiversity Paradox and Proposed Resolutions

The high diversity of coexisting species in most ecosystems has been a major puzzle for more than 50 years. Hutchinson first articulated the so-called Paradox of the Plankton in 1961 [3] for the case of marine ecosystems: why do many species of plankton that feed on the same nutrients coexist, instead of one species outcompeting all the others?

This latter expectation has been formulated precisely as the so-called competitive exclusion principle [6]. As the premise of the paradox, it has been validated, for example, in Ref. [29], which demonstrates that in a system where species with different traits feed on and compete for the same resource, clusters of organisms emerge as a result of the competition exclusion.

The Paradox of the Plankton is not limited to marine ecosystems, but has been generalized to terrestrial systems and expressed as the biodiversity paradox [4, 5].

The various tentative resolutions of the paradox can be divided into two classes [5, 30, 31]. In the first, it's argued that the competitive exclusion principle applies to a fixed point equilibrium state, while the ecosystem fails, due to temporal or/and spatial factors, to reach such an equilibrium. For example, the time needed for the system to reach equilibrium might be much longer than the time over which the system undergoes significant changes in its boundary conditions, such as weather [32]. Also, spatial heterogeneity can increase the global diversity of the system by maintaining local patches that each obey the competitive exclusion principle but globally support the coexistence of multiple species [33, 34] (for another perspective, see [35]). In other words, the system does not reach a global equilibrium state due to spatial dispersion. In the second class of resolutions, interactions such as predation, in conjunction with competitive exclusion, promote the coexistence of species through time-dependent or stochastic steady states [8, 36–38]. One widely celebrated example of this behavior is the continual succession of different community members known as the "Kill the Winner" (KtW) dynamics [8, 9, 39]. We will discuss details of the KtW model and demonstrate its breakdown in the presence of the demographic stochasticity in Chapter 3. And in Chapter 4, we will show that the coupling of KtW and coevolution together are able to recover the diversity of the stochastic system.

2.2 Methods of Characterizing Species Diversity in Ecology

The diversity of a system not only refers to the richness, *i.e.* the total number, of species, but also the evenness of the population distribution. There are several ways to quantify this concept, such as by calculating the diversity index and by looking at the abundance distribution. I briefly introduce these conventional methods in this section, and explain some useful connections between them that are not usually made explicitly in the literature.

2.2.1 Diversity Index

If the population of each species is known in an ecosystem, with p_i being the population fraction of species i in the system and R being the richness, then the diversity index can be defined in the following general form [7].

$$ {}^qD = \frac{1}{\sqrt[q-1]{\sum_{i=1}^{R} p_i p_i^{q-1}}} = \left(\sum_{i=1}^{R} p_i^q \right)^{-\frac{1}{q-1}} , \qquad q > 1, \tag{2.1} $$

where q refers to the order of the diversity. We observe that the expression in the bracket is the mean of p^{q-1}, denoted as $M_{q-1} \equiv \langle p^{q-1} \rangle$, under the given population distribution $\{p_i\}$. And the diversity ${}^qD = 1/M_{q-1}^{1/(q-1)}$ is interpreted as the effective number of species that would give the same M_{q-1} assuming equally distributed populations. There are several special cases: ${}^1D = 1/R$; $\lim_{q\to 1} {}^qD = \exp(S)$, with S being the Shannon entropy $S \equiv -\sum_{i=1}^{R} p_i \ln p_i$ [40, 41]; and ${}^2D$ is the reciprocal of the Simpson index $\lambda \equiv \sum_{i=1}^{R} p_i^2$ [42].

Despite its convenience and wide usage, the above index itself is not sufficient to reveal all the features of the diversity. To be comprehensive, ecologists look at the distribution of the species population, or abundance, directly. There are two broadly used distributions, called the species-abundance distribution (SAD) and the rank-abundance distribution (RAD).

2.2.2 Abundance Distributions

The species-abundance distribution (SAD) quantifies, how many species S, usually in the same trophic level, have a certain population size (abundance) A. It has the same interpretation as the population size distribution. The rank-abundance distribution (RAD) is obtained by sorting in descending order the abundance values of all species, assigning rank $r = 1$ to the most abundant species, $r = 2$ to the second most abundant one, *etc.*, and then plotting the abundance A against the rank r. It shows straightforwardly the species richness and evenness in the system.

The two distributions are not independent. Instead, RAD can be conveniently derived from the SAD. Let S_A be the number of species that consist of A individuals, and thus describe the species abundance distribution. The rank, r_A, of a species with abundance A can then be calculated as follows.

$$r_A = \sum_{A' \geq A}^{+\infty} S_{A'}. \tag{2.2}$$

The key idea of the equation is that the rank of a species is equal to the total number of species that are more or equally abundant. Approximate the summation with integration and the above equation becomes

$$r_A \to r(A) = \int_A^{+\infty} S(A')dA'. \tag{2.3}$$

The inverse function $A(r)$ then gives the functional form of the abundance-rank relation.

In particular, suppose that the species abundance distribution is power-law with exponent $-n$,

$$S_A = C_1 A^{-n}, \tag{2.4}$$

where C_1 is the normalization factor such that $\int_1^{+\infty} S_A dA = R$. Then we can derive $r(A)$ following Eq. (2.3) for $n > 1$ as below [43]:

$$r(A) = C_1 \int_A^{+\infty} A'^{-n} dA' = \frac{C_1}{n-1} A^{-(n-1)}. \tag{2.5}$$

Therefore the abundance-rank equation is

$$A = C_2 r^{-\frac{1}{n-1}}, \tag{2.6}$$

with $C_2 = [(n-1)/C_1]^{-1/(n-1)}$. This is the power-law rank-abundance distribution, with exponent $-1/(n-1)$. For $n = 1$, the integral in Eq. (2.3) gives a logarithmic function,

$$r(A) = C_1 \int_A^{A_1} A'^{-1} dA' = C_1 \ln \frac{A_1}{A}, \tag{2.7}$$

with A_1 being the highest species abundance. The abundance-rank relation is then

$$A = A_1 e^{-\frac{r}{C_1}}. \tag{2.8}$$

This is the exponential rank-abundance distribution.

2.2.3 Rank-abundance Distribution in Ecosystems

The most famous RAD-type distribution probably is the Zipf's law [44], which states that the frequency of a certain word in a given language is inversely proportional to its rank, *i.e.* $A \propto r^{-1}$. Zipf's law is ubiquitous in many languages and also found in other non-linguistic systems, such as the population ranks of cities, income ranks, and so on [43].

In ecology, ecosystems also share common functional forms of SADs or RADs. First, the general qualitative trend of the RAD is uniform in all sampled ecosystems. It typically consists of two parts: the high abundance segment composed of a few core species, and the main body containing the majority of species and extending to the rare biosphere with low abundance. Although rare species have nearly undetectably small abundances at a certain temporal point, they can contribute to the resilience of the system in later composition turnovers [45]. Second, many ecosystems share the same quantitative asymptotic behaviors at large rank. Both exponentially decaying [46–48] and power-law [49–54] RADs have been observed in a broad range of systems. Data are constantly being generated, as the sequencing technology proceeds, for example, by the Tara Oceans, a project sampling microbes on the global scale [55–59]. It seems that the preliminary data are consistent with the broad characterization of marine virus and bacteria species exhibiting power-law decaying $A(r)$, whereas terrestrial and microbiomes exhibit an exponential tail. Although it is tempting to speculate that the form arises due to a combination of density-dependent birth-death processes coupled with turbulence, as I have attempted in unfinished work, this remains an unproven speculation at present.

2.2.4 Models in the Literature

Although SAD and RAD give straightforward characterization of the diversity, they are both macroscopic features and veil the underlying microscopic rules followed by species. There have been intense efforts to reverse engineer the microscopic processes based on the SAD and RAD. I briefly introduce two end member theories [60] here.

The unified neutral theory [61] assumes that species are functionally equivalent. They undergo birth, death, migration processes in a completely random and independent manner, yet with the same rates. The difference in their relative abundances is purely due to fluctuations. The unified neutral theory has been criticized ever since its proposal, since its assumption is against the observation and it has failed to match many of the observed distributions [62–66]. Nevertheless it is widely accepted as a null hypothesis.

The opposite end member is the niche apportionment models, first proposed in Refs. [67, 68] and then expanded in Refs. [69–72]. This family of models assigns species to occupy their specific niches and thus to break up the niche space or available resources. The niche occupation behavior determines the relative abundance. Each species has its unique feature and niche, in contrast to the functional equivalence in the neutral theory. Examples of testing the niche theory are in Refs. [73, 74].

In reality, ecosystems are found on a spectrum between these two end members, *e.g.* in Ref [48]. And there have been efforts to unify the two types of theories [60, 75, 76].

Although the neutral and niche theories are radically distinct in their premises, they can generate the same RAD behavior [77–79]. Therefore, it's not conclusive to infer whether the underlying microscopic process is neutral or niche-based, judging from the observed SAD and RAD.

Chapter 3

Breakdown of the "Kill the Winner" Hypothesis in the Presence of Demographic Stochasticity

The "Kill the Winner" hypothesis [8, 9] is an attempt to address the biodiversity paradox. It has been frequently revisited and expanded in the context of marine systems [39, 80], and is related to the Janzen-Connell hypothesis [81, 82] for tree biodiversity. It argues that host-specific predators control the population of each prey, preventing a winner from emerging and thus maintaining the coexistence of all species in the system. It is seen in both natural ecosystems as well as some laboratory systems such as chemostats [45, 83].

The original calculations assume that the system is spatially homogeneous and use continuous biomass to describe the population. However, the continuous variables, which are allowed to become arbitrarily small, can not capture effects induced by the finite population size, such as extinction events [84, 85]. The fact that the population size is integer-valued leads inexorably to shot noise, referred to in the ecological context as demographic stochasticity.

In the rest of this chapter, we explore the effect of demographic stochasticity on the KtW paradigm and demonstrate that the stochasticity causes the coexistence steady state in the deterministic KtW model to break down through a cascade of extinctions, leading to a loss of diversity. This work has been published as a part of Ref. [26].

3.1 Original "Kill the Winner" Hypothesis

The original "Kill the Winner" hypothesis [8, 9] was proposed to explain the coexistence of bacteria and plankton, which both consumed the same limited chemical resource in the ocean. The basic idea is that the coexistence of competitors is maintained by their predators that prevent any winners from taking over. The plankton community generally has a lower efficiency of resource usage than the bacteria. They remain in the system, only because a protozoan consumes the bacteria non-selectively and thus limits the bacterial population, leaving room for the plankton to thrive. Inside the bacterial community, different strains have distinct growth rates. They coexist, with no dominating winners, due to host-specific viruses controlling the corresponding strains. This results in two layers of coexistence, nested like Russian dolls [39]: the coexistence of bacteria and plankton as the first layer, and the coexistence of all bacterial strains as the second.

In a later paper [39], the bacteria and plankton communities in the original model were generalized to two groups: the competition specialists and the defense specialists. The former (bacteria) have a high resource consumption efficiency but are susceptible to predators, while the latter are resistant to the predation but poor at resource usage. Predators exert pressure on the competition specialists and prevent them from dominating over the defense specialists; the defense specialists are limited by the total amount of available resource.

The original KtW model was formulated as deterministic Lotka-Volterra type equations for the species biomass concentrations [8, 9]. The high diversity of the system is exhibited in the steady state where multiple species coexist with positive biomass values.

Here, we reinterpret the original equations, which were about biomass concentrations, in terms of number densities in Eq. (3.1), and later compare the results with those of its stochastic version.

$$\dot{B}_i = b_i B_i R - p_i V_i B_i - gPB_i - r_i B_i, \tag{3.1a}$$

$$\dot{V}_i = p_i V_i B_i - d_i V_i, \tag{3.1b}$$

$$\dot{P} = gP \sum_i B_i - d_p P, \tag{3.1c}$$

$$\dot{A} = b_A AR - r_A A, \tag{3.1d}$$

$$\dot{R} = -\sum_i b_i B_i R - b_A AR + \sum_i d_i V_i + \sum_i r_i B_i + d_p P + r_A A. \tag{3.1e}$$

The dot operator stands for the time derivative. B_i and V_i are, respectively, the densities of the ith bacterial and viral strains. P is the density of the protozoan, A of the plankton, and R of the resource. Bacteria have strain-specific growth rate b_i and death rate r_i. Viruses of strain i attack their specific bacterial hosts with rate p_i and decay with rate d_i. The protozoan hunts all bacteria nonselectively, with rate g, and dies with rate d_p. Plankton face no predation and have a growth rate of b_A and death rate of r_A.

The set of Eq. (3.1) is redundant in the sense that $\sum_i \dot{B}_i + \sum_i \dot{V}_i + \dot{P} + \dot{A} + \dot{R} = 0$. This reflects that the total number density of all categories is constant, assumed to be 1, shown in Eq. (3.2). We have set this condition to mirror the constraint in the original model that the entire biomass is conserved. With this fixed carrying capacity, this model thus is a type of “urn” model.

$$\sum_i B_i + \sum_i V_i + P + A + R = \text{const.} \tag{3.2}$$

There is a caveat in Eqs. (3.1) and (3.2). While the biomass of a system is truly conserved, the total number density usually is not. For example, at the lysis of a bacterial cell due to viral infection, out of each bacterium emerges $O(10) \sim O(100)$ viruses. This ratio is called the burst size. While the number density dramatically increases, the total biomass stays the same. In order to use Eq. (3.2) to model the carrying capacity, we have to ignore the fact that viruses have a large burst size. By setting aside the burst size, Eq. (3.1) generally can not capture the fact that viruses are 10 times more abundant than bacteria. But doing this will make it easy to write down the stochastic version using an urn model for comparison. At this moment, we focus on demonstrating the difference between the deterministic and stochastic situations, and ignoring the burst size does not matter for this purpose. Later, when we generalize the model, we will break the urn and then add back the burst size.

Following the original model, we further assume the bacteria growth rates are ordered in the way below:

$$b_1 > b_2 > b_3 > \dots. \tag{3.3}$$

We can solve for the steady state by setting the time derivatives in Eq.(3.1) to zero and calculating all densities. We are particularly interested in the nontrivial steady state where most species coexist and have positive densities. In that state, there exists m bacterial and $m-1$ viral strains together with the protozoan, plankton and resource. The analytical expressions are as follows.

$$\dot{A} = 0 \implies R^* = \frac{r_A}{b_A} \tag{3.4a}$$

$$\dot{P} = 0 \implies \sum_{i=1}^{m} B_i^* = \frac{d_p}{g} \tag{3.4b}$$

$$\dot{V_i} = 0 \implies B_i^* = \frac{d_i}{p_i}, i = 1, 2, \dots, m-1 \tag{3.4c}$$

$$\implies B_m^* = \frac{d_p}{g} - \sum_{i=1}^{m-1} \frac{d_i}{p_i} \tag{3.4d}$$

$$V_m^* = 0 \tag{3.4e}$$

$$\dot{B_m} = 0 \implies P^* = \frac{1}{g}(b_m R^* - r_m) \tag{3.4f}$$

$$\dot{B_i} = 0 \implies V_i^* = \frac{1}{p_i}(b_i R^* - r_i - gP^*), i = 1, 2, \dots, m-1 \tag{3.4g}$$

$$\implies V_i^* = \frac{1}{p_i}\left[(b_i - b_m)R^* - (r_i - r_m)\right], i = 1, 2, \dots, m-1 \tag{3.4h}$$

$$\implies A^* = 1 - P^* - R^* - \sum_{i=1}^{m} B_i^* - \sum_{i=1}^{m} V_i^* \tag{3.4i}$$

The starred variables in the above expressions stand for the steady state values. Eq. (3.4b) indicates that the protozoan controls the population of the entire bacterial community. And Eq. (3.4c) demonstrates that each viral strain constrains the corresponding bacterial strain. Together, the number of existing bacterial strains m is determined by predation behaviors of both the protozoan and viruses, as shown in Eq. (3.4d).

To show the dynamical behavior of the deterministic model Eq. (3.1), we assign a set of illustrative parameters, and numerically evolve the equations using the Runge-Kutta (RK4) method. Assume there are six bacterial strains with birth rates $\boldsymbol{b} = (0.1, 0.09, 0.08, 0.07, 0.06, 0.05)$. Let other parameters be strain independent and $p_i \equiv p = 0.1$, $d_i \equiv d = 0.01$, $g = 0.1$, $d_p = 0.055$, $b_A = 0.02$, $r_i = r_A \equiv r = 0.001$. Then the steady state densities are given by the following values:

$$R^* = 0.05, \tag{3.5a}$$

$$P^* = 0.015, \tag{3.5b}$$

$$A^* = 0.31, \tag{3.5c}$$

$$\boldsymbol{B}^* = (0.1, 0.1, 0.1, 0.1, 0.1, 0.05), \tag{3.5d}$$

$$\boldsymbol{V}^* = (0.025, 0.02, 0.015, 0.01, 0.005, 0). \tag{3.5e}$$

Note that the viral populations are ordered *s.t.* $V_1^* > V_2^* > V_3^* > V_4^* > V_5^* > V_6^* = 0$. The left column of Fig. 3.1 shows the time series obtained from the deterministic model. The system is initially perturbed away from the steady state. All species densities, which are predicted to be positive, undergo strong oscillations and slowly decay toward the steady state values. The system has a high diversity throughout time despite the fluctuations.

3.2 Stochastic Version of the Original KtW Model

We can write down the individual-level reactions (3.6) that correspond to the Lotka-Volterra type rate equations Eq. (3.1). X_i stands for a bacterium of the ith strain, Y_i for a virus of the ith strain, Z for a protozoan individual, S for a plankton organism, and E for a resource quantum.

$$X_i + E \xrightarrow{b_i} 2X_i, \qquad X_i + Y_i \xrightarrow{p_i} 2Y_i, \tag{3.6a}$$

$$X_i + Z \xrightarrow{g} 2Z, \qquad S + E \xrightarrow{b_A} 2S, \tag{3.6b}$$

$$B_i \xrightarrow{r_i} E, \qquad Y_i \xrightarrow{d_i} E, \tag{3.6c}$$

$$Z \xrightarrow{d_p} E, \qquad S \xrightarrow{r_A} E. \tag{3.6d}$$

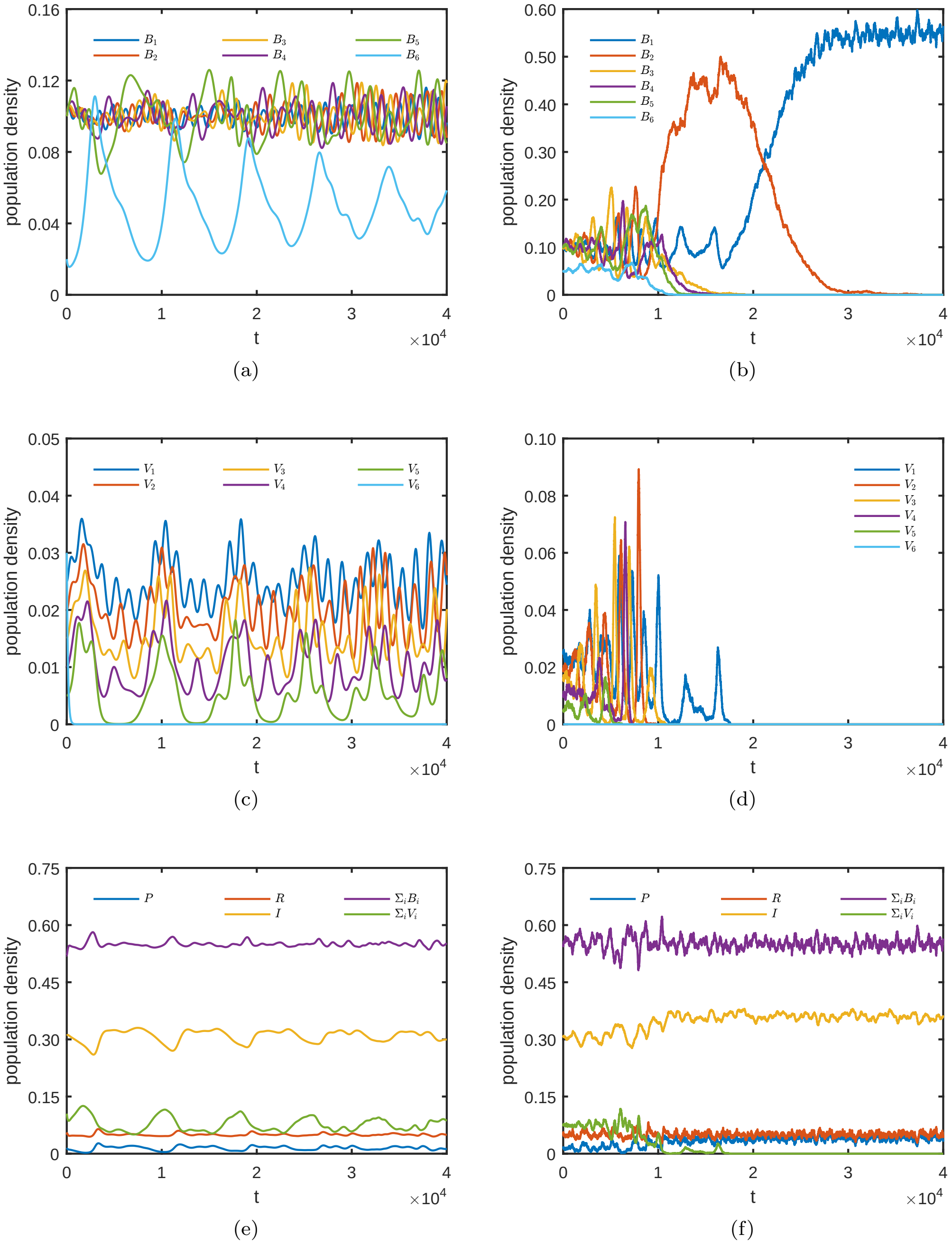


Figure 3.1: Time series of species population densities, with the deterministic result on the left and stochastic one on the right. The first row is the populations of bacterial strains, the second row the viral strains, and the third row the protozoan, resource, plankton, bacterial total and viral total. In the deterministic version, the system is initially perturbed away from the steady state by setting $B_6 = 0.02$, $V_6 = 0.03$ and everything else at their steady state values. Despite of the large oscillations, all species, except V_6, coexist, in accordance with the prediction by the deterministic model. In the stochastic version, the system size is $C = 10000$ and species are initiated with the exact steady state values. Demographic noise drives species to deviate from their steady state, and some populations drift to zero. Eventually, the entire viral community vanishes, while one winner emerges from the bacterial community. The coexistence state is destroyed. Figures are adopted from the published work Ref. [26].

Results of numerical simulations of the above stochastic model with the Gillespie algorithm [86] are shown in the right column of Fig. 3.1. In the stochastic model, species go extinct one after another, destroying the coexistence state. And eventually only one bacterial strain remains with the protozoan and plankton, while the viral community is completely annihilated. For the particular time series presented in the right column of Fig. 3.1, the order in which species became extinct was $V_5 \to V_4 \to V_2 \to V_3 \to B_6 \to B_5 \to B_4 \to B_3 \to V_1 \to B_2$. The general trend was determined by the fact that bacterial growth rates were ordered in the way such that B_6 had the lowest growth rate and B_1 the highest, and V_5 was the closest to extinction and V_1 the farthest. This order is reflected in Eq. (3.5). It's noticeable that V_2 went extinct before V_3 did. This resulted in B_2 being freed and driving other strains with lower growth rates to extinction until later being outcompeted by B_1. These large fluctuations and temporal dominance of one certain strain are also a consequence of stochasticity and cannot be captured by the deterministic model.

3.3 Generalized "Kill the Winner" Model

The original KtW model includes interactions on both the species level (bacteria, plankton and the protozoan) and the strain level (bacterial and viral strains), which makes it mathematically difficult to tackle. We realize that the key component of the KtW hypothesis is that for each resource competitor there is a corresponding predator that can prevent it from becoming a dominant winner. The Russian doll-like hierarchy is hence not essential for the basic idea. Thus we focus on only a single layer in the original model, the bacterial and viral strains, and ignore the multilevel structure. The KtW model is in this way simplified and generalized to a system of m pairs of prey (bacteria) and predators (viruses). The individual reactions are as follows.

$$X_i \xrightarrow{b_i} 2X_i, \qquad X_i + X_j \xrightarrow{e_{ij}} X_j, \tag{3.7a}$$

$$Y_i + X_i \xrightarrow{p_i} (\beta_i + 1)Y_i, \qquad Y_i \xrightarrow{d_i} \emptyset. \tag{3.7b}$$

All rates are positive. $i, j = 1, 2, \dots, m$ are strain indices. Bacterial individuals X_i, have strain-specific growth rate b_i. They compete with each other for an implicit resource with strength e_{ij}. Viruses of the ith strain Y_i, infect the corresponding host X_i with rate p_i and burst size β_i, and decay to nothing $\emptyset$ with rate d_i. These reactions form the minimal generalized KtW model. In this set of reactions, the carrying capacity is modeled by the competition among bacterial strains, instead of as an explicit urn. The total population is no longer forced to be a constant and we are thus able to use a realistic viral burst size.

3.3.1 Mean-field Solution of the General KtW Model

Reactions (3.7) have the following mean-field rate equations.

$$\dot{B}_i = b_i B_i - \sum_{j=1}^{m} e_{ij} B_i B_j - p_i B_i V_i, \tag{3.8a}$$

$$\dot{V}_i = \beta_i p_i B_i V_i - d_i V_i. \tag{3.8b}$$

B_i and V_i represent the densities of the ith bacterial and viral strains, respectively. We set e_{ij} to a constant value e for simplicity.

Eq. (3.8) has a nonzero steady state as shown below.

$$B_i^* = \frac{d_i}{\beta_i p_i}, \qquad V_i^* = \frac{1}{p_i}\left(b_i - e\sum_{j=1}^{m} B_j^*\right). \tag{3.9}$$

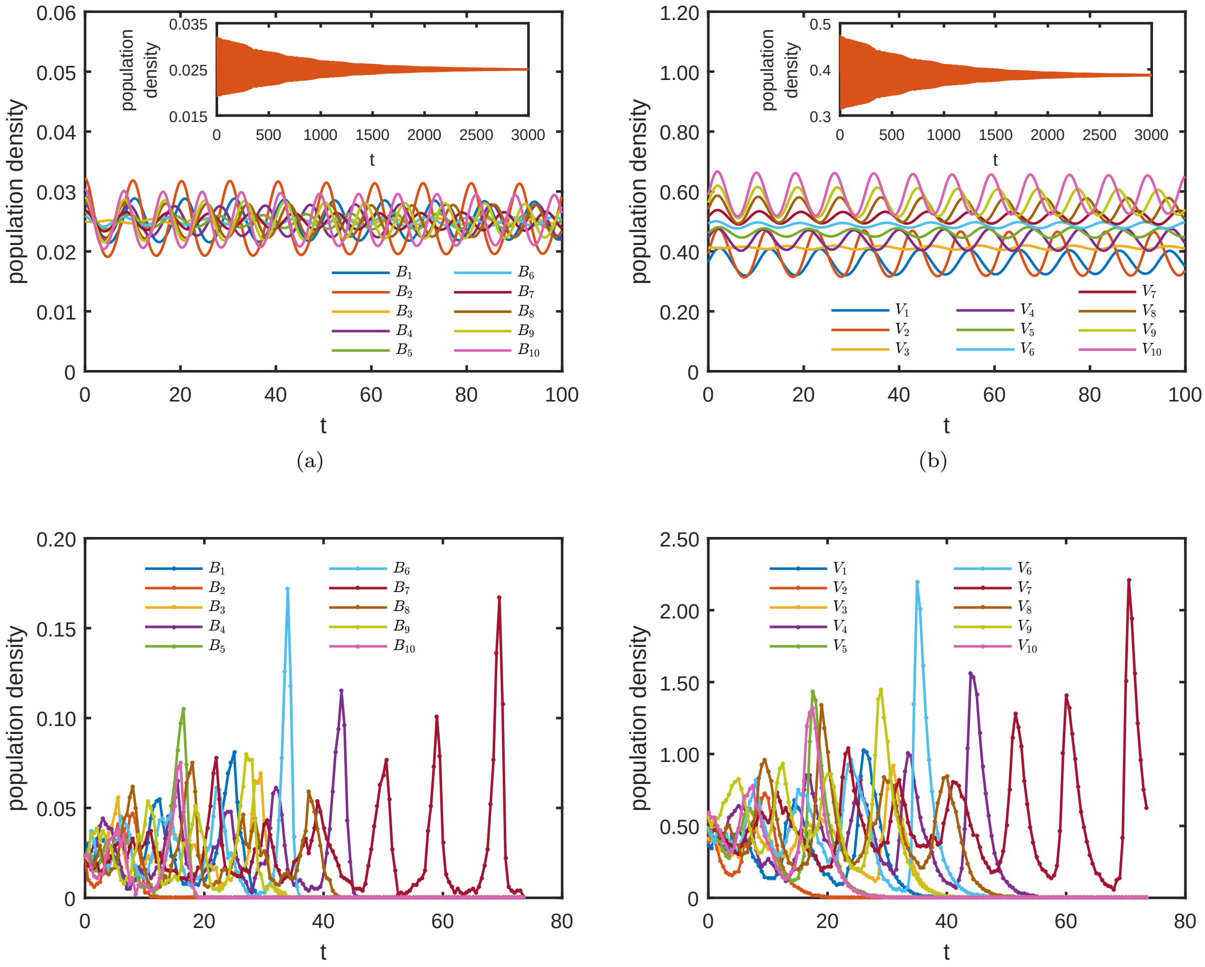
We require all B_i^* and V_i^* to be positive, which limits the parameters to satisfy $b_i > e\sum_{j=1}^{m} d_j/\beta_j p_j, \forall i$. We have conducted the linear stability analysis, which will be discussed later in the subsection, and found that the steady state Eq. (3.9) is exponentially stable, as long as the quantity $x_i \equiv \beta_i p_i^2 B_i^* V_i^* = d_i(b_i - e\sum_{j=1}^{m} d_j/\beta_j p_j)$ is distinct for each i. The steady state can be either a focus or node, depending on whether the eigenvalues of the linear stability matrix have nonzero imaginary parts or not. The parameters used in this chapter result in the steady state being a focus, but the conclusion also applies to the node case.

In Fig. 3.2, we show in the first row the time series of prey and predator densities obtained from a numerical evolution of Eq. (3.8) for $m = 10$ pairs of bacteria and phages. Species densities are initially perturbed away from the steady state. As shown in the figure insets, species densities decay back to the steady state at long times, confirming the result of the linear stability analysis. The oscillatory behavior at short time scales demonstrates the steady state to be a focus.

Linear Stability of the Steady State of the Generalized KtW Model

The Jacobian matrix of Eq. (3.8) is $2m \times 2m$, and can be written in the form of a block matrix as below.

$$J = \begin{pmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{pmatrix}. \tag{3.10}$$



(a)

(b)

(c)

(d)

Figure 3.2: Population density time series obtained from the generalized KtW model, with 10 bacterium-phage pairs. The left column is for bacteria and the right for viruses. The first row shows the result from a numerical evolution of the deterministic generalized KtW equations, with species densities initially perturbed randomly away from the steady state. The parameters are $\boldsymbol{b} = (0.75, 0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2)$, $p_i \equiv p = 2$, $\beta_i \equiv \beta = 10$, $d_i \equiv d = 0.5$, and $e_{ij} \equiv e = 0.1$. Densities undergo oscillatory decay toward the steady state. The insets show the long time behavior, which demonstrates that the steady state is a focus. For readability, only the decays of B_2 and V_2 are shown. The second row presents a typical simulation result of the stochastic version of the generalized KtW model, using the same set of parameters. The system size is $C = 1000$ and populations are initialized with the steady state values. The oscillatory decay behavior is destroyed by the demographic noise. And the system eventually collapses after all bacterial strains become extinct. Figures are adopted from the published work Ref. [26].

The four blocks are all $m \times m$ matrices with their elements defined as follows.

$$\mathcal{A}_{ij} = \frac{\partial B_i}{\partial B_j}, \qquad \mathcal{B}_{ij} = \frac{\partial B_i}{\partial V_j}, \qquad \mathcal{C}_{ij} = \frac{\partial V_i}{\partial B_j}, \qquad \mathcal{D}_{ij} = \frac{\partial V_i}{\partial V_j}. \tag{3.11}$$

The four matrices can be evaluated by observing that

$$\frac{\partial B_i}{\partial B_j} = \begin{cases} b_i - 2eB_i - e\sum_{j\neq i} B_j - p_i V_i, & j = i \\ -eB_i, & j \neq i \end{cases} \tag{3.12a}$$

$$\frac{\partial B_i}{\partial V_j} = \begin{cases} -p_i B_i, & j = i \\ 0, & j \neq i \end{cases} \tag{3.12b}$$

$$\frac{\partial V_i}{\partial B_j} = \begin{cases} \beta_i p_i V_i, & j = i \\ 0, & j \neq i \end{cases} \tag{3.12c}$$

$$\frac{\partial V_i}{\partial V_j} = \begin{cases} \beta_i p_i B_i - d_i, & j = i \\ 0, & j \neq i \end{cases} \tag{3.12d}$$

Substitute the steady state values B_i^* and V_i^* into the above expressions, and we have

$$\mathcal{A}^* = \begin{pmatrix} -eB_1^* & -eB_1^* & \cdots & -eB_1^* \\ -eB_2^* & -eB_2^* & \cdots & -eB_2^* \\ \vdots & \vdots & \ddots & \vdots \\ -eB_m^* & -eB_m^* & \cdots & -eB_m^* \end{pmatrix}, \tag{3.13a}$$

$$\mathcal{B}^* = \begin{pmatrix} -p_1 B_1^* & & & \\ & -p_2 B_2^* & & \text{\huge 0} \\ \text{\huge 0} & & \ddots & \\ & & & -p_m B_m^* \end{pmatrix}, \tag{3.13b}$$

$$\mathcal{C}^* = \begin{pmatrix} \beta_1 p_1 V_1^* & & & \\ & \beta_2 p_2 V_2^* & & \text{\huge 0} \\ \text{\huge 0} & & \ddots & \\ & & & \beta_m p_m V_m^* \end{pmatrix}, \tag{3.13c}$$

$$\mathcal{D}^* = 0. \tag{3.13d}$$

The characteristic equation of the Jacobian matrix J is given by

$$\det(J-\lambda I)=\det\begin{pmatrix}\mathcal{A}^*-\lambda\mathcal{I} & \mathcal{B}^*\\ \mathcal{C}^* & \mathcal{D}^*-\lambda\mathcal{I}\end{pmatrix}=0, \tag{3.14}$$

where λ is the eigenvalue, I is the $2m\times 2m$ identity matrix, and $\mathcal{I}$ is the $m\times m$ identity matrix. Since the two diagonal matrices $\mathcal{C}^*$ and $(\mathcal{D}^*-\lambda\mathcal{I})$ commute with each other, we have the following equation

$$\det\begin{pmatrix}\mathcal{A}^*-\lambda\mathcal{I} & \mathcal{B}^*\\ \mathcal{C}^* & \mathcal{D}^*-\lambda\mathcal{I}\end{pmatrix}=\det\big((\mathcal{A}^*-\lambda\mathcal{I})(\mathcal{D}^*-\lambda\mathcal{I})-\mathcal{B}^*\mathcal{C}^*\big). \tag{3.15}$$

We can define $-eB_i^*\equiv a_i$, then matrix $\mathcal{A}$ becomes

$$\mathcal{A}^*=\begin{pmatrix}a_1 & a_1 & \cdots & a_1\\ a_2 & a_2 & \cdots & a_2\\ \vdots & \vdots & \ddots & \vdots\\ a_m & a_m & \cdots & a_m\end{pmatrix}, \tag{3.16}$$

and further

$$(\mathcal{A}^*-\lambda\mathcal{I})(\mathcal{D}^*-\lambda\mathcal{I})=\begin{pmatrix}\lambda^2-a_1\lambda & -a_1\lambda & \cdots & -a_1\lambda\\ -a_2\lambda & \lambda^2-a_2\lambda & \cdots & -a_2\lambda\\ \vdots & \vdots & \ddots & \vdots\\ -a_m\lambda & -a_m\lambda & \cdots & \lambda^2-a_m\lambda\end{pmatrix}. \tag{3.17}$$

We can also straightforwardly calculate that

$$\mathcal{B}^*\mathcal{C}^*=\begin{pmatrix}-\beta_1p_1^2B_1^*V_1^* & & & \\ & -\beta_2p_2^2B_2^*V_2^* & & \text{\Large 0}\\ \text{\Large 0} & & \ddots & \\ & & & -\beta_mp_m^2B_m^*V_m^*\end{pmatrix}. \tag{3.18}$$

With the above two equations, we then have

$$
\begin{aligned}
&(\mathcal{A}^* - \lambda\mathcal{I})(\mathcal{D}^* - \lambda\mathcal{I}) - \mathcal{B}^*\mathcal{C}^* \\
&= \begin{pmatrix}
\lambda^2 - a_1\lambda + \beta_1 p_1^2 B_1^* V_1^* & -a_1\lambda & \cdots & -a_1\lambda \\
-a_2\lambda & \lambda^2 - a_2\lambda + \beta_2 p_2^2 B_2^* V_2^* & \cdots & -a_2\lambda \\
\vdots & \vdots & \ddots & \vdots \\
-a_m\lambda & -a_m\lambda & \cdots & \lambda^2 - a_m\lambda + \beta_m p_m^2 B_m^* V_m^*
\end{pmatrix}.
\end{aligned}
\tag{3.19}
$$

Our goal is to calculate the determinant of the above matrix to obtain the characteristic equation. We can first simplify the matrix with linear transformations. Subtract column 1 from column i, $\forall i > 1$, and we have a new matrix named $\mathcal{M}$ as follows.

$$
\mathcal{M} = \begin{pmatrix}
\lambda^2 - a_1\lambda + \beta_1 p_1^2 B_1^* V_1^* & -\lambda^2 - \beta_1 p_1^2 B_1^* V_1^* & -\lambda^2 - \beta_1 p_1^2 B_1^* V_1^* & \cdots & -\lambda^2 - \beta_1 p_1^2 B_1^* V_1^* \\
-a_2\lambda & \lambda^2 + \beta_2 p_2^2 B_2^* V_2^* & 0 & \cdots & 0 \\
-a_3\lambda & 0 & \lambda^2 + \beta_3 p_3^2 B_3^* V_3^* & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-a_m\lambda & 0 & 0 & \cdots & \lambda^2 + \beta_m p_m^2 B_m^* V_m^*
\end{pmatrix}. \tag{3.20}
$$

Note that matrix $\mathcal{M}$ has nonzero elements only in the first row, in the first column and on the diagonal. With this special structure, its determinant can be calculated below.

$$
\det(\mathcal{M}) = \prod_{i=1}^{m} \mathcal{M}_{ii} + \sum_{k=2}^{m} (-1)^{2k-3} \mathcal{M}_{k1}\mathcal{M}_{1k} \prod_{i\neq 1,k}^{m} \mathcal{M}_{ii} \tag{3.21a}
$$

$$
= \prod_{i=1}^{m} \mathcal{M}_{ii} - \sum_{k=2}^{m} \mathcal{M}_{k1}\mathcal{M}_{1k} \prod_{i\neq 1,k}^{m} \mathcal{M}_{ii} \tag{3.21b}
$$

$$
\begin{aligned}
=&(\lambda^2 - a_1\lambda + \beta_1 p_1^2 B_1^* V_1^*)(\lambda^2 + \beta_2 p_2^2 B_2^* V_2^*) \cdots (\lambda^2 + \beta_m p_m^2 B_m^* V_m^*) \\
&- (\lambda^2 + \beta_1 p_1^2 B_1^* V_1^*)(a_2\lambda)(\lambda^2 + \beta_3 p_3^2 B_3^* V_3^*) \cdots (\lambda^2 + \beta_m p_m^2 B_m^* V_m^*) \\
&- (\lambda^2 + \beta_1 p_1^2 B_1^* V_1^*)(\lambda^2 + \beta_2 p_2^2 B_2^* V_2^*)(a_3\lambda) \cdots (\lambda^2 + \beta_m p_m^2 B_m^* V_m^*) \\
&- (\lambda^2 + \beta_1 p_1^2 B_1^* V_1^*)(\lambda^2 + \beta_2 p_2^2 B_2^* V_2^*)(\lambda^2 + \beta_3 p_3^2 B_3^* V_m^*) \cdots (a_m\lambda).
\end{aligned}
\tag{3.21c}
$$

Rearrange the terms and we arrive at the following expression.

$$\begin{aligned}\det(J-\lambda I) = \det(\mathcal{M}) =& (\lambda^2+\beta_1 p_1^2 B_1^* V_1^*)(\lambda^2+\beta_2 p_2^2 B_2^* V_2^*)\cdots(\lambda^2+\beta_m p_m^2 B_m^* V_m^*)\\ &-(a_1\lambda)(\lambda^2+\beta_2 p_2^2 B_2^* V_2^*)(\lambda^2+\beta_3 p_3^2 B_3^* V_3^*)\cdots(\lambda^2+\beta_m p_m^2 B_m^* V_m^*)\\ &-(\lambda^2+\beta_1 p_1^2 B_1^* V_1^*)(a_2\lambda)(\lambda^2+\beta_3 p_3^2 B_3^* V_3^*)\cdots(\lambda^2+\beta_m p_m^2 B_m^* V_m^*)\\ &-(\lambda^2+\beta_1 p_1^2 B_1^* V_1^*)(\lambda^2+\beta_2 p_2^2 B_2^* V_2^*)(a_3\lambda)\cdots(\lambda^2+\beta_m p_m^2 B_m^* V_m^*)\\ &-(\lambda^2+\beta_1 p_1^2 B_1^* V_1^*)(\lambda^2+\beta_2 p_2^2 B_2^* V_2^*)(\lambda^2+\beta_3 p_3^2 B_3^* V_m^*)\cdots(a_m\lambda).\end{aligned} \tag{3.22}$$

Let

$$f_i \equiv \lambda^2+\beta_i p_i^2 B_i^* V_i^* \equiv \lambda^2 + x_i, \tag{3.23}$$

where $x_i = \beta_i p_i^2 B_i^* V_i^*$. Then the characteristic equation is simplied as

$$\det(J-\lambda I) = \prod_{i=1}^{m} f_i - \sum_{i=1}^{m} a_i\lambda \prod_{j\neq i}^{m} f_j = 0. \tag{3.24}$$

In order to determine the stability of the steady state, we don't necessarily need the exact λ values that satisfy the above equation, but only need to know whether their real parts are positive or negative. The Routh-Hurwitz criterion [87, 88] can be applied to determine how many roots of a polynomial have negative real parts. But it's not trivial to carry out the calculation for Eq. (3.24), which is of degree $2m$, when m is large. Here, I show a simple method to determine the stability of the steady state.

First, consider the case where all x_i are distinct from one another. Then we can prove that for any eigenvalue λ satisfying Eq. (3.24), we always have $f_i(\lambda) \neq 0$, $\forall i$. In fact, if we assume $\exists i^*$, s.t. $f_{i^*} = 0$, then $\lambda^2 = -x_{i^*}$. Since x_i are different from each other, we have $f_j \neq 0, \forall j \neq i^*$. Therefore, the term $a_{i^*}\lambda\prod_{j\neq i^*}^{m} f_j \neq 0$, and with the first term being zero, Eq. (3.24) cannot be satisfied. This brings out the contradiction, and our assumption is wrong. Thus $f_i(\lambda) \neq 0, \forall i$. And we now can divide both sides of Eq. (3.24) by $\prod_{i=1}^{m} f_i$ to get the following equation.

$$1 - \sum_{i=1}^{m} \frac{a_i\lambda}{f_i} = 0. \tag{3.25}$$

Now let

$$\lambda = \alpha + \mathrm{i}\beta, \tag{3.26}$$

where i is the imaginary unit, and substitute it into Eq. (3.25), then we arrive at an equation of the following form.

$$1 + \sum_{j=1}^{m} \operatorname{Re}\gamma_j + \mathrm{i}\sum_{j=1}^{m} \operatorname{Im}\gamma_j = 0, \tag{3.27}$$

where Re and Im stand for the real and imaginary parts of a complex number, respectively. The above equation requires that

$$1 + \sum_{j=1}^{m} \operatorname{Re}\gamma_j = 0, \tag{3.28a}$$

$$\sum_{j=1}^{m} \operatorname{Im}\gamma_j = 0. \tag{3.28b}$$

And γ_j is given by the following expressions.

$$\operatorname{Re}\gamma_j = -\frac{\alpha a_j(\alpha^2+\beta^2+x_j)}{(\alpha^2-\beta^2+x_j)^2+4\alpha^2\beta^2}, \tag{3.29a}$$

$$\operatorname{Im}\gamma_j = -\frac{\beta a_j(-\alpha^2-\beta^2+x_j)}{(\alpha^2-\beta^2+x_j)^2+4\alpha^2\beta^2}. \tag{3.29b}$$

We now can obtain the following equation,

$$1 - \alpha\sum_{j=1}^{m} \frac{a_j(\alpha^2+\beta^2+x_j)}{(\alpha^2-\beta^2+x_j)^2+4\alpha^2\beta^2} = 0. \tag{3.30}$$

Since $a_j = -eB_j^* < 0$ and $x_j > 0$, we conclude that $\alpha < 0$. This applies to any λ that is a root of Eq. (3.24). In other words, all eigenvalues have negative real parts and the steady state Eq. (3.9) is locally exponentially stable.

Back to our condition that x_i is distinct from one another. If this is not satisfied, then f_i can be zero for some value(s) of i, and $\exists i = k$, s.t. $\lambda = \pm\mathrm{i}\sqrt{x_k}$ are two roots of Eq. (3.24). Due to these pure imaginary eigenvalues, the steady state is not exponentially stable.

Note that

$$x_i = \beta_i p_i^2 B_i^* V_i^* = d_i(b_i - e\sum_{i=1}^{m} \frac{d_i}{\beta_i p_i}), \tag{3.31}$$

and we can easily select parameters such that the steady state is exponentially stable.

In the first row of Fig. 3.2, we show the time series of prey and predator densities from a numerical evolution of Eq. (3.8) for $m = 10$ pairs of bacteria and phages, with the parameters $\boldsymbol{b} = (0.75, 0.8, 0.85, 0.9,$

0.95, 1, 1.05, 1.1, 1.15, 1.2), $p_i \equiv p = 2$, $\beta_i \equiv \beta = 10$, $d_i \equiv d = 0.5$, and $e_{ij} \equiv e = 0.1$. The steady state is

$$B_i^* = 0.025, \quad \forall i, \tag{3.32a}$$

$$\boldsymbol{V}^* = (0.3625, 0.3875, 0.4125, 0.4375, 0.4625, 0.4875, 0.5125, 0.5375, 0.5625, 0.5875). \tag{3.32b}$$

Species densities are initially perturbed away from the steady state by a small random amount. As shown in the figure insets, species densities decay back to the steady state in the long time, confirming the result of the linear stability analysis. The oscillatory behavior on the short time scale is due to the imaginary part of the eigenvalues of the linear stability matrix.

3.3.2 Stochastic Simulation of the General KtW Model

To reveal the effect of demographic noise, we also conduct the Gillespie stochastic simulation [86] of the corresponding individual level reactions (3.7) with the same parameter set as that used in the deterministic equations to generate the first row of Fig. 3.2. The resultant species density time series are shown in the second row of Fig. 3.2.

In contrast to the deterministic behavior of oscillatory decay, species go extinct in a short time. Bacterial strains become extinct due to random fluctuation; this consequentially triggers the extinction of the corresponding viral strains, which die due to a lack of food. The number of existing species monotonically decreases in the process, and the system diversity undergoes a cascade.

3.4 Conclusion

We have compared the deterministic and stochastic versions of the KtW model, both the original and the generalized ones. While the deterministic model predicts a stable steady state of coexistence, the demographic stochasticity present in a system with finite populations induces an extinction cascade and leads the coexistence state to break down. The reason for the breakdown lies in the fact that species populations in the stochastic model are all finite, and the probability of the population reaching zero due to random fluctuation is always nonzero.

In Chapter 4, we will explore how to the revalidate the KtW theoretical framework, by introducing the coevolution of predators and prey.

Chapter 4

Coevolution Maintains Diversity in the KtW Model

We have shown in Chapter 3 that demographic stochasticity causes the coexistence steady state in the KtW model to break down. Ecosystems have evolved many potential mechanisms to get around the path to extinction, as introduced in Chapter 2. Here, we discuss one possibility: prey and predator coevolve with each other so that fit mutants are constantly being introduced into the system, thus preventing the elimination of the species. Specifically, prey improve their phenotypic traits (*e.g.* strengthening the shell) to escape from predators, and predators also adjust their corresponding traits (*e.g.* sharpening the claws) to catch prey. This coevolutionary arms race has been well-documented in many systems [89–97]. Previous theoretical studies focused on the dynamics of the traits of prey and predator groups [98–101], and the structure of the predation network [102]. The coevolution can generally be divided into two modes: the gene-for-gene mode, where predators can catch all prey with traits greater/smaller than a certain value, and the matching-allele or lock-and-key mode, where predators can only eat prey with a specific trait value. Here, we study how coevolution affects the diversity of the KtW model, whose host-specific predation fits in the lock-and-key picture. This work has been published as a part of Ref. [26].

4.1 Stochastic Coevolving-KtW Model

We modify the stochastic generalized KtW model (3.7) by adding in the following two sets of reactions to describe mutations of the prey X_i and predator Y_i, respectively, from strain i to $i \pm 1$.

$$X_i \xrightarrow{\mu_1/2} X_{i\pm1}, \qquad Y_i \xrightarrow{\mu_2/2} Y_{i\pm1}. \tag{4.1}$$

We assume that the mutation rates are strain independent and one individual can mutate into its two neighbor strains with the same rate, $\mu_1/2$ for bacteria and $\mu_2/2$ for viruses. We set the boundary condition to be that mutations out of the index set $\{1, 2, \ldots, m\}$ are ignored. We will refer to Eqs. (3.7) and (4.1) together as the coevolving KtW (CKtW) model.

For sufficiently high mutation rates, the absorbing extinction state in the generalized KtW model can be avoided, in the sense that a strain can reemerge as mutants generated from its neighbor relatives after its population drops to zero. Therefore, mutation can stimulate a flux of population through different strains and promote coexistence.

4.2 Coexistence in the CKtW Model

We quantify the diversity of the system in the CKtW model using the Shannon entropy [40, 41],

$$S = -\sum_{i=1}^{m} f_i \ln f_i, \tag{4.2}$$

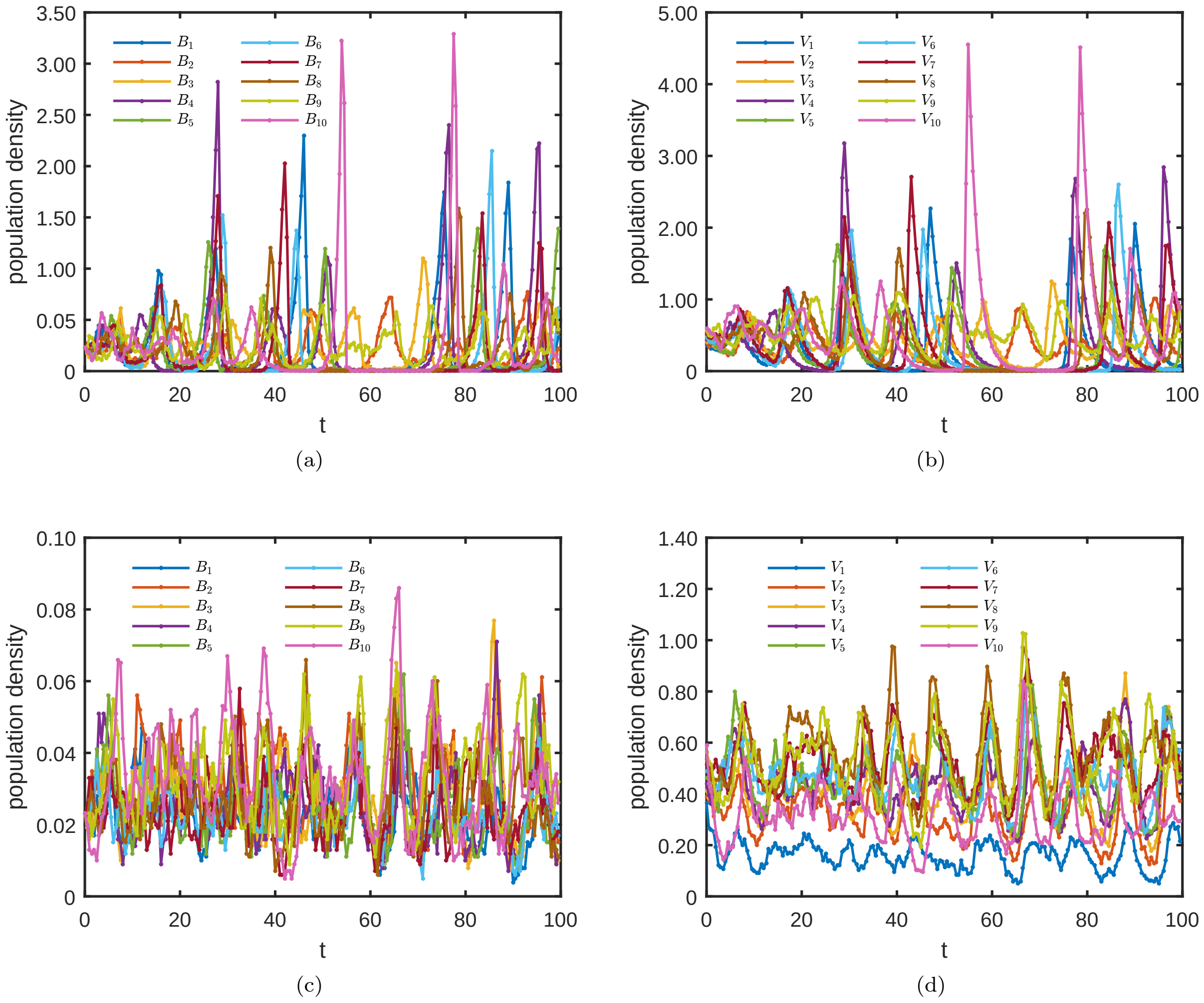
where f_i is the fraction of the ith bacterial (viral) strain in the entire bacterial (viral) community. The expression reaches the maximum, when all strains coexist at their deterministic steady state Eq. (3.9), and the minimum 0, when only one strain exists. We score $S = -1$, if either the bacterial or viral community goes extinct.

We present population density time series in Fig. 4.1, and the dependence of prey diversity on the mutation rates in Fig. 4.2. We set $\mu_1 = \mu_2 \equiv \mu$ for simplification. The diversity of the prey community for a certain simulation replicate is calculated at the end of the diversity time series shown in the inset of Fig. 4.2, after the system has gone through the transient region. We then average the diversity over 100 replicates for each parameter set to obtain the main figure of Fig. 4.2. Although in principle, species in a stochastic system will always go extinct at a time exponentially long depending on the population size [85], this extinction time scale is not relevant in our simulation, and we thus focus on the system state in the long steady region before the eventual collapse.

For small enough mutation rate (population time series not shown), the entire community can become extinct before mutants can emerge, and the system still collapses, demonstrated by the diversity time series of $\mu = 0$ in the inset of Fig. 4.2, as in the generalized KtW model. This corresponds to region I in Fig. 4.2.

For intermediate mutation rates, most strains stay near extinction, driven by the demographic noise, while some mutants can grow to be dominant if they happen to confront only a few predators when first emerging. Subsequently, the predator population expands, feeding on the dominating winners, thus reducing the winner population, and allowing the next dominator to grow. In this way, we see that winner populations spike alternatively in the time series, as in the first row of Fig. 4.1. As illustrated in region II in Fig. 4.2, near the onset μ value of coexistence, the diversity has a large deviation and is very sensitive to the mutation rate. The large deviation is also seen in the diversity time series corresponding to $\mu = 0.015$ in the inset.

For large mutation rate, the coevolution-driven population flow is fast enough to compensate for the



(a)

(b)

(c)

(d)

Figure 4.1: Population density time series in the stochastic coevolving KtW model. The left column is for bacteria and the right for viruses. The system size is $C = 1000$, and the mutation rates are set to be equal, $\mu_1 = \mu_2 \equiv \mu$. Other rates are the same as those in Fig. 3.2. The upper row is obtained from a typical stochastic simulation of the coevolving KtW model. The system size is $C = 1000$, and the mutation rates are $\mu = 0.015$. Populations undergo winner alternation in the presence of the low mutation rate. The lower row is from the same model as in the first row with a high mutation rate $\mu = 1$. Strains coexist, with small fluctuations around the steady state. Figures are adopted from the published work Ref. [26].

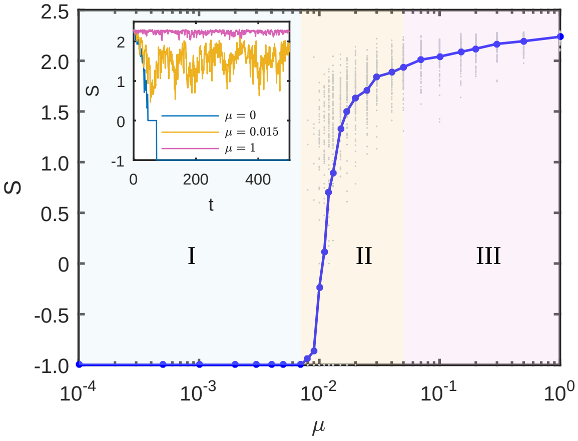


Figure 4.2: The main figure shows the prey diversity S, defined in the main text, as a function of the mutation rate $\mu_1 = \mu_2 \equiv \mu$. For each value of μ, we conduct 100 replicates and calculate the diversity values at the end of the simulations, represented by the gray dots with the blue one being their mean. The inset shows diversity time series at mutation rates from the three regions, with $\mu = 0, 0.015$, and 1, respectively. For this particular set of parameters, the mean-field generalized KtW equations give equal bacterial strain concentration at the steady state, and the maximum diversity in the corresponding CKtW model is $\ln m$. The figure is adopted from the published work Ref. [26].

demographic fluctuations. All strains remain near the steady state, and no one can win over others, as shown by the population time series in the second row of Fig. 4.1. The diversity slowly approaches the maximum, with small deviations, as demonstrated in region III in Fig. 4.2, as well as in the time series of $\mu = 1$ in the inset.

For extremely large mutation rate (figures not shown), we can not view the mutation as a perturbation to the ecological population dynamics. Species populations deviate from the mean-field steady state Eq. (3.9). Specifically, under the boundary condition in our model, in which the mutation out of the species space $\{1, 2, \ldots, m\}$ are effectively individual death, the population leaks through the boundary and eventually reaches zero at extremely large mutation rates.

According to the above discussion, we see three phases of dynamics, as sketched in Fig. 4.3, the extinction phase at low mutation rate, the winner-alternating phase at intermediate mutation rate, and the coexisting phase at high mutation rate.

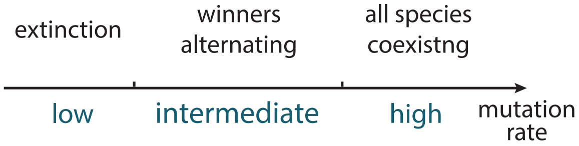


Figure 4.3: A descriptive phase diagram of the dynamics, with the mutation rate as the tuning parameter. The figures is adopted from the published work Ref. [26].

Open system model

In the above analyses, we have pre-assigned a fixed number of predator-prey pairs, m, in the system. A more realistic approach is to let the system be open and evolve by itself to establish however many species there can be.

As mutants take on new traits, the population spreads in the trait space. This expansion usually is associated with a trade-off in the fitness [98]: the further the trait is from the origin, the lower the growth rate becomes. We model this trade-off effect, assuming a 1-D trait space, by setting up M species and assigning the highest birth rate to the species with index $M/2$, and decreasing the birth rate as the species index goes from $M/2$ to 1 and and from $M/2$ to M. The species with index $M/2$ is at the center of the trait space and then is the origin of the trait expansion. Species 1 and M have the lowest birth rates that are almost 0, and further mutation of the two will result in mutants with negative birth rates, which can not grow and are thus excluded from the model. The species space $\{1, 2, \ldots, M\}$ contains all possible species that can potentially exist in the system. However, under conditions of resource limitation, formulated by the competition strength e, only a few with relatively high growth rates, out of M, can eventually be established in the system. The number of species that manage to thrive corresponds to m in the previous models.

Specifically, we set $M = 20$ distinct pairs of preys and predators. The prey birth rates are $\boldsymbol{b} = (0.05,$ $0.15,$ $0.25,$ $0.35,$ $0.45,$ $0.55,$ $0.65,$ $0.75,$ $0.85,$ $0.95,$ $1,$ $0.9,$ $0.8,$ $0.7,$ $0.6,$ $0.5,$ $0.4,$ $0.3,$ $0.2,$ $0.1)$. The 11th species has the highest birth rate and is the origin of trait expansion. Mutations of the first and last species generating mutants with negative birth rates are excluded from the model. Other parameters are $p_i \equiv p = 2$, $\beta_i \equiv \beta = 10$, $d_i \equiv d = 0.5$, and $e_{ij} \equiv e = 1$. The individual level reactions have the same form as Eqs. (3.7) and (4.1), with index $i = 1, 2, \ldots, M$.

In the mean-field situation, the carrying capacity allows the coexistence of 13 pairs, with indices from 5 to 17, while the remaining seven species are forbidden. In the presence of demographic stochasticity, mutants can emerge in the forbidden region in the species space, although they can not develop a significant population size, limited by the high competition with other individuals. The number of coexisting pairs m can be greater that the value 13 predicted by the mean-field calculation, and varies with time.

As shown in the prey population time series in Figs. 4.4 (a) and (b), a small mutation rate results in the alternation of dominating winners, and a large mutation rate generates coexistence with much smaller fluctuations. Figures 4.4 (c) and (d) show the distribution of prey population across all species as a function of the distance to the winner defined as the most abundant strain. The red bar graph stands for a snapshot at a certain moment in the steady region, and the blue one represents the average of the distribution over a long steady interval. It's clear that a winner stands out at low mutation rate, while no one is significantly dominant at high mutation rate. Figure 4.4(e) shows the dependence of the prey diversity, defined as the Shannon entropy, on the mutation rate. The three regions as seen in the CKtW model with fixed number of pairs are recovered.

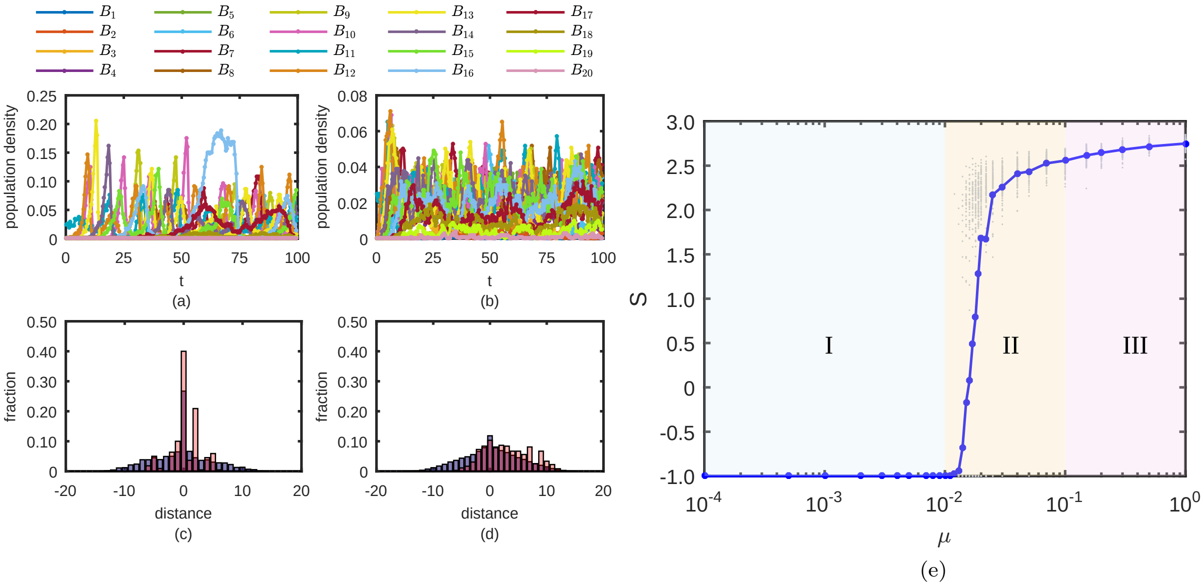


Figure 4.4: Simulation results of the CKtW model with the number of coexisting species limited by the carrying capacity. Parameters used are $\boldsymbol{b} = (0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)$, $p_i \equiv p = 2$, $\beta_i \equiv \beta = 10$, $d_i \equiv d = 0.5$, and $e_{ij} \equiv e = 1$. The system size is $C = 1000$. Population is initiated in the fittest species and expands in the species space. (a) and (b) are prey population time series for small mutation rate $\mu_1 = \mu_2 \equiv \mu = 0.02$ and large mutation rate $\mu_1 = \mu_2 \equiv \mu = 0.5$, respectively. At the small mutation rate, winners alternate with time and the population is localized to the winner species. At the large mutation rate, all species coexist and the population distribution is roughly uniform in the mean-field allowed region, with some mutants leaked into the forbidden species. (c) and (d) show the prey population distribution across the species as a function of index distance from the winner strain, constructed from (a) and (b), respectively. The red bar graph is calculated at $t = 499.5$, after the transient regime. The blue one is the distribution averaged over 501 snapshots uniformly sampled between $t = 249.5$ and $t = 499.5$. For reference, the mean-field steady state predicts that species with indices from 5 to 17 coexist with equal abundance and that other species have zero population. The center bar at 0 distance is the population fraction of the most abundant strain. It's clear that a winner dominates at the low mutation rate but not at the high one. (e) The dependence of prey Shannon entropy on the mutation rate, defined in the same way as in Eq. (4.2). At low mutation, the system collapses due to extinction; at intermediate mutation, diversity increases rapidly with the rate; at high mutation, diversity stays near the maximum given by the deterministic steady state. Figures are adopted from the published work Ref. [26].

4.3 Discussion

In the intermediate and fast mutation regions of the CKtW model, the ecological and evolutionary dynamics are coupled to each other and occur on the same time scale. This type of coupling can most easily be observed in microbial systems, in which organisms have a high mutation frequency [22, 103, 104]. Recent work has shown clearly the existence of genomic islands, where genomes of different strains vary in loci that are believed to be associated with phage resistance [105]. Both host-specific predation and mutation are important in generating the observed diversity of the bacterial genome. The minimal CKtW model can in principle describe the diversity in the above system. Also, the CKtW model may be tested by conducting an experiment of bacterium-phage coevolution, with the coevolving rate being tuned by inducible mutations.

In addition to inevitable simplification of biological details, both the generalized KtW and the coevolving KtW models assume that the system is well mixed, ignoring any spatial dispersion. Consequently, they can not capture the reservoir effect [106] present in an ecosystem, which means that for any local community, organisms in its surrounding environment can move into it, keeping it supplied and refreshed. Specifically, even if a species goes extinct in a local community, it can be reseeded there by the surrounding reservoir. Well-mixed models should be thought of as describing not the entire system, but a much smaller correlation volume, in which local demographic stochasticity can be significant [107, 108].

Also, the models only consider pair-wise predator-prey interaction, while predation in an ecosystem usually happens in food webs, which can result in more complicated dynamics than that in the pair-wise case. Therefore, a spatially extended stochastic model of a food web would serve as a better quantitative description for real ecosystems. Nevertheless, the overall behavior of the generalized KtW and the coevolving KtW models, i.e. extinction driven by demographic noise and coexistence maintained by coevolution, only depend on the quantization of population and the fact that coevolution constantly creates fit mutants to avoid extinction. Therefore, we expect our results to persist in spatial food web models, even though quantitative features, such as time scales, will strongly depend on the specific network architecture and spatial heterogeneity.

4.4 Conclusion

We have proposed a stochastic model that couples the generalized KtW hypothesis and the coevolution of predators and prey. We have shown that the coevolution generically avoids the extinction cascade induced in the KtW framework by the demographic noise and maintains the diversity of the ecosystem, even in the absence of spatial extension. Our results strongly suggest that diversity reflects the dynamical interplay between ecological and evolutionary processes, and is driven by how far the system is from an equilibrium ecological state (as could be quantified by deviations from detailed balance). The surprisingly deep role of demographic stochasticity uncovered here is consistent with earlier demonstrations that individual-level minimal models capture a wide variety of ecological phenomena, including large-amplitude persistent population cycles [1], anomalous phase shifts due to the emergence of mutant sub-populations [22, 109], spatial patterns [2, 107] and even reversals of the direction of selection [110] without requiring overly detailed modeling of inter-species interactions.

Part II

Dynamics at the Genomic Scale

Chapter 5

Introduction to Repetitive Elements in Eukaryotic Genomes

Repetitive elements, or repeats, are DNA sequences that are present in multiple copies in a genome. They occupy a significant amount of the genome, but usually are non-coding. Repetitive elements generally include two categories [12]: tandem repeats located adjacently to each other, and interspersed repeats scattered all over the genome.

Tandem repeats are found in all sequenced species genomes, not only prevalent in eukaryotes, but also widely existing in bacteria [111, 112] and viruses [113, 114]. They were first discovered as the satellite bands of the density-gradient centrifugal separation of DNA molecules [115]. The repeated segment in the sequence is called the unit. Depending on the length of the unit, they are divided into microsatellites (unit length < 10 nucleotides) [116, 117], minisatellites (unit length ≥ 10 nucleotides) [118, 119] and sometimes megasatellites (unit length ≥ 135 nucleotides). The fact that tandem repeats are adjacent to one another indicates their duplication mechanisms to be local. The addition or loss of a full unit usually occurs by strand-slippage replication and recombination [120–124]. Tandem repeats locate both in coding and non-coding regions in the genome. Their expansion and contraction can induce mutations in the coding and regulatory sequences, resulting in both deleterious and advantageous mutations [125, 126]. Tandem repeats have high mutation rates compared with other DNA sequences and are highly polymorphic from one individual to another [127]. This property has promoted tandem repeats to be widely used in DNA fingerprinting, lineage analysis and gene mapping.

Interspersed repeats are usually resultant from transposon activity. Transposons, or transposable elements (TEs), are DNA sequences that can migrate from site to site in the host genome. Some transposons, known as retrotransposons [128, 129], make copies of themselves during the transposition, by reversely transcribing RNA intermediates. This results in the growth of their copy numbers. Other transposons, called DNA transposons [130, 131], excise themselves directly out of the original positions and insert into new sites. Their copy numbers stay unchanged in the transposition process,, but can be increased if a homologous or sister chromosome is used as the template for DNA break repair afterwards. Transposons are regarded as a major driver of adaptation and evolution [132], since they can induce both beneficial and deleterious transformations in the host genome, by inserting into encoding or regulation sequences, or causing misaligned pairing and unequal crossovers of chromosomes. In most cases, the modifications are disadvantageous to the host. For example, L1 elements can insert into the factor VIII gene on the human X chromosome and cause hemophilia A disease [133].

The discovery of repetitive elements provides a resolution to the C-value paradox [10], which reflects the irrelevancy between the amount of DNA in a haploid (the C-value) and the complexity of the organism. Here's an example manifesting the paradox: the maize genome has 2.3 billion base pairs [134], while the human 3.2 billion [11]; still, even with the comparable genome size, human is much more functionally complex than maize. The resolution to the paradox lies in the fact that the genome contains a large amount of non-coding sequences, many of which are repetitive elements.

Then, do these non-coding sequences generate any benefits to the host organisms at all? Why do organisms carry such a big resource-consuming burden in their genome? Although some repetitive elements crucially function as regulatory sequences, for others this seems not to be the case. There has been a long debate over the topic whether the repeats are just parasitic "selfish" "junk" DNA [17, 135–137] or whether they have functions and evolutionary roles that are yet poorly understood [138, 139]. We will attempt to explore this issue by looking at the diversity of repetitive elements in Chapter 8.

In the rest of this chapter, I will focus on the transposable elements, introduce a pair of retrotransposons in the human genome, and review previous models to describe transposon dynamics.

5.1 Transposable Elements

Transposable elements, also known as "jumping genes", were discovered in 1950 in the maize genome [140, 141]. They widely exist in most genomes in all three domains of life. Especially they take up a large fraction of eukaryotic genomes. For example, the Initial Human Genome Project has revealed that roughly 45% of the human genomic sequence originates from transposons [11].

There are two classifications of transposons following two criteria [13]: autonomous transposons *vs.* non-autonomous transposons, and DNA transposons *vs.* retrotransposons.

Autonomous transposons are elements that encode all needed enzymes and thus have a complete mechanism system for the transposition. Non-autonomous transposon cannot encode all necessary enzymes and must rely on the enzymes produced by other elements.

DNA transposons cut themselves out of the genome and reintegrate to the genome at other sites. This is a "cut-and-paste" route. Transposons in bacteria and archaea mostly belong to this group. Retrotransposons produce RNA intermediates first and undergo a reverse transcription to complete the transposition. This "copy-and-paste" route results in an increase of the element number in the genome. The retrotransposons can be further divided into two types: LTR-transposons that have a long terminal repeat (LTR) structure and non-LTR transposons that do not. Retrotransposons primarily exist in eukaryotes. The counterpart in bacteria is the group II intron retroelements [142, 143]. Group II introns usually perform an accurate insertion into a very specific target gene, a process known as retrohoming. In the absence of the target gene, they can also insert into a random site, with a much lower rate, and complete a retrotransposition [144]. Group II introns are found in roughly only 30% of sequenced bacteria and usually have very small copy numbers, rarely exceeding 10 copies [145].

Why do retroelements not proliferate in bacteria as the retrontransposons do in eukaryotes? To address this question, we collaborate with Professor Thomas Kuhlman's group. Together, we have designed and conducted experiments and theoretical modeling on the growth defect induced by retroelements in the bacterial genome. We induced human retrontransposons into bacteria cells and observed that the invasion led to significant reduction in the population growth rate and eventually cell death. Our model showed that this negative impact prevented the prevalence of retroelements in the bacterial genomes. We hypothesized that eukaryotes must have evolved methods to get around the growth defect associated with the transposons. This had indications on the emergence of spliceosome, nuclear membrane and linear chromosomes [28]. Although I participated in this work, the main results were obtained by K. Michael Martini and so are not reported in this thesis.

To fight against the expansion of transposons, cells have evolved several methods of defense [146], including RNA interference [147–149], chromatin modifications and DNA methylation [150, 151].

5.2 LINE-1 and Alu Elements in the Human Genome

Among the non-LTR retrotransposons, we are specially interested in the following two families: the long interspersed nuclear elements (LINEs) and the short interspersed nuclear elements (SINEs). The former are autonomous, and the latter non-autonomous, with SINEs relying on the machinery encoded by LINEs to spread.

LINEs typically are over 5000 base pairs (bp) in length, while SINEs are usually shorter than 500 bp [152]. We take LINE-1 (L1) and Alu elements as representatives from the two families, respectively. They are both very prevalent in primates. In the human genome, L1 is the only active autonomous transposon. Among the 500000 L1 copies, which take up 17% of the genome [11], only 7000 copies are complete and only

80–100 are active [153]. Alu elements contribute to 11% of the human genome with about 1500000 copies [11]. L1 elements help SINEs like Alu and SVA to transpose [154]. We focus on the L1-Alu pair and discuss the details of their interaction.

5.2.1 Structure of Human LINE-1 Element

A complete L1 sequence consists of four segments[155], as shown in Fig. 5.1(a): an untranslated region (UTR) containing a pol II promoter [156] and an antisense promoter [157], two open reading frames ORF1 and ORF2 [133], and a poly-adenine (poly-A) region. ORF1 encodes a RNA-binding protein. ORF2 encodes a protein with both DNA endonuclease [158] and reverse transcriptase [159, 160]. These proteins are necessary and sufficient to complete the transposition [161]. Recent research [162] has revealed the existence of an antisense ORF0 upstream of ORF1, but its exact function remains unclear.

5.2.2 Structure of Human Alu Element

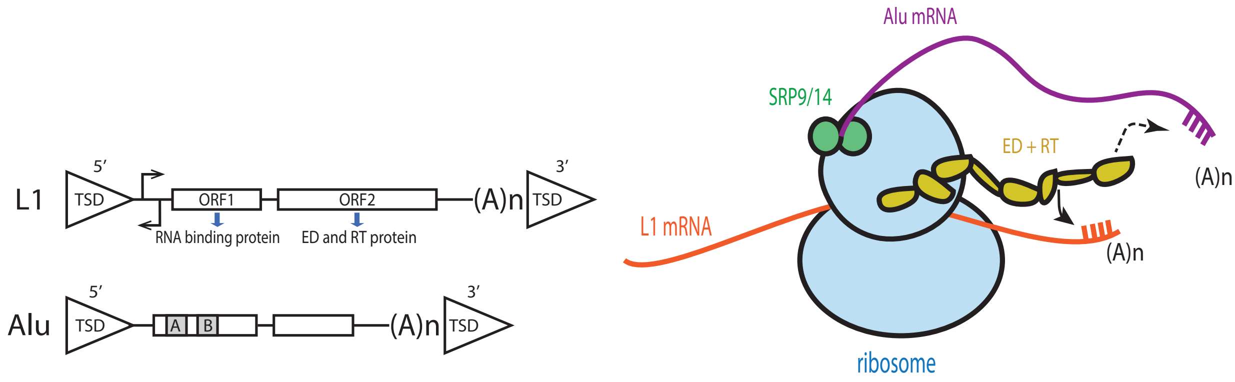
An Alu element does not have any ORFs and cannot encode proteins to complete the transposition. It has a pol III promoter, a non-coding segment containing two monomers and a poly-A region [163].

Alu elements are believed to originate from the 7SL-RNA [164–167], which is the RNA component of the sequence recognition particle (SRP) that leads the translocation of nascent peptides [168, 169]. The 7SL-RNA contains an Alu-domain and an S-domain. The Alu-domain combines with two sequence recognition proteins named SRP9 and SRP14 to form a complex [170]. This complex attaches to the ribosome sequence recognition factor binding site. S-domain then binds with the recognition sequence on the nascent peptide. This entire complex of SRP, ribosome and peptide then is targeted to the signal site on the endoplasmic recultum. In this way, with the Alu-domain held to the ribosome, and the S-domain to the peptide, the 7SL-RNA helps translocate the nascent peptide to the endoplasmic recultum for further processing.

Alu elements emerged by losing the 7SL-RNA S-domain and acquiring a tandem Alu-domain followed with a poly-A tail [164]. As a result, they reserve the ability to form a ribonucleoprotein particle (RNP) with SRP9 and SRP14 and to cling to the ribosome. This provides them with the opportunity of hijacking the L1 proteins at the assembly factory [171].

5.2.3 Dependence of Alu on LINE-1

As sketched in Fig. 5.1(b), when a protein is produced at a ribosome coded by an L1 mRNA, it tends to bind with that particular mRNA, presumably by recognizing its poly-A tail [172], and later reversely transcribes it into the genome. This is known as the *cis*-preference of L1 elements [173]. However, if an Alu mRNA



(a)

(b)

Figure 5.1: (a) The structure of L1 and Alu elements. L1 has a pol II promoter (the right-pointing arrow) and an antisense promoter (the left-pointing arrow) followed by two open reading frames (ORFs) encoding, respectively, a RNA binding protein and a protein that consists of an endonuclease (ED) and a reverse transcriptase (RT). Alu is composed of two non-coding monomers, with the left one bearing A and B boxes (the shaded area in the figure) as the pol III promoter. L1 and Alu elements share similar poly-A tails and are both flanked by target site duplicates (TSDs). (b) The *cis* and *trans* effects of L1 elements. When an ED+RT protein is translated at the ribosome, it *cis*-preferentially attaches to the L1 mRNA that codes it, indicated by the solid arrow. An Alu mRNA can combine with two signal recognition proteins SRP9 and SPR14, and then attach to the ribosome. The nascent ED+RT protein then can *trans*-bind to the Alu mRNA, which has a similar poly-A tail (indicated by the dashed arrow), presumably with a similar probability to that of binding to the L1 mRNA. Figures are adopted from the published work Ref. [27].

attaches to the same ribosome, then it can bind with the nascent protein by faking the L1 mRNA poly-A tail [174]. In this way, Alu elements steals the transposition machinery designed by L1 elements [171, 175]. This is known as the *trans*-effect of L1 elements [173].

Due to various regulation pathways, transposon activity rate is low, usually on the order of $10^{-4} \sim 10^{-6}$ per element per generation [176, 177]. In the human genome, L1 transposition events happen once in every 20 - 200 births, and Alu events occur once in every 20 births [178–180].

5.3 Modeling the Transposon Dynamics

Several theoretical approaches have been proposed to study the dynamics of transposons. Population genetics models [181–186] were first developed to describe the equilibrium distribution of transposons in a population. Recent developments view the genome as an ecosystem, with genetic elements of different types playing the role of individuals from different species [16, 187–191]. In the case of non-autonomous transposons, a mean-field model [188, 189] describes their parasitic relationship with an autonomous transposon, viewing the transposons at predator and prey species. A decaying oscillatory mode has being predicted for a certain parameter range.

However, there are two major drawbacks in these models. First, they did not account for the molecular level interactions between transposable elements. The dynamic behavior turns out to be sensitively dependent on these details. Second, the models used continuous variables for the element copy numbers and ordinary differential equations to describe the system, which could not handle the fact that copy numbers are finite integers. As known in ecology as the demographic noise, finite population size usually causes fluctuations in the population dynamics, preventing it from reaching the mean-field result [1]. To study such a system, a stochastic model, instead of the mean-field one, should be used. The copy number fluctuations are large in a cell, since the number of active (expressed) transposons is usually of order ten to a hundred [153]. Thus, the next generation of transposon models needs to take into account molecular details and stochasticity. We will develop and solve such a stochastic model in Chapter 6, taking the L1-Alu pair as a model system.

Chapter 6

Stochastic Predator-Prey Dynamics of Transposons in the Human Genome

I describe in this chapter a minimal individual-level model for the population dynamics of a pair of autonomous and non-autonomous transposons, using LINE-1 and Alu elements in the human genome as an model system. We begin with interactions between the transposon pair, and then use techniques from statistical mechanics to derive and solve stochastic differential equations. Our model predicts that demographic stochasticity generates persistent and noisy oscillations in the copy numbers of the transposons, similar to the predator-prey quasi-cycles, with a characteristic time scale that is much longer than the cell replication time, indicating that the state of the predator-prey oscillator is stored in the genome and transmitted to successive generations. Our work builds upon recent results that have shown how demographic stochasticity in ecosystems, where population size is integer-valued and locally finite, can lead to minimal models of persistent population cycles [1] or spatial patterns [2, 107, 192–194] without extra assumptions about the details of predation. This work has been published as Ref. [27].

6.1 Minimal Model for Transposon Dynamics

In Chapter 5, I have introduced the detailed interactions on the molecular level between LINE-1 and Alu elements. Here we discard all details about how proteins are made and how complexes are formed, and develop a minimal model for the dynamics of LINE-1 and Alu. Although this model is developed for the L1-Alu pair, we believe that the idea and the qualitative results can be applied to all autonomous/non-autonomous pairs with minor modifications.

Reactions, with the corresponding forward rates, describing behaviors of individuals from each chemical species are shown in Eq. (6.1), where L stands for an active LINE, S for an active SINE, and R_L for the complex of the ribosome, LINE mRNA and nascent protein. $\emptyset$ stands for null. Deactivated transposons do not participate in the transposition events and thus are excluded from the model.

$$L \xrightarrow{b_R} L + R_L, \qquad R_L \xrightarrow{b_L} L, \qquad R_L + S \xrightarrow{b_S/V} 2S, \tag{6.1a}$$

$$R_L \xrightarrow{d_R} \emptyset, \qquad L \xrightarrow{d_L} \emptyset, \qquad S \xrightarrow{d_S} \emptyset. \tag{6.1b}$$

An L element encodes the complex R_L at the rate b_R. The complex R_L reversely transposes to produce a new L element at the rate b_L, if there is no interruption. S element hijacks the complex R_L to duplicate itself at the rate b_S/V, where V is the system size. The complex R_L decays at the rate d_R. L and S elements are deactivated, at the rates d_L and d_S, respectively. We assume the system is well mixed because the mixing of reactants occurs constantly within the cell lifetime, and so is faster than the reactions.

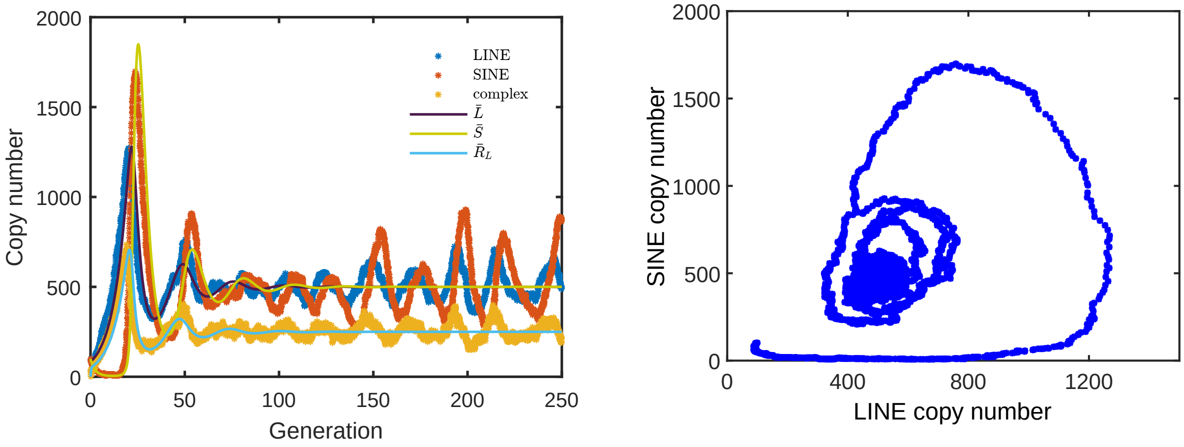
6.2 Numerical Simulation

Eqs. (6.1) look like and can actually be understood as chemical reactions. Since each reaction fires independently and randomly, the evolution of the system starting with a certain initial condition is a stochastic process. Under the assumption of a well-mixed system, we can use the Gillespie algorithm [86] to exactly evolve the stochastic system over time.

For the model Eq. (6.1), the time series of element copy numbers obtained from Gillespie simulations are plotted in Fig. 6.1(a). The solid lines are the numerical integration, using the Runge-Kutta method (RK4), of the deterministic mass rate equation of the reactions, which will be discussed in detail in a later section. In the simulation, after the initial transition, the copy numbers fluctuate persistently around the deterministic values. A noisy periodic oscillation is clearly present. The circular envelope of the trajectory on the L-S plane shown in Fig. 6.1(b) indicates a phase difference of roughly $\pi/2$, with SINE lagging LINE.

These noisy cycles and $\pi/2$-phase lag feature can be qualitatively understood in the following way. When the SINE copy number increases, LINEs will have more complexes stolen and their transposition rate thus is reduced; with the unchanged deactivation rate , the LINE copy number decreases. Consequently, fewer complexes will be made and SINE transposition rate drops, which also causes the SINE copy number to shrink. Now more complexes are available; thus LINE transposition rate recovers and so does its copy number. SINEs then follow and grow back in the copy number soon afterwards. This cycle persistently goes on. This qualitative picture is similar to the predator-prey interaction in the ecosystem, with SINEs being the predators hunting LINEs.

In the following sections, I will analyze in detail the features of the oscillation, and explain that the noisy cycles of the SINE and LINE pair in this model are stimulated by the demographic stochasticity and has



(a)

(b)

Figure 6.1: Results of a typical stochastic simulation with illustrative parameters $b_R = 2$, $b_L = 1$, $b_S = 1$, $d_R = 2$, $d_L = 0.5$, $d_S = 0.5$, and the system size $V = 500$. (a) The copy numbers of active LINEs, SINEs and ribosome/L-mRNA/protein complexes as a function of time, in the unit of a cell generation. Solid lines are obtained by evolving the deterministic equations and show oscillatory decay toward steady values. Discrete dots are generated by a numerical simulation with the same parameters. Copy numbers fluctuate around the deterministic steady state, demonstrating quasi-cycles with period ~ 25 generations. Demographic noise induces quasi-cycles by constantly stimulating the deterministic oscillation mode. (b) The trajectory on the L-S plane. The circular envelope indicates a phase difference of roughly $\pi/2$. Figures are adopted from the published work Ref. [27].

the same origin as the quasi-cycles in the predator-prey system.

6.3 Analytical Calculation

In this section, I will derive equations for the stochastic system Eq. (6.1), and calculate both the mean field and stochastic features of the dynamics. The analytical results are compared to the numerical simulations.

Let the copy numbers of active LINEs, SINEs and complexes be N_L, N_S and N_R. We can write down the master equation for the model Eq. (6.1), about the probability $P(N_L, N_S, N_R)$ of the system being in the state (N_L, N_S, N_R).

$$\begin{aligned}\frac{d}{dt}P(N_L, N_S, N_R) =& \Big\{(\mathcal{E}^-_{R_L} - 1)N_L b_R + (\mathcal{E}^+_{R_L}\mathcal{E}^-_L - 1)N_R b_L \\ & (\mathcal{E}^+_{R_L}\mathcal{E}^-_S - 1)N_R N_S \frac{b_S}{V} + (\mathcal{E}^+_{R_L} - 1)N_R d_R \\ & (\mathcal{E}^+_L - 1)N_L d_L + (\mathcal{E}^+_S - 1)N_S d_S\Big\}P \end{aligned} \tag{6.2}$$

The raising and lowering operators $\mathcal{E}^{\pm}_X$ change the copy number N_X of species X by ± 1 respectively, where

X can be L, S and R_L. The above master equation is an exact interpretation of the stochastic system, and gives the time evolution of the probability distribution in the state space.

The state status represented by (N_L, N_S, N_R) is discrete, since the copy numbers can only be integers. In order to make it mathematically more tractable, we define the copy number concentrations L, S and R_L, respectively, so that with the system size being V, $N_L = VL$, $N_S = VS$ and $N_R = VR_L$. Using (L, S, R_L) as the state status, we can rewrite the master equation as follows, about the probability $\mathcal{P}(L, S, R_L)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{P}(L,S,R_L) =& V\Big\{(\mathcal{E}^-_{R_L}-1)b_R L + (\mathcal{E}^+_{R_L}\mathcal{E}^-_L - 1)b_L R_L \\ &+ (\mathcal{E}^+_{R_L}\mathcal{E}^-_S - 1)b_S R_L S + (\mathcal{E}^+_{R_L}-1)d_R R_L \\ &+ (\mathcal{E}^+_L - 1)d_L L + (\mathcal{E}^+_S - 1)d_S S\Big\}\mathcal{P},\end{aligned} \tag{6.3}$$

with the raising and lowering operators given by

$$\mathcal{E}^{\pm}_X f(X) \equiv f(\frac{N_X \pm 1}{V}) \approx f(X) \pm \frac{1}{V}\partial_X f + \frac{1}{2V^2}\partial^2_X f, \tag{6.4}$$

where f is an arbitrary function of the concentration X, and X stands for L, S or R_L.

Substituting the expansions of operators Eq. (6.4) into the master equation Eq. (6.3), and saving terms up to order $O(V^{-1})$, we obtain the following Fokker-Planck equation of the model.

$$\begin{aligned}\frac{d}{dt}\mathcal{P}(L,S,R_L) = & \partial_{R_L}\left(-b_R L + b_L R_L + b_S R_L S + d_R R_L\right)\mathcal{P} \\ &+ \partial_L\left(-b_L R_L + d_L L\right)\mathcal{P} + \partial_S\left(-b_S R_L S + d_S S\right)\mathcal{P} \\ &+ \frac{1}{2V}\partial^2_{R_L}\left(b_R L + b_L R_L + b_S R_L S + d_R R_L\right)\mathcal{P} \\ &+ \frac{1}{2V}\partial^2_L\left(b_L R_L + d_L L\right)\mathcal{P} + \frac{1}{2V}\partial^2_S\left(b_S R_L S + d_S S\right)\mathcal{P} \\ &+ \frac{1}{2V}\partial_{R_L}\partial_L\left(-2b_L R_L\right)\mathcal{P} + \frac{1}{2V}\partial_{R_L}\partial_S\left(-2b_S R_L S\right)\mathcal{P}\end{aligned} \tag{6.5}$$

The above equation is non-linear with multiplicative noise, in the continuous state space. We follow a further standard procedure, known as the van Kampen system size expansion [195], to simplify it by expanding the state status into the deterministic part $\bar{L}$, $\bar{S}$ and $\bar{R}_L$, and the stochastic part, ξ, η and θ.

$$L = \bar{L} + \frac{\xi}{\sqrt{V}}, \quad S = \bar{S} + \frac{\eta}{\sqrt{V}}, \quad R_L = \bar{R}_L + \frac{\theta}{\sqrt{V}}. \tag{6.6}$$

Now the state status can be represented by (ξ, η, θ). Let

$$\Pi(\xi, \eta, \theta) = \mathcal{P}(L, S, R_L). \tag{6.7}$$

Substitute the expansion Eq. (6.6) into the Fokker-Planck equation Eq. (6.5) and compare different orders of V on both sides.

On the left hand side,

$$\frac{d}{dt}\mathcal{P} = \partial_t \Pi - \sqrt{V}\frac{d\bar{L}}{dt}\partial_\xi \Pi - \sqrt{V}\frac{d\bar{S}}{dt}\partial_\eta \Pi - \sqrt{V}\frac{d\bar{R}_L}{dt}\partial_\theta \Pi. \tag{6.8}$$

On the right hand side (RHS),

$$\begin{aligned}\text{RHS} = &\sqrt{V}\left(-b_R\bar{L} + b_L\bar{R}_L + b_S\bar{S}\bar{R}_L + d_R\bar{R}_L\right)\partial_\theta\Pi + \sqrt{V}\left(-b_L\bar{R}_L + d_L\bar{L}\right)\partial_\xi\Pi + \sqrt{V}\left(-b_S\bar{S}\bar{R}_L + d_S\bar{S}\right)\partial_\eta\Pi \\ &+ \partial_\theta\left(-b_R\xi + b_L\theta + b_S\bar{R}_L\eta + b_S\bar{S}\theta + d_R\theta\right)\Pi + \partial_\xi\left(-b_L\theta + d_L\xi\right)\Pi + \partial_\eta\left(-b_S\bar{R}_L\eta - b_S\bar{S}\theta + d_S\eta\right)\Pi \\ &+ \frac{1}{2}\partial_\theta^2\left(b_R\bar{L} + b_L\bar{R}_L + b_S\bar{S}\bar{R}_L + d_R\bar{R}_L\right)\Pi + \frac{1}{2}\partial_\xi^2\left(b_L\bar{R}_L + d_L\bar{L}\right)\Pi + \frac{1}{2}\partial_\eta^2\left(b_S\bar{S}\bar{R}_L + d_S\bar{S}\right)\Pi \\ &+ \frac{1}{2}\partial_\theta\partial_\xi\left(-2b_L\bar{R}_L\right)\Pi + \frac{1}{2}\partial_\theta\partial_\eta\left(-2b_S\bar{S}\bar{R}_L\right)\Pi + O(V^{-1/2})\end{aligned} \tag{6.9}$$

Matching the two sides to order $O(\sqrt{V})$, we obtain the deterministic, or mean field, equations.

$$\frac{d\bar{L}}{dt} = b_L\bar{R}_L - d_L\bar{L}, \tag{6.10a}$$

$$\frac{d\bar{S}}{dt} = b_S\bar{S}\bar{R}_L - d_S\bar{S}, \tag{6.10b}$$

$$\frac{d\bar{R}_L}{dt} = b_R\bar{L} - b_L\bar{R}_L - b_S\bar{S}\bar{R}_L - d_R\bar{R}_L. \tag{6.10c}$$

By matching $O(1)$ terms, we obtain the linearized Fokker-Planck equation about $\Pi(\xi, \eta, \theta)$.

$$\begin{aligned}\frac{\partial\Pi}{\partial t} = &\partial_\theta\left(-b_R\xi + b_L\theta + b_S\bar{R}_L\eta + b_S\bar{S}\theta + d_R\theta\right)\Pi + \partial_\xi\left(-b_L\theta + d_L\xi\right)\Pi + \partial_\eta\left(-b_S\bar{R}_L\eta - b_S\bar{S}\theta + d_S\eta\right)\Pi \\ &+ \frac{1}{2}\partial_\theta^2\left(b_R\bar{L} + b_L\bar{R}_L + b_S\bar{S}\bar{R}_L + d_R\bar{R}_L\right)\Pi + \frac{1}{2}\partial_\xi^2\left(b_L\bar{R}_L + d_L\bar{L}\right)\Pi + \frac{1}{2}\partial_\eta^2\left(b_S\bar{S}\bar{R}_L + d_S\bar{S}\right)\Pi \\ &+ \frac{1}{2}\partial_\theta\partial_\xi\left(-2b_L\bar{R}_L\right)\Pi + \frac{1}{2}\partial_\theta\partial_\eta\left(-2b_S\bar{S}\bar{R}_L\right)\Pi\end{aligned} \tag{6.11}$$

Using Itô's Lemma [196], we can write down the linearized Langevin equations for ξ, η and θ, directly from the first order derivative terms in the above Fokker-Planck equation.

$$\frac{d\xi}{dt} = b_L\theta - d_L\xi + r(t), \tag{6.12a}$$
$$\frac{d\eta}{dt} = b_S\bar{R}_L\eta + b_S\bar{S}\theta - d_S\eta + s(t), \tag{6.12b}$$
$$\frac{d\theta}{dt} = b_R\xi - b_L\theta - b_S\bar{R}_L\eta - b_S\bar{S}\theta - d_R\theta + h(t). \tag{6.12c}$$

$r(t)$, $s(t)$ and $h(t)$ are noises in ξ, η and θ, respectively. The correlations between these noises are given by the second order derivative terms in the Fokker-Planck equation Eq. (6.11).

$$\langle h(t)h(t')\rangle = \delta(t-t')(b_R\bar{L} + b_L\bar{R}_L + b_S\bar{S}\bar{R}_L + d_R\bar{R}_L), \tag{6.13a}$$
$$\langle r(t)r(t')\rangle = \delta(t-t')(b_L\bar{R}_L + d_L\bar{L}), \tag{6.13b}$$
$$\langle s(t)s(t')\rangle = \delta(t-t')(b_S\bar{S}\bar{R}_L + d_S\bar{S}), \tag{6.13c}$$
$$\langle h(t)r(t')\rangle = \delta(t-t')(-b_L\bar{R}_L), \tag{6.13d}$$
$$\langle h(t)s(t')\rangle = \delta(t-t')(-b_S\bar{S}\bar{R}_L), \tag{6.13e}$$
$$\langle r(t)s(t')\rangle = 0. \tag{6.13f}$$

The above linear Langevin equations describe the fluctuations of concentrations around the determinicstic trajectory.

In the rest of the section, I will analyze in detail the behaviors of the deterministic and stochastic parts. Together, they give the complete dynamics of the model.

6.3.1 Steady States of the Deterministic Equations

The deterministic equations Eq. (6.10) have three steady states, given below, which satisfy $d\bar{L}/dt = 0$, $d\bar{S}/dt = 0$ and $d\bar{R}_L/dt = 0$.

$$(\bar{L}^*, \bar{S}^*, \bar{R}_L^*)_1 = (0, 0, 0), \tag{6.14a}$$
$$(\bar{L}^*, \bar{S}^*, \bar{R}_L^*)_2 = (L_2, 0, R_{L2}), \tag{6.14b}$$
$$(\bar{L}^*, \bar{S}^*, \bar{R}_L^*)_3 = (L_3, S_3, R_{L3}). \tag{6.14c}$$

We are interested in the third one with all copy numbers being non-zero. It's straightforward to derive the expression of the steady state.

$$L_3 = \frac{d_S b_L}{b_S d_L}, \tag{6.15a}$$

$$S_3 = \frac{b_R b_L - b_L d_L - d_R d_L}{b_S d_L}, \tag{6.15b}$$

$$R_{L3} = \frac{d_S}{b_S}. \tag{6.15c}$$

We may further require all the three copy numbers to be positive to be physically meaningful.

We first analyze the linear stability of the above state, as the demographic noise constantly perturbs the system away from it. The Jacobian matrix of Eq. (6.10) evaluated at $(\bar{L}^*, \bar{S}^*, \bar{R}_L^*)_3$ is

$$J_3 = \begin{pmatrix} -b_S S_3 - d_R - b_L & b_R & -b_S R_{L3} \\ b_L & -d_L & 0 \\ b_S S_3 & 0 & 0 \end{pmatrix}, \tag{6.16}$$

with the following characteristic equation

$$\lambda^3 + \left(\frac{b_L b_R}{d_L} + d_L\right)\lambda^2 + \left[b_L\left(b_R - d_L\right) - d_L d_R\right]\frac{d_S}{d_L}\lambda + d_S\left[b_L\left(b_R - d_L\right) - d_L d_R\right] = 0, \tag{6.17}$$

where λ is the eigenvalue.

For this degree-3 polynomial equation, we can apply the Routh-Hurwitz criterion [87, 88] to find the condition for it to have all three roots with negative real parts. The Routh-Hurwitz table for Eq. (6.17) is

α_3	α_1
α_2	α_0
β_1	0
γ_1	0

$$\alpha_3 = 1, \qquad \alpha_2 = \frac{b_L b_R}{d_L} + d_L, \qquad \alpha_1 = \left[b_L\left(b_R - d_L\right) - d_L d_R\right]\frac{d_S}{d_L}, \tag{6.18a}$$

$$\alpha_0 = d_S\left[b_L\left(b_R - d_L\right) - d_L d_R\right], \qquad \beta_1 = \frac{\alpha_2\alpha_1 - \alpha_3\alpha_0}{\alpha_2}, \qquad \gamma_1 = \frac{\beta_1\alpha_0 - \alpha_2 \cdot 0}{\beta_1}. \tag{6.18b}$$

α_i, $i = 0, 1, 2, 3$, is the coefficient of the λ^i term. Elements in row n are obtained by cross-multiplying elements in rows $n-1$ and $n-2$, for $n = 3, 4$. According to the criterion, the sufficient and necessary condition for all three roots of Eq. (6.17) to have negative real parts is

$$\alpha_3 > 0, \quad \alpha_2 > 0, \quad \beta_1 > 0, \quad \gamma_1 > 0. \tag{6.19}$$

It is straightforward to prove that the sufficient and necessary condition for the above inequalities to be true is

$$L_3 > 0, \quad S_3 > 0 \tag{6.20}$$

In other words, the physically meaningful coexistence steady state is always stable. Referring to Eq. (6.15), we obtain the following stable condition

$$b_R b_L - b_L d_L - d_R d_L > 0 \tag{6.21}$$

6.3.2 Oscillatory Mode of the Deterministic Trajectory

The linear stability analysis demonstrates that the physically meaningful steady state with positive copy numbers of the deterministic equation Eq. (6.10) is always exponentially stable. Still, it's not clear from the calculation whether the steady state is a node or a focus, which is determined by whether the imaginary parts of the Jacobian matrix eigenvalues are zero or not.

Here, we reduce the original three-body deterministic equations to two-body ones and calculate their eigenvalues of the linear stability matrix. This reduction is done via an adiabatic limit, similar to the derivation of the Michaelis-Menten mechanism [197]. Specifically, we set $d\bar{R}_L/dt = 0$, so that

$$\bar{R}_L = \frac{b_R \bar{L}}{b_L + b_S \bar{S} + d_R} \tag{6.22}$$

Let

$$\alpha = \frac{b_L + d_R}{b_R}, \quad \beta = \frac{b_S}{b_R}, \tag{6.23}$$

then

$$\bar{R}_L = \frac{\bar{L}}{\alpha + \beta \bar{S}}. \tag{6.24}$$

Substitute the above expression into Eq. (6.10) and obtain the two-body equations as follows.

$$\frac{d\bar{L}}{dt} = \frac{b_L \bar{L}}{\alpha + \beta \bar{S}} - d_L \bar{L}, \tag{6.25a}$$

$$\frac{d\bar{S}}{dt} = \frac{b_S \bar{S} \bar{L}}{\alpha + \beta \bar{S}} - d_S \bar{S}. \tag{6.25b}$$

Generally, setting $d\bar{R}_L/dt = 0$ and substituting the steady solution of $\bar{R}_L$ into other equations require the dynamics of $\bar{R}_L$ to be much faster than those of other chemical species so that it can be viewed as reaching equilibrium instantaneously. This separation of time scales is not necessarily present in our model. Still, we use the Michaelis-Menten mechanism as an approximation and will show later that it gives reasonable results.

The coexistence steady state of the reduced equations are $(\bar{L}^*, \bar{S}^*) = (L_3, S_3)$. The Jacobian matrix at the steady state is

$$J = \begin{pmatrix} 0 & -\frac{b_L \beta L_3}{(\alpha+\beta S_3)^2} \\ \frac{b_S S_3}{\alpha+\beta S_3} & -\frac{b_S \beta L_3 S_3}{(\alpha+\beta S_3)^2} \end{pmatrix} \equiv \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix}, \tag{6.26}$$

where

$$\tilde{A} = 0, \quad \tilde{B} = -\frac{b_L \beta L_3}{(\alpha + \beta S_3)^2}, \quad \tilde{C} = \frac{b_S S_3}{\alpha + \beta S_3}, \quad \tilde{D} = -\frac{b_S \beta L_3 S_3}{(\alpha + \beta S_3)^2}. \tag{6.27}$$

The characteristic equation is

$$\lambda^2 - (\tilde{A} + \tilde{D})\lambda + (\tilde{A}\tilde{D} - \tilde{B}\tilde{C}) = 0, \tag{6.28}$$

with eigenvalues

$$\lambda = \frac{1}{2}\left[(\tilde{A} + \tilde{D}) \pm \sqrt{(\tilde{A} + \tilde{D})^2 - 4(\tilde{A}\tilde{D} - \tilde{B}\tilde{C})}\right]. \tag{6.29}$$

Both eigenvalues have negative real parts for any physically meaningful states with $L_3 > 0$ and $S_3 > 0$, and the corresponding steady state thus is always exponentially stable. This result agrees with the analysis on the three-body equation.

When $(\tilde{A} + \tilde{D})^2 < 4(\tilde{A}\tilde{D} - \tilde{B}\tilde{C})$, which corresponds to the following condition

$$d_S < \frac{4 b_L d_R d_L}{b_L(b_R - d_L) - d_R d_L}, \tag{6.30}$$

the eigenvalues have nonzero imaginary parts. In this case, when perturbed away from the steady state, $\bar{L}$ and $\bar{S}$ will undergo an oscillatory decay back. The decay rate is equal to the absolute value of the real parts of the eigenvalues, with the characteristic time being

$$\tau = \frac{2}{\tilde{A} + \tilde{D}}. \tag{6.31}$$

The oscillation angular frequency is equal to the absolute value of the imaginary parts, with the period being

$$\begin{aligned} T =& \frac{4\pi}{\sqrt{4(\tilde{A}\tilde{D} - \tilde{B}\tilde{C}) - (\tilde{A} + \tilde{D})^2}} \\ =& \frac{4\pi}{\sqrt{\frac{4d_S d_L}{b_R}\left[b_R - d_L - \frac{d_L}{b_L} d_R\right] - \frac{d_S^2}{b_R^2}\left[b_R - d_L - \frac{d_L}{b_L} d_R\right]^2}}. \end{aligned} \tag{6.32}$$

In Fig. 6.1(a), the solid lines are obtained by numerically integrating the deterministic equations Eq. (6.10), using the Runge-Kutta method (RK4). For the demonstrated parameter set, the system starts with an arbitrary initial condition, decays with oscillations toward the steady state, and stays at the steady state afterwards. This is in agreement with the above linear stability analyses. Furthermore, we have verified numerically that the imaginary part of the two-body linear stability matrix eigenvalues provides a reasonable estimate for the angular frequency of the stochastic cycles. Specifically, for the parameters used to generate Fig. 6.1 and Fig. 6.2, the eigenvalue imaginary part is 0.2330, and the Gillespie simulation gives a peak angular frequency of 0.23 generation^{-1}, measured from the oscillation power spectra shown in Fig. 6.2.

6.3.3 Power Spectra of the Fluctuations

In this subsection, I will focus on the stochastic part of the system and calculate the power spectra of the fluctuations.

Starting with the linearized Langevin equations Eq. (6.12), we perform on both sides the Fourier transform defined below:

$$\tilde{\chi}(\omega) = \frac{1}{2\pi}\int \chi(t) e^{-i\omega t}\mathrm{d}t, \quad \chi(t) = \int \tilde{\chi}(\omega) e^{i\omega t}\mathrm{d}t, \tag{6.33}$$

where $\chi(t)$ stands for an arbitrary function and $\tilde{\chi}(\omega)$ stands for its Fourier transform. We obtain the following set of linear equations.

$$i\omega\tilde{\xi}(\omega) = b_L\tilde{\theta} - d_L\tilde{\xi} + \tilde{r}(\omega), \tag{6.34a}$$

$$i\omega\tilde{\eta}(\omega) = b_S\bar{R}_L\tilde{\eta} + b_S\bar{S}\tilde{\theta} - d_S\tilde{\eta} + \tilde{s}(\omega), \tag{6.34b}$$

$$i\omega\tilde{\theta}(\omega) = b_R\tilde{\xi} - b_L\tilde{\theta} - b_S\bar{R}_L\tilde{\eta} - b_S\bar{S}\tilde{\theta} - d_R\tilde{\theta} + \tilde{h}(\omega), \tag{6.34c}$$

with the correlations below:

$$\langle\tilde{h}(\omega_1)\tilde{h}(\omega_2)\rangle = \frac{1}{2\pi}\delta(\omega_1-\omega_2)(b_R\bar{L}+b_L\bar{R}_L+b_S\bar{S}\bar{R}_L+d_R\bar{R}_L) \equiv \frac{1}{2\pi}\delta(\omega_1-\omega_2)A, \tag{6.35a}$$

$$\langle\tilde{r}(\omega_1)\tilde{r}(\omega_2)\rangle = \frac{1}{2\pi}\delta(\omega_1-\omega_2)(b_L\bar{R}_L+d_L\bar{L}) \equiv \frac{1}{2\pi}\delta(\omega_1-\omega_2)B, \tag{6.35b}$$

$$\langle\tilde{s}(\omega_1)\tilde{s}(\omega_2)\rangle = \frac{1}{2\pi}\delta(\omega_1-\omega_2)(b_S\bar{S}\bar{R}_L+d_S\bar{S}) \equiv \frac{1}{2\pi}\delta(\omega_1-\omega_2)C, \tag{6.35c}$$

$$\langle\tilde{h}(\omega_1)\tilde{r}(\omega_2)\rangle = \frac{1}{2\pi}\delta(\omega_1-\omega_2)(-b_L\bar{R}_L) \equiv \frac{1}{2\pi}\delta(\omega_1-\omega_2)D, \tag{6.35d}$$

$$\langle\tilde{h}(\omega_1)\tilde{s}(\omega_2)\rangle = \frac{1}{2\pi}\delta(\omega_1-\omega_2)(-b_S\bar{S}\bar{R}_L) \equiv \frac{1}{2\pi}\delta(\omega_1-\omega_2)E, \tag{6.35e}$$

$$\langle\tilde{r}(\omega_1)\tilde{s}(\omega_2)\rangle = 0, \tag{6.35f}$$

where A, B, C, D, and E defined above are constants independent of ω.

With the above equations and noise correlations functions, $P_{\chi_1\chi_2}(\omega) \equiv \langle\tilde{\chi}_1(\omega)\tilde{\chi}_2(-\omega)\rangle$ can be calculated, for $\tilde{\chi}_1$ and $\tilde{\chi}_2$ being any of $\tilde{\xi}$, $\tilde{\eta}$ and $\tilde{\theta}$. In particular, when $\tilde{\chi}_1$ and $\tilde{\chi}_2$ correspond to the same chemical species, $P_{\chi_1\chi_1}(\omega)$ is equal to the power spectrum of the corresponding fluctuation. The steps of deriving $P_{\chi_1\chi_2}(\omega)$ are briefly written down as follows.

From Eqs. (6.34a) and (6.34b), it's found that

$$\tilde{\xi} = \frac{b_L\tilde{\theta}+\tilde{r}}{i\omega+d_L} \equiv \frac{b_L\tilde{\theta}+\tilde{r}}{M}, \tag{6.36a}$$

$$\tilde{\eta} = \frac{b_S\bar{S}\tilde{\theta}+\tilde{s}}{i\omega+d_S-b_S\bar{R}_L} \equiv \frac{b_S\bar{S}\tilde{\theta}+\tilde{s}}{N}, \tag{6.36b}$$

where

$$M = i\omega+d_L, \quad N = i\omega+d_S-b_S\bar{R}_L. \tag{6.37}$$

Substitute the above equations into Eq. (6.34c), and we have

$$\begin{aligned}\left(i\omega+b_L+b_S\bar{S}+d_R\right)\tilde{\theta} &= b_R\tilde{\xi}-b_S\bar{R}_L\tilde{\eta}+\tilde{h}\\ &= b_R\frac{b_L}{M}\tilde{\theta}+b_R\frac{1}{M}\tilde{r}-b_S\bar{R}_L\frac{b_Sy}{N}\tilde{\theta}-b_S\bar{R}_L\frac{1}{N}\tilde{s}+\tilde{h}\end{aligned} \tag{6.38}$$

Let

$$Q = i\omega+b_L+b_S\bar{S}+d_R. \tag{6.39}$$

Then we can rearrange the equation to be

$$\left(QMN-Nb_Rb_L+Mb_S^2\bar{S}\bar{R}_L\right)\tilde{\theta} = Nb_R\tilde{r}-Mb_S\bar{R}_L\tilde{s}+MN\tilde{h}. \tag{6.40}$$

Define

$$F = QMN - Nb_R b_L + Mb_S^2 \bar{S}\bar{R}_L, \tag{6.41}$$

then

$$\begin{aligned}\tilde{\theta}(\omega) &= \frac{Nb_R}{F}\tilde{r}(\omega) - \frac{Mb_S\bar{R}_L}{F}\tilde{s}(\omega) + \frac{MN}{F}\tilde{h}(\omega) \\ &\equiv a\tilde{r}(\omega) + b\tilde{s}(\omega) + c\tilde{h}(\omega),\end{aligned} \tag{6.42}$$

with

$$a = \frac{Nb_R}{F}, \quad b = -\frac{Mb_S\bar{R}_L}{F}, \quad c = \frac{MN}{F}. \tag{6.43}$$

Substitute the above equations back to Eqs. (6.36a) and (6.36b), then we have the following expressions

$$\tilde{\xi} = \left(\frac{b_L}{M}a + \frac{1}{M}\right)\tilde{r} + \frac{b_L}{M}b\tilde{s} + \frac{b_L}{M}c\tilde{h} \equiv a_1\tilde{r} + b_1\tilde{s} + c_1\tilde{h}, \tag{6.44a}$$

$$\tilde{\eta} = \frac{b_S\bar{S}}{N}a\tilde{r} + \left(\frac{b_S\bar{S}}{N}b + \frac{1}{N}\right)\tilde{s} + \frac{b_S\bar{S}}{N}c\tilde{h} \equiv a_2\tilde{r} + b_2\tilde{s} + c_2\tilde{h}, \tag{6.44b}$$

where

$$a_1 = \frac{b_L}{M}a + \frac{1}{M}, \qquad b_1 = \frac{b_L}{M}b, \qquad c_1 = \frac{b_L}{M}c \tag{6.45a}$$

$$a_2 = \frac{b_S\bar{S}}{N}a, \qquad b_2 = \frac{b_S\bar{S}}{N}b + \frac{1}{N}, \qquad c_2 = \frac{b_S\bar{S}}{N}c. \tag{6.45b}$$

With the above equations of $\tilde{\xi}(\omega)$ and $\tilde{\eta}(\omega)$, we finally arrive at the following results.

$$\begin{aligned}P_{\xi\xi}(\omega) &\equiv \langle\tilde{\xi}(\omega)\tilde{\xi}(-\omega)\rangle \\ &= \langle\Big(a_1(\omega)\tilde{r}(\omega) + b_1(\omega)\tilde{s}(\omega) + c_1(\omega)\tilde{h}(\omega)\Big)\Big(a_1(-\omega)\tilde{r}(-\omega) + b_1(-\omega)\tilde{s}(-\omega) + c_1(-\omega)\tilde{h}(-\omega)\Big)\rangle \\ &= [a_1(\omega)a_1(-\omega)]\frac{B}{2\pi} + [b_1(\omega)b_1(-\omega)]\frac{C}{2\pi} + [c_1(\omega)c_1(-\omega)]\frac{A}{2\pi} \\ &\quad + [a_1(\omega)c_1(-\omega) + a_1(-\omega)c_1(\omega)]\frac{D}{2\pi} + [b_1(\omega)c_1(-\omega) + b_1(-\omega)c_1(\omega)]\frac{E}{2\pi},\end{aligned} \tag{6.46}$$

$$\begin{aligned}
P_{\eta\eta}(\omega) &\equiv \langle \tilde{\xi}(\omega)\tilde{\xi}(-\omega)\rangle \\
&= \langle \Big(a_2(\omega)\tilde{r}(\omega) + b_2(\omega)\tilde{s}(\omega) + c_2(\omega)\tilde{h}(\omega)\Big) \Big(a_2(-\omega)\tilde{r}(-\omega) + b_2(-\omega)\tilde{s}(-\omega) + c_2(-\omega)\tilde{h}(-\omega)\Big)\rangle \\
&= [a_2(\omega)a_2(-\omega)]\,\frac{B}{2\pi} + [b_2(\omega)b_2(-\omega)]\,\frac{C}{2\pi} + [c_2(\omega)c_2(-\omega)]\,\frac{A}{2\pi} \\
&\quad + [a_2(\omega)c_2(-\omega) + a_2(-\omega)c_2(\omega)]\,\frac{D}{2\pi} + [b_2(\omega)c_2(-\omega) + b_2(-\omega)c_2(\omega)]\,\frac{E}{2\pi},
\end{aligned} \tag{6.47}$$

$$\begin{aligned}
P_{\xi\eta}(\omega) &\equiv \langle \tilde{\xi}(\omega)\tilde{\eta}(-\omega)\rangle \\
&= \langle \Big(a_1(\omega)\tilde{r}(\omega) + b_1(\omega)\tilde{s}(\omega) + c_1(\omega)\tilde{h}(\omega)\Big) \Big(a_2(-\omega)\tilde{r}(-\omega) + b_2(-\omega)\tilde{s}(-\omega) + c_2(-\omega)\tilde{h}(-\omega)\Big)\rangle \\
&= [a_1(\omega)a_2(-\omega)]\,\frac{B}{2\pi} + [b_1(\omega)b_2(-\omega)]\,\frac{C}{2\pi} + [c_1(\omega)c_2(-\omega)]\,\frac{A}{2\pi} \\
&\quad + [a_1(\omega)c_2(-\omega) + a_2(-\omega)c_1(\omega)]\,\frac{D}{2\pi} + [b_1(\omega)c_2(-\omega) + b_2(-\omega)c_1(\omega)]\,\frac{E}{2\pi}.
\end{aligned} \tag{6.48}$$

Further simplification shows that both functions $P_{\xi\xi}(\omega)$ and $P_{\eta\eta}(\omega)$, the power spectra of the LINE and SINE fluctuations respectively, have numerators being fourth order polynomials of ω and denominators being sixth order polynomials of ω. Asymptotically, the power spectra have a tail in the form of ω^{-2}. Figure 6.2 shows a comparison between the power spectra obtained from stochastic simulations and the analytic calculation, which demonstrates a satisfactory agreement. The peak value gives a representative angular frequency of the noisy oscillations.

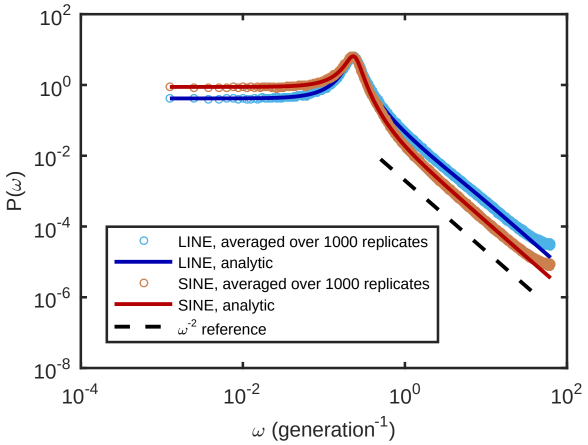


Figure 6.2: Power spectra of the LINE and SINE concentration fluctuations. Circles stand for the power spectra obtained by averaging over 1000 stochastic simulation replicates. Solid lines stand for the analytically calculated spectra. The dash line is a reference function $\sim \omega^{-2}$. Parameters are $b_R = 2$, $b_L = 1$, $b_S = 1$, $d_R = 2$, $d_L = 0.5$, $d_S = 0.5$, $V = 500$. The peak angular frequency is equal to 0.23 generation^{-1}, corresponding to a period of 27 generations. The straight tail, in log-log scale, has a slope of -2, indicating a ω^{-2} asymptotic behavior. The figure is adopted from the published work Ref. [27].

We can further show that $P_{\xi\eta}(\omega)$ is related to the correlation function of $\xi(t)$ and $\eta(t)$. Define the correlation function as

$$\Xi(\tau) \equiv \int \xi(t)\eta(t+\tau)\,\mathrm{d}t. \tag{6.49}$$

Then its Fourier transform is

$$\tilde{\Xi}(\omega) = 2\pi\tilde{\xi}(-\omega)\tilde{\eta}(\omega). \tag{6.50}$$

Therefore, $\langle\tilde{\Xi}(\omega)\rangle = 2\pi P^*_{\xi\eta}(\omega)$. $*$ here refers to the complex conjugate. Observe that the complex argument of $\tilde{\xi}(-\omega)\tilde{\eta}(\omega)$ is equal to the opposite of the phase lag ϕ of $\eta(t)$ (SINE) to $\xi(t)$ (LINE). Therefore ϕ is related to $\tilde{\Xi}(\omega)$ by the following equation

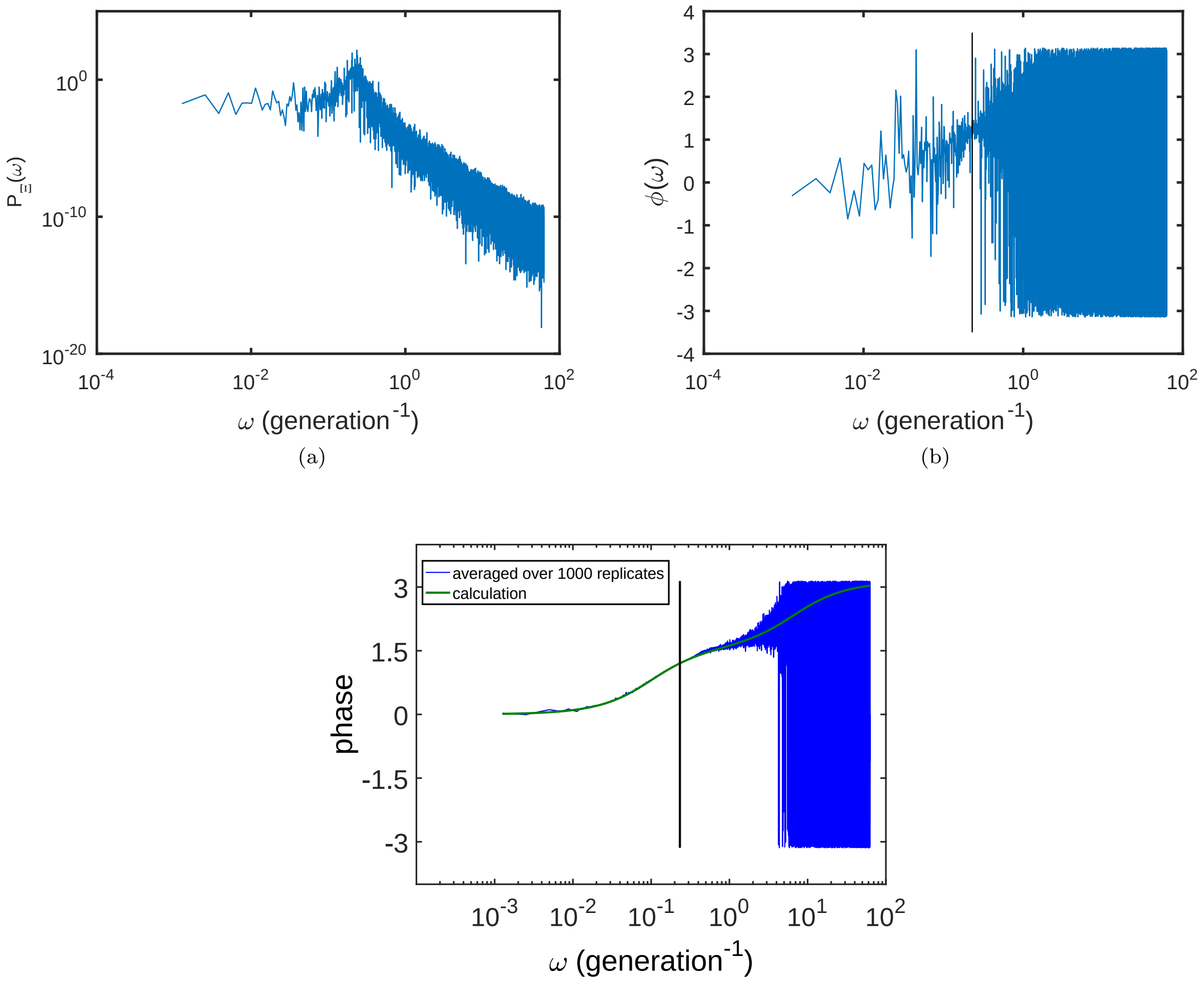
$$\phi(\omega) = -\arg\tilde{\Xi}(\omega). \tag{6.51}$$

Figure 6.3(a) shows the Fourier power spectrum of the correlation function of a typical simulation. It has a peak at the same position as the oscillation power spectra in Fig. 6.2. Figures 6.3(b) and (c) show the phase lag of SINE to LINE as a function of ω. The phase lag at the oscillation peak frequency is approximately 1.2 rad. This measured phase lag is close to the visual estimation $\pi/2$ based on the time series in Fig. 6.1. In Fig. 6.3(c), the averaged phase over replicates as a function of ω agrees with the analytical result. The smeared tail at large ω is due to numerical errors and can be further reduced by averaging over even more replicates.

6.4 Noise-induced Quasi-cycles

I have shown both the deterministic and stochastic dynamics of the model. In summary, on the mean field level, the system returns exponentially fast to the steady state once perturbed; this indicates the long term values of all copy numbers to be constants without any fluctuations. On the stochastic level, copy numbers fluctuate persistently with a dominant oscillatory mode, and the oscillation frequency can be conveniently estimated by that of the deterministic oscillatory mode.

The two parts are drastically distinct, and yet closely connected. Essentially, the persistent stochastic cycles are induced by the fact that the demographic noise constantly stimulates the trajectory of the system to go away from the deterministic decay path and restarts the oscillatory mode over and over again. Specifically, the deterministic trajectory is smooth and assumes continuous changes of the system state. However, the copy numbers of elements are finite integers and the system state thus is discrete. When the system follows the mean field trajectory toward the steady state, it cannot exactly step onto the smooth curve. Instead, it almost always overshoots or undershoots, because of the discreteness. As a result, the system can never



(a)

(b)

(c)

Figure 6.3: (a) The power spectrum of the correlation function obtained from a typical simulation, with the peak at $\omega = 0.23$ generation^{-1}. (b) The phase lag of SINE to LINE as a function of the angular frequency calculated from (a) using Eq. 6.51. (c) The comparison between the analytical calculation and the numerical simulation. The phase difference measured from the correlation function spectrum is averaged over 1000 replicates. The smeared tail is due to numerical errors. In (b) and (c), the black vertical line indicates $\omega = 0.23$ generation^{-1}. The corresponding value of ϕ is approximately 1.2 rad. Parameters used in the simulation are $b_R = 2$, $b_L = 1$, $b_S = 1$, $d_R = 2$, $d_L = 0.5$, $d_S = 0.5$, and $V = 500$.

truly reach the steady state, but can only wander around it.

If the steady state is a focus with an oscillatory mode deterministically, then the stochastic fluctuation resets the amplitude of the oscillation and prevents it from decaying to zero, resulting in an observed noisy oscillation with roughly the same frequency. This type of noise-induced noisy oscillation is called a quasi-cycle and has been discussed previously in the predator-prey system[1, 198]. Its characteristic feature is the asymptotic power-law tail in the fluctuation power spectrum of the form ω^{-2}. The noisy oscillations observed in the minimal model of SINE-LINE interaction are also quasi-cycles, indicated by the ω^{-2} power-law tail in Fig. 6.2.

6.5 Seeking SINE-LINE Dynamics in Real Genomes

I have shown that the minimal individual-level model of the SINE-LINE interaction predicts the existence of noise-induced quasi-cycles in the element copy numbers. Then we wonder: can the quasi-cycles be observed in real genomes?

For the human genome, transposition rates of L1 and Alu elements measured by the mutation accumulation method are of order 1 in $O(10) \sim O(100)$ births [178–180]. The deactivation rates have a lower limit set by the base pair point mutation rate, which is roughly 10^{-8} per base pair per generation [199, 200]. These rates seem to be too slow to generate any experimentally detectable dynamical behaviors. However, this estimate only accounts for fixed mutations that are not lethal, and thus underestimates the actual mutation rates. In a recent experiment [201] on real-time transposition events in living bacterium cells, the actual transposition rate of the DNA transposon directly observed was 10^3 times higher than that obtained by the mutation accumulation method. Moreover, the point mutation rate can be raised by a factor of 10^2 by deactivating the base pair mismatch repair machinery [202]. Thus, for a single-cell experiment rather than a large population, the relevant estimate is: $b_R = 2$, $b_L = 1 \times 10^{-2}$, $b_S = 1 \times 10^{-2}$, $d_R = 1$, $d_L = 1 \times 10^{-2}$, $d_S = 1 \times 10^{-2}$, with units being generation^{-1}. The resultant quasi-cycle period should be roughly 1×10^3 generations. Such oscillations could potentially be observed by integration of the LINE and SINE elements into a host microbial cell, *E. coli* for example, and using novel reporter techniques [201, 203]. Our recent work in Professor Thomas Kuhlman's lab has achieved successful integration of human LINE-1 elements into *E. coli* and *B. subtilis* bacterial cells [28]. The idea of engineering bacteria to demonstrate SINE-LINE quasi-cycles is very promising.

On the other hand, since the transposon events happen slowly on the population, they can potentially leave traces in the evolution history of the species. This leads to another perspective to look for the SINE-

LINE quasi-cycles. That is to look into the history of the species genome recorded by the molecular clock, and count the copy number of elements at different ages to reconstruct the dynamical history of the elements. This idea is explained in detail in Chapter 7.

6.6 Conclusion

In conclusion, we have developed a minimal stochastic model of SINE-LINE interaction, based on the molecular mechanisms of the human LINE-1 and Alu elements, and shown that there exist persistent, noise-induced quasi-cycles in the element copy numbers, which are potentially observable.

By viewing SINEs as predators that feed on LINEs, we have shown that the dynamics of transposons can fruitfully be analyzed using an analogy to ecological models, equipped with tools from statistical physics.

Chapter 7

Looking for SINE-LINE Quasi-cycles in Genomic History

We have demonstrated in Chapter 6 that the parasitic interaction between SINE and LINE pair can result in quasi-cycles in their copy numbers, with a period on the evolutionary time scale. In this chapter, we report a search for the predicted quasi-cycles in the genomic history of the species coelacanth.

7.1 Molecular Clock

For a certain DNA element, the base pairs have been undergoing point mutations since its emergence. The older the element, the more alternations to the sequence. Accordingly, given two elements, by looking at their sequence divergence level, we can deduce their relative ages. In this sense, the sequence itself acts as a clock that times its own history. This idea, known as the molecular clock [204–206], also applies to the protein amino acid sequence.

In practice, in order to convert the divergence level to actual time, we must refer to a certain mutation model. The Jukes and Cantor 1969 (JC69) model [207] and the Kimura (K80) model [208] are two of the most popular ones.

The JC69 model assumes a constant point substitution rate μ for all four types of nucleobases. The conversion from the divergence percentage p, to the Jukes-Cantor distance d between two elements is given by the following equation.

$$d = \nu = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right). \tag{7.1}$$

$\nu = \frac{3}{4}\mu t$, and has the interpretation of the expected number of replacement of a nucleobase during time t. The Jukes-Cantor distance d is proportional to the separation time between the two elements under comparison.

The K80 model takes into account that the transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, and $T \leftrightarrow G$) have different rates, both being constant, with a ratio κ. The conversion relation is given below.

$$K = -\frac{1}{2}\ln\left((1-2p-q)\sqrt{1-2q}\right), \tag{7.2}$$

where K is named the Kimura distance, and p and q are, respectively, the percentages of sites with transitional and transversional differences. Both JC69 and K80 models assume the four bases are equally frequent.

Since the mutation rate is unknown and is usually species dependent, the molecular clock is often calibrated by referring to the fossil records to match historical benchmarks [206], so that the above Jukes-Cantor or Kimura distance can be mapped to real time.

The assumption of a constant mutation rate is not necessarily true over the evolutionary time. A relaxed molecular clock with changing mutation rates has been developed [209].

In the rest of the chapter, we will use the JC69 molecular clock model and study the TE dynamics recorded in the genome. Genomic data are provided by our collaborator, Assistant Professor Oleg Simakov at University of Vienna.

7.2 Periodic Expansions of Transposons

The molecular clock has been widely applied to map the age distribution for transposable elements and further study their historical dynamics [11]. Researchers have found so-called periodic expansion of transposons in several species, cichlid [210], coelacanth [14] and hydra [211], to name a few. This periodicity happens on the evolutionary time scale. For example, the period of the cichlid transposon age distribution is roughly $10-20$ million years, calibrated with respect to the fossil record [212]. It's natural to relate the expansion to external factors like environmental changes. However, we would like to explore whether these cycles have any interpretations in terms of the intrinsic SINE-LINE interaction.

Among the species that have periodic transposon expansions, the "living fossil", coelacanth, stands out. This lobe-finned fish species evolved into the current form about 400 million years ago and has remained roughly the same ever since [15]. Despite the lack of change in the phenotype, the genome of coelacanth has been constantly evolving, but at a considerably low rate compared with other vertebrates, with transposons being especially active [213–215]. The slow phenotypic evolution reflects a small external selection pressure. We thus deduce that the change in the genome might be largely a result of intrinsic dynamics due to element interactions, rather than of external factors. This makes the coelacanth genome an ideal system in which to look for SINE-LINE quasi-cycles. Figure 7.1 shows the age distribution of all transposons in the coelacanth genome. Arrows mark the periodic expansion events.

In the next section, we will test in detail whether or not the observed periodic expansion of transposons in the coelacanth genome is due to the SINE-LINE interaction.

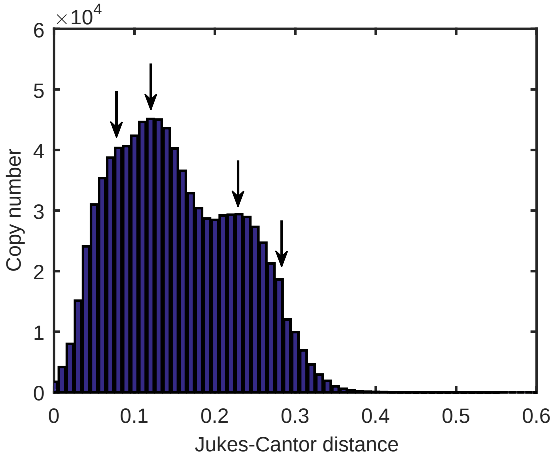


Figure 7.1: Age distribution of all transposons in the coelacanth distribution. Young elements that insert recently are on the left end of the graph, with small Jukes-Cantor distance, while old elements are on the right end. Vertical arrows point out the periodic expansion events. The figure is reproduced from Ref. [14], with data provided by Professor Oleg Simakov.

7.3 Looking for SINE-LINE Quasi-cycles in the Coelacanth Genome

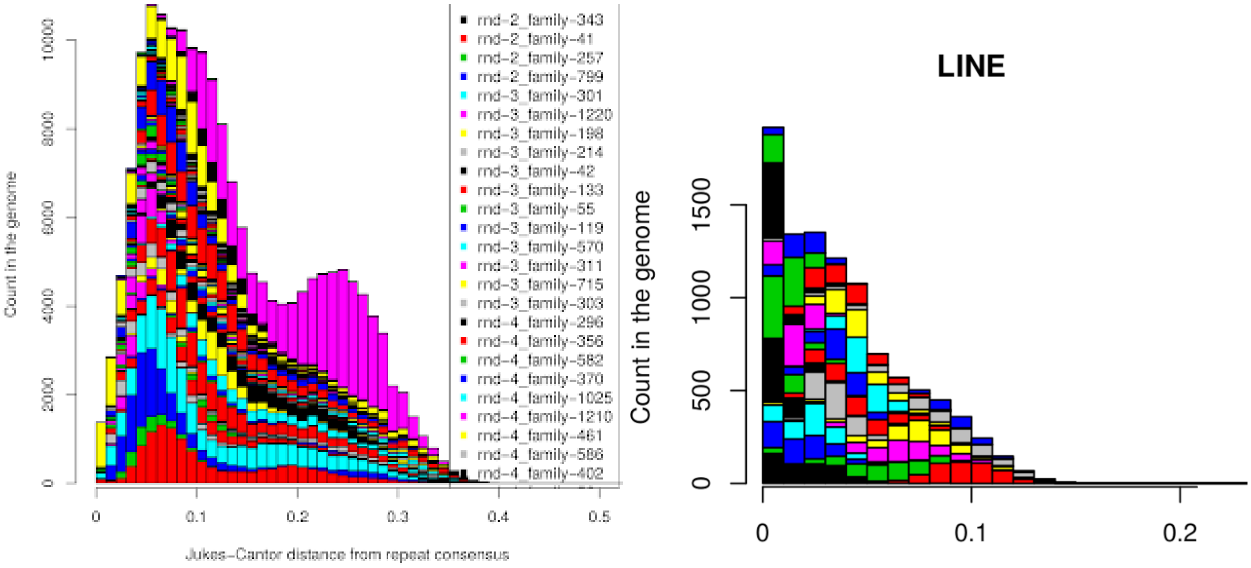
Figure 7.1 shows the age distribution of all transposons in the coelacanth genome, which consists of several categories and many families. According to Ref. [14], these transposons occupy 25% of the genome. Among them, there is a LINE-SINE pair, CR1 (the autonomous LINE) and Deu (the non-autonomous SINE). CR1 has a genome coverage of 2.9% and Deu 1.8%. In the rest of the section, we will exam whether there exist quasi-cycles in the copy numbers of the CR1-Deu pair.

7.3.1 Methods of Constructing the Age Distribution

There are different methods of constructing the age distribution. They differ in the way of computing the reference sequence. And as will be discussed later, the resultant distribution is highly impacted by the method. Here we introduce main ideas of two methods: the consensus sequence method and the phylogeny method.

Consensus Sequence Method

In the framework of the molecular clock, to calculate the accurate age, ideally every element in a certain family should be compared to the ancestor sequence, which, however, is usually unknown. In practice, the ancestor sequence is approximated by a consensus sequence [11], which is a weighted average of all elements in the current genome. Since transposons are deactivated due to the accumulation of point mutations, the active elements should be the most similar to the ancestor and should be assigned large weights. However, it's not trivial to identify active transposons in the first place. The accuracy of the consensus sequence depends on the specific set of weights. Also, when a family contains several subfamilies that independently inserted in the genome at different times, the consensus sequence of the family would have finite distances from each of the subfamilies. This induces an artifact: an active element of a certain subfamily, which should have age 0, now is assigned a nonzero age due to a finite distance to the consensus. This artifact is manifested in the peak at a nonzero age value in the age distribution , as shown in the coelacanth LINE age distribution in Fig. 7.2(a).



(a)

(b)

Figure 7.2: (a) Age distribution of all LINEs in the coelacanth genome obtained from the consensus sequence method. The leftmost peak at a non-zero age manifests the artifact that the consensus sequence has finite distances to all active elements in the subfamilies. (b) Age distribution of all LINEs in the coelacanth genome obtained from the neighbor-joining method. Each subfamily has its own consensus sequence, and the artifact in (a) is removed. Both figures are provided by Professor Oleg Simakov.

Phylogeny Method

The phylogeny method first builds a phylogenetic tree for elements in a certain transposon family. The purpose is to resolve subfamilies that independently invaded the genome. An example of the transposon phylogenetic tree is shown in Fig. 7.3. Each major branch corresponds to a subfamily, and the tips are the active elements. Next, for each subfamily, a consensus sequence is constructed using the active elements. After that, the age distribution of that subfamily can be calculated.

In this way, we avoid representing elements from different subfamilies with the same consensus sequence and remove the artifact mentioned above.

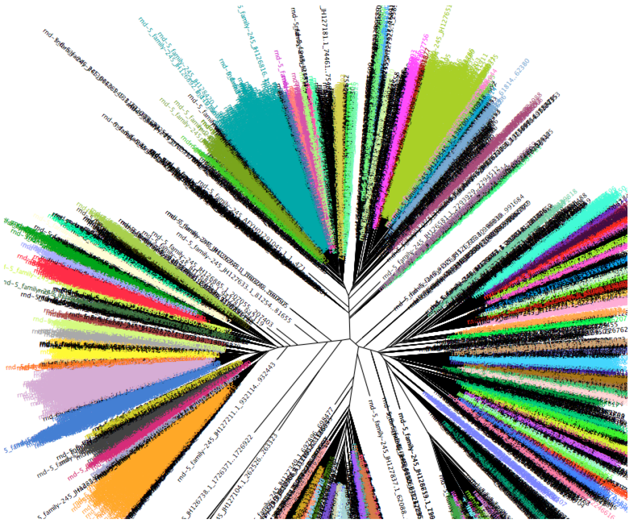


Figure 7.3: Phylogenetic tree of transposon family rnd-5_family-245. Subfamilies are revealed as the branches in the star-like graph. Elements on the tips are the active ones. The figure is provided by Professor Oleg Simakov.

Our collaborator Professor Oleg Simakov has developed a fast algorithm, based on the neighbor-joining method, to construct the transposon phylogenetic tree and calculate the age distribution accordingly, with the following procedure.

1. Input genome data into RepeatModeller, to find main classes of repeats: SINEs, LINEs, LTRs, DNA transposons, *etc.*.

2. Run BLASTN against the genome to find all the repeat loci.

3. Take all BLASTN-identified loci, remove CpG sites (since they are fast evolving), merge them into one file and run BLASTN of those loci against each other.

4. For each locus, find all of its BLASTN-alignments, and calculate the Jukes-Cantor distance between any two repeats. This results in a distance matrix of the repeats.

5. Run a simple neighbor-joining algorithm. In the distance matrix, find the closest pair, namely Locus X and Locus Y, with the smallest distance. Record the distance, merge X and Y into one node, and adjust the distances of all other loci to that node as the smaller distance to either X or Y. This gives a new distance matrix.

6. Repeat Step 5. until no nodes are left to merge.

The distance recorded in the neighbor-joining method represents the time at which an element bifurcates into two due to point mutations.

The age distribution of LINEs in the coelacanth genome calculated via the neighbor-joining method is plotted in Fig. 7.2(b). Compared with Fig. 7.2(a), the artificial peak at a nonzero age is removed.

In the following sections, we will use data primarily generated by the consensus sequence method, for convenient comparison with literature.

7.3.2 Data Analyses Based on the Consensus Sequence Method

We have made several attempts to look for quasi-cycles in the LINE-SINE pair of coelacanth, known as the CR1-Deu pair. Before showing the specific analyses, we need to point out two caveats. First, as demonstrated in Fig. 7.2, the age distribution strongly depends on the underlying construction method. The existence or nonexistence of quasi-cycles should be tested with different construction methods for solidity. Second, since the age series is short, with few "periods" present, the periodicity analysis will not be very reliable, unless there is a strong signal. With the two issues, the following analyses are developed as a procedure. They help build a comprehensive understanding of the age distribution, although they don't provide strong conclusions yet.

Behaviors of Single Transposon Families

Our first step is to inspect all families of CR1 and Deu. The data are provided by Professor Oleg Simakov. There are 26 CR1 families and 6 Deu families. Their age distributions are shown in Fig. 7.4. There are two main messages in the figure. First, different families do not behave in phase. Second, there are dominating families whose copy numbers are much higher than others.

We then look at the most abundant CR1 and Deu families, whose age distributions are shown as bold lines with dots in Fig. 7.5(a). This pair of CR1 and Deu families do not have the exact $\pi/2$ phase difference as predicted by the SINE-LINE stochastic model, although Deu peaks in general do appear later in time than CR1 ones.

Besides the CR1 and Deu families, we also look at another 5 abundant TE families whose maximal copy numbers exceed 2000 at some age. Their age distributions are shown as thin lines with crosses in Fig. 7.5(a). To our surprise, these transposons, although irrelevant to the SINE-LINE interaction, have oscillations too, many in phase with each other. Also, these TE families have similar decay behaviors at the tails near $d_{JC} = 0.3$. We think this tail similarity indicates that the cutoff in the age distribution around $d_{JC} = 0.5$, also shown in Fig. 7.1, is due to an instrumental filter that acts on all families. This filter exists simply

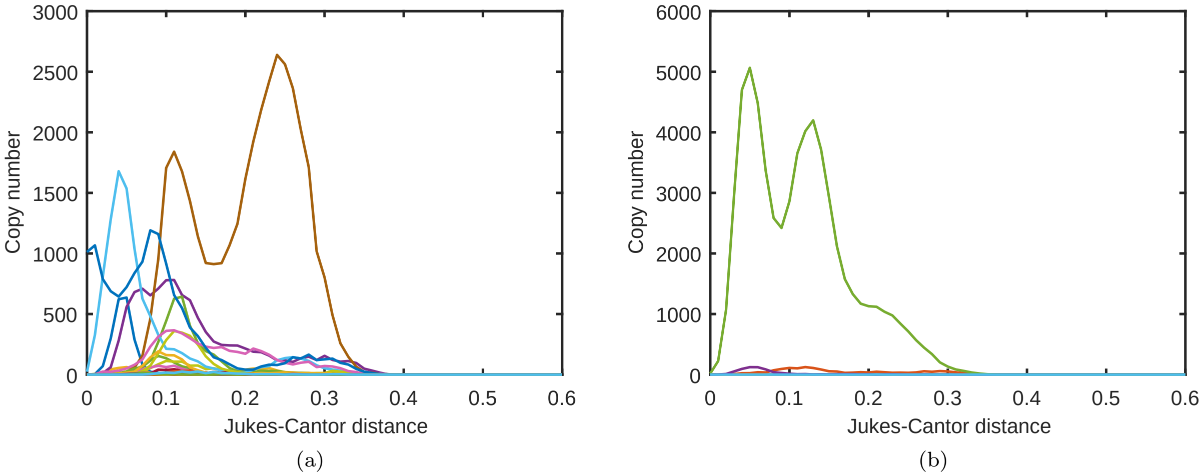


Figure 7.4: (a) Age distributions of 26 coelacanth CR1 families. (b) Age distributions of 6 coelacanth Deu families. Figures are produced with data provided by Professor Oleg Simakov.

because elements beyond a certain age have too many base substitutions to be recognizable. Especially, if the decays were due to intrinsic loss rate, then the length of the tail would positively depend on the number of elements. However, as shown in Fig. 7.4 and Fig. 7.5(a), elements with different copy numbers more or less end at the same cutoff. This cutoff in the age distribution makes it hard to date very far back into the history. With a high mutation rate in transposons, $d_{JC} = 0.5$ usually corresponds to ~ 50 million years.

The Fourier power spectra of the abundant TE families are shown in Fig. 7.5(b). Despite the limited length of data, the Fourier spectra show one physical peak around $f = 7$ for many of the 7 families. Still, there are not any strong signals of the predator-prey dynamics between CR1 and Deu. These two families even peak at different frequencies on the spectra, CR1 at $f \approx 6.5$ and Deu at $f \approx 10.5$. Again, the Fourier transform is not sufficient to give convincing information because the age series is too short.

Age Distributions of Transposon Superfamilies

Figure 7.5 raises a question: can transposons from other families have similar oscillations with the SINE and LINE pair? In fact, if there is a carrying capacity in the genome, then other TE families compete with the SINE-LINE pair for resources, and should have oscillations that are anti-phase with the pair. If there's is no such capacity, then other TEs should be independent of the SINE-LINE pair. We here test this argument.

First, we add up all 26 CR1 families into a CR1 superfamily and all 6 Deu families into a Deu superfamily. The superfamily age distributions are shown in Fig. 7.6. Still, the predator-prey relationship is not clear on the superfamily level. The CR1 and Deu superfamilies rather seem in phase. Next, we calculate DNA TE and LTR superfamilies in the same manner, and also superpose (CR1 + Deu). We expect DNA TE and LTR

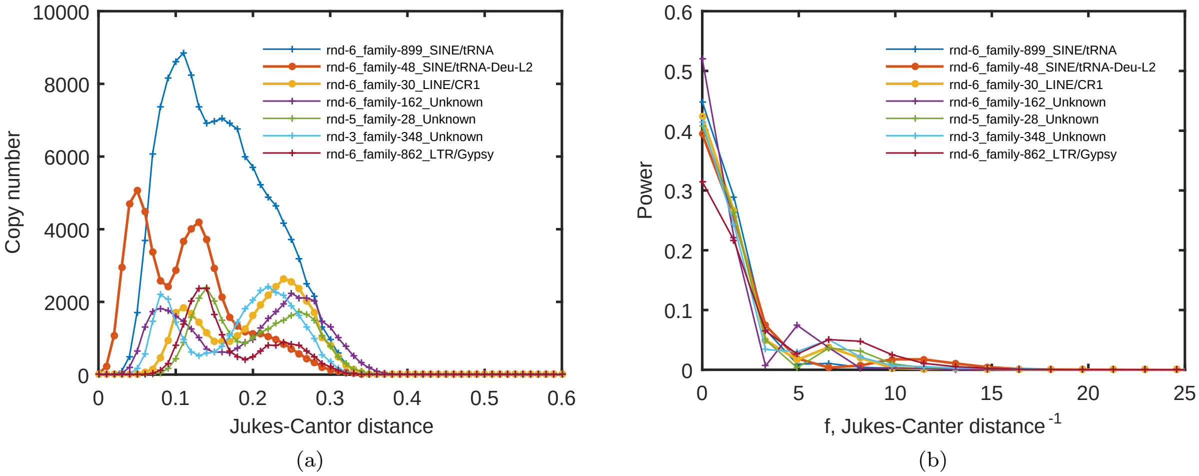


Figure 7.5: (a) Age distributions of the 7 most abundant TE families in the coelacanth genome. Their copy numbers all have maxima that are greater than 2000. (b) Fourier spectra of the 7 TE families in (a). In both figures, the bold lines with dots are for the CR1 and Deu pair, and the thin lines with crosses are for other families. Many families have oscillations, with similar frequencies and phases. Figures are produced with data provided by Professor Oleg Simakov.

to be either independent of or anti-phase with (CR1 + Deu). As seen in Fig. 7.6, the DNA TE superfamily does not have any cycles, and appears independent of others, as expected. The LTR superfamily is almost in phase with (CR1 + Deu). This is against with our expectation, and instead indicates that the dynamics of different TE superfamilies are, to some extent, synchronized.

Cross Correlations of Transposon Families

So far, we have not seen strong evidence for the quasi-cycles in CR1 and Deu. By looking at single families in Fig. 7.5 and superfamilies in Fig. 7.6, it's possible to miss the signal if one Deu family simultaneously depends on several CR1 families. We thus further examine the cross correlations of all pairs of Deu and CR1 families. Let x be the age series of a Deu family and y of a CR1 family. Then the correlation between x and y, $C_{xy}(m)$ as a function of the shift distance m, is given by the following expression.

$$C_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n, & m \geq 0 \\ C_{yx}(-m), & m < 0 \end{cases} \tag{7.3}$$

Under this definition, we expect $C_{\text{SINE,LINE}}$ to have a peak at a negative correlation time, if the pair has a predator-prey relationship.

With the coelacanth genome data, we calculate correlations for the 156 pairs formed by the 6 Deu families

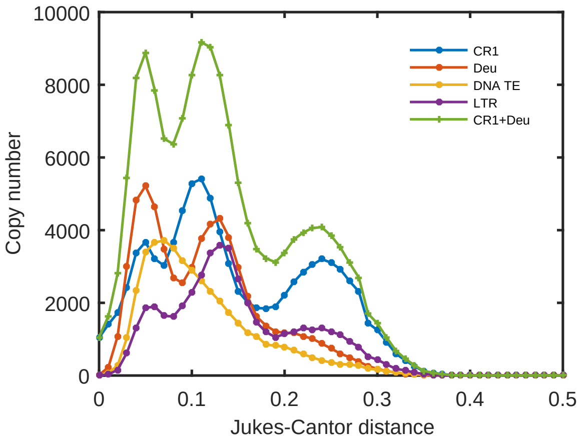


Figure 7.6: Age distributions of TE superfamilies in the coelacanth genome. The predator-prey relationship between CR1 and Deu is not clear on the superfamily level. LTR and DNA TE superfamilies appear synchronized, instead of anti-phase, with (CR1 + Deu). The figure is produced with data provided by Professor Oleg Simakov.

and 26 CR1 families. For each correlation, we find the m that makes the correlation maximal, and record it as the correlation time Δ_0. Figure 7.7 shows the distribution of Δ_0 for the Deu and CR1 families. Although there exist negative correlation times as expected from the interaction, the positive ones are against our argument.

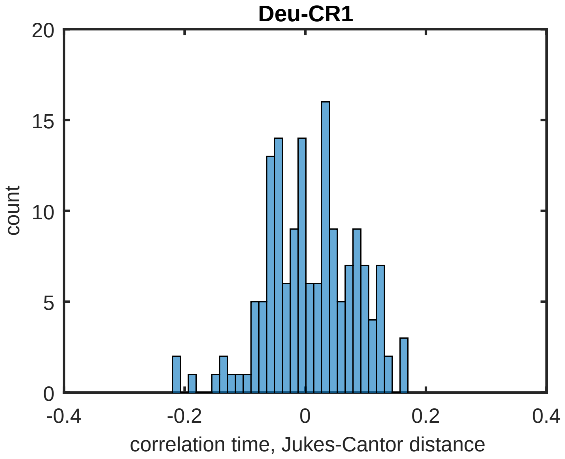


Figure 7.7: Distribution of correlation time Δ_0 between Deu and CR1 families. A negative correlation time is expected if the pair follows the predator-prey dynamics. The figure is produced with data provided by Professor Oleg Simakov.

We further follow the same procedure and calculate the correlation time between any two families from X and Y superfamilies, with X and Y being SINE, LINE, LTR, or DNA TE. The distribution of the correlation time is shown in Fig. 7.8. Compared with others, the SINE-LINE correlation time distribution does not have a significant outstanding feature, and is not sufficient to distinguish any specific dynamics of SINE and

LINE. Also, many pairs have a correlation time of 0, which means they are roughly in phase. Even for the pairs with nonzero correlation times, there is still ambiguity in interpreting the correlation, since the age series is too short and could be highly biased by stochasticity.

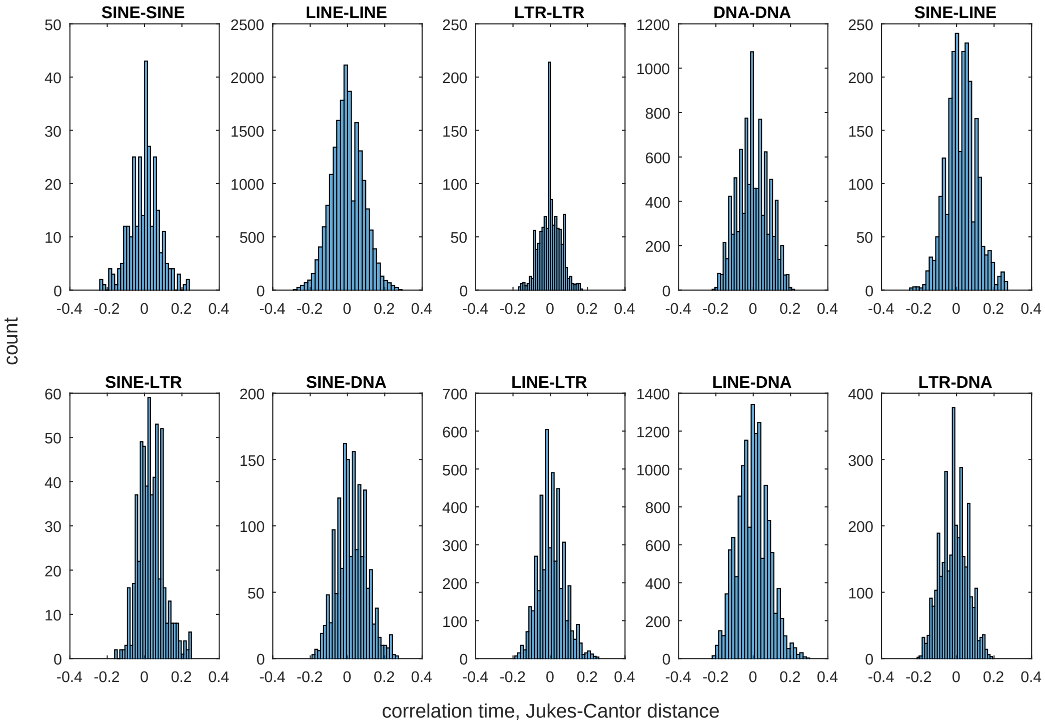


Figure 7.8: Distributions of the correlation time Δ_0 between two families, one from superfamily X, the other from Y. X and Y can be SINE, LINE, LTR and DNA TE, as noted by the title of each figure. A negative correlation time is expected if the pair follows the predator-prey dynamics. The SINE-LINE Δ_0 distribution does not have a significant feature to be distinguishable from others. Figures are produced with data provided by Professor Oleg Simakov.

7.3.3 Discussion

Based on the above analyses, we do not find convincing evidence for the SINE-LINE quasi-cycles between Deu and CR1 in the coelacanth genome. There are several issues that should be explored in detail in future research.

First, the shape of the age distribution depends on the underlying construction method. This is clearly shown in Fig. 7.2 comparing the consensus sequence and the neighbor-joining phylogeny methods. In particular, the construction method can artificially affect the phase difference of the SINE-LINE pair. This induces difficulty in investigating the predator-prey phase relationship between SINE and LINE.

Second, the age distribution data are too short to generate reliable Fourier power spectra and correlation calculation. This is the main reason why we could not reach a solid conclusion.

Third, based on the consensus sequence method, all TE families have oscillations that are roughly in phase. In particular, CR1 and Deu do not appear as prey and predators. There are several potential factors behind the observation. 1. Stochasticity has washed away the predator-prey phase relationship. 2. New CR1 and Deu elements randomly invaded the genome, disrupting the dynamics. 3. TE families were subject to strong external driving force(s), which potentially drove all elements to synchronize.

7.4 Theoretical Modeling of the Transposon Age Distribution

Besides mining the data to look for SINE-LINE quasi-cycles in the actual data, we also develop a model to theoretically investigate whether the actually quasi-cycles can be recorded by the molecular clock and what the age distribution should look like. In this section, we elaborate the model and discuss some related issues.

7.4.1 Model

The stochastic model in Chapter 6 predicts how the active elements should behave. To obtain the age distribution, we just need to introduce the mutation of sequences, which acts as the molecular clock. The individual level reactions are shown below.

$$L_0 \xrightarrow{\frac{b_L}{\alpha+\beta S_0/C}} 2L_0, \qquad S_0 + L_0 \xrightarrow{\frac{b_S}{\alpha+\beta S_0/C}} 2S_0 + L_0, \tag{7.4a}$$

$$L_i \xrightarrow{d_L} \emptyset, \qquad S_i \xrightarrow{d_S} \emptyset, \tag{7.4b}$$

$$L_i \xrightarrow{\mu_L} L_{i+1}, \qquad S_i \xrightarrow{\mu_S} S_{i+1}. \tag{7.4c}$$

L_i represents an individual of LINEs that have undergone i point mutations in the sequence. We also use L_i as the copy number of the corresponding elements. The same interpretation applies to S_i. C is the system size. The rates in the growth reactions (7.4a) are taken from the reduced two-body model Eq. (6.25) in Chapter 6. α and β here are tunable parameters. It's assumed that only the elements L_0 and S_0, which do not have any base pair substitutions, are actively duplicating. The decay rates d_L and d_S in reactions (7.4b) are assumed to be constant for all LINEs and SINEs, respectively. Reactions (7.4c) describe the accumulation of one point mutation in the sequence. The mutation rates of the LINE and SINE sequences are, respectively, $\mu_L \equiv \hat{\nu} M_L$ and $\mu_S \equiv \hat{\nu} M_S$, with $\hat{\nu}$ being the substitution rate of a single base pair, and M_L and M_S being the sequence lengths, respectively.

With the above reactions, the index i represents the number of base pair substitutions of element X_i, and i/M_X has exactly the same interpretation as the Jukes-Cantor distance. Here X can be L or S. $L_0(t)$ and $S_0(t)$ are the time series that follow the LINE-SINE dynamics discussed in Chapter 6. The sets $\{L_0, L_1, \ldots, L_i, \ldots\}$ and $\{S_0, S_1, \ldots, S_i, \ldots\}$ give the age distributions of LINEs and SINEs, respectively, with the subscripts proportional to the ages. They ideally should be the mirror images of $L_0(t)$ and $S_0(t)$, with reversed time.

Reactions (7.4) serve as a minimal model, and ignore many biological details. For example, if the consecutive substitutions $A \to G \to A$ happen to an L_0 element, they will leave the element unchanged and still active; but in the model, they are counted as accumulating 2 substitutions and the element will become L_2, which is deactivated. Still we think the model gives qualitatively correct results, since the above type of consecutive substitutions are rare.

7.4.2 Numerical Results

We simulate the reactions (7.4) using the Gillespie algorithm [86]. Although the quasi-cycles are seen with a wide range of parameters, as discussed in Chapter 6, it still requires specific parameter tuning in order to match the period with the observed cycles in the coelacanth genome. Below are several important factors to consider for choosing the parameters.

1. Due to the specific form of the growth rates in reactions (7.4a), the dynamical behavior is sensitive to the noise. Even though the steady state is exponentially stable, the decay toward it could be very slow compared with the oscillation mode. And demographic stochasticity easily kicks the trajectories of $L_0(t)$ or $S_0(t)$ to extinction.

2. The oscillation period of $L_0(t)$ and $S_0(t)$ should be large. The resolution of the age distribution graph is set by $1/M_X$, X being L or S. If the oscillation is too fast, it can not be resolved.

3. The mutation rates μ_L and μ_S should be large. The dynamics in the time series L_0 and S_0 is recorded by mutations to L_i and S_i, $i > 0$. The mutation rates are similar to the refresh rate of a voltmeter, which sets the response time to the input voltage. The signal can not be reliably measured, if it has a characteristic time shorter than the response time.

4. The oscillation amplitudes in $L_0(t)$ and $S_0(t)$ should be large. A signal with a small amplitude would be washed out by the decay tail of the mutation expansion from small i to large i.

5. d_L and d_S should be small. Otherwise, the decay of L_i and S_i would be so fast that no oscillations in $L_0(t)$ and $S_0(t)$ could be recorded.

Figure 7.9 shows the result of a fine tuned simulation, with Fig. 7.9(a) being the time series of active elements $L_0(t)$ and $S_0(t)$, and Fig. 7.9(b) being the age distribution recorded by the molecular clock. It should be pointed out that we do not have an instrumental filter to get rid of old elements beyond a certain age, and the tails in Fig. 7.9(b) is due to the expansion from small i to large i as a result of mutation.

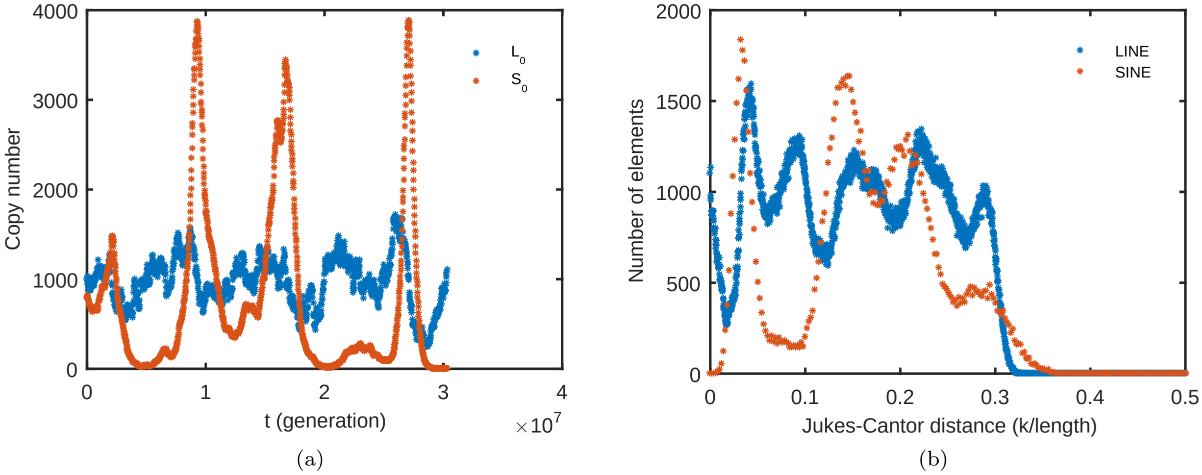


Figure 7.9: (a) Time series of active elements L_0 and S_0. Time flows from the left to the right. SINEs turn extinct, meaning that no active elements remain, around $t \sim 3.1 \times 10^7$. (b) Age distribution recorded by L_i and S_i. Time flows from the right to the left. This is the time reversed version of (a), with smaller amplitudes and smoother curves. Fine features are lost during the recording. Parameters used in this simulation are $M_L = 5000$, $M_S = 500$, $b_L = 5.02 \times 10^{-5}$, $b_S = 1 \times 10^{-5}$, $\alpha = 1$, $\beta = 0.01$, $\hat{\nu} = 1 \times 10^{-8}$, $d_L = 0$, $d_S = 0$, and $C = 2000$. Initially, $L_0(0) = 1000$ and $S_0(0) = 800$. The oldest element has $d_{JC} \approx 0.35$, which corresponds to an age of $d_{JC}/\hat{\nu} \approx 3.5 \times 10^7$. This is roughly the duration of the time series.

Major peaks in the oscillations of $L_0(t)$ and $S_0(t)$ are successfully recorded in the age distribution. And the phase relationship is reserved: LINE peaks precede the SINE ones. However, we also notice significant difference between the two figures. First, the oscillation amplitudes in the age distribution are much smaller than those in the time series. Second, fine features in the time series are smoothed out in the age distribution.

We will discuss these issues in the next section. We will see why and how the signal is distorted during the recording process, and what type of signal can be faithfully documented by the molecular clock.

7.4.3 Molecular Clock as a Low-pass Filter

We have noticed the smoothing effect of the molecular clock in Fig. 7.9. In fact, this effect can cause not only reduced amplitudes but also skipped peaks. Figure 7.10 represents a commonly observed results from the simulation. In the time series, there are 6 very sharp peaks in $S_0(t)$. However, in the age distribution, only 3 SINE peaks are recorded, and they are very shallow.

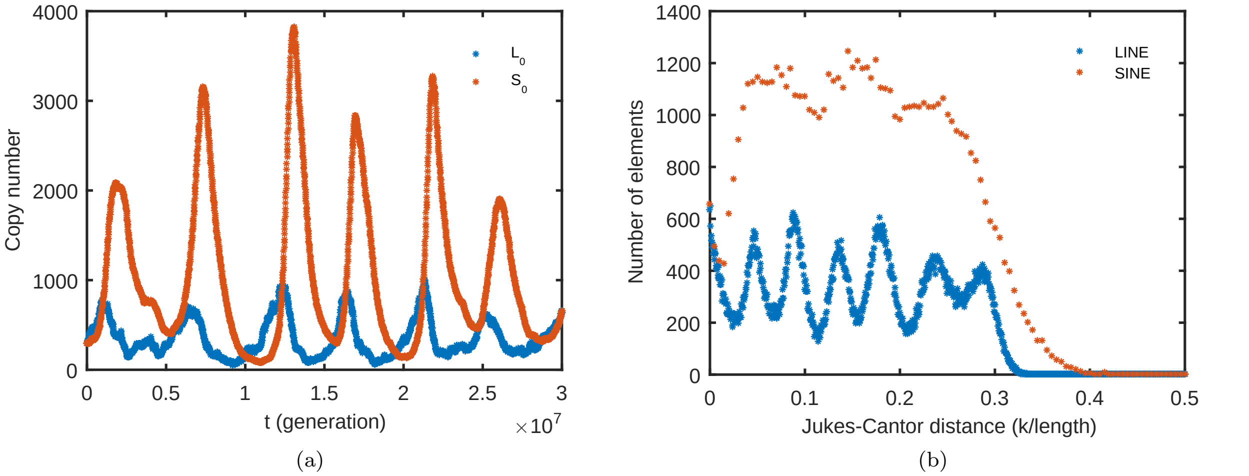


Figure 7.10: (a) Time series of active elements L_0 and S_0. (b) Age distribution recorded by L_i and S_i. Parameters used in this simulation are $M_L = 2000$, $M_S = 200$, $b_L = 2.1 \times 10^{-5}$, $b_S = 6 \times 10^{-6}$, $\alpha = 1$, $\beta = 0.05$, $\hat{\nu} = 1 \times 10^{-8}$, $d_L = 0$, $d_S = 0$, and $C = 1000$. Initially, $L_0(0) = 300$ and $S_0(0) = 300$. The oldest element has $d_{JC} \approx 0.4$, which corresponds to an age of $d_{JC}/\hat{\nu} \approx 4 \times 10^7$. This is roughly the duration of the time series.

Now, we derive analytically how the smoothing effect of the molecular clock comes into play. Let $L_0(t)$ be the time series of the active element copy number. We are interested in $L_i(t)$, which is the copy number time series of elements with i substitutions. The age distribution at time t^* is obtained by plotting $L_i(t^*)$ *vs.* i.

The mean-field rate equation for $L_1(t)$, based on reactions (7.4c) assuming $d_L = 0$, has the following form.

$$\frac{d}{dt} L_1 = \mu L_0 - \mu L_1. \tag{7.5}$$

This equation can be solved by multiplying both sides by $\exp(\mu t)$ and integrating over t. The result is

$$L_1(t) = L_0(t) - [L_0(0) - L_1(0)]e^{-\mu t} - e^{-\mu t} \int_0^t \dot{L}_0(\tau) e^{\mu\tau} d\tau. \tag{7.6}$$

The dot operator in the integral stands for derivative.

Let's first assume

$$L_0(t) = A \sin(\omega t + \phi). \tag{7.7}$$

We will later generalized the calculation to an arbitrary $L_0(t)$. Then the third term in Eq. (7.6) can be exactly calculated, as shown below.

$$\begin{aligned} I_3(t) &\equiv e^{-\mu t}\int_0^t \dot{L}_0(\tau)e^{\mu\tau}d\tau \\ &= \frac{\omega A}{\omega^2+\mu^2}\big[\omega\sin(\omega t+\phi)+\mu\cos(\omega t+\phi)-e^{-\mu t}(\omega\sin\phi+\mu\cos\phi)\big]. \end{aligned} \tag{7.8}$$

Let

$$\cos\theta = \frac{\mu}{\sqrt{\omega^2+\mu^2}}, \quad \sin\theta = \frac{\omega}{\sqrt{\omega^2+\mu^2}}. \tag{7.9}$$

Then we have

$$I_3(t) = A\sin\theta\big[\cos(\omega t+\phi-\theta) - e^{-\mu t}\cos(\phi-\theta)\big]. \tag{7.10}$$

And $L_1(t)$ is simplified as

$$L_1(t) = A\cos\theta\sin(\omega t+\phi-\theta) - [L_0(0)-L_1(0)]e^{-\mu t} + Ae^{-\mu t}\sin\theta\cos(\phi-\theta). \tag{7.11}$$

At large t, we can ignore the two decay terms and obtain

$$L_1(t) \to A\cos\theta\sin(\omega t+\phi-\theta). \tag{7.12}$$

Now let $A_1 = A\cos\theta$, and $L_1(t) = A_1\sin(\omega t+\phi-\theta)$. Follow the above steps, and we obtain $L_2(t)$, which satisfies

$$\frac{d}{dt}L_2 = \mu L_1 - \mu L_2. \tag{7.13}$$

At large t, $L_2(t)$ has the following form

$$L_2(t) \to A_1\cos\theta\sin(\omega t+\phi-2\theta). \tag{7.14}$$

Repeat the procedure recursively, and we obtain the following asymptotic expression for any $L_k(t)$.

$$L_k(t) \xrightarrow{t\to+\infty} A\left(\frac{\mu}{\sqrt{\omega^2+\mu^2}}\right)^k \sin(\omega t+\phi-k\theta). \tag{7.15}$$

Now we can fix the time $t=t^*$, and look at the $L_k(t^*)$ *vs.* k for the age distribution.

$$L_k(t^*) = A\left(\frac{\mu}{\sqrt{\omega^2+\mu^2}}\right)^k \sin(-k\theta+\omega t^*+\phi). \tag{7.16}$$

This is an exponentially decaying oscillatory function of the age k. The period of the oscillation mode is given by

$$T_k = \frac{2\pi}{\theta}, \tag{7.17}$$

with θ determined by Eq. (7.9). Since it takes $\Delta t = \Delta k/\mu$ to accumulate Δk base substitutions, the above period measured in k, corresponds to the following time interval.

$$T = \frac{T_k}{\mu} = \frac{2\pi}{\mu\theta}. \tag{7.18}$$

For large $\mu \gg \omega$, the decay factor $\cos\theta$ is roughly 1, and the amplitude of $L_k(t^*)$ is approximately equal to A, the amplitude of $L_0(t)$. More accurately,

$$\cos\theta = \frac{\mu}{\sqrt{\omega^2+\mu^2}} = \frac{1}{\sqrt{\left(\frac{\omega}{\mu}\right)^2+1}} \tag{7.19}$$

$$= 1 - \frac{1}{2}\left(\frac{\omega}{\mu}\right)^2 + O\left(\left(\frac{\omega}{\mu}\right)^3\right) \tag{7.20}$$

Also, $\cos\theta = 1 - \theta^2/2 + O(\theta^3)$, therefore we have

$$\theta \approx \frac{\omega}{\mu}, \tag{7.21}$$

and

$$T_k = 2\pi\frac{\mu}{\omega}, \quad T = \frac{T_k}{\mu} = \frac{2\pi}{\omega}. \tag{7.22}$$

Note that the period T in the age distribution is identical to that of the input $L_0(t)$ time series Eq. (7.7). So, in the limit of $\mu \gg \omega$, the age distribution $L_k(t^*)$ *vs.* k is a lossless reflection of the time series $L_0(t)$, faithfully recording its amplitude and period.

In the other limit, $\mu \ll \omega$, however, $\cos\theta \approx 0$, and $\theta \approx \pi/2$. As a result, $L_k(t^*)$ decays significantly with k, which means the amplitude information of $L_0(t)$ cannot be preserved. Also, $T_k = 2\pi/\theta = 4$, independent of ω. Therefore, the periodicity information is lost as well. The age distribution thus is not an accurate record of the input time series.

Now, consider an arbitrary time series $L_0(t)$, which can be expanded, shown below, as a superposition of different oscillation modes.

$$L_0(t) = \sum_{i=0}^{\infty} A_i \sin(\omega_i t + \phi_i). \tag{7.23}$$

Then, we can derive the age distribution to be

$$L_k(t^*) \xrightarrow{t^*\to+\infty} \sum_{i=0}^{\infty} \left(\frac{\mu}{\sqrt{\omega_i^2+\mu^2}}\right)^k A_i \sin(-k\theta_i + \omega_i t^* + \phi_i), \tag{7.24}$$

with θ_i given by

$$\cos\theta_i = \frac{\mu}{\sqrt{\omega_i^2+\mu^2}}, \quad \sin\theta_i = \frac{\omega_i}{\sqrt{\omega_i^2+\mu^2}}. \tag{7.25}$$

The molecular clock assumes a fixed mutation rate μ, which cannot be tuned. For a mode in $L_0(t)$ with $\omega_i \ll \mu$, its amplitude and period can be documented in the age distribution $L_k(t^*)$. For a mode with $\omega_i \gg \mu$, however, it will be lost in the recording process. Therefore, the molecular clock acts as a low-pass filter. With the mutation rate being proportional to the element length, the SINE has a smaller mutation rate than the LINE's, and thus suffers more information loss, as shown in Fig. 7.10.

7.4.4 Discussion

Noise-induced quasi-cycles usually have a small relative fluctuation size $\sim 1/\sqrt{N}$, compared with the average population N. They thus may not be clearly recorded by the molecular clock, due to the amplitude decay. Furthermore, the ω^{-2} tail in the power spectrum of the quasi-cycles will not be stored in the age distribution, since modes with large ω are filtered out. Both factors make it difficult to identify quasi-cycles from the age distribution. On one hand, the quasi-cycles may not be successfully documented, if they come with high frequency modes or small amplitudes. On the other hand, even if there are oscillations in the age distribution, we may not be able to rule out other possible mechanisms because of the absence of the signature ω^{-2} tail.

Besides, since the observed peaks in the coelacanth TE age distribution are deep, as shown in Fig. 7.4 and Fig. 7.5, they are likely to result from dynamics with more dramatic amplitudes than the quasi-cycles.

7.5 Conclusion

I have reported an attempt to look for SINE-LINE quasi-cycles in the transposon history of the coelacanth genome, recorded by the molecular clock. The SINE-LINE quasi-cycle is a potential origin of the periodic expansion. However we did not find strong evidence in the genomic data nor an effective way to identify the cycles in theoretical modeling. Other mechanisms, such as external environmental changes, can still be responsible for the periodicity. A future research direction is to compare quantitatively and systematically the age distributions of different species, to identify species-specific signals.

Chapter 8

Diversity of Repetitive Elements in the Genome

I have shown in previous chapters that certain transposons interact with each other in similar ways to that of species in ecosystems. To explore this further, we now consider the diversity of all families of repetitive elements in the genome. Specifically, we borrow the metric called rank-abundance distribution (RAD) from ecology, introduced in Chapter 2, as a characterization of the diversity.

The RAD is widely used in ecology to characterize the diversity of species in a given ecosystem. To obtain the distribution, we first sort the species abundances (populations) in descending order and then plot the abundance against rank. The RAD straightforwardly illustrates the richness and evenness of species. In this chapter, we adopt the RAD metric to study the diversity of repetitive elements in the genome.

8.1 Rank-abundance Distribution of All Repetitive Elements

We examine genomes of 46 species, with data downloaded from RepeatMasker.org and processed by Professor Oleg Simakov. We count in all repetitive elements, including transposons and tandem repeats, to calculate their rank-abundance distribution, with abundance defined as the element copy number. Figure 8.1 demonstrates the RADs of 6 sample species in double logarithmic scale. Colors distinguish the categories of elements. RADs for the remaining 40 species are shown in Fig. 8.2.

There are several features in these RAD plots. First, there exist both dominating and rare families, similar to that observed in ecosystems. Second, categories present in the genome are species-dependent. Third, the RAD functional form is not universal across species. Fourth, some species, especially those with simple repeats on the RAD tail, have power-law asymptotic abundance-rank relation.

8.2 Rank-abundance Distribution of Different Repeat Categories

Now, we plot the RAD of each repeat category, to look for any category-specific features. The result for a sample species *Alligator mississippiensis* (American alligator) is shown in Fig. 8.3. The RAD of simple repeats is especially interesting, since it appears as a straight line in the double logarithmic plot indicating

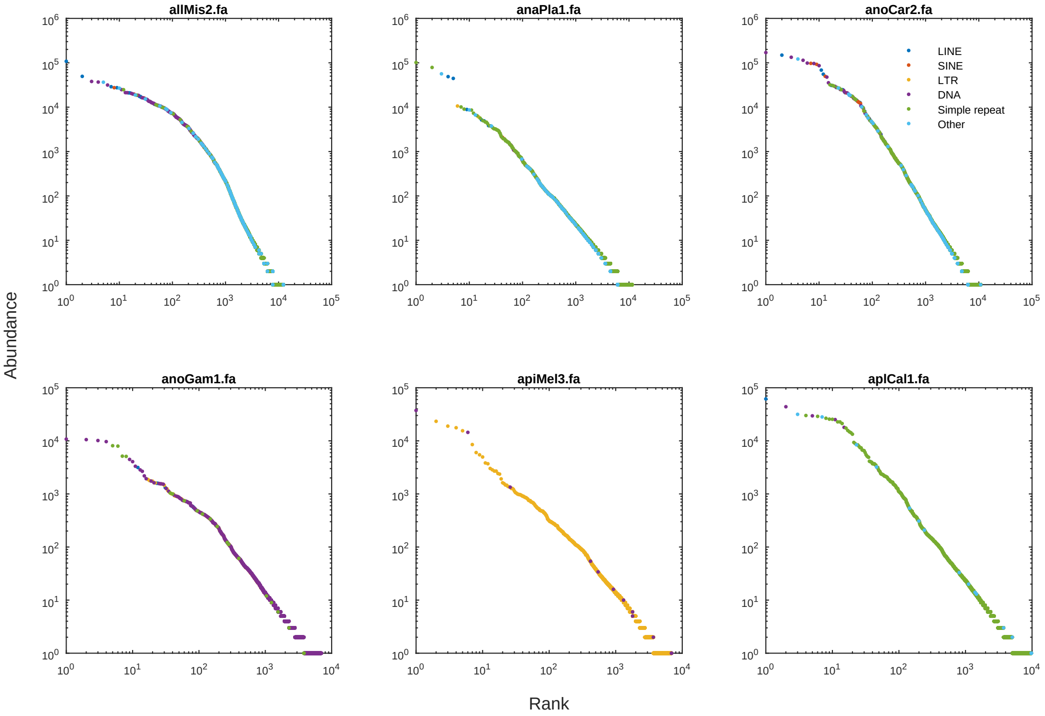


Figure 8.1: The RADs of repetitive elements in 6 sample genomes. The color refers to the category of the element. Figures are produced with data downloaded from RepeatMasker.org and processed by Professor Oleg Simakov.

a power-law behavior. It should be pointed out that the copy number of simple repeats are defined as the number of segments. So, one $(\mathrm{ACTG})_6$ segment and one $(\mathrm{ACTG})_8$ segment are counted as two copies of the same simple repeat family with the common unit (ACTG).

We further calculate the RAD of simple repeats in all other species, and obtain 46 RAD plots, presented in Fig. 8.4. In order to compare genomes with different repeat richnesses, we normalize the abundance and rank by dividing them by their maxima, $\tilde{A} \equiv A/A_{max}$ and $\tilde{r} \equiv r/r_{max}$, respectively. These RADs all appear power-law over several orders of magnitude, with similar slopes in the double logarithmic scale.

We calculate the RAD power-law exponent by linear fitting $\log_{10} \tilde{A}$ against $\log_{10} \tilde{r}$ for the slope. We use data in the range $0.005 < \tilde{r} < 0.2$. Figure 8.5 shows the scatter plot of the 46 exponents, revealing a cluster in the range from -1.2 to -1.5.

8.3 Comparison with Previous Observations

Different power-law abundance distributions of genomic elements have been recorded in literature.

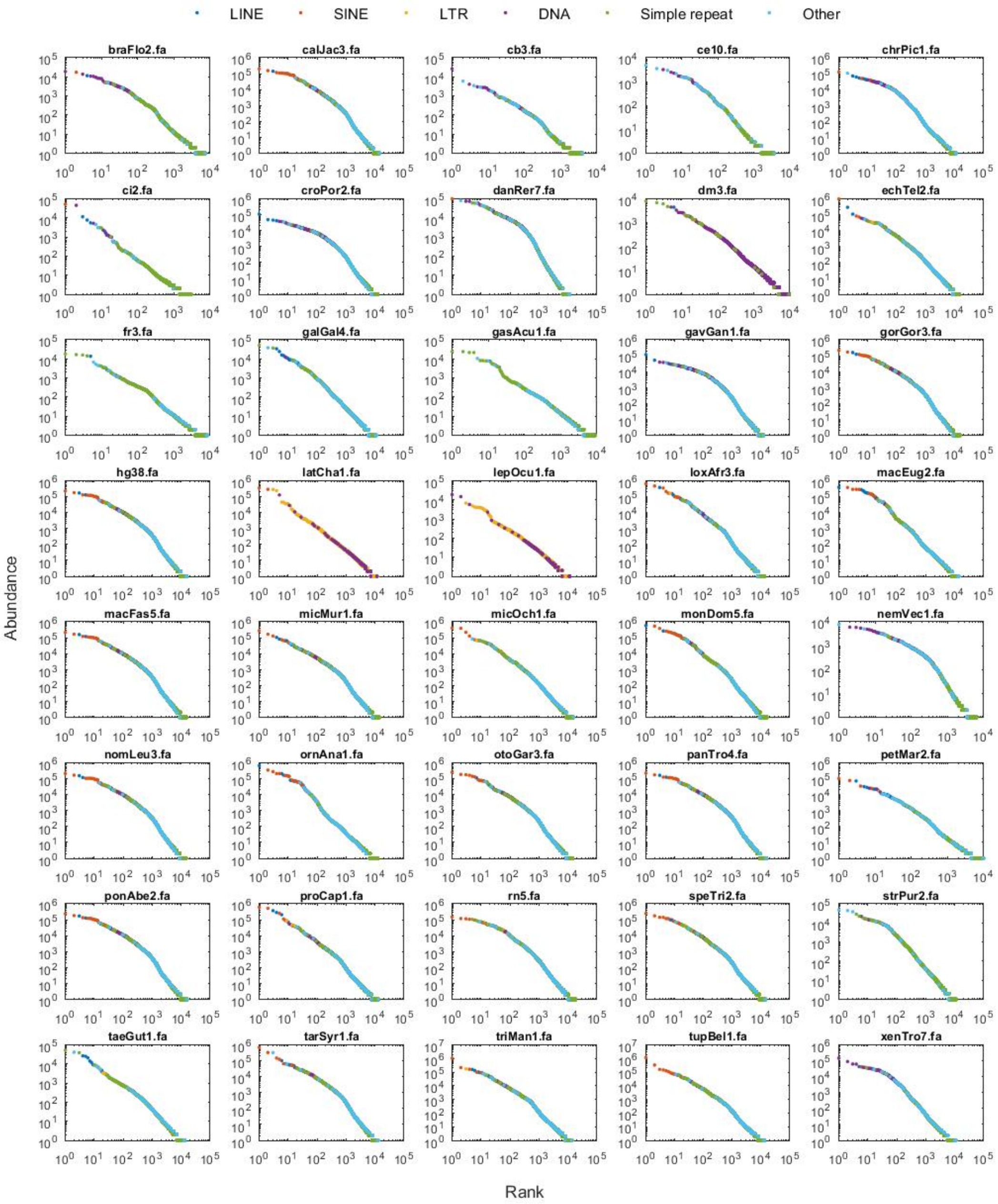


Figure 8.2: The RADs of repetitive elements in the remaining 40 sampled genomes. The color refers to the category of the element. Figures are produced with data downloaded from RepeatMasker.org and processed by Professor Oleg Simakov.

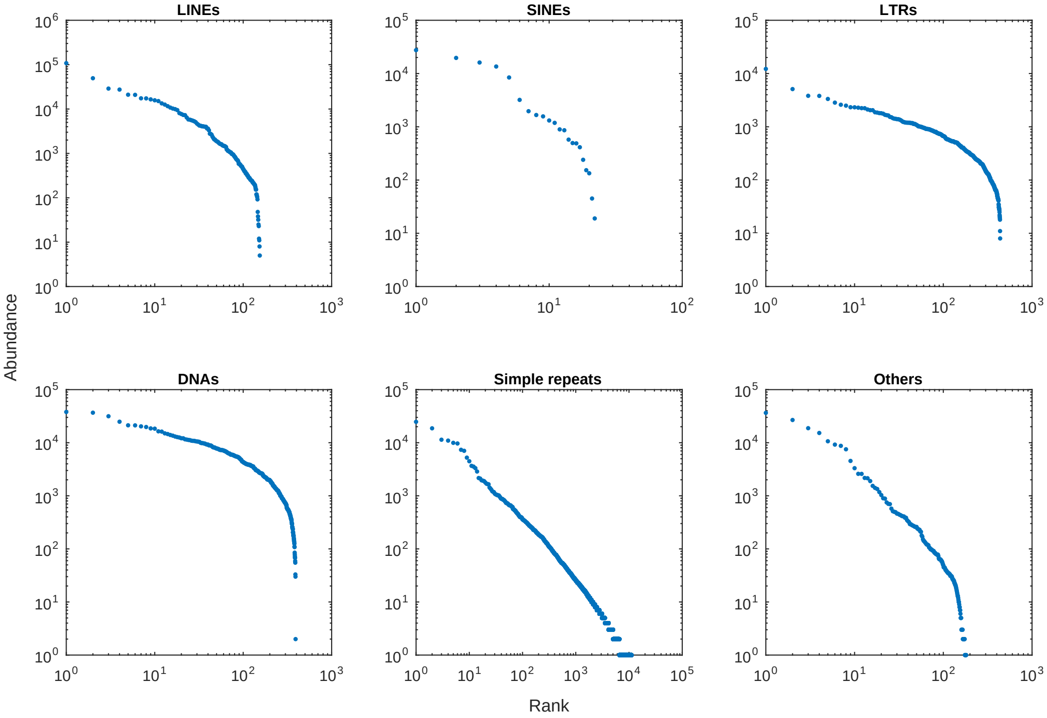


Figure 8.3: The RADs of repeat categories in the genome of *Alligator mississippiensis*. Figures are produced with data downloaded from RepeatMasker.org and processed by Professor Oleg Simakov.

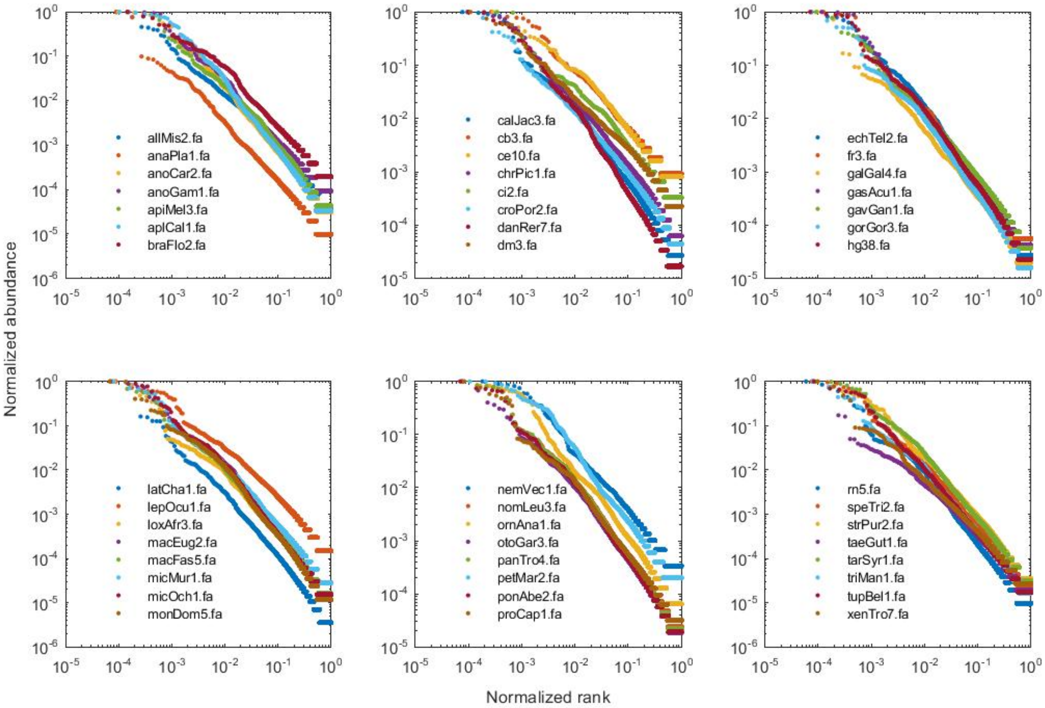


Figure 8.4: The normalized RADs of simple repeats of all 46 sampled genomes. They all have power-law behaviors with similar exponents. Figures are produced with data downloaded from RepeatMasker.org and processed by Professor Oleg Simakov.

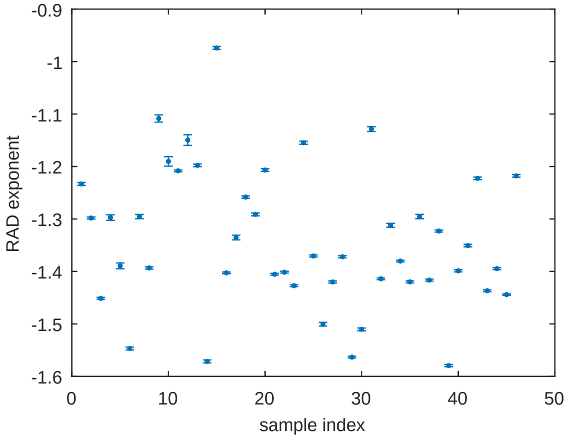


Figure 8.5: The scatter plot of the power-law exponents in Fig. 8.4. The exponents are calculated by performing linear fit with $\log_{10}\tilde{A}$ and $\log_{10}\tilde{r}$, in the data range $0.005 < \tilde{r} < 0.2$. The error bar of each dot represents the 95% confidence interval of the exponent.

The first type involves the so-called n-mers, with $4 \leq n \leq 40$. A n-mer is defined as a segment of n consecutive bases. Researchers scan the entire genome of a certain species to count the number of occurrences of each n-mer and then calculate the frequency distribution. This distribution is power-law and further lead to a power-law rank-frequency distribution similar to the Zipf's law [216–218], due to the rational of Eq. (2.5). The authors interpret the power-law rank-frequency distribution as a linguistic feature of the sequence. The n-mers are artificially constructed, and different from the simple repeats, which naturally emerge as independent genomic elements.

It is also found that gene family size and protein family size both are distributed in a power-law way [219–222]. The family here is defined as a cluster of genes or proteins that have sequences within a similarity range. These family size distributions tend to have large exponents and correspond to RADs with shallower slopes on the double logarithmic scale than those in Fig. 8.4. The gene/protein families are similar in concept to the repeat families, but are under strong selection pressures, which may lead them to obey different distributions. Models based on a birth-death process have been proposed to explain the power-law gene and protein family size distributions [219, 220, 223–225]. However, they generally require finely tuned parameters or rate functions in order to produce the desired exponents.

8.4 Discussion

I have presented the rank-abundance distribution of repetitive elements as a novel method of representing the diversity. Especially, we observe power-law RADs of simple repeats, with similar exponents across genomes.

Although it is hard to interpret the exact microscopic evolutionary process that results in the observed RADs, it is promising to deduce the generic factors by developing minimal models as well as by experimentally exploring the molecular interactions between repeat families. In particular, we need to understand if these RADs reflect predominantly niche or neutral processes. If the latter is favored, then this might be consistent with the hypothesis that non-coding repeats are essentially junk DNA with no functional significance. If not, then this could be evidence for the functional significance in the genome of repetitive elements. We hope that future study can help resolve this debate over the "junk" DNA.

Part III

Dynamics at the Evolutionary Scale

Chapter 9

Introduction to Niche Construction Theory

Evolution is known as the change of heritable phenotypes of organisms over generations. It happens on all levels of biological organization. Some examples are the change of allele frequency (eg. fixation of a certain gene) on the molecular level, the development of an individual trait (eg. viral resistance) on the organism level, and the emergence of a collective behavior (eg. eusociality) on the species or population level.

The evolutionary process can be generally divided into two stages. In the first, mutation and gene migration create phenotypic variations among individuals in the population. In the second, natural selection and genetic drift determine how the phenotypic variations change with time. In reality, the two stages coexist without a well-defined temporal boundary. As a result, some biological units, being a certain allele, organism or population, turn extinct, while others survive and increase in number.

In this chapter, I briefly review the main ideas of natural selection, and introduce the more recently developed niche construction theory, which emphasizes the influence of organisms on their environment.

9.1 Natural Selection

The theory of natural selection was first reported by Alfred Russel Wallace and Charles Darwin [18], and then systematically elaborated in the book *On the Origin of Species* by Darwin. The main process contains two parts. First, various phenotypes are created by mutation or genetic migration. These phenotypic variations must be heritable. Second, if the phenotypes are associated with the capability to survive or reproduce in the background environment of the population, then after generations, advantageous phenotypes will increase in population fraction or be selected for, and disadvantageous ones will decrease or be selected against. As a result, only significantly fit phenotypes survive.

In natural selection, organisms evolve and adapt to the selection pressure exerted by the environment, and the environment itself is treated as a boundary condition that is adiabatically constant during the entire process. This can be formally expressed by the following equations.

$$\frac{dO}{dt} = f(O, E), \tag{9.1a}$$

$$\frac{dE}{dt} = g(E), \tag{9.1b}$$

where O stands for the organism and E for the environment. f and g are system-specific functions that describe the evolution of the organism and environment, respectively. Specially, the environment is independent of the organisms.

9.2 Niche Construction

The niche of a species refers to its position in the ecosystem. It involves the environmental resources that the species relies on, including the geographic configuration, the climate, *etc.*, and the interactions with other species in the same ecosystem, represented by the species' position in the food web and their dynamical history.

9.2.1 Niche Construction as an Evolutionary Process

The phenomenon that organisms modify the environment and thus create new niches is termed niche construction [19]. Natural selection theory treats niche construction simply as a special phenotype of some organisms. Since the 1980s, this conventional view point has been challenged, and the emphasis on niche construction as a key factor in evolution has been promoted [19, 226–229]. In niche construction theory, organisms can shape the environment they live in, change the selection pressure, and, as a result, reroute their own evolutionary path.

A good example is the Great Oxidation Event on the earth [230] caused by the emergence of photosynthetic cyanobacteria. During the event, the oxygen concentration of the earth atmosphere increased from virtually zero to 2% $\sim$ 4% and finally reached the current value 20%. By changing the oxygen level, the cyanobacteria created a new niche and subsequently influenced the evolution of the entire biosphere.

9.2.2 Ecosystem Engineering

In ecology, a similar feedback of organisms on the environment has been observed, known as ecosystem engineering [231, 232]: the engineer organisms create, modify, maintain or destroy their habitat. Some commonly known engineers are earthworms, which modify the soil in terms of both physical morphology and chemical composition, beavers, which build dams and completely change the landscape, and humans, whose activities, such as agriculture and mining, dramatically impact the biosphere. This feedback mechanism is viewed as the bridge connecting ecology and evolutionary biology [233, 234].

The main difference between niche construction and ecosystem engineering lies in the fact that the former emphasizes the effect on the long-term evolutionary consequence, while the latter focuses on the temporary remodel of the habitat.

9.2.3 Theoretical Models

In niche construction theory, Eq. (9.1) are rewritten formally as follows [235],

$$\frac{\mathrm{d}O}{\mathrm{d}t} = f(O, E), \tag{9.2a}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = g(O, E). \tag{9.2b}$$

The functional response g of the environment now depends both on the environment itself and the organisms. And the time dependence is on the evolutionary scale.

Previous works applied population dynamics models [236–239] and population genetics models [227] to study the effect of niche construction or ecosystem engineering on organism populations and evolution. Instead of implicitly embedding the influence of environment in the parameters, these models incorporated the environment as an explicit variable to which the organisms dynamically feed back.

9.2.4 Criticisms on the Niche Construction Theory

Niche construction theory remains controversial [240–242]. Critics focus on the fact that niche construction can be viewed, without losing any explanatory power, as a special trait resulting from natural selection, and that it does not provide answers which can not be reached in the framework of natural selection. The advocates for the theory emphasize that viewing the niche construction as a separate process helps develop an accurate evolutionary history of the species and reveals properties that may be hidden in the natural selection picture.

In Chapter 11, I will explicitly incorporate niche construction into the evolution of phylogenetic trees in order to answer the question of whether or not niche construction leaves an imprint on evolution, as represented by the structure of phylogenetic trees.

Chapter 10

Introduction to the Topology of Phylogenetic Trees

Phylogenetics studies the evolutionary relations of a group of organisms. By evaluating the similarity of heritable traits, including DNA sequences and phenotypes, it can be inferred whether two organisms originate from a common ancestor and when the speciation happened. The reconstructed evolutionary history is represented by the so-called phylogenetic tree. The nodes on a phylogenetic tree stand for species, with the external ones or leaves being the actual species that are observed and the internal ones being hypothetical species that are inferred based on the similarity and the embedded evolutionary process. When a tree is rooted, the top node, or the root, represents the inferred common ancestor of all nodes in the tree. A node can bifurcate or multifurcate into more than two descendants at a speciation event.

There are several metrics used in the literature to describe the scaling behavior associated with the topology of trees [23, 243, 244]. In this chapter, we introduce one of the methods [23] to characterize the topology of phylogenetic trees and show that they have a universal and unique scaling behavior.

10.1 Characterizing the Tree Topology

For an arbitrary rooted tree, we define the depth d of a node as the number of edges on the path from the root to the node, and the height h of a node as the number of edges on the longest path from the node to a leaf. Then we have $d(\text{root}) = 0$ and $h(\text{leaf}) = 0$. We use the height of the root as the height of the tree H.

Ref. [23] introduces a quantitative metric to characterize the topology of a tree. We will focus on the phylogenetic trees, although this metric has been applied to many other trees and networks and revealed interesting scaling behaviors [245–247]. We rephrase it in this section.

First, we define quantity A for an arbitrary node i on the tree as the size, or number of nodes, of the subtree S_i rooted at node i. For a binary tree, with the child nodes named left and right, $A(i)$ can be calculated recursively as follows.

$$\begin{cases} A(\text{leaf}) = 1, \\ A(i) = 1 + A(i \to \text{left}) + A(i \to \text{right}). \end{cases} \tag{10.1}$$

Next, we define quantity C for node i as the cumulative size of the subtree S_i,

$$C(i) \equiv \sum_{j \in S_i} A(j). \tag{10.2}$$

Alternatively, define d_{ij} as the number of edges from node i to node j, or the depth of node j in the subtree S_i. Then the above equation is equivalent to

$$C(i) = \sum_{j \in S_i} (d_{ij} + 1) = \sum_{j \in S_i} d_{ij} + A(i). \tag{10.3}$$

Divide both sides by $A(i)$, and we have

$$\frac{C(i)}{A(i)} = \langle d(i) \rangle + 1, \tag{10.4}$$

with $\langle d(i) \rangle$ being the average depth of nodes in the subtree S_i. $C(i)$ can also be calculated recursively, shown below.

$$\begin{cases} C(\text{leaf}) = 1, \\ C(i) = A(i) + C(i \to \text{left}) + C(i \to \text{right}). \end{cases} \tag{10.5}$$

Since $i \to$ left and $i \to$ right are symmetric in the above recursion, we deduce that mirroring the tree rooted at i, by switching the left and right subtrees, preserves C for the node i.

Now, for every node in a given tree, we can calculate its A and C, and further obtain the relation $C(A)$.

If the tree is multifurcating, instead of binary, then the addition in the recursive calculations Eq. (10.1) and Eq. (10.5) should run over all children of i, as indicated below.

$$A(i) = 1 + \sum_{j \in \text{Children}(i)} A(j), \tag{10.6}$$

$$C(i) = A(i) + \sum_{j \in \text{Children}(i)} C(j), \tag{10.7}$$

where Children(i) is the set of node i's immediate children.

In the rest of this section, we calculate $C(A)$ for two extreme cases of binary trees and demonstrate that it can be used to characterize the shape of the tree.

10.1.1 Topology of Completely Balanced Binary Trees

A binary tree is said to be completely balanced when all its levels are fully filled, meaning that at depth d, there are 2^d nodes. An example is shown in Fig. 10.1(a). On this type of tree, every non-leaf node has two child nodes and every leaf has the same depth. For an arbitrary node i, its depth and height have the following relation,

$$d(i) = H - h(i). \tag{10.8}$$

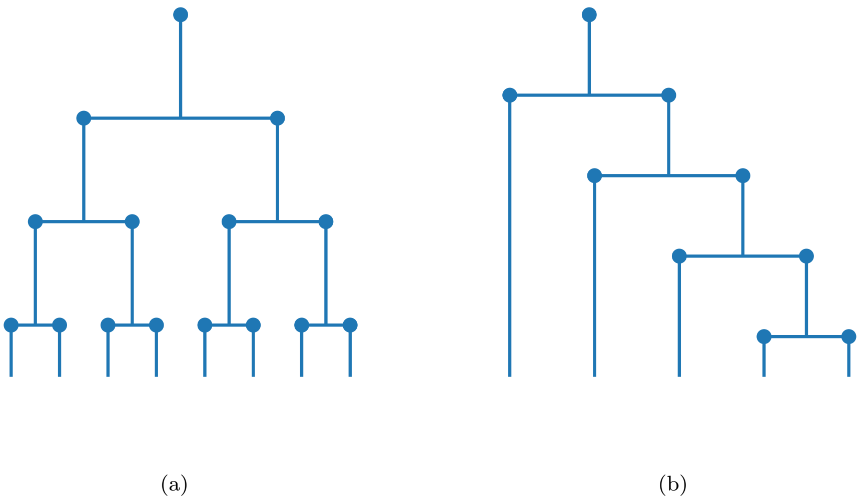


Figure 10.1: (a) A completely balanced tree of height 3 with 8 leaves. The dots represent nodes. Each non-leaf node has both left and right children. At depth d, the number of nodes is 2^d. (b) A completely imbalanced tree of height 4. For each non-leaf node, the left child is always a leaf and only the right child may continue branching.

Since all nodes on the same layer have the same subtree size, $A(i)$ of node i is determined by the height of the node. Therefore, we can rewrite the recursive equation Eq. (10.1) of $A(i)$ in terms of A_h.

$$\begin{cases} A_0 = 1, \\ A(i) = A_h = 1 + A_{h-1} + A_{h-1}. \end{cases} \tag{10.9}$$

From the above equation, it's straightforward to see that $A_h + 1 = 2(A_{h-1} + 1) = 2^h(A_0 + 1) = 2^{h+1}$. Therefore we have the expression of A_h below.

$$A_h = 2^{h+1} - 1. \tag{10.10}$$

We can calculate $C(i)$ by using its definition Eq. (10.2) and observing that the number of nodes at height h is $n_h = 2^{H-h}$. For a node i at height h, we have

$$C(i) = C_h = \sum_{j \in S_i} A(j) = \sum_{h'=0}^{h} A_{h'} n_{h'} = \sum_{h'=0}^{h} \left(2^{h'+1} - 1\right) 2^{h-h'}. \tag{10.11}$$

And the expression of C_h is simplified to be

$$C_h = 2^{h+1} h + 1. \tag{10.12}$$

Based on Eqs. (10.10) and (10.12), we can solve for the relation of $C(A)$ as follows.

$$C = \left[\frac{\ln(A+1)}{\ln 2} - 1\right](A+1) + 1. \tag{10.13}$$

Asymptotically $C(A) \sim A \ln A$ at large A.

10.1.2 Topology of Completely Imbalanced Binary Trees

Another extreme case of the binary tree, a completely imbalanced tree, is shown in Fig. 10.1(b). For all non-leaf nodes, the left children are always leaves and the bifurcation only happens on the right branch, leading the tree to be extremely right-biased.

$A(i)$ is again associated with the height h of the node, and the recursive equation Eq. (10.1) can be rewritten below, by observing that $A(i \to \text{left}) = A_0$.

$$\begin{cases} A_0 = 1, \\ A(i) = A_h = 1 + A_0 + A_{h-1}. \end{cases} \tag{10.14}$$

We can see that $A_h = A_{h-1} + 2 = A_0 + 2h$, and

$$A_h = 2h + 1. \tag{10.15}$$

The recursive cumulative size C can also be rewritten in terms of h.

$$\begin{cases} C_0 = 1, \\ C(i) = C_h = A_h + C_0 + C_{h-1}. \end{cases} \tag{10.16}$$

We then have the following series of equations.

$$C_h - C_{h-1} = A_h + 1, \tag{10.17a}$$

$$C_{h-1} - C_{h-2} = A_{h-1} + 1, \tag{10.17b}$$

$$\dots,$$

$$C_1 - C_0 = A_1 + 1. \tag{10.17c}$$

Adding up the left and right hand sides, respectively, of the above set of equations, we have $C_h - C_0 = \sum_{h'=1}^{h} A_{h'} + h$. Together with Eq. (10.15), we arrive at the following expression of C_h.

$$C_h = h^2 + 3h + 1. \tag{10.18}$$

And the $C(A)$ relation can be further derived, as shown below.

$$C = \frac{A^2}{4} + A - \frac{1}{4}. \tag{10.19}$$

The asymptotic behavior is $C(A) \sim A^2$ at large A.

Furthermore, due to the symmetry of $i \to$ left and $i \to$ right in the recursive equation Eq. (10.5), as long as there is one and only one child node branching for every non-leaf node, the $C(A)$ relation has the same form as Eq. (10.19).

10.2 Topology of Phylogenetic Trees

We have seen from the calculation in the previous section that the $C(A)$ relation is determined by the topological structure of the tree. At large A, $C(A) \sim A \ln A$ for a completely balanced binary tree, while $C(A) \sim A^2$ for a completely imbalanced binary tree.

An actual phylogenetic tree is partially imbalanced, and the $C(A)$ scales in between the above two extrema. Several reports [23–25, 248] have found that $C(A) \sim A^\eta$, with $\eta \approx 1.4$. Especially, in Ref. [23], the researchers examined systematically a large set of phylogenetic trees both inter- and intra-species, and found a universal power-law asymptotic scaling of $C(A)$, with the exponent $\eta = 1.44$. Ref. [249] argues that the measurement methodology is influenced by bias due to uneven speciation rates, choice of taxa, choice of outgroups for the trees. However it does not explain how these effects could lead to power-law behavior of tree topology.

It has also been observed that the diversification rate on a phylogenetic tree declines over time [250–252]. In other words, as the speciation event proceeds from the root to the leaves, the rate decreases. This phenomenon is usually interpreted as due to the niche space being filled up and the carrying capacity being approached. However, the premise that the niche space has a fixed capacity is not necessarily correct.

10.3 Models for Phylogenetic Tree Topology

There have been many theoretical models on the evolution of a phylogenetic tree. See Ref. [252–255] for comprehensive reviews. The equal-rates-Markov (ERM) model and the proportional-to-distinguishable-arrangements (PDA) model are among the most popular ones.

The ERM model was first developed by Yule in 1924 [256], and later expanded in literature [257–259]. ERM assumes that all extant species on the tree have the same speciation rate. Despite being not always realistic, this simple model is usually used as a null hypothesis for the evolutionary process of the tree. The resultant tree, however, is less imbalanced than the observed ones. In fact, most local branching models invariably give rise to $C \sim A \ln A$ for large A, because they are essentially random walks at large A.

The original PDA model [260–262] did not involve any rules of the growth of the tree. The model assumes that for a given tree size, all tree topologies are equally likely to appear. The tree then is a result of recursively sampling the topology for the subtrees. Evolutionary processes that correspond to the PDA model were developed later [263, 264].

There were also attempts to directly address the scaling behaviors of the phylogenetic tree [243, 265–267]. But no solid conclusion has been reached yet. We will report in Chapter 11 our exploration to develop evolutionary process models for the observed scaling behavior. Our goal was to see if in principle it is possible for phylogenetic trees to exhibit power-law scaling of topological measure. Specially, we propose to explicitly incorporate niche construction in the evolutionary process and examine how it contributes to the topology of the phylogenetic tree.

Chapter 11

Effect of Niche Construction on the Topology of Phylogenetic Trees

As introduced in previous chapters, the effect of niche construction on evolution is not sufficiently appreciated, and the process that leads to the universal scaling of the phylogenetic tree is not well explained. In this chapter, I develop models that explicitly include simple caricatures of niche construction in the evolutionary process of a phylogenetic tree, and explore how this incorporation impacts the topology of the tree. In particular, we will see that the power-law behavior of $C(A)$ reported in natural phylogenetic trees can be recapitulated from the statistical effects of niche construction. We will see that the critical behavior is induced by a singularity that models the deactivation of nodes in the tree. Due to the generality of this effect, it is not impossible that the observed power-law scaling reflects the imprint of such a broad class of processes as niche construction. Indeed, our results echo the conclusions of an earlier analysis by O'Dwyer [248], which considered a specific metric characteristic of phylogenetic trees: the edge-length distribution. O'Dwyer's work suggested that the observed scale invariance in this quantity could not arise from neutral models of ecological communities. Our focus on niche construction provides a specific example of his conclusion, although we have not yet tested whether our results can also account for the edge-length distribution that he has reported from the data.

11.1 Niche Inheritance Model

When a new species emerges from its parent, it occupies a novel niche. An intuitive way to describe the process is to associate a species with a certain niche value, and let its children inherit the niche with an amount of fluctuation. We introduce the Niche Inheritance Model below.

11.1.1 Ingredients of the Niche Inheritance Model

We assign each species node three attributes: the amount of available niche n, the speciation rate r, and the extinction probability e. The speciation rate is treated as an increasing function of niche $r(n)$ in the sense that the more available niche there is, the more likely it is for a speciation event to be successful. It also implicitly wraps up all ecological interactions among species. For simplicity, we set

$$r(n) = \begin{cases} n, & n \geq 0 \\ r_\epsilon, & n < 0 \end{cases} \tag{11.1}$$

The extinction rate e can be positively related to r due to the fact that frequent speciation leads to heavy competition among species. We will discuss $e(r)$ in the next section.

Let the parent node be represented as (n_0, r_0, e_0). We first sample the time interval till its speciation based on a Poisson process with the rate r_0. Then forward time to the speciation moment. Let the parent diversify into two children (n_1, r_1, e_1) and (n_2, r_2, e_2). We treat the branching to be binary, because a multifurcation can be viewed as a coarse-grained bifurcation. The niche sizes n_1 and n_2 are inherited from the parent with fluctuations due to the construction/destruction, as expressed below:

$$n_1 = n_0 + \Delta n_1, \tag{11.2a}$$

$$n_2 = n_0 + \Delta n_2. \tag{11.2b}$$

The fluctuations Δn_i, $i = 1, 2$, are assumed to be generated by the following distribution:

$$\frac{\Delta n_i}{n_0} \sim \mathcal{N}(\mu_n, \sigma_n^2), \tag{11.3}$$

where $\mathcal{N}(\mu_n, \sigma_n^2)$ stands for a normal distribution with mean μ_n and variance σ_n^2. r_i and e_i are next calculated accordingly. For each child node, test whether it goes extinct or remains to bifurcate later. The test is done by drawing a uniformly distributed random number in $[0, 1]$, and comparing it with the extinction probability e. The child goes extinct and is removed from the tree, if the random number is smaller than e. All inferred nodes and branches dependent on the extinct child are also pruned away.

In a numerical simulation, we start with a root node and evolve the tree based on the above rules until it reaches a certain size. A and C for each node on the tree are then calculated following their recursive definitions Eq. (10.1) and Eq. (10.5), respectively.

11.1.2 Existence of the Absorbing Boundary

In the above framework of the Niche Inheritance Model, there exists a boundary case when $r_\epsilon = 0$, which means that nodes with negative niches will never bifurcate. We discuss qualitatively here the difference between a zero and nonzero r_ϵ and leave more quantitative details to Section 11.3.

Imagine the left node starts by chance with a larger n and thus a higher r than the right one. If $r_\epsilon \neq 0$ and all succeeding nodes are able to branch, then the right node, by fluctuation, will eventually gain a descendant with high r, and the left node will gain a descendant with low r. The two subtrees in general undergo the same random process and are symmetric. Therefore, at a long time scale, the entire tree is balanced.

However, if $r_\epsilon = 0$, once a node gets a negative niche, it is deactivated and will not be able to contribute any descendants in the evolutionary process. This eliminates the possibility of leveling up the general growth between the left and right subtrees. Leaf nodes that have long life times like those in the completely imbalanced case in Fig. 10.1(b) will emerge in the tree. Therefore, $r_\epsilon = 0$ drives the asymmetry of the tree and leads it to be imbalanced. We will refer to the case of $r_\epsilon = 0$ as the absorbing boundary, since it effectively removes bifurcating species from the tree.

In actual evolution, we observe species that seem not to be changing phenotypically while their relatives actively diversify, for example, the "living fossil" species coelacanth, which we have discussed in Chapter 7. It therefore is reasonable to use the absorbing boundary as a simplified starting point of analyzing the Niche Inheritance Model. In the rest of this chapter, we will first focus on the edge case of $r_\epsilon = 0$ to discuss the effect of niche construction on the tree topology, and then examine the effect of a finite r_ϵ.

11.2 Effect of Niche Construction on Tree Topology with the Absorbing Boundary

From the qualitative argument in the previous section, we already see that $r_\epsilon = 0$ can induce imbalance in the tree. But is this sufficient? Are there any other factors which are crucially indispensable in this model? The answer lies in the niche construction strength, represented by σ_n.

Based on Eq. (11.3), we see that σ_n tunes the probability for the child node to be associated with a negative niche value. Therefore, it determines how often the nodes hit the absorbing boundary. When $\sigma_n = 0$, the niche does not fluctuate and all nodes have the same value of niche as well as the same speciation rate. In this situation, the Niche Inheritance Model is equivalent to the Yule process [256]. We expect the resultant tree to be balanced with a significant symmetry. When σ_n is large, however, the access to the absorbing boundary is frequent. There will be many nodes turning inactive and many branches being terminated during the evolution. The effect of imbalance exerted by the absorbing boundary is therefore switched on and becomes visible.

In the rest of this section, we will present numerical simulations and a mean-field calculation to demonstrate the effect of niche construction on the topology of the phylogenetic trees, under the absorbing boundary of $r_\epsilon = 0$.

11.2.1 Topology of Trees with Extinctive Nodes

First, we let species nodes be extinctive, with the following probability,

$$e(r) = \frac{r}{r + R_0}. \tag{11.4}$$

This effectively limits the bifurcation rate of the tree, since nodes with high speciation rates become extinct and are removed.

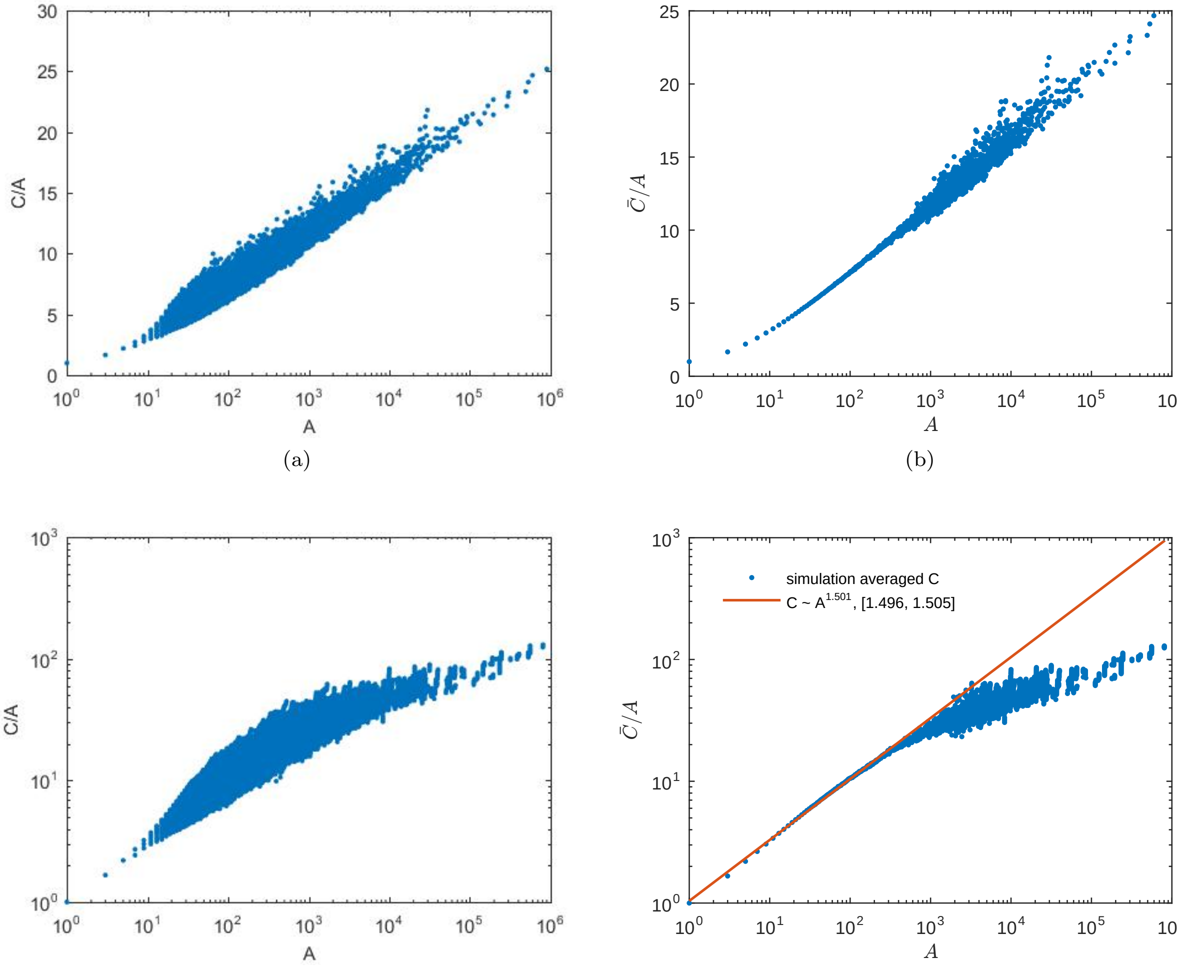
C(A) at Zero and Strong Niche Construction Strengths

We demonstrate in Fig. 11.1 the $C(A)$ relations in two extreme cases: zero niche construction ($\sigma_n = 0$) in the first row, and strong niche construction ($\sigma_n = 2$) in the second. In the left column of Fig. 11.1, each dot represents the (C, A) pair of a node. If two nodes have subtrees of the same size A but different topologies, then they will most likely host different values of C (except if one tree can be transformed to the other by mirroring the left and right branches). For a given size, there can be many subtrees of distinct topologies. Therefore, we usually have multiple C values associated with the same A, especially when A is not too small. However, if A is large, then there may only be few subtrees that have reached the desired size, and C hence comes in few values. Especially, when A is equal to the size of the phylogenetic tree, there is only one topology present, that of the tree itself, and C is single valued. The (C, A) pair now is associated with the root.

In the right column of Fig. 11.1, we average the C values corresponding to the same A, and present the resulted $\bar{C}$ *v.s.* A. At small A, there are many samples of subtrees and $\bar{C}$ reasonably represents the expected value. This is indicated by the thin and smooth region in the $\bar{C}$-A graph. However, at large A, there are few subtrees present, and the tree topology is thus heavily undersampled. $\bar{C}$ then does not reflect the true expected value. This is illustrated as the broad scattered region in the $\bar{C}$-A graph.

As discussed at the beginning of this section, for $\sigma_n = 0$, we expect the tree to be balanced with a $C(A) \sim A \ln A$ asymptotic behavior. This is verified in the first row of Fig. 11.1. Notice that the scale is linear-logarithmic and the $A \ln A$ behavior is illustrated by the dots scattering along a straight line.

When there is a significant niche construction effect or a large σ_n, we expect the tree to be imbalanced with $C(A)$ deviating from the balanced scaling. This is demonstrated in the second row of Fig. 11.1. Instead



(a)

(b)

(c)

(d)

Figure 11.1: (a) $C(A)$ calculated for a typical tree generated by the extinctive Niche Inheritance Model, with $\sigma_n = 0$. The dots scatter along a straight in the linear-logarithmic scale, indicating $C/A \sim \ln A$. (b) C is averaged over values with the same A, for the tree in (a). The smeared region at large A is due to a lack of subtree samples. (c) Typical $C(A)$ calculated with $\sigma_n = 2$. The scale is double logarithmic. (d) Averaged $C(A)$ for the tree in (c). Fitting the well-averaged region, $A < 200$, to a power function $C \sim A^\eta$ gives an exponent of $\eta = 1.501$, with the 95% confidence interval being $[1.496, 1.505]$. The red line stands for the fitted function. Other parameters for both sets of simulations are $r_\epsilon = 0$, $\mu_n = 0$, $R_0 = 10$, and $n_0 = 1$ for the root node.

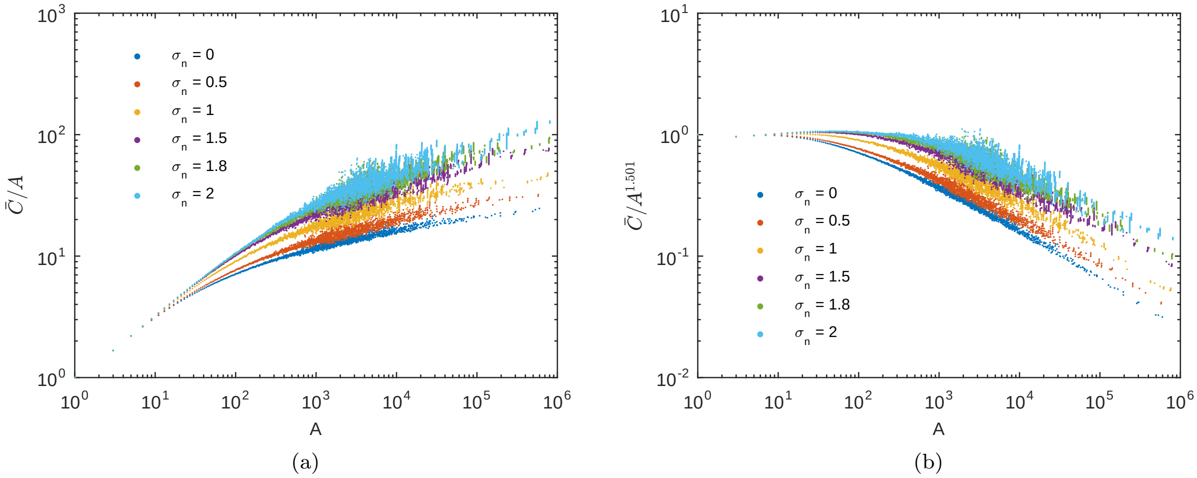


Figure 11.2: (a) Dependence of averaged $C(A)$ on σ_n, with $r_\epsilon = 0$, in the extinctive Niche Inheritance Model. As σ_n increases, the apparent power-law region of $\bar{C}(A)$ also stretches. Other parameters are $\mu_n = 0$, $R_0 = 10$, and $n_0 = 1$ for the root node. (b) is the compensated plot of (a). The vertical axis is $\bar{C}/A^\eta$, and the value of η is taken from the fit in Fig. 11.1(d). The flat region of each curve corresponds to the power-law regime.

of $A \ln A$, in the range comparable with the observed data [23], $A < 200$, $C(A)$ falls roughly along a straight line in the double logarithmic scale, indicating a power-law behavior, $C(A) \sim A^\eta$. A fitting to the power law function gives the exponent of $\eta \approx 1.50$.

C(A) at Intermediate Niche Construction Strengths

Knowing the end behaviors, we now move one step further to calculate $C(A)$ for intermediate values of σ_n. The results are presented in Fig. 11.2(a), as the averaged $\bar{C}(A)$ curves at different values of σ_n. Judging from the smooth regions at small A of $\bar{C}$-A curves, as σ_n increases from 0 to 2, the $\bar{C}(A)$ relation transitions from $A \ln A$ to an apparent power-law behavior. In the compensated plots in Fig. 11.2(b), the power-law regime is illustrated as the flat region at small A for each curve. It grows in width as σ_n increases. At the large A end, we do not have a strong conclusion, since the data is not adequately averaged.

Next, we will carry out a mean-field calculation on a simplified model with $e = 0$ and attempt to understand the dependence of $C(A)$ on σ_n.

11.2.2 Mean-field Calculation on the Non-extinctive Niche Inheritance Model

The extinction probability in the previous subsection depends positively on the speciation rate and thus induces a bias toward small rates in the evolutionary process. Here, we are interested in a mathematically simpler version without such a bias. This can be done by setting the extinction probability to be constant for all nodes. Furthermore, since any nonzero constant e can be mapped to $e = 0$ by effectively offsetting the speciation rates to $r(1 - e)$, we only need to look at the simplest situation with

$$e = 0. \tag{11.5}$$

It should be pointed out that removing the bound on the values of niche and speciation rate will result in the speciation rate growing exponentially large in a short time, since the niche of a child changes proportionally to its parent's niche as in Eq. (11.3). This is not biologically meaningful. Still, this simplified model can be handled mathematically from a mean-field point of view, and provides insights to the non-extinctive model, as will be discussed later.

In the rest of this subsection, we conduct a mean-field theory calculation for the non-extinctive Niche Inheritance Model to derive the dependence of $C(A)$ on σ_n. We again work in the presence of the absorbing boundary $r_\epsilon = 0$, so that a certain number of nodes will turn inactive and not branching during the evolutionary process. We will discuss the applicability of the mean-field assumption at the end of this subsection.

Deactivation Probability of Nodes

Comparing trees with different topologies, we observe the following facts. In a completely balanced binary tree, the leaf nodes are all active and can branch. In a completely imbalanced binary tree, only one child can branch and the other is inactive. A phylogenetic tree should lie somewhere in between the two extreme cases. If we define the deactivation probability of a leaf node as q, then

$$\begin{cases} q = 0, & \text{completely balanced binary tree,} \\ q = 0.5, & \text{completely imbalanced binary tree.} \end{cases} \tag{11.6}$$

Dependence of C(A) on the Deactivation Probability

Suppose there are n_d leaf nodes, when the tree evolves to depth d. Then on average, $n_d q$ of the leaves will turn inactive, and each of the remaining $n_d(1-q)$ nodes will branch into two leaves at depth $d+1$. Therefore, we have a recursive relation for n_d,

$$n_{d+1} = 2n_d(1 - q). \tag{11.7}$$

The general expression for n_d is then calculated to be

$$\begin{cases} n_0 = 1, \\ n_1 = 2, \\ n_d = n_1 a^{d-1}, \end{cases} \tag{11.8}$$

with $a = 2(1-q)$ as the average number of active children of one parent node. The full parameter range is $0 \leq q \leq 1$ and correspondingly $0 \leq a \leq 2$. However, for $1/2 < q \leq 1$ and $0 \leq a < 1$, the tree can not grow to a significant size, We thus exclude this situation from the consideration.

The subtree size A of a node at depth D is given by

$$A = \sum_{d=0}^{D} n_d. \tag{11.9}$$

The average depth of nodes in the subtree is

$$\langle d \rangle = \frac{\sum_{d=0}^{D} d n_d}{A}. \tag{11.10}$$

Following Eq. (10.4), we calculate C of the node as

$$C = A(\langle d \rangle + 1). \tag{11.11}$$

With Eq. (11.8), we obtain the following explicit expressions of A and C in terms of D, for $0 \leq q < 1/2$ and $1 < a \leq 2$.

$$A = 1 + 2\frac{a^D - 1}{a - 1}, \tag{11.12}$$

$$C = \frac{2}{a-1}[(D+1)a^D - 1] - \frac{2a(a^D - 1)}{(a-1)^2} + A. \tag{11.13}$$

The $C(A)$ relation is further given as follows, by eliminating D from the above two equations.

$$C(A) = (A-1)\log_a\left[\frac{(A-1)(a-1)}{2} + 1\right] + \frac{2}{a-1}\log_a\left[\frac{(A-1)(a-1)}{2} + 1\right] + (A-1) + \frac{a-A}{a-1}. \tag{11.14}$$

For the completely balanced tree with $a = 2$, $C(A)$ is reduced to

$$C(A) = (A+1)\log_2(A+1) - A, \tag{11.15}$$

which is equivalent to Eq. (10.13).

As $q \to 1/2$ and $a \to 1$, we can derive the limit form of $C(A)$, given below, using L'Hôpital's rule.

$$C(A) \to \frac{A^2+1}{2}. \tag{11.16}$$

Despite the different functional form, it has the same asymptotic scaling $C(A) \sim A^2$ at large A as Eq. (10.19).

A similar mean-field calculation has been performed in Ref. [265]. The authors reported the same qualitative result as what we have derived above, but with a slightly different model.

Dependence of C(A) on the Niche Construction Strength

Now we analyze the relationship between the parameter σ_n in the Niche Inheritance Model and the deactivation probability q.

Based on Eq. (11.2) and Eq. (11.3), the child node turns inactive if $n_0 + n_0 x < 0$, where x is the random number drawn from the distribution $\mathcal{N}(\mu_n, \sigma_n^2)$ to characterize the niche construction effect. Therefore, we have the following equation to link q and σ_n.

$$q = \text{Prob}(r = 0) = \text{Prob}(x < -1) = \frac{1}{2}\text{erfc}\left(\frac{1}{\sqrt{2}\sigma_n}\right). \tag{11.17}$$

So, for a given σ_n, we can compute q using the above equation, and then calculate $C(A)$ following Eq. (11.14). Figure 11.3 shows the $C(A)$ relations with different values of σ_n. When σ_n is finite, $C(A)$ always approaches $A \ln A$ when A is large. This can also be derived from Eq. (11.14). Despite the $A \ln A$ asymptote, there exists a range at small A, in which $C(A)$ can be approximated as a power function, as indicated by the reference line of $A^{1.5}$. As $\sigma_n \to +\infty$, $q \to 1/2$ and the asymptotic behavior approaches $C(A) \sim A^2$.

Comparing the mean-field result Fig. 11.3 with the simulation presented in Fig. 11.2, we conclude that the analytical calculation qualitatively captures the trend observed in the simulation, even though it is based on a simplified non-extinction version.

In the presence of the absorbing boundary $r_\epsilon = 0$, the strength of the niche construction effect strongly impacts the topology of the phylogenetic tree. When there is no construction, species are effectively the same and have the same bifurcation rate. The resultant tree is balanced, and $C(A) \sim A \ln A$.

As the construction effect gets strong, $C(A)$ develops an apparent power-law regime at small A, before transitioning to the $A \ln A$ asymptote. And the width of the regime increases with the construction strength. We point out that, in the data obtained from actual phylogenetic trees in Ref. [23], the tree size A is usually

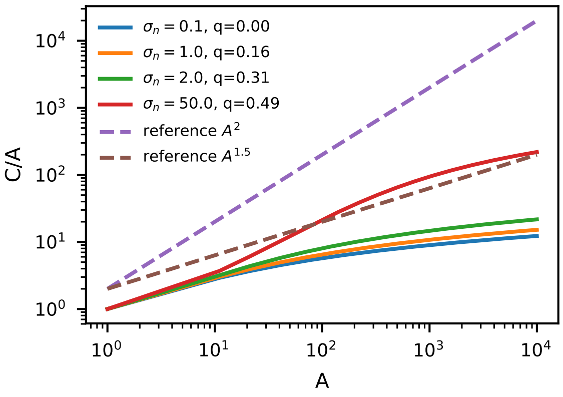


Figure 11.3: Mean-field analytical $C(A)$ at different values of σ_n. For finite σ_n, $C(A)$ always approaches $A \ln A$, but has a range at small A that resembles a power-law behavior, parallel to the reference line of $A^{1.5}$ in the double logarithmic scale. As $\sigma_n \to +\infty$, $q \to 1/2$ and the asymptotic behavior approaches $C(A) \sim A^2$.

limited to $O(100)$. In this range of A, our model successfully yields a power-law $C(A)$ relation.

Comments on the Mean-field Calculation

There are two main caveats in the above calculation. First, the calculation does not account for any stochasticity in the process. It is applicable to an averaged situation, since n_d is the expected number of nodes at depth d. Second, the calculation is only correct with an infinite growth time of the tree. In order to use the recursion relation Eq. (11.7), all active nodes at depth d have to be able to complete the branching process. This premise can always be achieved if the growth is terminated at $T = +\infty$. However, if T is finite, then there will always be some active nodes that will not branch before the termination. This effectively leads to a larger deactivation probability than the constant q. This effect is significant for nodes with small speciation rates, which can occur at any depth. Therefore, the effective deactivation probability $\tilde{q}_d$ should be dependent on the distribution of speciation rate at depth d.

Nevertheless, the mean-field calculation succeeds in describing the qualitative behavior of $C(A)$ at different σ_n. However, it does not truly explain the origin of scale-invariant behavior. As is well-known from the theory of critical phenomena, non-trivial power-law scaling arises from singularities in limit processes [268]. We turn to this question next.

11.3 Singularity Induced by the Absorbing Condition

In the previous sections, we have shown that with $r_\epsilon = 0$, the niche construction strength strongly impacts the $C(A)$ relation. Specifically, as σ_n increases, the tree becomes imbalanced, and $C(A)$ develops an apparent power-law at small A, before transitioning to $A \ln A$ at large A,

The imbalance induced by a large niche construction effect is crucially related to the condition $r(n) = r_\epsilon = 0$ for $n < 0$, which means that nodes stop branching when $n < 0$. As has been discussed in previous sections, the absorbing boundary is the origin of the imbalance of the phylogenetic tree. A large σ_n then ensures that a finite number of nodes will reach the absorbing boundary and be cut off the tree. Together, we see the development of a power-law scaling.

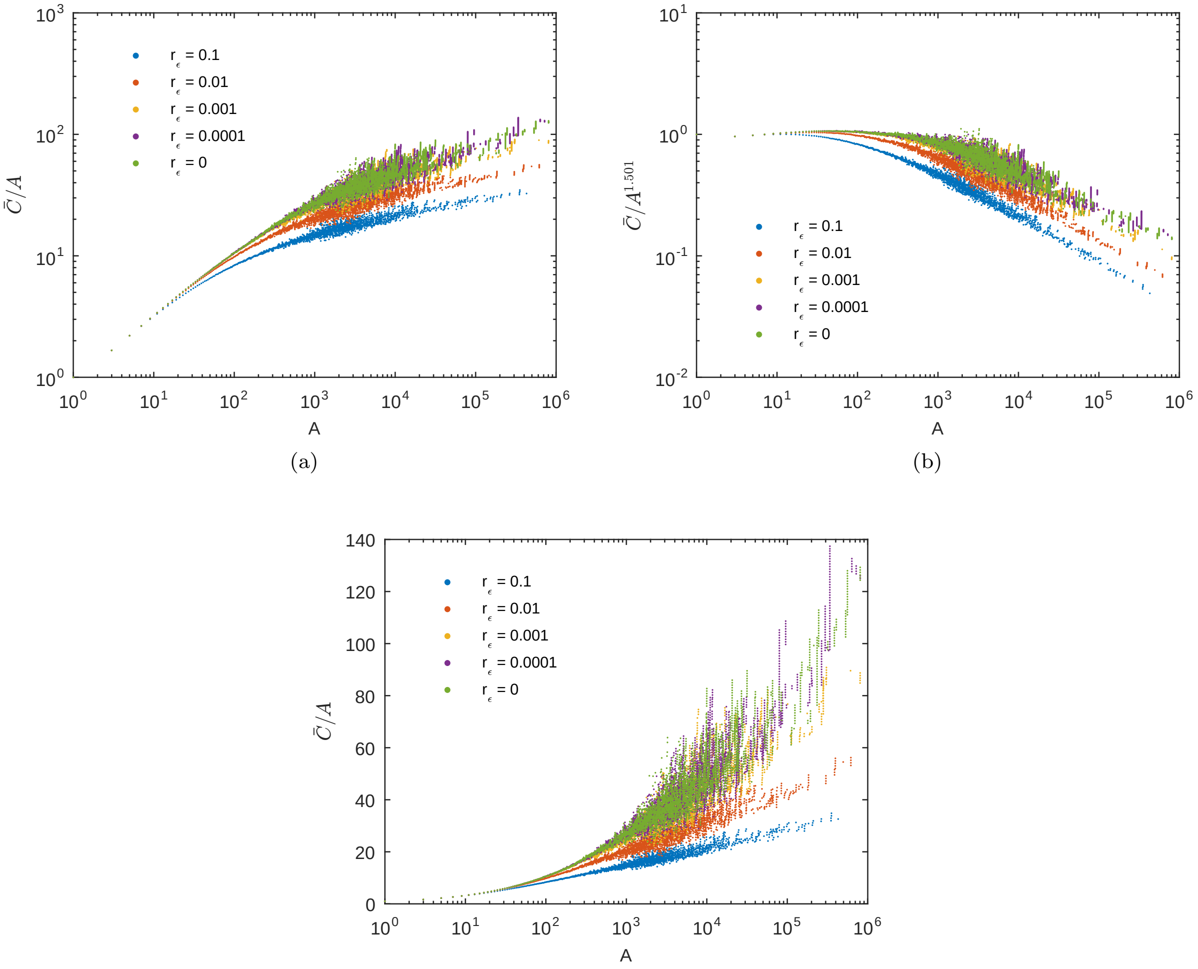
11.3.1 Effect of a Relaxed Absorbing Boundary

One question arises here, how crucial is the absorbing boundary to the imbalance power-law scaling? Or can r_ϵ be relaxed to a finite value?

The short answer is yes, if the tree has a finite growing time T. In this case, when r_ϵ is finite but still small enough *s.t.* $1/r_\epsilon \gg T$, then very few nodes with negative niches will be able to complete the bifurcation before the termination of the tree growth. So, effectively, a small r_ϵ acts as an absorbing boundary as well. As r_ϵ increases, the deactivation effect due to a finite T only acts on nodes near the tips. The symmetry between the left and right branches is gradually restored. Therefore, at large r_ϵ, the tree becomes balanced.

In the above argument, we have implied that nodes are able to reach the $n < 0$ region, so that r_ϵ can play a role in the evolution. For all the analyses in the rest of the section, we work in the framework of the extinctive Niche Inheritance Model and apply a large niche construction effect with $\sigma_n = 2$.

In Fig. 11.4, we show the dependence of $\bar{C}(A)$ on r_ϵ for trees terminated at a finite size $A \approx 10^6$, with $\sigma_n = 2$. When $r_\epsilon = 0$, we observe the apparent power-law regime of $\bar{C}(A)$ in the well-averaged range of A. This is demonstrated as the segment of straight line under the double logarithmic scale in Fig. 11.4(a) and the compensated plot Fig. 11.4(b). We have discussed the dependence of the power-law regime on σ_n in the previous section. As r_ϵ increases, the apparent power-law region of $\bar{C}(A)$ reduces in range. Eventually, a behavior of $A \ln A$ becomes significant in the entire range of A at a large r_ϵ, as illustrated by the straight line with $r_\epsilon = 0.1$ under the linear-logarithmic scale in Fig. 11.4(c).



(a)

(b)

(c)

Figure 11.4: (a) Dependence of averaged $C(A)$ on r_ϵ, with $\sigma_n = 2$, in the extinctive Niche Inheritance Model. As r_ϵ approaches zero, $\bar{C}(A)$ has a longer and longer apparent power-law range. Other parameters are $\mu_n = 0$, $R_0 = 10$, and $n_0 = 1$ for the root node. (b) The compensated plot of (a). The vertical axis is $\bar{C}/A^\eta$, and the value of η is taken from the fit in Fig. 11.1(d). The flat region of each curve corresponds to the power-law regime. (c) The same data as in (a) plotted in the linear-logarithmic scale. At large A, C/A appears as a straight line, indicating $C(A) \sim A \ln A$. This behavior is clear for large r_ϵ.

11.3.2 Critical Scaling at the Absorbing Boundary

Although the behavior of $\bar{C}(A)$ at large A in Fig. 11.4 is not well represented due to undersampling, we conjecture that, for a nonzero r_ϵ, $C(A)$ consists of two distinct asymptotic limits: at small A, $C(A) \sim A^\eta$; at large A, $C(A) \sim A \ln A$. The crossover happens around the transition point $A = A_T$. In fact, in Fig. 11.4, the curve corresponding to $r_\epsilon = 0.01$ has both a significant power-law region and a smooth $A \ln A$ region, before the issue of undersampling smears the data.

With this conjecture, we observe that A_T divides $C(A)$ into two regimes and that the transition point A_T increases as r_ϵ approaches 0. We claim that the dependence of A_T on r_ϵ is critical. Then, in terms of a phase transition language, there exists a crossover scaling function $F(x)$, with

$$x = r_\epsilon^a A, \tag{11.18}$$

such that

$$C(A, r_\epsilon) = A^\eta F(x). \tag{11.19}$$

The functional form of $F(x)$ should accommodate the fact that

$$C(A) \sim \begin{cases} A^\eta, & \text{small } A, \\ A, & \text{large } A. \end{cases} \tag{11.20}$$

Here, we have ignored the $\ln A$ correction at large A. We require that $F(x)$ has the following asymptotic behaviors.

$$F(x) \to \begin{cases} \text{const}, & \text{small } A \text{ and } x \to 0, \\ x^{1-\eta}, & \text{large } A \text{ and } x \to +\infty. \end{cases} \tag{11.21}$$

If the function $F(x)$ exists and our claim about the critical scaling is correct, then different data sets corresponding to different r_ϵ values should collapse onto the same curve, when plotted as C/A^η *vs.* $x = r_\epsilon^a A$. Indeed, in Fig. 11.5, we show the data collapse obtained by tuning a and η. For the presented eight data sets, $a = 2.5$ and $\eta = 1.55$ gives the best data collapse. At large x, the data show a shallower tail than the expected $x^{1-\eta}$ asymptote. We presume that this is because of the $\ln A$ factor carried along in the simulation data, but not accounted for in the crossover scaling argument.

The data collapse indicates a critical behavior of $C(A)$ as r_ϵ approaches 0. Notice that the value of η found from the data collapse is slightly different from the one obtained via fitting in Fig. 11.1(d). From the function $F(x)$, obtained in the data collapse, we can read off the value x_T at which $F(x)$ crosses over from

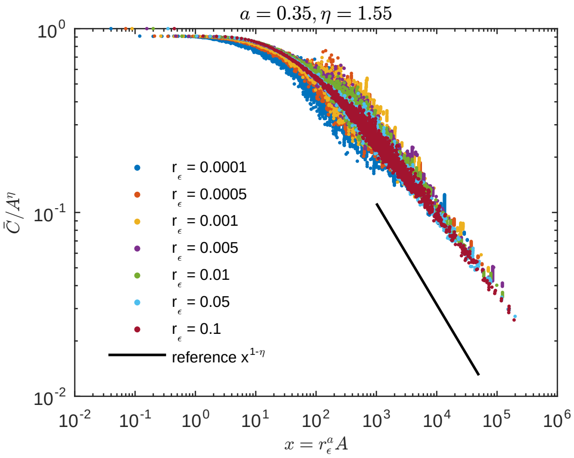


Figure 11.5: Critical scaling of $C(A)$ as r_ϵ decreases, indicated by the data collapse. We conduct simulations for eight values of r_ϵ ranging from 0 to 0.1, across four orders of magnitude. By tuning η and a, we reach a data collapse of the eight data sets. $a = 2.5$ and $\eta = 1.55$ gives the best result, shown in the figure in the double logarithmic scale. The tail is shallower than the expected $x^{1-\eta}$ behavior, which is indicated by the straight reference line, because the simulation data incorporate a $\ln A$ factor. Other parameters for all eight sets are $\sigma_n = 2$, $\mu_n = 0$, $R_0 = 10$, and $n_0 = 1$ for the root node.

a constant to $x^{1-\eta}$. In Fig. 11.5, this value is $x_T \approx 3.85$. Then, for an arbitrary r_ϵ, we can calculate A_T as follows:

$$A_T = x_T r_\epsilon^{-a}. \tag{11.22}$$

As $r_\epsilon \to 0$, $A_T \to +\infty$ and the power-law scaling of $C(A) \sim A^\eta$ expands to the entire range of A. We do not yet know if the scaling laws and scaling functions are universal, and if not, what are the relevant or marginal operators in the branching process that control the scaling laws.

11.4 Conclusion

I have presented a model to explain the observed universal scaling of phylogenetic trees. We incorporate niche construction as an explicit evolutionary process in the tree growth. By analyzing the Niche Inheritance Model, we make two major conclusions. First, a large niche construction effect, together with the absorbing boundary, leads to an apparent power-law regime in the tree topology. This is in the same range of A as observed in actual phylogenetic trees [23]. Second, the establishment of the power-law $C(A)$ relation is a critical phenomenon. We demonstrate this by analyzing the crossover of $C(A)$ from A^η at small A to $A \ln A$ at large A, with a r_ϵ-dependent threshold, reflecting a singular behavior in the niche construction model as $r_\epsilon \to 0$.

Our model has simple rules for the evolution of the tree. The significance is that there is a local in time interplay between the speciation rate and niche availability and that this can generate a critical behavior in $C(A)$ because of the singularity induced by the cutoff of $r_\epsilon = 0$ at negative n.

There are several issues that require further investigation. First, we have predicted a scaling form for the crossover point A_T as a function of r_ϵ, separating the power law and the $A \ln A$ regions. Actual phylogenetic trees, however, have small sizes that do not exceed A_T. Therefore, it is not possible to detect the crossover to $A \ln A$, and it remains unclear whether or not actual phylogenetic trees follow the critical scaling. Second, the exponent of the power-law behavior in our model is close to, but not exactly equal to the reported values. A detailed study should be done to show how the exponent in our model depends on other details in the evolutionary process. Third, our model is surely not the only explanation for the observed scaling behavior, but it shows that one must search for singular effects if a power-law $C(A)$ is to be recovered. Fourth, our model does not capture the decreasing cladogenesis rate that has been observed in actual phylogenetic trees. It remains to be examined whether incorporating sophisticated mechanisms to account for the realistic cladogenesis rate reduction would change the scaling and how.

References

[1] A. J. McKane and T. J. Newman, Physical Review Letters **94**, 218102 (2005).

[2] T. Butler and N. Goldenfeld, Physical Review E **80**, 030902 (2009).

[3] G. E. Hutchinson, The American Naturalist **95**, 137 (1961).

[4] J. S. Clark, M. Dietze, S. Chakraborty, P. K. Agarwal, I. Ibanez, S. LaDeau, and M. Wolosin, Ecology Letters **10**, 647 (2007).

[5] J. B. Wilson, New Zealand Journal of Ecology **13**, 17 (1990).

[6] G. Hardin, Science **131**, 1292 (1960).

[7] M. O. Hill, Ecology **54**, 427 (1973).

[8] T. Thingstad and R. Lignell, Aquatic Microbial Ecology **13**, 19 (1997).

[9] T. F. Thingstad, Limnology and Oceanography **45**, 1320 (2000).

[10] C. A. Thomas Jr, Annual Review of Genetics **5**, 237 (1971).

[11] International Human Genome Sequencing Consortium, Nature **409**, 860 (2001).

[12] G.-F. Richard, A. Kerrest, and B. Dujon, Microbiology and Molecular Biology Reviews **72**, 686 (2008).

[13] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman, Nature Reviews Genetics **8**, 973 (2007).

[14] D. Chalopin, S. Fan, O. Simakov, A. Meyer, M. Schartl, and J.-N. Volff, Journal of Experimental Zoology Part B: Molecular and Developmental Evolution **322**, 322 (2014).

[15] L. Cavin and G. Guinot, Frontiers in Ecology and Evolution **2**, 49 (2014).

[16] J. F. Brookfield, Nature Reviews Genetics **6**, 128 (2005).

[17] W. F. Doolittle, Proceedings of the National Academy of Sciences **110**, 5294 (2013).

[18] C. Darwin and A. Wallace, Zoological Journal of the Linnean Society **3**, 45 (1858).

[19] F. J. Odling-Smee, in *The Role of Behavior in Evolution*, edited by H. C. Plotkin (MIT Press, 1988) pp. 73–132.

[20] T. Yoshida, L. E. Jones, S. P. Ellner, G. F. Fussmann, and N. G. Hairston, Nature **424**, 303 (2003).

[21] S. P. Ellner, M. A. Geber, and N. G. Hairston, Ecology letters **14**, 603 (2011).

[22] H.-Y. Shih and N. Goldenfeld, Physical Review E **90**, 050702 (2014).

[23] E. A. Herrada, C. J. Tessone, K. Klemm, V. M. Eguíluz, E. Hernández-García, and C. M. Duarte, PlOS One **3**, e2757 (2008).

[24] P. Jeraldo Maldonado, *Computational Approaches to Stochastic Systems in Physics and Biology*, Ph.D. thesis, University of Illinois at Urbana-Champaign (2012).

[25] N. Goldenfeld, RNA biology **11**, 248 (2014).

[26] C. Xue and N. Goldenfeld, Physical Review Letters **119**, 268101 (2017).

[27] C. Xue and N. Goldenfeld, Physical Review Letters **117**, 208101 (2016).

[28] G. Lee, N. A. Sherer, N. H. Kim, E. Rajic, D. Kaur, N. Urriola, K. M. Martini, C. Xue, N. Goldenfeld, and T. E. Kuhlman, "Testing the retroelement invasion hypothesis for the emergence of the ancestral eukaryotic cell," (2018), in review.

[29] T. Rogers, A. J. McKane, and A. G. Rossberg, EPL (Europhysics Letters) **97**, 40008 (2012).

[30] S. Roy and J. Chattopadhyay, Ecological Complexity **4**, 26 (2007).

[31] P. Chesson, Annual Review of Ecology and Systematics **31**, 343 (2000).

[32] R. Levins, The American Naturalist **114**, 765 (1979).

[33] P. Richerson, R. Armstrong, and C. R. Goldman, Proceedings of the National Academy of Sciences **67**, 1710 (1970).

[34] S. A. Levin, The American Naturalist **108**, 207 (1974).

[35] J. W. Fox, Trends in Ecology & Evolution **28**, 86 (2013).

[36] M. Scheffer, S. Rinaldi, J. Huisman, and F. J. Weissing, Hydrobiologia **491**, 9 (2003).

[37] J. Roughgarden and M. Feldman, Ecology **56**, 489 (1975).

[38] K. Vetsigian, R. Jajoo, and R. Kishony, PLoS Biology **9**, e1001184 (2011).

[39] C. Winter, T. Bouvier, M. G. Weinbauer, and T. F. Thingstad, Microbiology and Molecular Biology Reviews **74**, 42 (2010).

[40] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[41] C. E. Shannon, Bell System Technical Journal **27**, 623 (1948).

[42] E. H. Simpson, nature **163**, 688 (1949).

[43] M. Newman, Contemporary Physics **46**, 323 (2005).

[44] G. K. Zipf, *The Psychology of Language* (Houghton-Mifflin, 1935).

[45] A. Fernández, S. Huang, S. Seston, J. Xing, R. Hickey, C. Criddle, and J. Tiedje, Applied and Environmental Microbiology **65**, 3697 (1999).

[46] P. J. DeVRIES, D. Murray, and R. Lande, Biological journal of the Linnean Society **62**, 343 (1997).

[47] H. Ter Steege, N. C. Pitman, D. Sabatier, C. Baraloto, R. P. Salomão, J. E. Guevara, O. L. Phillips, C. V. Castilho, W. E. Magnusson, J.-F. Molino, *et al.*, Science **342**, 1243092 (2013).

[48] P. Jeraldo, M. Sipos, N. Chia, J. M. Brulc, A. S. Dhillon, M. E. Konkel, C. L. Larson, K. E. Nelson, A. Qu, L. B. Schook, F. Yang, B. A. White, and N. Goldenfeld, Proceedings of the National Academy of Sciences **109**, 9692 (2012).

[49] D. Morse, N. Stork, and J. Lawton, Ecological Entomology **13**, 25 (1988).

[50] E. Siemann, D. Tilman, and J. Haarstad, Journal of Animal Ecology **68**, 824 (1999).

[51] I. Perfecto and J. Vandermeer, Environmental Entomology **42**, 38 (2013).

[52] J. C. Stearns, M. D. Lynch, D. B. Senadheera, H. C. Tenenbaum, M. B. Goldberg, D. G. Cvitkovitch, K. Croitoru, G. Moreno-Hagelsieb, and J. D. Neufeld, Scientific Reports **1**, 170 (2011).

[53] T. Pommier, P. R. Neal, J. M. Gasol, M. Coll, S. G. Acinas, and C. Pedrós-Alió, Aquatic Microbial Ecology **61**, 221 (2010).

[54] C. Pedrós-Alió, M. Potvin, and C. Lovejoy, Progress in Oceanography **139**, 233 (2015).

[55] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, . Tara Oceans Coordinators, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, P. Weissenbach, Jean andWincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork, Science **348**, 1261359 (2015).

[56] E. Villar, G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B. Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, F. d'Ortenzio, J.-P. Gattuso, N. Grima, L. Guidi, C. N. Hill, O. Jahn, J.-L. Jamet, H. Le Goff, C. Lepoivre, S. Malviya, E. Pelletier, J.-B. Romagnan, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, T. Vannier, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, . Tara Oceans Coordinators, S. G. Acinas, P. Bork, E. Boss, C. de Vargas, G. Gorsky, H. Ogata, S. Pesant, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, and D. Iudicone, Science **348**, 1261447 (2015).

[57] J. R. Brum, J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, C. Cruaud, C. De Vargas, J. M. Gasol, G. Gorsky, A. C. Gregory, L. Guidi, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, B. T. Poulos, S. M. Schwenck, S. Speich, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, . Tara Oceans Coordinators, P. Bork, C. Bowler, S. Sunagawa, P. Wincker, E. Karsenti, and m. B. Sullivan, Science **348**, 1261498 (2015).

[58] C. De Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, . Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, J. Pesant, Stephane Raes, M. E. Sieracki, S. Speich, L. Stemman, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti, Science **348**, 1261605 (2015).

[59] G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. D. Meester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, . Tara Oceans Corordinators, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemman, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, and J. Raes, Science **348**, 1262073 (2015).

[60] D. Gravel, C. D. Canham, M. Beaudet, and C. Messier, Ecology Letters **9**, 399 (2006).

[61] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*, Vol. 32 (Princeton University Press, 2001).

[62] B. J. Enquist, J. Sanderson, and M. D. Weiser, Science **295**, 1835 (2002).

[63] B. J. McGill, Nature **422**, 881 (2003).

[64] B. J. McGill, B. A. Maurer, and M. D. Weiser, Ecology **87**, 1411 (2006).

[65] M. Dornelas, S. R. Connolly, and T. P. Hughes, Nature **440**, 80 (2006).

[66] S. R. Connolly, M. A. MacNeil, M. J. Caley, N. Knowlton, E. Cripps, M. Hisano, L. M. Thibaut, B. D. Bhattacharya, L. Benedetti-Cecchi, R. E. Brainard, A. Brandt, F. Bulleri, K. E. Ellingsen, S. Kaiser, I. Kröncke, K. Linse, E. Maggi, T. D. O'Hara, L. Plaisance, G. C. B. Poore, S. K. Sarkar, K. K. Satpathy, U. Schückel, A. Williams, and R. S. Wilson, Proceedings of the National Academy of Sciences **111**, 8524 (2014).

[67] R. H. MacArthur, Proceedings of the National Academy of Sciences **43**, 293 (1957).

[68] R. MacArthur, The American Naturalist **94**, 25 (1960).

[69] G. Sugihara, The American Naturalist **116**, 770 (1980).

[70] M. Tokeshi, Journal of Animal Ecology **59**, 1129 (1990).

[71] M. Tokeshi, in *Advances in Ecological Research*, Vol. 24 (Elsevier, 1993) pp. 111–186.

[72] M. Tokeshi, Oikos **75**, 543 (1996).

[73] J. Fargione, C. S. Brown, and D. Tilman, Proceedings of the National Academy of Sciences **100**, 8916 (2003).

[74] D. Mouillot, M. George-Nascimento, and R. Poulin, Journal of Animal Ecology **72**, 757 (2003).

[75] T. J. Matthews and R. J. Whittaker, Ecology and Evolution **4**, 2263 (2014).

[76] T. Zillio and R. Condit, Oikos **116**, 931 (2007).

[77] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan, Nature **424**, 1035 (2003).

[78] J. Harte, Nature **424**, 1006 (2003).

[79] R. A. Chisholm and S. W. Pacala, Proceedings of the National Academy of Sciences **107**, 15821 (2010).

[80] S. Maslov and K. Sneppen, Scientific Reports **7**, 39642 (2017).

[81] D. H. Janzen, The American Naturalist **104**, 501 (1970).

[82] J. H. Connell, "On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees," in *Dynamics of Population*, edited by P. J. den Boer and G. R. Gradwell (PUDOC, 1971) pp. 298–312.

[83] S. Louca and M. Doebeli, Environmental microbiology **19**, 3863 (2017).

[84] O. Ovaskainen and B. Meerson, Trends in Ecology & Evolution **25**, 643 (2010).

[85] O. Gottesman and B. Meerson, Physical Review E **85**, 021140 (2012).

[86] D. T. Gillespie, Journal of Computational Physics **22**, 403 (1976).

[87] E. J. Routh, *A Treatise on the Stability of a Given State of Motion: Particularly Steady Motion* (Macmillan and Company, 1877).

[88] A. Hurwitz, Mathematische Annalen **46**, 273 (1895).

[89] J. Bergelson, M. Kreitman, E. A. Stahl, and D. Tian, Science **292**, 2281 (2001).

[90] S. T. Chisholm, G. Coaker, B. Day, and B. J. Staskawicz, Cell **124**, 803 (2006).

[91] C. Person, Canadian Journal of Botany **37**, 1101 (1959).

[92] J. N. Thompson and J. J. Burdon, Nature **360**, 121 (1992).

[93] A. Buckling and P. B. Rainey, Nature **420**, 496 (2002).

[94] A. Buckling and P. B. Rainey, Proceedings of the Royal Society of London B: Biological Sciences **269**, 931 (2002).

[95] L. Råberg, E. Alacid, E. Garces, and R. Figueroa, Ecology and Evolution **4**, 4775 (2014).

[96] P. Luijckx, H. Fienberg, D. Duneau, and D. Ebert, Current Biology **23**, 1085 (2013).

[97] C. Adema and E. Loker, Developmental & Comparative Immunology **48**, 275 (2015).

[98] U. Dieckmann, P. Marrow, and R. Law, Journal of Theoretical Biology **176**, 91 (1995).

[99] U. Dieckmann and R. Law, Journal of Mathematical Biology **34**, 579 (1996).

[100] A. Agrawal and C. M. Lively, Evolutionary Ecology Research **4**, 91 (2002).

[101] J. S. Weitz, H. Hartman, and S. A. Levin, Proceedings of the National Academy of Sciences of the United States of America **102**, 9535 (2005).

[102] S. J. Beckett and H. T. Williams, Interface Focus **3**, 20130033 (2013).

[103] B. J. Bohannan and R. E. Lenski, Ecology **78**, 2303 (1997).

[104] T. Yoshida, S. P. Ellner, L. E. Jones, B. J. Bohannan, R. E. Lenski, and N. G. Hairston Jr, PLoS Biology **5**, e235 (2007).

[105] F. Rodriguez-Valera, A.-B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pašić, T. F. Thingstad, F. Rohwer, and A. Mira, Nature Reviews Microbiology **7**, 828 (2009).

[106] A. Shmida and S. Ellner, Plant Ecology **58**, 29 (1984).

[107] U. C. Täuber, Journal of Physics A: Mathematical and Theoretical **45**, 405002 (2012).

[108] T. Butler and N. Goldenfeld, Physical Review E **84**, 011112 (2011).

[109] T. Yoshida, L. E. Jones, S. P. Ellner, G. F. Fussmann, and N. G. Hairston, Nature **424**, 303 (2003).

[110] G. W. A. Constable, T. Rogers, A. J. McKane, and C. E. Tarnita, Proceedings of the National Academy of Sciences , 201603693 (2016).

[111] P. Supply, E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht, Molecular Microbiology **36**, 762 (2000).

[112] M. Bichara, J. Wagner, and I. Lambert, Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **598**, 144 (2006).

[113] N. Delaroque, D. G. Müller, G. Bothe, T. Pohl, R. Knippers, and W. Boland, Virology **287**, 112 (2001).

[114] M. Chen, Z. Tan, G. Zeng, and Z. Zeng, Infection, Genetics and Evolution **12**, 1452 (2012).

[115] S. Kit, Journal of Molecular Biology **3**, 711 (1961).

[116] Y.-C. Li, A. B. Korol, T. Fahima, A. Beiles, and E. Nevo, Molecular Ecology **11**, 2453 (2002).

[117] Y.-C. Li, A. B. Korol, T. Fahima, and E. Nevo, Molecular Biology and Evolution **21**, 991 (2004).

[118] A. R. Wyman and R. White, Proceedings of the National Academy of Sciences **77**, 6754 (1980).

[119] G. Vergnaud and F. Denoeud, Genome Research **10**, 899 (2000).

[120] F. Pâques, W.-Y. Leung, and J. E. Haber, Molecular and Cellular Biology **18**, 2045 (1998).

[121] H. Fan and J.-Y. Chu, Genomics, Proteomics & Bioinformatics **5**, 7 (2007).

[122] G.-F. Richard and F. Pâques, EMBO Reports **1**, 122 (2000).

[123] A. Bhargava and F. Fuentes, Molecular Biotechnology **44**, 250 (2010).

[124] J. C. Kim and S. M. Mirkin, Current Opinion in Genetics & Development **23**, 280 (2013).

[125] R. Gemayel, M. D. Vinces, M. Legendre, and K. J. Verstrepen, Annual Review of Genetics **44**, 445 (2010).

[126] Y. Kashi and D. G. King, TRENDS in Genetics **22**, 253 (2006).

[127] A. J. Jeffreys, V. Wilson, and S. L. Thein, Nature **314**, 67 (1985).

[128] C. Vitte and O. Panaud, Cytogenetic and Genome Research **110**, 91 (2005).

[129] J. S. Han, Mobile DNA **1**, 15 (2010).

[130] A. B. Hickman and F. Dyda, Chemical Reviews **116**, 12758 (2016).

[131] M. Muñoz-López and J. L. García-Pérez, Current Genomics **11**, 115 (2010).

[132] H. H. Kazazian, Jr., Science **303**, 1626 (2004).

[133] B. A. Dombroski, S. L. Mathias, E. Nanthakumar, A. F. Scott, and H. H. Kazazian, Jr., Science **254**, 1805 (1991).

[134] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, *et al.*, Science **326**, 1112 (2009).

[135] S. Ohno, in *Brookhaven Symposia in Biology*, Vol. 23 (1972) pp. 366–370.

[136] L. E. Orgel and F. H. Crick, Nature **284**, 604 (1980).

[137] W. F. Doolittle and C. Sapienza, Nature **284**, 601 (1980).

[138] C. Biémont and C. Vieira, Nature **443**, 521 (2006).

[139] ENCODE Project Consortium *et al.*, Nature **489**, 57 (2012).

[140] B. McClintock, Proceedings of the National Academy of Sciences **36**, 344 (1950).

[141] B. McClintock, Genetics **38**, 579 (1953).

[142] A. M. Lambowitz and S. Zimmerly, Annual Review of Genetics **38**, 1 (2004).

[143] A. M. Pyle, Annual Review of Biophysics **45**, 183 (2016).

[144] B. Cousineau, S. Lawrence, D. Smith, and M. Belfort, Nature **404**, 1018 (2000).

[145] A. Lambowitz and M. Belfort, Microbiology Spectrum **3**, MDNA3 (2015).

[146] R. K. Slotkin and R. Martienssen, Nature Reviews Genetics **8**, 272 (2007).

[147] D. Gebert and D. Rosenkranz, Wiley Interdisciplinary Reviews: RNA **6**, 687 (2015).

[148] A. A. Aravin, G. J. Hannon, and J. Brennecke, Science **318**, 761 (2007).

[149] G. J. Hannon, Nature **418**, 244 (2002).

[150] Z. D. Smith and A. Meissner, Nature Reviews Genetics **14**, 204 (2013).

[151] M. M. Suzuki and A. Bird, Nature Reviews Genetics **9**, 465 (2008).

[152] M. F. Singer, Cell **28**, 433 (1982).

[153] B. Brouha, J. Schustak, R. M. Badge, S. Lutz-Prigge, A. H. Farley, J. V. Moran, and H. H. Kazazian, Jr., Proceedings of the National Academy of Sciences **100**, 5280 (2003).

[154] R. E. Mills, E. A. Bennett, R. C. Iskow, and S. E. Devine, Trends in Genetics **23**, 183 (2007).

[155] D. V. Babushok and H. H. Kazazian, Jr., Human Mutation **28**, 527 (2007).

[156] G. D. Swergold, Molecular and Cellular Biology **10**, 6718 (1990).

[157] M. Speek, Molecular and Cellular Biology **21**, 1973 (2001).

[158] Q. Feng, J. V. Moran, H. H. Kazazian, Jr., and J. D. Boeke, Cell **87**, 905 (1996).

[159] S. L. Mathias, A. F. Scott, H. H. Kazazian, Jr., J. D. Boeke, and A. Gabriel, Science **254**, 1808 (1991).

[160] S. R. Richardson, A. J. Doucet, H. C. Kopera, J. B. Moldovan, J. L. Garcia-Pérez, and J. V. Moran, Microbiology Spectrum **3**, MDNA3 (2015).

[161] J. V. Moran, S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke, and H. H. Kazazian, Jr., Cell **87**, 917 (1996).

[162] A. M. Denli, I. Narvaiza, B. E. Kerman, M. Pena, C. Benner, M. C. Marchetto, J. K. Diedrich, A. Aslanian, J. Ma, J. J. Moresco, *et al.*, Cell **163**, 583 (2015).

[163] M. A. Batzer and P. L. Deininger, Nature Reviews Genetics **3**, 370 (2002).

[164] E. Ullu and C. Tschudi, Nature **312**, 171 (1984).

[165] N. Okada, M. Hamada, I. Ogiwara, and K. Ohshima, Gene **205**, 229 (1997).

[166] V. Ahl, H. Keller, S. Schmidt, and O. Weichenrieder, Molecular Cell **60**, 715 (2015).

[167] J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius, and J. Schmitz, Trends in Genetics **23**, 158 (2007).

[168] P. Walter and G. Blobel, Nature **299**, 691 (1982).

[169] Y. Nyathi, B. M. Wilkinson, and M. R. Pool, Biochimica et Biophysica Acta (BBA)-Molecular Cell Research **1833**, 2392 (2013).

[170] O. Weichenrieder, K. Wild, K. Strub, and S. Cusack, Nature **408**, 167 (2000).

[171] M. Dewannieux, C. Esnault, and T. Heidmann, Nature Genetics **35**, 41 (2003).

[172] A. J. Doucet, J. E. Wilusz, T. Miyoshi, Y. Liu, and J. V. Moran, Molecular Cell **60**, 728 (2015).

[173] W. Wei, N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag, H. H. Kazazian, Jr., J. D. Boeke, and J. V. Moran, Molecular and Cellular Biology **21**, 1429 (2001).

[174] J. D. Boeke, Nature Genetics **16**, 6 (1997).

[175] A. M. Weiner, Current Opinion in Cell Biology **14**, 343 (2002).

[176] A. Sousa, C. Bourgard, L. M. Wahl, and I. Gordo, Biology Letters **9**, 20130838 (2013).

[177] X. Maside, C. Bartolome, S. Assimacopoulos, and B. Charlesworth, Genetics Research **78**, 121 (2001).

[178] N. A. Rosenberg, A. G. Tsolaki, and M. M. Tanaka, Theoretical Population Biology **63**, 347 (2003).

[179] C. R. L. Huang, K. H. Burns, and J. D. Boeke, Annual Review of Genetics **46**, 651 (2012).

[180] R. Cordaux, D. J. Hedges, S. W. Herke, and M. A. Batzer, Gene **373**, 134 (2006).

[181] B. Charlesworth and D. Charlesworth, Genetical Research **42**, 1 (1983).

[182] B. Charlesworth, P. Sniegowski, and W. Stephan, Nature **371**, 215 (1994).

[183] C. H. Langley, J. F. Brookfield, and N. Kaplan, Genetics **104**, 457 (1983).

[184] J. F. Brookfield and R. M. Badge, Genetica **100**, 281 (1997).

[185] A. Le Rouzic and P. Capy, Genetics **174**, 785 (2006).

[186] A. Le Rouzic and G. Deceliere, Genetical Research **85**, 171 (2005).

[187] S. Venner, C. Feschotte, and C. Biémont, Trends in Genetics **25**, 317 (2009).

[188] A. Le Rouzic, S. Dupas, and P. Capy, Gene **390**, 214 (2007).

[189] A. Le Rouzic, T. S. Boutin, and P. Capy, Proceedings of the National Academy of Sciences **104**, 19375 (2007).

[190] F. Serra, V. Becher, and H. Dopazo, PLOS One **8**, e63915 (2013).

[191] S. Linquist, K. Cottenie, T. A. Elliott, B. Saylor, S. C. Kremer, and T. R. Gregory, Molecular Ecology **24**, 3232 (2015).

[192] T. Biancalani, D. Fanelli, and F. Di Patti, Physical Review E **81**, 046215 (2010).

[193] B. Houchmandzadeh, Journal of Biosciences **39**, 249 (2014).

[194] H. Fort, Entropy **15**, 5237 (2013).

[195] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd ed. (Elsevier Science, 2007).

[196] K. Itô, Proceedings of the Imperial Academy **20**, 519 (1944).

[197] L. Michaelis and M. L. Menten, Biochem. z **49**, 352 (1913).

[198] M. Pineda-Krch, H. J Blok, U. Dieckmann, and M. Doebeli, Oikos **116**, 53 (2007).

[199] M. W. Nachman and S. L. Crowell, Genetics **156**, 297 (2000).

[200] J. C. Roach, G. Glusman, A. F. A. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas, Science **328**, 636 (2010).

[201] N. H. Kim, G. Lee, N. A. Sherer, K. M. Martini, N. Goldenfeld, and T. E. Kuhlman, Proceedings of the National Academy of Sciences **113**, 7278 (2016).

[202] M. Elez, A. W. Murray, L.-J. Bi, X.-E. Zhang, I. Matic, and M. Radman, Current Biology **20**, 1432 (2010).

[203] T. E. Kuhlman, (2017), private communication.

[204] E. Zuckerkandl and L. Pauling, *Horizons in Biochemistry*, edited by M. Kasha and B. Pullman (Academic Press, 1962) pp. 180 – 225.

[205] E. Margoliash, Proceedings of the National Academy of Sciences **50**, 672 (1963).

[206] S. Kumar, Nature Reviews Genetics **6**, 654 (2005).

[207] T. H. Jukes and C. Cantor, *Evolution of Protein Molecules* (Academic Press, 1969) pp. 21 –132.

[208] M. Kimura, Journal of Molecular Evolution **16**, 111 (1980).

[209] A. J. Drummond, S. Y. Ho, M. J. Phillips, and A. Rambaut, PLoS biology **4**, e88 (2006).

[210] D. Brawand, C. E. Wagner, Y. I. Li, M. Malinsky, I. Keller, S. Fan, O. Simakov, A. Y. Ng, Z. W. Lim, E. Bezault, *et al.*, Nature **513**, 375 (2014).

[211] J. A. Chapman, E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, T. Rattei, P. G. Balasubramanian, J. Borman, D. Busam, *et al.*, Nature **464**, 592 (2010).

[212] M. J. Genner, O. Seehausen, D. H. Lunt, D. A. Joyce, P. W. Shaw, G. R. Carvalho, and G. F. Turner, Molecular Biology and Evolution **24**, 1269 (2007).

[213] M. Naville, D. Chalopin, D. Casane, P. Laurenti, and J.-N. Volff, Mobile Genetic Elements **5**, 55 (2015).

[214] M. Naville, D. Chalopin, and J.-N. Volff, PLOS One **9**, e114382 (2014).

[215] C. T. Amemiya, T. P. Powers, S. J. Prohaska, J. Grimwood, J. Schmutz, M. Dickson, T. Miyake, M. A. Schoenborn, R. M. Myers, F. H. Ruddle, *et al.*, Proceedings of the National Academy of Sciences **107**, 3622 (2010).

[216] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Physical Review Letters **73**, 3169 (1994).

[217] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein, Genome Biology **3**, 1 (2002).

[218] S. S. Sindi, B. R. Hunt, and J. A. Yorke, Physical Review E **78**, 061912 (2008).

[219] M. Huynen and E. van Nimwegen, Molecular Biology and Evolution **15**, 583 (1998).

[220] J. Qian, N. M. Luscombe, and M. Gerstein, Journal of Molecular Biology **313**, 673 (2001).

[221] D. B. Searls, Nature **420**, 211 (2002).

[222] A. Magner, W. Szpankowski, and D. Kihara, Scientific Reports **5**, 8166 (2015).

[223] W. J. Reed and B. D. Hughes, Mathematical Biosciences **189**, 97 (2004).

[224] T. Hughes and D. A. Liberles, Gene **414**, 85 (2008).

[225] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, BMC Evolutionary Biology **2**, 1 (2002).

[226] R. C. Lewontin, in *Evolutoin from Molecules to Men*, edited by D. S. Bendall (Cambridge University Press, 1983).

[227] K. N. Laland, F. J. Odling-Smee, and M. W. Feldman, Proceedings of the National Academy of Sciences **96**, 10242 (1999).

[228] F. J. Odling-Smee, K. N. Laland, and M. W. Feldman, *Niche Construction: the Neglected Process in Evolution*, 37 (Princeton University Press, 2003).

[229] K. Laland, B. Matthews, and M. W. Feldman, Evolutionary Ecology **30**, 191 (2016).

[230] A. Bekker, H. Holland, P.-L. Wang, D. Rumble, H. Stein, J. Hannah, L. Coetzee, and N. Beukes, Nature **427**, 117 (2004).

[231] C. G. Jones, J. H. Lawton, and M. Shachak, Oikos **69**, 373 (1994).

[232] A. Hastings, J. E. Byers, J. A. Crooks, K. Cuddington, C. G. Jones, J. G. Lambrinos, T. S. Talley, and W. G. Wilson, Ecology Letters **10**, 153 (2007).

[233] G. Barker and J. Odling-Smee, in *Entangled Life* (Springer, 2014) pp. 187–211.

[234] J. Odling-Smee, D. H. Erwin, E. P. Palkovacs, M. W. Feldman, and K. N. Laland, The Quarterly Review of Biology **88**, 3 (2013).

[235] K. N. Laland, J. Odling-Smee, and M. W. Feldman, Nature **429**, 609 (2004).

[236] E. Gilad, J. von Hardenberg, A. Provenzale, M. Shachak, and E. Meron, Journal of Theoretical Biology **244**, 680 (2007).

[237] K. Cuddington, W. G. Wilson, and A. Hastings, The American Naturalist **173**, 488 (2009).

[238] W. Gurney and J. Lawton, Oikos **76**, 273 (1996).

[239] D. C. Krakauer, K. M. Page, and D. H. Erwin, The American Naturalist **173**, 26 (2009).

[240] K. N. Laland and K. Sterelny, Evolution **60**, 1751 (2006).

[241] M. Gupta, N. Prasad, S. Dey, A. Joshi, and T. Vidya, Journal of Genetics **96**, 491 (2017).

[242] M. W. Feldman, J. Odling-Smee, and K. N. Laland, Journal of genetics **96**, 505 (2017).

[243] E. Hernandez-Garcia, M. Tuğrul, E. Alejandro Herrada, V. M. Eguiluz, and K. Klemm, International Journal of Bifurcation and Chaos **20**, 805 (2010).

[244] C. Colijn and G. Plazzotta, Systematic Biology **67**, 113 (2017).

[245] J. R. Banavar, A. Maritan, and A. Rinaldo, Nature **399**, 130 (1999).

[246] A. Masucci, Physica A: Statistical Mechanics and its Applications **390**, 4652 (2011).

[247] A. Herrada, V. M. Eguíluz, E. Hernández-García, and C. M. Duarte, BMC Evolutionary Biology **11**, 155 (2011).

[248] J. P. O'Dwyer, S. W. Kembel, and T. J. Sharpton, Proceedings of the National Academy of Sciences **112**, 8356 (2015).

[249] C. R. Altaba, PlOS One **4**, e4611 (2009).

[250] S. Nee, A. O. Mooers, and P. H. Harvey, Proceedings of the National Academy of Sciences **89**, 8322 (1992).

[251] D. L. Rabosky and I. J. Lovette, Evolution **62**, 1866 (2008).

[252] H. Morlon, Ecology Letters **17**, 508 (2014).

[253] A. O. Mooers and S. B. Heard, Quarterly Review of Biology **72**, 31 (1997).

[254] D. J. Aldous, Statistical Science **16**, 23 (2001).

[255] M. Blum and O. François, Systematic Biology **55**, 685 (2006).

[256] G. U. Yule, Philosophical Transactions of the Royal Society of London. Series B **213**, 21 (1924).

[257] D. G. Kendall, The Annals of Mathematical Statistics **19**, 1 (1948).

[258] E. Harding, Advances in Applied Probability **3**, 44 (1971).

[259] L. L. Cavalli-Sforza and A. W. Edwards, Evolution **21**, 550 (1967).

[260] D. E. Rosen, Systematic Zoology **27**, 159 (1978).

[261] J. S. Rogers, Evolution **48**, 2026 (1994).

[262] D. Aldous, in *Random Discrete Structures* (Springer, 1996) pp. 1–18.

[263] M. Steel and A. McKenzie, Mathematical Biosciences **170**, 91 (2001).

[264] I. Pinelis, Proceedings of the Royal Society of London B: Biological Sciences **270**, 1425 (2003).

[265] M. Stich and S. Manrubia, The European Physical Journal B **70**, 583 (2009).

[266] S. Keller-Schmidt, M. Tuğrul, V. M. Eguíluz, E. Hernández-García, and K. Klemm, Physical Review E **91**, 022803 (2015).

[267] S. Keller-Schmidt and K. Klemm, Advances in Complex Systems **15**, 1250043 (2012).

[268] N. D. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, 1992).