The Pennsylvania State University
The Graduate School

SUBSTITUTION MARKOV CHAINS WITH APPLICATIONS TO MOLECULAR EVOLUTION

A Dissertation in
Mathematics
by
David Koslicki

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2012

The dissertation of David Koslicki was reviewed and approved* by the following:

Manfred Denker
Professor of Mathematics
Dissertation Advisor, Chair of Committee

Omri Sarig
Associate Professor of Mathematics
Theodore R. and Edlyn Racoosin Chair

Yakov Pesin
Distinguished Professor of Mathematics

Kateryna Makova
Associate Professor of Biology

John Roe
Professor of Mathematics
Head of the Department of Mathematics

*Signatures are on file in the Graduate School.

Abstract

This dissertation defines, develops, and applies the theory of substitution Markov chains, thereby making rigorous the intuitive concept of a "random substitution" (repeatedly replacing subletters/subwords with different, randomly chosen letters/words). A substitution Markov chain (abbreviated SMC) is special class of countable-state Markov chain. Several contributions are contained in this dissertation: first, an SMC is defined and then frequencies of subwords are shown to converge almost surely. I then investigate the boundary theory for this class of Markov chains (which is fundamentally different than previously considered classes of countable state Markov chains) and calculate two examples of the Martin boundary. I then present a technique for modifying an SMC that transforms it into a more classical, irreducible Markov chain, and verify that frequencies of subwords still converge in this case.

I then demonstrate how a particular class of SMC's provide a natural framework to accurately model molecular evolution. Not only can a diverse set of mutational events be modeled with this class of substitution Markov chain, but the theorems regarding convergence of subword frequencies can be utilized to develop an alignment-free method of parameter estimation for this model. It is difficult to overstate the advantage of an alignment-free parameter estimation technique.

The last two chapters of this dissertation represent the initiation of a novel line research concerned with the adaptation of tools from symbolic dynamics and thermodynamic formalism to the study of DNA sequences. Having given a firm foundation to the non-trivial connections between molecular evolution, countable state Markov chains, and symbolic dynamics, I take the perspective of treating a DNA sequence as a concatenation of dynamical systems and then apply the analytic techniques from symbolic dynamics to classify the corresponding biological phenomena. Considered here are topological entropy and pressure. This perspective allows for contributions to be made in diverse genomic analysis problems such as intron/exon classification, a-priori coding sequence density estimation, quantification of inter- and intra-species synonymous codon bias, and unsupervised classification of short reads of DNA as coding or non-coding.

Table of Contents

List of Figures

List of Tables

List of Symbols

$A(w)$	The Ancestors of w, p. 41, Definition 3.1.5
$(\mathcal{A}^*, P)$	Substitution Markov chain (SMC) with state space $\mathcal{A}^*$ and transition operator P, p. 12, Definition 2.3.1
$CDS(i, n, x)$	The proportion of coding sequences on human chromosome i, in a window (contiguous subsequence) of length $4^n + n - 1$, starting at nucleotide x, p. 126, Definition 7.3.2
D	A (usually countably infinite) matrix, p. 77, Definition 4.3.7
$D(w)$	The descendants of w, p. 41, Definition 3.1.4
$D_{\downarrow}((r_n)\|[a, b])$	Downward crossings, p. 51, Definition 3.2.1
Σ_n	The n-th coordinate process associated to a substitution Markov chain, p. 13, Definition 2.3.2
Σ	Short-hand for an SMC, p. 13
$\Sigma_n^{(m)}$	The m-th induced SMC, p. 17, Definition 2.3.12
$F(\cdot)$	The column vector of subword occurrences of w, p. 18, Definition 2.3.15
$G(\cdot, \cdot)$	The Green's function (or operator), p. 42, Definition 3.1.7
$H^{(m)}$	A (usually countably infinite) matrix, p. 76, Definition 4.3.4
$H_{top}(w)$	The topological entropy of the finite (or infinite) word w, p. 102, Definition 6.1.5
$I^{(m)}$	A (usually countably infinite) matrix, p. 76, Definition 4.3.5

$K(\cdot,\cdot)$ The Martin kernel, p. 47, Definition 3.1.20

$L(P)$ The set of all subwords that can appear in a SMC: the "Language", p. 16, Definition 2.3.10

$\mathcal{M}(P) = \mathcal{M}$ The Martin boundary, p. 49, Definition 3.1.24

$(\mathcal{M}, \nu^1)$ The Poisson boundary, p. 57, Definition 3.3.1

$M_{\Sigma_n}^{(m)}$ The m-th substitution matrix, p. 18, Definition 2.3.14

${}_R M_{\Sigma}^{(m)}$ The substitution matrix associated to an SMC(R), p. 76, Definition 4.3.2

μ_ψ The equilibrium measure associated to the potential ψ, p. 142, Definition 7.5.2

ψ A potential, a kind of real valued function on $\mathcal{A}^*$, p. 122, Definition 7.2.2

φ_{hs} The median, or canonical, potential for coding sequence density estimation in the human genome, p. 132

$p_u(n)$ The complexity function, p. 34, Definition 2.5.2

$P(w, \psi)$ The topological pressure of w associated to the potential ψ, p. 123, Definition 7.2.3

$P^{\text{hs}}(i, n, x, \varphi)$ The topological pressure of a window of length $4^n + n - 1$, starting at nucleotide x, on human chromosome i, associated to the potential $\log(\varphi)$ p. 127

$\pi(x, y)$ The geodesic connecting x and y, p. 67, Definition 3.4.11

SMC(R) A reversible substitution Markov chain, p. 75, Definition 4.2.2

SMC(I) A substitution Markov chain with insertions, p. 87, Definition 5.3.1

SMC(R|I|D) A reversible substitution Markov chain with insertions and deletions, p. 88, Definition 5.3.3

$\hat{X}(P)$ The Martin compactification of the Markov chain (X, P), p. 48, Definition 3.1.19

$\ll_n$	An order on $\mathcal{A}^*$ induced by P^n, p. 41, Definition 3.1.3
$x \wedge y$	The confluent, or most recent common ancestor, of x and y, p. 67, Definition 3.4.13

Acknowledgments

This work was partially supported by the National Science Foundation grants DMS-1008538 and DMS-1120622.

I would also like to acknowledge the variety of support and resources provided by:

- The Pennsylvania State University
- The University of California at Los Angeles, Institute for Pure and Applied Mathematics (UCLA IPAM)
- Georg-August-Universität Göttingen
- Drexel University

Dedication

This dissertation is dedicated:

- To my parents Mark and Diane Koslicki for instilling me with a sense of wonder about God's creation, encouraging me to always ask questions, and equipping me with the tools to answer them.
- To Manfred Denker for his keen mathematical insights, his abundant provision of opportunity, and his valued friendship. One could not ask for a better advisor.
- To Naomi Sakuma for her competitiveness in helping initiate this thesis.
- To Henna Pajulammi for her friendship, empathetic encouragement, and keeping me sane during the final chapters of this dissertation.
- To Kurt and Wendy Vinhage, for their singular, unwavering help in time of need.
- To Greta Kocol who inspired my initial interest in mathematics and demonstrated how to effectively and entertainingly teach it.
- To the many excellent mathematicians and academics I have had the opportunity to learn from. An incomplete and unordered list includes Omri Sarig, Kateryna Makova, John Roe, James Sellers, Sergei Tabachnikov, George Andrews, Kris Wysocki, and Francesca Chiaromonte.

Prologue

Biology and mathematics have historically maintained a one-way relationship, albeit a productive and fruitful one-way relationship. Biology takes from mathematics ideas and concepts and utilizes them for data analysis. Mathematics gives to biology applications of its otherwise theoretical results. This one-way relationship is undergoing a radical shift: biological problems, insight, and data are giving rise to genuinely new mathematics.

Such a paradigm shift has happened before in the physical sciences. For example, long was it viewed that mathematics was an "analysis service provider" for physics. More recently though, physics has made novel contributions to mathematics. Take, for example, the field of Gibbsian statistical physics. Physicists working in this area created a class of equations (called "canonical ensembles") that mathematicians (in particular, Ya. Sinai and D. Ruelle) were able to modify and utilize in developing principles for choosing invariant measures for a wide class of dynamical systems. String theory, and its contributions to the study of Calabi-Yau manifolds, is another example of physics generating new mathematics.

I believe that such a reciprocal relationship is beginning to play out between mathematics and biology. Consider, for example, the advent of next-generation sequencing technology, whose rise in throughput (and fall in cost) dwarfs that of Moore's law. Such massive amounts of data (with the type of data being novel as well) has left many biologist stymied: the mathematical frameworks and structures in which to properly analyze and understand such data are nearly non-existent. It is the mathematician's opportunity to provide such theoretical structures and make fundamental contributions to both fields by taking advantage of this interplay between biology and mathematics.

Such a contribution is made in this dissertation via the introduction of the theoretical structures referred to as substitution Markov chains. This dissertation studies these structures from both mathematical and biological perspectives and gives an example of the new synergistic relationship between these two fields of science.

Chapter 1

Introduction and main results

This dissertation bridges the fields of mathematics and biology through defining, developing, and applying the theory of substitution Markov chains. A substitution Markov chain (abbreviated SMC) is special class of countable-state Markov chain which naturally arises when investigating randomizations of deterministic substitutions. In this introductory chapter, I will review the main results of each chapter, as well as attempt to summarize the biological and mathematical significance of these results.

1.1 Chapter 2

In chapter 2, I set up the basic definitions involved in studying random substitutions. A special class countable state Markov chain (henceforth called a *substitution Markov chain* or SMC) is defined.

Mathematical significance The main result of this chapter is the utilization of induced substitution matrices (definition 2.3.14) to prove the almost-sure convergence of subword frequencies associated to a primitive random substitution (theorem 2.4.10) regardless of the initial distribution. I also provide an algorithm with which the subword frequencies can be explicitly calculated (section 2.4.15.1).

The setup is similar to [79], but I am able to show almost sure convergence of sub*word* frequencies (as opposed to just sub*letter* frequencies in [79]) and the proof technique circumvents the complicated second moment estimations contained in [79].

The chapter concludes with demonstrating that the complexity function (number of n-length subwords) converges in expectation on an SMC.

Biological significance Chapter 2 lays the necessary mathematical groundwork that will be further developed in chapter 4, and then finally results in the general model of molecular evolution presented in chapter 5.

1.2 Chapter 3

Mathematical significance In this chapter, I investigate the Martin boundary associated to a substitution Markov chain. Typically, one considers irreducible Markov chains ([23], [20], [113], [21], [22], [114]) or else *standard* initial distributions (see [28] and Chapter 3). However, the substitution Markov chains under consideration are not irreducible (nor is any non-empty subset of the state-space), and the initial distributions under consideration (point-masses) are not *standard* as in the definition of [28]. I thus devote the first half of chapter 3 to prove analogs to the classical convergence theorems (convergence to the boundary in theorem 3.2.6 and the integral representation theorem in theorem 3.2.9) and compare these to the classical theorems (when the Markov chain can be divided into irreducible classes or when standard measures are being considered).

I then explicitly calculate the Martin boundary for two classes of examples. In the first, the Martin boundary is shown to be homeomorphic to the unit interval, and in the second class of examples, the Martin boundary is shown to be homeomorphic to a Cantor-like Stone space.

Biological significance The biological applications and significance of the Martin boundary of a substitution Markov chain is still under investigation. Since the Martin boundary aids in the classification of SMC's and a certain

class of SMC's model molecular evolution (see chapter 5), it is expected that the Martin boundary will aid in species classification (or other phylogenetically motivated problems).

1.3 Chapter 4

In this chapter I associate to an SMC a reversible, countable-state, irreducible Markov chain and investigate its properties (including subword frequency convergence). I refer to this reversible Markov chain as an SMC(R).

Mathematical significance Defining the reversible Markov chain SMC(R) allows us to circumvent the problem of reducibility and non-standard initial distributions. This frees us from needing to consider the root of the substitution Markov chain (or the initial distribution). Furthermore, by relating the substitution matrices of the irreducible and the completely reducible Markov chains, I am able to show in theorem 4.3.12 that subword frequencies converge almost surely in the reversible case as well. The Martin boundaries of the two examples Σ_{eg1} and Σ_{eg2} remain unchanged when these SMC's are made reversible.

Biological significance This chapter lays more of the mathematical groundwork necessary for the model of molecular evolution presented in chapter 5. Creating a reversible Markov chain associated to an SMC has the effect of allowing transitions between sequences by not only *inserting* subwords, but also by *deleting* subwords. When used to model the evolution of a DNA sequence, this allows for the incorporation of mutational events such as insertions and deletions, as well as single nucleotide substitutions.

1.4 Chapter 5

In this chapter, I present a comprehensive new framework (based on substitution Markov chains) for handling biologically accurate models of molecular evolution.

Mathematical significance From a purely mathematical perspective, this chapter is an application of substitution Markov chains. The Markov chains under consideration are a particular subclass of SMC(R) that allow for insertions of subwords as well as deletions of subwords. I denote such a chain as an SMC(R|I|D).

Biological significance This model of molecular evolution provides a systematic framework for studying models that implement heterogeneous rates, neighboring dependencies, conservation of reading frame, differing rates of insertion and deletion, as well as customizable parametrization of the probabilities and types of substitutions, insertions, and deletions. Since this model is based on a subclass of SMC(R) that allows for insertions as well as deletions, I call this model an SMC(R|I|D).

I utilize the convergence of subword frequencies (theorems 2.4.10 and 4.3.12) from chapters 2 and 4 to develop an alignment-free parameter estimation technique. This alignment-free technique circumvents many of the nuanced issues related to alignment-dependent parameter estimation.

While substitution models of molecular evolution have been well studied and developed, the inclusion of insertions and deletions (indels) into biologically accurate models has enjoyed less success. As remarked in [10], a robust and comprehensive understanding of probabilistic indel analysis (and its relationship to sequence alignment) is still lacking. A number of models that include substitutions, insertions and deletions have been proposed ([11], [71], [72], [102], [101]), but there has yet to be developed a comprehensive mathematical structure in which biologically accurate models can be developed. In fact, it is this lack of a well-studied mathematical structure that leads to the analytic intractability of some proposed indel models (as mentioned in [71]). This lack of a unifying structure not only gives rise to a variety of non-biologically motivated constructs (such as "fragments" [102], [101] and the embedding of a given sequence into an infinite sequence [71]), but also leads to difficulties in comparing models, their assumptions, and their applicability.

The model is discrete time (it can be viewed as a generalization of a stochastic grammar) and the biological assumptions are clearly stated: First, I assume

that in one time unit, besides substitutions, only one mutation event (deletion, insertion, inversion, etc.) is allowed. Secondly, I assume reversibility, though this assumption can easily be relaxed. These are the only inherent assumptions in this model. Further assumptions can be made if, for example, one insists on using the HKY ([46]) model as the underlying substitution model.

1.5 Chapter 6

Having shown in chapter 2 that the complexity function converges in expectation on an SMC and due to the usefulness of SMC's in modeling molecular evolution, it is natural to expect topological entropy to be useful in genomic analysis. This line of reasoning is pursued in this chapter by investigating a definition of topological entropy for finite strings that is adapted for genomic analysis (definition 6.1.5).

Mathematical significance Topological entropy is a useful and well-known tool in symbolic dynamics. Surprisingly, there seem to be no previous attempts to modify the definition for use with finite sequences (i.e. a finitely observed symbolic dynamical system). I provide such a definition in this chapter. I then show that all the salient properties of topological entropy are retained in this definition, that it is a statistically consistent estimator, and that it converges in expectation on an SMC. Finally, I calculate the expected value of topological entropy of finite sequences produced by an arbitrary Bernoulli scheme.

Biological significance It is universally recognized that the most difficult issue in implementing entropy techniques is the convergence problem due to finite sample effects ([105], [57]). A few different approaches to circumvent these problems with topological entropy and adapt it to *finite* length sequences have been attempted before. For example, in [104], linguistic complexity (the fraction of total subwords to total possible subwords) is utilized to circumvent finite sample problems. This though leads to the observation that the complexity/randomness of intron regions is *lower* than the complexity/randomness

of exon regions. However, in [16] it is found that the complexity of randomly produced sequences is *higher* than that of DNA sequences, a result one would expect given the commonly held notion that intron regions of DNA are free from selective pressure and so evolve more randomly than do exon regions.

Our definition of topological entropy for finite strings allows for the comparison of entropies of sequences of differing lengths, a property no other implementation of topological entropy has been able to incorporate. I also calculate the expected value of the topological entropy in proposition 6.1.9 to precisely draw out the connections between topological entropy and information content. I then apply this definition to the human genome in section 6.3 to observe that the entropy of intron regions is in fact lower than that of exon regions in the human genome as one would expect. I then provide evidence in section 6.3.3 indicating that this definition of topological entropy can be used to detect sequences that are under selective pressure.

1.6 Chapter 7

Due to successfully defining, analyzing, and applying the notion of topological entropy for finite strings in chapter 6, in this chapter we aim to similarly develop a notion of topological *pressure* for finite strings. This chapter is joint work with Daniel J. Thompson.

Mathematical significance This and the previous chapter, as far as the authors are aware, represent the first attempt to define finitary approximations to analytic techniques from symbolic dynamics and thermodynamics. Our definition of topological pressure for finite sequences is a statistically consistent estimator of topological pressure as it is classically defined. We also show that analogues of classical theorems (variational principle, Gibbs property, etc.) hold for our definition and that it converges in expectation under an SMC. Finally, we demonstrate that for a given non-negative potential, we can explicitly calculate the associated equilibrium measure.

Biological significance In this chapter, we demonstrate how topological pressure can be utilized to recover the distribution of coding sequences across the human and mouse genomes. This technique can shed light on the issue of mammalian codon bias, where it is recognized that a complete understanding has not yet been achieved [14, 47, 83]. We conclude the chapter by developing a measure of coding potential based on primary sequence composition.

The advantage of utilizing topological pressure is its relative simplicity. The definition is entirely combinatorial and implicitly takes account of important considerations such as neighboring dependencies, different choices of reading frame, autocorrelation, background codon frequencies, and GC content.

Since topological pressure can be interpreted as a weighted measure of complexity, in section 7.3.11 we give a detailed analysis of the pattern of codon weights that predict the distribution of coding sequences across the human genome. We also compare synonymous codon usage between species and its relationship with CDS density estimation in section 7.4. This is a first step in the investigation of interspecies codon usage via topological pressure.

In section 7.5 we demonstrate how topological pressure naturally gives rise to a measure (called an equilibrium measure). This measure can be used to analyze sequences that are orders of magnitude shorter than those on which pressure is utilized. This represents a strategy in which large scale information (pressure) can be utilized to extract information at a much smaller scale (measure of a sequence).

The development of robust techniques that detect the coding potential of short sequences is an important area of research [18, 34, 40, 44, 62, 63, 90, 109] with applications to sequence annotation as well as gene prediction. It has been recognized that measures of coding potential based on single sequence nucleotide composition [63, p.i281] are an important part of the problem of differentiating between short reads of coding and non-coding sequences and are complementary to the very effective comparative/database techniques developed in, for example, [109]. We contribute to this area of research by showing that a certain equilibrium measure can be used to distinguish between randomly selected introns and exons in the human genome.

1.7 Symbolic dynamics, ergodic theory, and molecular evolution

The last two chapters represent the initiation of a novel line research concerned with the adaptation of tools from symbolic dynamics, ergodic theory, and thermodynamics to the study of genomic analysis. Chapters 1 through 5 place on a firm foundation the non-trivial connections between molecular evolution, countable state Markov chains, and symbolic dynamics. Chapters 6 and 7 take the perspective of treating a DNA sequence as a concatenation of symbolic dynamical systems and then applies techniques from ergodic theory and thermodynamics to classify the corresponding biological phenomena. The techniques under consideration (entropy and pressure) are shown to converge under the previously developed model of molecular evolution (an SMC(R|I|D)), so confidence can be maintained that the given techniques remain robust under mutational influences. This is strong evidence for the validity and utility of employing these finite approximations of tools from symbolic dynamics to the study of DNA sequences. I expect it will be fruitful to adapt a more diverse set of analytic techniques from symbolic dynamics to the study of DNA sequences. Furthermore, after fixing parameter values of an SMC(R|I|D), I expect relevant biological results to be produced from investigating the ergodic properties of the resulting countable state Markov chain.

Chapter 2

Substitution Markov chains

In this chapter we set up the basic definitions involved in studying random substitutions. The main result of this chapter is the utilization of induced substitution matrices (definition 2.3.14) to prove the almost-sure convergence of subword frequencies associated to a random substitution (theorem 2.4.10). We also provide an algorithm with which the subword frequencies can be explicitly calculated (section 2.4.15.1).

To mathematically motivate the problem, we first review deterministic substitutions and the previous randomization attempts.

2.1 Deterministic substitutions

Deterministic substitutions are used to construct (symbolic) dynamical systems by means of iterating a substitution rule on the letters of a finite alphabet. A classic example is the Thue-Morse sequence [84]

$$0110100110010110\cdots$$

which can be obtained by iterating the substitution $0 \to 01$, $1 \to 10$ repeatedly on the letter 0. In general, a substitution is defined as follows: We let $\mathcal{A}$ be a finite length alphabet and let $\mathcal{A}^*$ denote the set of all finite length words formed via concatenation from $\mathcal{A}$. Then a *substitution* σ is an application from $\mathcal{A}$ into the set $\mathcal{A}^*$ whose domain is then extended to $\mathcal{A}^*$ by concatenation:

$\sigma(WV) = \sigma(W)\sigma(V)$. The symbolic dynamical system that arises from this substitution is then the orbit closure (with the metric $d(u,v) = 2^{-\min\{i:u_i \neq v_i\}}$) under the shift map on a word obtained by the above algorithm.

The first comprehensive approach to substitution dynamical systems was Queffélec [84] in which the author considered substitutions of constant length

$$|\sigma(a_i)| = m \forall\ a_i \in \mathcal{A}$$

and proved that under mild conditions that the entropy of the associated dynamical system is always zero. She then utilized the spectrum of the dynamical system to classify such systems. She also showed the the topological entropy of such substitution sequences is always zero. The method of characterization was via the eigenvalues of the *substitution (or incidence) matrix* M_σ where $(M_\sigma)_{i,j} = |\sigma(a_j)|_{a_i}$ where $a_i, a_j \in \mathcal{A}$ and $|u|_v$ denotes the number of occurrences of the subword v in u (allowing overlaps). The random analog of such matrices will play an important role in this thesis.

While deterministic substitutions are a well-studied concept, very little is known regarding randomizations of deterministic substitutions.

2.2 Previous attempts to randomize deterministic substitutions

There are two randomizations of a deterministic substitutions in the literature: *S-Adic substitutions*, and *Substitutions Aléatoires Itérées* (or simply *random substitutions*).

2.2.1 S-Adic substitutions

As originally defined (see [33] for further detail), S-Adic substitutions are infinite, random compositions of deterministic substitutions
"$\lim_{n\to\infty}$"$\sigma_1\sigma_2\cdots\sigma_n(a)$, the concept of which being introduced to characterize minimal symbolic dynamical systems with linear subword complexity. S-Adic substitutions are almost a random dynamical system as can be seen by

rephrasing the definition implicit in proposition 5 of [33]:

Definition 2.2.2 (S-Adic Substitution)**.** *Given a probability preserving transformation (X, B, μ, θ), a finite family $F = \{\sigma_i\}_{i\in I}$ of substitutions on some alphabet A with a function $f : X \to I$, an S-Adic substitution is then given by the transformation $\theta_f^n : X \times X \to X \times A^*$ defined by*

$$\theta_f^n(x, y) = (\theta^n(x), \sigma_{f(\theta^n x)} \circ \cdots \circ \sigma_{f(\theta x)} \circ \sigma_{f(x)}(y))$$

After the concept of S-Adic substitutions was introduced by Ferenczi in [33], the authors of [26], and [108] investigated S-Adic substitutions and their relation to systems of linear complexity, linear recurrence, and their relationship to Bratteli diagrams respectively. We note that [108] uses S-Adic systems to construct a dynamical system that is minimal but not uniquely ergodic.

2.2.3 Random substitutions

The second randomization of deterministic substitutions in the literature is that of Peyriére ([81], [79], [80], [82]) who wrote a series of overlapping articles on what he called random substitutions. Peyriére took as his motivation a construction of Mandelbrot (referred to by Peyriére as an M-System, or a Random Beadset, or a Random Substitution). These articles give a definition of a random substitution and investigate the convergence of the frequencies of subwords under various conditions (they also investigate an associated kind of graph). The main approach used in these papers is to use Doob's martingale convergence theorem or an estimation of second moments to obtain the desired results.

After Peyriére's papers, the concept of random substitutions lay dormant. The only other paper to date that cites Peyriére's work was [1], a short paper published in a physics journal in which the authors run some numerical estimates and obtain a few results regarding the Fourier spectra of the shift map on a sequence generated by a specific kind of random substitution on a two letter alphabet.

We take as our starting point a modified version of Peyriére's definition of a random substitution As will be seen, the definition of a random substitu-

tion as contained in this dissertation will generalize and incorporate not only Peyriére's definition, but also that of an S-Adic substitution.

2.3 Definitions

To emphasize our perspective that a random substitution is simply a certain class of countable-state Markov chain, we will henceforth utilize the term *substitution Markov chain* or *SMC*. As mentioned, we define substitution Markov chains similarly to [79]. For the following, we need:

1. $\mathcal{A} = \{a_1, \ldots, a_t\}$ a finite set of ordered symbols referred to as an *alphabet*
2. For each letter $a \in \mathcal{A}$, (Ω_a, P_a) a finite (non-empty) probability space
3. For each letter $a \in \mathcal{A}$, a function $g_a : \Omega_a \to \mathcal{A}^*$

Let $\mathcal{A}^n$ denote all words of length n formed from letters of $\mathcal{A}$. Let $\mathcal{A}^* = \cup_{n\geq 1}\mathcal{A}^n$ be the set of finite length words formed from $\mathcal{A}$.

We now define the Markov chain representing an SMC.

Definition 2.3.1 (Substitution Markov Chain (SMC))**.** *A substitution Markov chain (with fixed $\mathcal{A}$, $\{(\Omega_a, P_a)\}_{a\in\mathcal{A}}$, and $\{g_a\}_{a\in\mathcal{A}}$) is a Markov chain $(\mathcal{A}^*, P)$ with state space $\mathcal{A}^*$ and transition operator P defined in the following way. For $u = b_1 \ldots b_n \in \mathcal{A}^*$ a word, we let $\Omega_u = \Omega_{b_1} \times \cdots \times \Omega_{b_n}$ and $P_u = P_{b_1} \times \cdots \times P_{b_n}$. We define $g_u : \Omega_u \to \mathcal{A}^*$ via concatenation of words: for $\omega = (\omega_1, \ldots, \omega_n) \in \Omega_u$, $g_u(\omega) = g_{b_1}(\omega_1) \ldots g_{b_n}(\omega_n)$. Now define P by*

$$P(u, v) = \sum_{\omega \in g_u^{-1}(v)} P_u(\omega)$$

Note that in definition 2.3.1 the sum $\sum_{\omega\in g_u^{-1}(v)} P_u(\omega)$ is taken over *all* pre-images $g_u^{-1}(v)$ whereas in [82],[80], [81], [79], the summation is over only a single element. While this difference does not effect the results obtained by Peyriére, if the summation is not taken over all pre-images, one is implicitly assuming that the underlying Markov chain has the structure of a tree which (as the following example shows) is not always the case.

We aim to utilize a number of probabilistic techniques in analyzing substitution Markov chains and so recast the definition in more probabilistic terms (and introduce new notation) now.

Definition 2.3.2 (SMC as a Random Variable)**.** *For $v \in \mathcal{A}$, by $\Sigma_n(v)$ we denote the n-th coordinate random variable associated to the Markov chain $(\mathcal{A}^*, P)$ with initial distribution unit mass on v.*

We choose such notation $\Sigma_n(v)$ to reflect the similarity in spirit to deterministic substitutions $\sigma^n(v)$ (even though the first is a random variable and the second a morphism).

Note that for Ω^∞ the trajectory space of $(\mathcal{A}^*, P)$, $\Sigma_n(v) : \Omega^\infty \to \mathcal{A}^*$. Thus for a particular trajectory $\omega \in \Omega^\infty$, step n, and initial word v, $\Sigma_n(v)(\omega) \in \mathcal{A}^*$. We will usually suppress the dependence of Σ_n on both $v \in \mathcal{A}^*$ and especially $\omega \in \Omega^\infty$.

We now introduce notation to concisely represent a substitution Markov chain and the associated spaces and probabilities.

Notation 2.3.1. Given $\mathcal{A} = \{a_i\}_{i\in I}$, $\{(\Omega_{a_i}, P_{a_i})\}_{a_i\in\mathcal{A}}$, and $\{g_{a_i}\}_{a_i\in\mathcal{A}}$, let $\{w_j^{a_i}\} = g_{a_i}(\Omega_{a_i})$ and $p_{a_i,w_j^{a_i}} = P_{a_i}^{-1}(w_j^{a_i})$. Then we introduce the notation

$$\Sigma : \begin{cases} a_1 \to \begin{cases} w_1^{a_1} \text{ with prob. } p_{a_1,w_1^{a_1}} \\ w_2^{a_1} \text{ with prob. } p_{a_1,w_2^{a_1}} \\ \vdots \end{cases} \\ a_2 \to \begin{cases} w_1^{a_2} \text{ with prob. } p_{a_2,w_1^{a_2}} \\ w_2^{a_2} \text{ with prob. } p_{a_2,w_2^{a_2}} \\ \vdots \end{cases} \\ \vdots \end{cases}$$

to represent the Markov chain $(\mathcal{A}^*, P)$ defined in 2.3.1 and to make clear the rule(s) used to develop the transition probabilities.

If no ambiguity arises, we will call any of $(\mathcal{A}^*, P)$, P (with the state space suppressed), Σ_n, or Σ (with dependence on n suppressed) a substitution

Markov chain and assume that $\mathcal{A} = \{a_i\}_{i\in I}$, $\{(\Omega_{a_i}, P_{a_i})\}_{a_i\in\mathcal{A}}$, and $\{g_{a_i}\}_{a_i\in\mathcal{A}}$ are all clearly defined from the context.

2.3.3 An example

We gather a few examples to elucidate the above definitions. First consider the substitution Markov chain

$$\Sigma : \begin{cases} a \to \begin{cases} ab \text{ with prob. } 1/2 \\ ba \text{ with prob. } 1/2 \end{cases} \\ \\ b \to b \end{cases}$$

For this SMC, our alphabet is on two letters, $\mathcal{A} = \{a, b\}$ and our probability spaces are $\Omega_a = \{1, 2\}$, $P_a(1) = 1/2, P_a(2) = 1/2$ and $\Omega_b = \{1\}$, $P_b(1) = 1$. Also, our functions g_a, g_b are defined as $g_a(1) = ab, g_a(2) = ba$ and $g_b(1) = b$. The random variable $\Sigma_n(a)$ develops (in n, for a particular trajectory) as follows: For $\Sigma_1(a)$, flip a fair coin to decide whether to replace a with ab or ba. Then for $\Sigma_2(a)$ take the resulting word (say ab) and then from left to right look at each letter individually and **independently**, flip a coin whenever you see an a to determine the appropriate replacement and always replace b with b, etc.

A portion of the Markov chain associated to this Σ_n would look like

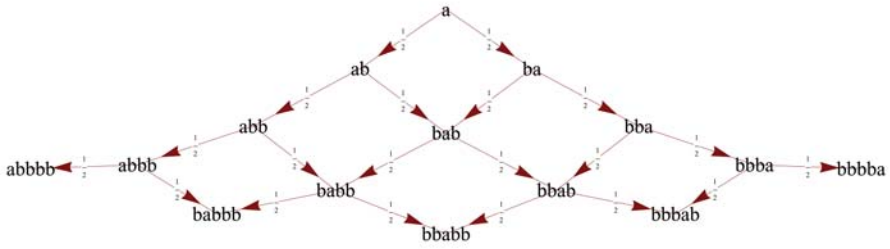


Figure 2.1. A Portion of an Example Substitution Markov Chain

Thus we have a sequence of $\mathcal{A}^*$ valued random variables Σ_n, $n \geq 0$ which represent the random position in $\mathcal{A}^*$ at time n. Equipping the trajectory

space $\Omega^\infty = (\mathcal{A}^*)$ with the probability measure defined by the Kolmogorov extension theorem

$$ {}_x(v_0 \dots v_n) := {}_x[\Sigma_0 = v_0, \Sigma_1 = v_1, \dots, \Sigma_n = v_n] \tag{2.1} $$

$$ = \delta_x(v_0) P(v_0, v_1) \dots P(v_{n-1}, v_n) \tag{2.2} $$

We also denote the associated expectation by ${}_x$. When taking the expectation ${}_x$, the distribution is then given by δ_x, that is unit mass on $x \in X$. As stated in definition 2.3.2, by $\Sigma_n(x)$ we assert that the Markov chain is given with initial distribution unit mass on x. Thus for any for any function $f : \mathcal{A}^* \to$, ${}_x f(\Sigma_n) = f(\Sigma_n(x))$ where the last expectation is taken over all Ω^∞. The n-step transition probabilities are

$$ P^{(n)}(w, v) := {}_w[\Sigma_n = v] $$

So for the infinite dimensional matrix P, $P^{(n)}(w, v)$ is the $(w, v)^{\text{th}}$ entry of the matrix power P^n.

There are a few nice, well known properties of the transition matrix P for a time-homogeneous Markov chain and we mention a few here.

Lemma 2.3.4. *The number $P^{(n)}(w, v)$ is the element at position (w, v) in the stochastic matrix power P^n. Also*

$$ P^{(m+n)}(w, v) = \sum_{x \in X} P^{(m)}(w, x) P^{(n)}(x, v) $$

It is important to note that the Markov chain associated to a substitution Markov chain is rarely irreducible. Since we will have further need of this fact, we prove now that the substitution Markov chains under consideration do not have irreducible Markov chains.

Definition 2.3.5. *A Markov chain (X, P) is said to be irreducible if for all $w, v \in X$ there is some integer n such that $P^n(w, v) > 0$.*

Lemma 2.3.6. *For any substitution Markov chain $(\mathcal{A}^*, P)$, if for $w, v \in \mathcal{A}^*$ such that $|w| > |v|$, then $P(w, v) = 0$.*

Proof. Let $w, v \in \mathcal{A}^*$ such that $N = |w| > |v| = n$, then for any $\omega = (\omega_1, \ldots, \omega_N) \in \Omega_w$,

$$|g_w(\omega)| = |g_{w_1}(\omega_1)| + \cdots + |g_{w_N}(\omega_N)| \geq N > n$$

Hence $g_w^{-1}(v) = \emptyset$ and so $P(w, v) = \sum_{\omega \in g_w^{-1}(v)} P_w(\omega) = 0$. □

Here, and throughout the rest of the paper, we use $|w|$ to represent the length of the word $w \in \mathcal{A}^*$.

Corollary 2.3.7. *Any substitution Markov chain $(\mathcal{A}^*, P)$ such that for some $a \in \mathcal{A}$ and $\omega \in \Omega_a$ such that $|g_a(\omega)| > 1$ is not irreducible.*

Proof. For $(\mathcal{A}^*, P)$ satisfying the hypotheses, by lemma 2.3.6, for any $v \in \mathcal{A}$, since $|v| = 1$, $P(g_a(\omega), v) = 0$. Fix, say, $v = a$ and arguing by induction assume that $P^m(g_a(\omega), a) = 0$. Then by lemma 2.3.4, $P^{m+1}(g_a(\omega), a) = \sum_{x \in \mathcal{A}} P^m(g_a(\omega), x) P(x, a)$. Now if $|x| > 1$, then by lemma 2.3.6, we have $P(x, a) = 0$. Now if $|x| = 1$, we've already shown that $P(g_a(\omega), x) = 0$, hence $P^{m+1}(g_a(\omega), a) = 0$. So by induction we have for each $n > 0$, $P^n(g_a(\omega), a) = 0$ and we have shown that the substitution Markov chain is not irreducible. □

In studying substitution Markov chains, combinatorial properties of the set of words $g_a(\Omega_a)$ for $a \in \mathcal{A}$ play a vital role in the structure of the Markov chain $(\mathcal{A}^*, P)$.

Definition 2.3.8 (Constant Length SMC)**.** *For the substitution Markov chain $(\mathcal{A}^*, P)$ as defined in 2.3.1, if for each $a \in \mathcal{A}$, $|g_a(\Omega_a)| = k$, we say that the $(\mathcal{A}^*, P)$ is a substitution Markov chain of constant length or a substitution Markov chain of length k.*

Definition 2.3.9 (Subwords)**.** *For $w, v \in \mathcal{A}^*$ by $|v|_w$ we denote the number of appearances of w as a subword of v with overlaps. More rigorously, for $w = w_0 \ldots w_n$ and $v = v_0 \ldots w_m$, $|v|_w = |\{j : v_j v_{j+1} \ldots v_{j+n-1} = w\}|$. Also, by $|v|$ we denote the length of the word v, so $|v| = m$.*

Definition 2.3.10 (Language)**.** *For a substitution Markov chain $(\mathcal{A}^*, P)$, let the language be defined by $L(P) = \{w : \exists a, v \in \mathcal{A}^*, n \in \quad \text{s.t.} |v|_w > 0, P^{(n)}(a, v) > 0\}$. Also, let $L^{(m)}(P) = L(P) \cap \mathcal{A}^m$.*

Should no ambiguity arise, we will write L for $L(P)$.

The following definition will be used to define the important substitution matrices. Recall that for an SMC $(\mathcal{A}^*, P)$, the n-th coordinate process $\Sigma_n(v)$ is a random variable $\Sigma_n(v) : \Omega^\infty \to \mathcal{A}^*$. We will use $\Sigma_n(v)$ to define a new n-th coordinate process $\Sigma_n^{(m)}(w)$ associated to a Markov chain defined on the state space $(\mathcal{A}^m)^*$. So in the newly defined Markov chain, the state-space will consist of words whose "letters" will be m-length words formed from $\mathcal{A}$.

Lemma 2.3.11. *For a word $v = v_1 \dots v_m \in \mathcal{A}^m$, and $\omega \in \Omega$ such that,*

$$\Sigma_n(v_1 \dots v_m)(\omega) = y_1 \dots y_{l_1}\ y_{l_1+1} \dots y_{l_1+l_2}\ \dots y_{l_1+\dots+l_m}$$

then

$$\begin{aligned} l_1 =& |\Sigma_n(v_1)(\omega)| \\ l_2 =& |\Sigma_n(v_2)(\omega)| \\ &\vdots \\ l_m =& |\Sigma_n(v_m)(\omega)| \end{aligned}$$

Proof. This is a direct consequence of definition 2.3.1 since we defined $g_u : \Omega_u \to \mathcal{A}^*$ via concatenation of words: for $(\omega_1, \dots, \omega_n) \in \Omega_u$, $g_u((\omega_1, \dots, \omega_n)) = g_{b_1}(\omega_1) \dots g_{b_n}(\omega_n)$. □

We can now define the induced substitution;

Definition 2.3.12 (Induced Substitution)**.** *For a substitution Markov chain Σ, define $\Sigma_n^{(m)}$ by the following: if for $v = v_1 \dots v_m \in \mathcal{A}^m$ and $\omega \in \Omega^\infty$ (the trajectory space)*

$$\Sigma_n(v_1 \dots v_m)(\omega) = y_1 \dots y_{l_1}\ y_{l_1+1} \dots y_{l_1+l_2}\ \dots y_{l_1+\dots+l_m}$$

with $l_i = |\Sigma_n(v_i)|$ as in lemma 2.3.11, we then define

$$\Sigma_n^{(m)}(v_1 \dots v_m)(\omega) = (y_1 \dots y_m)(y_2 \dots y_{m+1}) \dots (y_{l_1} \dots y_{l_1+m-1})$$

So $\Sigma_n^{(m)}$ as defined is a random variable on the same sample space as Σ_n and the "letters" produced by $\Sigma_n^{(m)}$ are the m-length subwords produced by Σ_n. Note that the last "letter" of $\Sigma_n^{(m)}(v_1 \dots v_m)(\omega)$ is $(y_{l_1} \dots y_{l_1+m-1})$ and this has y_{l_1} the last letter produced by $\Sigma_n(v_1)(\omega)$.

Just as we introduced notation in 2.3.1 to succinctly describe the spaces and probabilities associated to an SMC, we will use $\Sigma^{(m)}$ in such a notational role as well (see the examples in section 2.4.11).

Note the following.

Lemma 2.3.13. *For $n \in$ and $v = v_1 \dots v_m \in \mathcal{A}^m$,*

$$|\Sigma_n(v_1)| = |\Sigma_n^{(m)}(v_1 \dots v_m)|$$

Proof. This is consequence of how we defined $\Sigma_n^{(m)}$ in definition 2.3.12: $|\Sigma_n(v_1)| = l_1 = |\Sigma_n^{(m)}(v_1 \dots v_m)|$. □

In particular, the above lemma implies that if Σ_n is of constant length k, so is $\Sigma_n^{(m)}$.

Definition 2.3.14 (Substitution Matrices)**.** *For a given substitution Markov chain Σ and integers m, n, define the m-th substitution matrix entry wise as follows: for $i, j \in L^{(m)}(P)$,*

$$\left(M_{\Sigma_n}^{(m)}\right)_{ij} = \quad {}_j|\Sigma_n^{(m)}|_i$$

For brevity, $M_\Sigma := M_{\Sigma_1}^{(1)}$. When subword frequencies are investigated, it will be beneficial to view $M_{\Sigma_1}^{(m)}$ column-wise as well.

Definition 2.3.15. *For $\mathcal{A}^* \cap L^{(m)}(P) = \{v_1, \dots, v_n\}$, define $F_m : \mathcal{A}^* \to \quad ^n$ as the column vector of subword occurrences $F_m(w) = (|w|_{v_1}, \dots, |w|_{v_n})^T$*

Then by definition of expectation,

$$M_{\Sigma_1}^{(m)} = \left[\int F_m(\Sigma_1(\omega)) d \quad {}_{v_1} \Big| \dots \Big| \int F_m(\Sigma_1(\omega)) d \quad {}_{v_n} \right] \tag{2.3}$$

Definition 2.3.16 (Matrix Irreducible Substitution)**.** *Call a substitution Markov chain* $(\mathcal{A}^*, P)$ ***matrix irreducible*** *if for every* $i, j \in L(P)$ *there exists an integer* $n = n(i,j) > 0$ *such that* $((M^{(1)}_{\Sigma_1})^n)_{ij} > 0$*.*

For simplicity, instead of writing the matrix power as $(M^{(1)}_{\Sigma_1})^n$, we will use the notation M^n_Σ.

Definition 2.3.17 (Primitive Substitution)**.** *Call a substitution Markov chain* $(\mathcal{A}^*, P)$ ***primitive*** *if there exists an integer* $n > 0$ *such that every entry of the matrix power* M^n_Σ *is strictly positive, that is* $M^n_\Sigma > 0$*.*

Note that the term "Primitive" is synonymous to "irreducible and aperiodic".

The name *primitive* is chosen since in non-negative matrix theory, a matrix M with the property that for some n, $M^n > 0$ is called primitive as well. The name *matrix irreducible* is also chosen since a matrix M is irreducible if for each i, j there is some integer $n = n(i,j)$ such that $M^n_{ij} > 0$. Confusion might arise in that two notions of irreducibility will be important. First, the Markov chain associated to a substitution Markov chain might be irreducible; in this case we call the substitution Markov chain *irreducible*. Second, the *substitution matrix* associated to a substitution Markov chain might be irreducible, and so in this case we call the substitution Markov chain *matrix irreducible*.

Recall the definition of the period of an irreducible matrix: $\gcd\{k \geq 1, M^k_{ii} > 0\}$ for any i and note that an irreducible matrix is primitive if and only if the period is 1. The primitivity property can be interpreted as saying that starting at any state that contains as a subletter $i \in \mathcal{A}$ on the Markov chain $(\mathcal{A}^*, P)$, then almost surely in n steps the letter j will appear as a subletter.

2.4 Properties of substitution matrices

One of the main tools used in analyzing substitution Markov chains will be the substitution matrices $M^{(m)}_\Sigma$ and so in this section a number of properties regarding these matrices will be derived. The main goal will be to show that subword frequencies exist almost surely. For the remainder of this section assume that for a substitution Markov chain $(\mathcal{A}^*, P)$ under consideration, there

exists a letter $a \in \mathcal{A}$ such that for some $\omega \in \Omega_a$, $g_a(\omega)$ begins with a. Note that this implies M_Σ has period 1 if it is irreducible. We also assume that there exists some $\alpha \in \mathcal{A}$ such that for some $\omega \in \Omega_\alpha$, $|g_\alpha(\omega)| > 1$. This ensures that the Markov chain $(\mathcal{A}^*, P)$ is not trivial (reduces to a finite chain on $\mathcal{A}$) and is also not irreducible. For simplicity, assume that any mentioned substitution Markov chain is primitive unless otherwise stated.

The exposition in this subsection closely follows that of [84] pages 87-97. The results in [84] apply only to deterministic substitutions so it is important to verify that the randomized versions of the results hold as well.

We recall the Perron-Frobenius theorem for primitive matrices (see [84] or even better [93]).

Theorem 2.4.1 (Perron-Frobenius)**.** *Let M be a primitive non-negative matrix. Then*

1. *M admits a strictly positive eigenvalue Λ such that for any other eigenvalue λ of M, $\Lambda > \lambda$.*

2. *There exists a strictly positive eigenvector corresponding to Λ.*

3. *Λ is a simple eigenvalue.*

Proof. See [93, Theorem 1.1]. □

Theorem 2.4.2. *For a primitive $n \times n$ matrix M, for e, f the positive right and left (column) eigenvectors corresponding to the dominant eigenvalue Λ normed such that $f \cdot e = 1$, then for λ_2 the second largest eigenvalue in norm with multiplicity m_2,*

1. *If $\lambda_2 \neq 0$, then as $k \to \infty$,*

$$M^k = \Lambda^k e f^t + O(k^{m_2 - 1}|\lambda_2|^k)$$

2. *If $\lambda_2 = 0$, then for $k \geq n - 1$,*

$$M^k = r^k e f^t$$

In either case, $\lim_{k\to\infty} \Lambda^{-k} M^k = ef^t$.

Proof. See [93, Theorem 1.2]. □

We now investigate the properties of the substitution matrices.

Lemma 2.4.3. *For a given substitution Markov chain* $(\mathcal{A}^*, P)$*,*

$$M_{\Sigma_n^{(1)}} = M_{\Sigma_1^{(1)}}^n$$

Proof. Recall that $M_{\Sigma_n} := M_{\Sigma_n}^{(1)}$. In the following calculation, $i, j, r_1, \ldots r_{n-1} \in \mathcal{A}$ and $w_1 \ldots w_n \in \mathcal{A}^*$. Calculating, we have

$$\begin{aligned}
(M_\Sigma^n)_{ij} &= \sum_{r_1,\ldots r_{n-1}} (M_{\Sigma_1})_{ir_1} (M_{\Sigma_1})_{r_1 r_2} \ldots (M_{\Sigma_1})_{r_{n-1} j} && (2.4)\\
&= \sum_{r_1,\ldots r_{n-1}} {}_{r_1}|\Sigma_1|_i \ {}_{r_2}|\Sigma_1|_{r_1} \cdots \ {}_{j}|\Sigma_1|_{r_{n-1}} && (2.5)\\
&= \sum_{r_1,\ldots r_{n-1}} \left(\sum_{w_1} P(r_1, w_1)|w_1|_i \right) \ldots \left(\sum_{w_n} P(j, w_n)|w_n|_{r_{n-1}} \right) && (2.6)\\
&= \sum_{\substack{r_1,\ldots r_{n-1} \\ w_1,\ldots,w_n}} P(r_1, w_1)|w_1|_i P(r_2, w_2)|w_2|_{r_1} \ldots P(j, w_n)|w_n|_{r_{n-1}} && (2.7)\\
&= \sum_{w_1,\ldots,w_n} |w_1|_i \left(\sum_{r_1} P(r_1, w_1)|w_2|_{r_1} \right) \ldots \\
&\qquad\qquad \times \left(\sum_{r_{n-1}} P(r_{n-1}, w_{n-1})|w_n|_{r_{n-1}} \right) P(j, w_n) && (2.8)\\
&= \sum_{w_1,\ldots w_n} P(j, w_n) P(w_n, w_{n-1}) \ldots P(w_2, w_1)|w_1|_i && (2.9)\\
&= \sum_{w_1} P^{(n)}(j, w_1)|w_1|_i && (2.10)\\
&= \sum_{w_1} {}_{j}[\Sigma_n = w_1]|w_1|_i && (2.11)\\
&= \int |\Sigma_n|_i d \ {}_{j} && (2.12)\\
&= \ {}_{j}|\Sigma_n|_i && (2.13)\\
&= (M_{\Sigma_n})_{ij} && (2.14)
\end{aligned}$$

□

Lemma 2.4.4. *If $M_{\Sigma_1}^{(1)}$ is primitive, then for each $n > 0$, $M_{\Sigma_1}^{(n)}$ is primitive as well.*

Proof. Let $A = A_0 \dots A_{n-1}, B \in L^{(n)}(P)$, we will show that for some integer t, $(M_{\Sigma_1}^{(n)})_{BA}^t > 0$. Since $B \in L^{(m)}(P)$, there exists $\alpha \in \mathcal{A}$ and an integer N_α such that ${}_\alpha|\Sigma_{N_\alpha}^{(1)}|_B > 0$. Now assuming that $M_{\Sigma_1}^{(1)}$ is primitive, there exists an integer N such that ${}_{A_0}|\Sigma_N^{(1)}|_\alpha > 0$. Thus, by definition of $\Sigma_1^{(n)}$ and lemmas 2.4.3 and 2.3.13, we have that

$$\begin{aligned}(M_{\Sigma_1}^{(n)})_{BA}^{N+N_\alpha} &= (M_{\Sigma_{N+N_\alpha}}^{(n)})_{BA} \\ &= {}_A|\Sigma_{N+N_\alpha}^{(n)}|_B \\ &= |\Sigma(A)_{N+N_\alpha}^{(n)}|_B \\ &\geq |\Sigma(A_0)_{N+N_\alpha}^{(1)}|_B > 0\end{aligned}$$

Thus the matrix $M_{\Sigma_1}^{(n)}$ is irreducible. Now since for some $\omega \in \Omega_a$, $g_a(\omega)$ begins with a (our main assumption of this section), then there is some $w \in L^{(n)}(P)$ and ω' such that $g_w(\omega')$ begins with w. Thus $M_{\Sigma_1}^{(n)}$ has period 1 and so is primitive. □

By theorem 2.4.1, for a primitive substitution Markov chain there exists a dominant eigenvalue henceforth denoted Λ.

Lemma 2.4.5. *For a primitive substitution Markov chain $(\mathcal{A}^*, P)$, the sequence of vectors $\frac{F_1(\Sigma_n(\alpha))}{\Lambda^n}$ converges almost surely to the strictly positive eigenvector of M_Σ corresponding to Λ with this eigenvector being independent of α.*

Proof. By 2.4.1, the matrix M_Σ can be decomposed into the sum of matrices $\Lambda P_\Lambda + N$ with P_Λ projection onto the one dimensional eigenspace corresponding to Λ an N such that $NP_\Lambda = P_\Lambda N = 0$. Thus $(M_\Sigma)^n = \Lambda^n P_\Lambda + N^n$. Recalling the definition of F_1 in 2.3.15, note that for $\alpha \in \mathcal{A}$, $M_{\Sigma_1} F_1(\alpha) = \int F_1(\Sigma_1) d\ {}_\alpha$ by 2.3, and similarly, for any $w \in \mathcal{A}^*$, $F_1(\Sigma_1(w)) = M_{\Sigma_1} F_1(w)$ where the

right hand side is a matrix/vector product. Thus for $\alpha \in \mathcal{A}$,

$$\frac{F_1(\Sigma_n(\alpha))}{\Lambda^n} = \frac{M_{\Sigma_n} F_1(\alpha)}{\Lambda^n} \tag{2.15}$$

$$= \frac{M_{\Sigma_1}^n F_1(\alpha)}{\Lambda^n} \tag{2.16}$$

$$= P_\Lambda(F_1(\alpha)) + \frac{N^n F_1(\alpha)}{\Lambda^n} \tag{2.17}$$

where line 2.16 is by lemma 2.4.3. Thus by line 2.17 we have that $\frac{F_1(\Sigma_n(\alpha))}{\Lambda^n}$ converges almost surely to the strictly positive eigenvector $v(\alpha) = P_\Lambda(F_1(\alpha))$ of M_Σ corresponding to Λ. Since the Perron-Frobenius eigenvector is unique, we have that the eigenvector is independent of α. □

In the process of proving lemma 2.4.5, the following has been shown.

Corollary 2.4.6. *$F_1(\Sigma_n(\alpha))$ is $_\alpha$ almost surely asymptotic to $P_\Lambda(F_1(\alpha))\Lambda^n + N^n F_1(\alpha)$.*

Lemma 2.4.7. *For a primitive substitution Markov chain $(\mathcal{A}^*, P)$ and any $\alpha \in \mathcal{A}$, the sequence of numbers $\frac{|\Sigma_{n+1}(\alpha)|}{|\Sigma_n(\alpha)|}$ converges to Λ.*

Proof. For $< \cdot, \cdot >$ the standard euclidean inner-product and $\quad$ the vector consisting only of ones, we have $|\Sigma_n(\alpha)| =< \quad F_1(\Sigma_n(\alpha)), \quad >$. Thus,

$$\frac{|\Sigma_{n+1}(\alpha)|}{|\Sigma_n(\alpha)|} = \frac{< \quad F_1\Sigma_{n+1}(\alpha), \quad >}{< \quad F_1\Sigma_n(\alpha), \quad >} \tag{2.18}$$

$$= \Lambda \frac{< \frac{F_1\Sigma_{n+1}(\alpha)}{\Lambda^{n+1}}, \quad >}{< \frac{F_1\Sigma_n(\alpha)}{\Lambda^n}, \quad >} \tag{2.19}$$

$$\xrightarrow{\alpha \ a.s.} \Lambda \frac{< v(\alpha), \quad >}{< v(\alpha), \quad >} \tag{2.20}$$

$$= \Lambda \tag{2.21}$$

where convergence in line 2.20 is by lemma 2.4.5 and $v(\alpha)$ is the strictly positive eigenvector as in lemma 2.4.5 as well. □

Note that in the process of proving lemma 2.4.7, we have shown that $|\Sigma_n(\alpha)|$ is $_\alpha$ almost surely asymptotic to $\Lambda^n < v(\alpha), \quad >$. Combining this

with corollary 2.4.6, we have the following corollary.

Corollary 2.4.8. *$\frac{F_1(\Sigma_n(\alpha))}{|\Sigma_n(\alpha)|}$ converges $_\alpha$ almost surely to $\frac{v(\alpha)}{<v(\alpha),>}$ the normalized strictly positive eigenvector of M_Σ corresponding to the eigenvalue Λ.*

The spectral properties of $M_\Sigma^{(m)}$ for various m are closely related. In fact, the next lemma shows the dominant eigenvalue is the same for all m. Later it will be shown that the non-zero eigenvalues of $M_\Sigma^{(2)}$ are the only non-zero eigenvalues. It is also shown later that that the eigenvectors of all $M_\Sigma^{(m)}$ corresponding to the dominant eigenvalue are linear transformations of the dominant eigenvector of $M_\Sigma^{(2)}$ as well.

Lemma 2.4.9. *For Λ the dominant eigenvalue of the primitive matrix $M_\Sigma^{(1)}$, then for any $m > 0$, Λ is the dominant eigenvalue of $M_\Sigma^{(m)}$ as well.*

Proof. By lemma 2.4.4, we know that $M_\Sigma^{(m)}$ is primitive and so has a dominant eigenvalue. Denote this eigenvalue by Λ_m. Now lemma 2.4.7, we have that for any $v = v_0 \dots v_{m-1} \in \mathcal{A}^m$

$$\lim_{n\to\infty} \frac{|\Sigma_{n+1}^{(m)}(v)|}{|\Sigma_n^{(m)}(v)|} = \Lambda_m$$

but we also have that $|\Sigma_n(v_0)| = |\Sigma_n^{(m)}(v)|$, thus for each m, $\Lambda_m = \Lambda$. □

The next proposition shows that if a substitution Markov chain is primitive, then the frequency of occurrence of a word $w \in L^{(m)}$ converges in expectation to the appropriate entry of the dominant eigenvector of $M_\Sigma^{(m)}$.

Theorem 2.4.10 (Convergence of Frequencies)**.** *For a primitive substitution Markov chain Σ, $\alpha \in \mathcal{A}$, and $w \in L^{(m)}(P)$ the sequence of real numbers*

$$\frac{|\Sigma_n(\alpha_0)|_w}{|\Sigma_n(\alpha_0)|}$$

converges $_\alpha$ almost surely to a limit whose value is independent of α_0.

Proof. It is sufficient to show that the quantity $\frac{F_m(\Sigma_n(\alpha_0))}{|\Sigma_n(\alpha_0)|}$ converges. By corollary 2.4.8 applied to $\Sigma^{(m)}$, we have that

$$\frac{F_m(\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1}))}{|\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1})|}$$

converges almost surely to a strictly positive vector not depending on $\alpha_0 \dots \alpha_{m-1}$ (and hence not on α_0). Now by definition, we have that

$$|\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1})| = |\Sigma_n(\alpha_0)|.$$

Furthermore, we have that

$$F_m(\Sigma_n(\alpha_0)) = F_m(\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1})) + O(m-1)$$

Hence, $F_m(\Sigma_n(\alpha_0))$ is asymptotic to $F_m(\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1}))$. Therefore

$$\lim_{n\to\infty} \frac{F_m(\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1}))}{|\Sigma_n^{(m)}(\alpha_0 \dots \alpha_{m-1})|} = \lim_{n\to\infty} \frac{F_m(\Sigma_n(\alpha))}{|\Sigma_n(\alpha)|}$$

□

2.4.11 Examples

We conclude this section with an example that
demonstrates the above definitions and propositions. Let Σ be the substitution Markov chain defined by

$$\Sigma : \begin{cases} a \to \begin{cases} aa \text{ with prob. } 1/2 \\ ab \text{ with prob. } 1/2 \end{cases} \\ \\ b \to ba \end{cases}$$

Then the induced substitution $\Sigma^{(2)}$ is

$$\Sigma^{(2)}: \begin{cases} (aa) \to \begin{cases} (aa)(aa) \text{ with prob. } 1/2 \\ (ab)(ba) \text{ with prob. } 1/2 \end{cases} \\ (ab) \to \begin{cases} (aa)(ab) \text{ with prob. } 1/2 \\ (ab)(bb) \text{ with prob. } 1/2 \end{cases} \\ (ba) \to (ba)(aa) \\ (bb) \to (ba)(ab) \end{cases}$$

Also, the first two substitution matrices are

$$M_\Sigma^{(1)} = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{bmatrix} 3/2 & 1 \\ 1/2 & 1 \end{bmatrix} \end{array}$$

and

$$M_\Sigma^{(2)} = \begin{array}{c} \\ (aa) \\ (ab) \\ (ba) \\ (bb) \end{array} \begin{array}{c} \begin{array}{cccc} (aa) & (ab) & (ba) & (bb) \end{array} \\ \begin{bmatrix} 1 & 1/2 & 1 & 0 \\ 1/2 & 1 & 0 & 1 \\ 1/2 & 0 & 1 & 1 \\ 0 & 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

With the eigenvalues of $M_\Sigma^{(1)}$ being $\Lambda = 2, \lambda_2 = 1/2$ and the eigenvalues of $M_\Sigma^{(2)}$ being $\Lambda = 2, \lambda_2 = 1, \lambda_3 = 1/2, \lambda_4 = -1/2$, and both matrices are primitive. The normalized Perron-Frobenius eigenvectors of M_Σ and $M_\Sigma^{(2)}$ are

$$v = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}$$

and

$$\frac{v_2}{< v_2, \quad >} = \begin{bmatrix} 2/5 \\ 4/15 \\ 4/15 \\ 1/15 \end{bmatrix}$$

respectively. Thus, for example, a.s. in the limit the frequency of occurrence of the number of a's is $2/3$, and the frequency of occurrence of the number of ab subwords is $4/15$.

2.4.12 Eigenvalues of the substitution matrices and frequency calculations

In the preceding example, note how two eigenvalues of $M_\Sigma^{(1)}$ and $M_\Sigma^{(2)}$ coincided. In this section we show how the eigenvalues of $M_\Sigma^{(2)}$ completely describe the eigenvalues of $M_\Sigma^{(m)}$ for $m \geq 2$ and give an algorithm to efficiently calculate the subword frequency of any given word.

The results contained in this subsection are essentially a randomized version of those contained in [84] pages 100-104. We include them here not only as results on their own, but they will become important to parameter estimation in the biological applications of substitution Markov chains.

Again, in this subsection we assume that for a substitution Markov chain $(\mathcal{A}^*, P)$ under consideration it is not only primitive, but also there exists a letter $\alpha \in \mathcal{A}$ such that for some $\omega \in \Omega_\alpha$, $g_\alpha(\omega)$ begins with α. We also assume that there exists some $\beta \in \mathcal{A}$ such that for some $\omega \in \Omega_\beta$, $|g_\beta(\omega)| > 1$. This last assumption guarantees that for a fixed integer m, an integer n can be found such that for any $v_0, v_1 \in \mathcal{A}$,

$$\forall \omega \in \Omega^\infty, \ |\Sigma_n(v_0)(\omega)| + m - 2 < |\Sigma_n(v_0)(\omega)| + |\Sigma_n(v_1)(\omega)| \tag{2.22}$$

Indeed, since Σ is assumed to primitive then there is some p such that $\Sigma_p(v_1)(\omega)$ has β as a subletter so $|\Sigma_{(n+1)p}(v_1)(\omega)| \geq 2^n$.

Now recalling the definition of $\Sigma^{(m)}$, if n is fixed to satisfy equation (2.22),

then for $v = v_0 v_1 \dots v_{m-1}$, $\Sigma_n^{(m)}(v)$ is entirely determined by $v_0 v_1$, the first two letters of the word v. Now lemma 2.4.5 implies that

$$|\Sigma_n(v_1)| \sim \Lambda^n ||P_\Lambda(F_1(v_1))||$$

almost surely w.r.t P_{v_1} as $n \to \infty$. Thus in terms of Λ, if for some $C > 1$ a large enough positive constant, the condition 2.22 on m, n is equivalent to

$$\Lambda^n > Cm \tag{2.23}$$

Now fix a particular substitution Markov chain and m, n integers satisfying 2.23. Define $\pi_2 : L^{(m)}(P) \to L^{(2)}(P)$ as the restriction to the first two letters:

$$\pi_2(v_0 v_1 \dots v_{m-1}) = v_0 v_1$$

Recall that for S a set, by S^* we mean the set consisting of all finite concatenations of elements of S. Now define $\tau_{2,m,n} : L^{(2)}(P) \times \Omega^\infty \to (L^{(m)})^*$ by the following: For $v_0 v_1 \in L^{(2)}(P)$ and $\omega \in \Omega^\infty$ such that

$$\Sigma_n(v_0 v_1)(\omega) = y_0 \dots y_{|\Sigma_n(v_0)(\omega)|-1} y_{|\Sigma_n(v_0)(\omega)|} \dots y_{|\Sigma_n(v_0 v_1)(\omega)|}$$

let

$$\tau_{2,m,n}(v_0 v_1)(\omega) = (y_0 \dots y_{m-1})(y_1 \dots y_m) \dots (y_{|\Sigma(v_0)(\omega)|-1} \dots y_{|\Sigma(v_0)(\omega)|+m-2})$$

We now extend $\tau_{2,m,n}$ to a map from $(L^{(2)})^* \times \Omega^\infty$ into $(L^{(m)})^*$ and π_2 to a map from $(L^{(m)})^*$ into $(L^{(2)})^*$ in the natural way. Then we have the following

Lemma 2.4.13.

$$\begin{aligned} \tau_{2,m,n} \circ \pi_2 &= \Sigma_n^{(m)} \\ \pi_2 \circ \tau_{2,m,n} &= \Sigma_n^{(2)} \\ \Sigma^{(m)} \circ \tau_{2,m,n} &= \tau_{2,m,n} \circ \Sigma^{(2)} \end{aligned} \tag{2.24}$$

Proof. Consider $\omega \in \Omega^\infty$ and $(v_0, \ldots, v_{m-1}) \in L^{(m)}$ such that

$$\begin{aligned}&\Sigma_n^{(m)}(v_0 \ldots v_{m-1})(\omega)\\ &= (y_0 \ldots y_{m-1})(y_1 \ldots y_m) \ldots (y_{|\Sigma(v_0)(\omega)|} \ldots y_{|\Sigma(v_0)(\omega)|+m-2})\end{aligned}$$

and

$$\begin{aligned}&\Sigma^{(m)}((y_0 \ldots y_{m-1})(y_1 \ldots y_m) \ldots (y_{|\Sigma(v_0)(\omega)|} \ldots y_{|\Sigma(v_0)(\omega)|+m-2}), \omega)\\ &= (\hat{y}_0 \ldots \hat{y}_{m-1}) \ldots (\hat{y}_{|\Sigma(v_0)(\omega)|} \ldots \hat{y}_{|\Sigma(v_0)(\omega)|+m-2})\end{aligned}$$

Then by the definitions of $\tau_{2,m,n}$, $\Sigma_n^{(m)}$, and π_2, we have that

$$\begin{aligned}\tau_{2,m,n}(\pi_2(v_0 \ldots v_{m-1}), \omega) &= \tau_{2,m,n}(v_0 v_1, \omega)\\ &= (y_0 \ldots y_{m-1})(y_1 \ldots y_m) \ldots\\ &\quad (y_{|\Sigma(v_0)(\omega)|} \ldots y_{|\Sigma(v_0)(\omega)|+m-2})\\ &= \Sigma_n^{(m)}(v_0 \ldots v_{m-1})(\omega)\end{aligned}$$

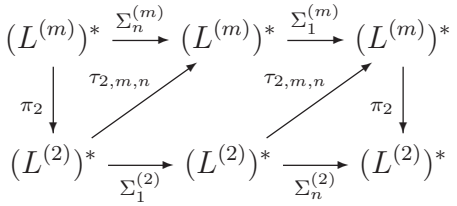
$$\begin{aligned}\pi_2(\tau_{2,m,n}(v_0 v_1, \omega)) &= \pi_2((y_0 \ldots y_{m-1})(y_1 \ldots y_m) \ldots\\ &\quad (y_{|\Sigma(v_0)(\omega)|} \ldots y_{|\Sigma(v_0)(\omega)|+m-2}))\\ &= (y_0 y_1)(y_1 y_2) \ldots (y_{|\Sigma(v_0)(\omega)|+m-3} y_{|\Sigma(v_0)(\omega)|+m-2})\\ &= \Sigma_n^{(2)}(v_0)(\omega)\end{aligned}$$

$$\begin{aligned}\Sigma^{(m)}(\tau_{2,m,n}(v_0 v_1, \omega), \omega) &= \Sigma^{(m)}((y_0 \ldots y_{m-1})(y_1 \ldots y_m) \ldots\\ &\quad (y_{|\Sigma(v_0)(\omega)|} \ldots y_{|\Sigma(v_0)(\omega)|+m-2}), \omega)\\ &= (\hat{y}_0 \ldots \hat{y}_{m-1}) \ldots (\hat{y}_{|\Sigma(v_0)(\omega)|} \ldots \hat{y}_{|\Sigma(v_0)(\omega)|+m-2})\\ &= \tau_{2,m,n}((y_0 y_1)(y_1 y_2) \ldots\\ &\quad (y_{|\Sigma(v_0)(\omega)|+m-3} y_{|\Sigma(v_0)(\omega)|+m-2}), \omega)\end{aligned}$$

$$= \tau_{2,m,n}(\Sigma^{(2)}(v_0v_1,\omega),\omega)$$

□

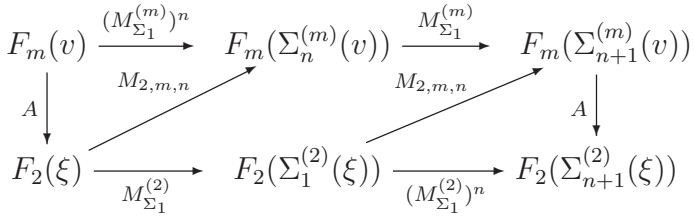
We can summarize 2.24 by the following commutative diagram



Now recall that $F_m(\Sigma^{(m)}(v)) = M_\Sigma^{(m)} F_m(v)$. Then define $M_{2,m,n}$ as the matrix of the expectation of $\tau_{2,m,n}$: For $v_0v_1 \in L^{(2)}(P)$ let

$$(M_{2,m,n})_{v_0v_1} = \quad \Omega^\infty(\tau_{2,m,n}(v_0v_1,\omega))$$

Then by the above definition, we have $M_{2,m,n}F_2(\pi_2(v)) = \quad F_m(\Sigma_n^{(m)}(v))$. Now denoting by A the matrix of the projection π_2 and $\xi = \pi_2(v)$, the above diagram becomes



Proposition 2.4.14. *For $m \geq 2$, the eigenvalues of the matrix $M_\Sigma^{(m)}$ are those of $M_\Sigma^{(2)}$ with perhaps in addition 0.*

Proof. Since the above commutative diagram implies that

$$M_{2,m,n}M_\Sigma^{(2)} = M_\Sigma^{(m)}M_{2,m,n}$$

then for any algebraic polynomial Q, we also have

$$M_{2,m,n}Q(M_\Sigma^{(2)}) = Q(M_\Sigma^{(m)})M_{2,m,n}$$

Similarly,

$$M_{2,m,n}Q(M_\Sigma^{(2)})A = Q(M_\Sigma^{(m)})(M_\Sigma^{(m)})^n$$

So if $Q(M_\Sigma^{(2)}) = 0$, then the polynomial $X \mapsto Q(X)X^n$ annihilates $M_\Sigma^{(m)}$ as well.

Also,

$$(M_\Sigma^{(2)})^n Q(M_\Sigma^{(2)}) = AQ(M_\Sigma^{(m)})(M_\Sigma^{(m)})^n$$

implies that the polynomial $X \mapsto Q(x)X^n$ annihilates the matrix $M_\Sigma^{(2)}$ if $Q(M_\Sigma^{(m)}) = 0$. Thus $M_\Sigma^{(2)}$ and $M_\Sigma^{(m)}$ have the same non-zero eigenvalues. □

Proposition 2.4.15. *If v_2 is an eigenvector of $M_\Sigma^{(2)}$ corresponding to the eigenvalue Λ, then $M_{2,m,n}(v_2)$ is an eigenvector of $M_\Sigma^{(m)}$ corresponding to the eigenvalue Λ*

Proof. This follows immediately from the equality $M_{2,m,n}M_\Sigma^{(2)} = M_\Sigma^{(m)}M_{2,m,n}$ which is derived from the above commutative diagram. □

2.4.15.1 Determining frequencies

From the above definitions and propositions, we derive the following algorithm to calculate the frequency of any $v \in L^{(m)}(P)$: First find the frequency of all pairs by calculating the Perron-Frobenius eigenvector of $M_\Sigma^{(2)}$, normalized so the sum of its entries is 1. Fix n satisfying 2.22. Then for every pair of letters $(\alpha\beta)$ which has positive frequency, enumerate the expected value of occurrences of v in $\Sigma_n(\alpha\beta)$ with the first letter of v (that is v_0) occurring in $\Sigma_n(\alpha)$. Note that the random variable $\Sigma_n(\alpha\beta)$ has only finitely many values (since the underlying Markov chain has finite range), so the expectation is straightforward to calculate. This resulting number is the $(v, (\alpha\beta)) \in L^{(m)} \times L^{(2)}$ entry of the matrix $M_{2,m,n}$. The desired frequency can then be found by examining the v-th entry of $M_{2,m,n}(v_2)$.

We conclude this section with an example that will hopefully clarify some of the above definitions. Recalling the substitution Markov chain

$$\Sigma : \begin{cases} a \to \begin{cases} aa \text{ with prob. } 1/2 \\ ab \text{ with prob. } 1/2 \end{cases} \\ \\ b \to ba \end{cases}$$

and its induced substitution $\Sigma^{(2)}$:

$$\Sigma^{(2)} : \begin{cases} (aa) \to \begin{cases} (aa)(aa) \text{ with prob. } 1/2 \\ (ab)(ba) \text{ with prob. } 1/2 \end{cases} \\ \\ (ab) \to \begin{cases} (aa)(ab) \text{ with prob. } 1/2 \\ (ab)(bb) \text{ with prob. } 1/2 \end{cases} \\ \\ (ba) \to (ba)(aa) \\ \\ (bb) \to (ba)(ab) \\ \\ \end{cases}$$

it was found that

$$\frac{v_2}{< v_2, \quad >} = \begin{bmatrix} 2/5 \\ 4/15 \\ 4/15 \\ 1/15 \end{bmatrix}$$

Say we wish to enumerate the frequencies of all words of length 3 that can occur as subwords of Σ. Then $m = 3$ and we can fix $n = 1$ since for this n, 2.22 is satisfied (the substitution is of constant length 2). Concentrating on the (aa) column of $M_{2,m,n}$ observe that $\Sigma_1(aa)$ is the random variable with values $\{aaaa, aaab, abaa, abab\}$ distributed uniformly. Thus the length 3 word aba occurs with first letter in $\Sigma_1(a)$ with an expected value of $1/4+1/4 = 1/2$. So the $(aba, (aa))$ entry of $M_{2,m,n}$ is $1/2$. Continuing in this fashion, we obtain that

$$M_{2,m,n} = \begin{bmatrix} 3/4 & 0 & 1/2 & 0 \\ 1/4 & 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 1 \\ 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 0 \end{bmatrix}$$

with the columns sequentially corresponding (left to right) to the words aa, ab, ba, bb and the columns corresponding sequentially (top to bottom) to the words $aaa, aab, aba, abb, baa, bab, bba$ (note for example, that the word bbb does not appear in Σ_n for any n). Thus to find the frequencies of the words $aaa, aab, aba, abb, baa, bab, bba$:

$$M_{2,m,n}(v_2) = M_{2,m,n} \begin{bmatrix} 6 \\ 4 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 13/2 \\ 11/2 \\ 6 \\ 2 \\ 7 \\ 1 \\ 2 \end{bmatrix}$$

Normalizing this vector so the sum of the entries is one, we obtain the frequencies of all words of length 3 that appear:

$$\begin{bmatrix} 13/60 \\ 11/60 \\ 1/5 \\ 1/15 \\ 7/30 \\ 1/30 \\ 1/15 \end{bmatrix}$$

It should be noted that for substitutions with many $a \in \mathcal{A}$ such that $|g_a(\omega)| > 1$, the number of elements with non-zero probability summed over in the expected value in the calculation of $M_{2,m,n}$ grows exponentially in n, thus exponentially in m (via the inequality 2.22). This algorithm of explicit calculation of frequency of words $|v| = m$ is thus suited for "small" m.

It is very important to note that by proposition 2.4.15, the Perron-Frobenius eigenvector of $M_\Sigma^{(m)}$ depends linearly on that of $M_\Sigma^{(2)}$. This will effect the number of independent parameters that can be chosen in the model of molecular evolution that is presented later in this dissertation.

2.5 Topological entropy

2.5.1 Convergence of topological entropy

Since the state space of a substitution Markov chain is $\mathcal{A}^*$, we can define a complexity function on the individual states. After defining an analogue of topological entropy, we show that under mild conditions it converges almost surely.

Definition 2.5.2 (Complexity Function)**.** *For $u \in \mathcal{A}^*$, let the complexity function $p_u(n)$ be defined as*

$$p_u(n) = |\{v \in \mathcal{A}^n, |u|_v > 0\}|$$

So the complexity function $p_u(n)$ represents the number of different subwords of length n occurring (with overlap) in u. The notion of a complexity function $p_u(n)$ is usually reserved for infinite words u, so we emphasize here that we allow u to be of finite length. We note that for u infinite, $p_u(n)$ is a non-decreasing function of n. We show that for a finite word u, once $p_u(n)$ decreases and attains the slope of -1, then it continues to do so at a slope of -1.

Proposition 2.5.3. *For $m \in$ and $u \in \mathcal{A}^m$, if n is such that $p_u(n+1) -$*

$p_u(n) = -1$ *then for all* $1 \leq i \leq m - n - 1$,

$$p_u(n + 1 + i) - p_u(n + i) = -1$$

Proof. For $u \in \mathcal{A}^m$, say $u = u_1 u_2 \dots u_m$ let

$$u^{(i)} := (u_1 u_2 \dots u_i)(u_2 \dots u_{i+1})(u_{m-i+1} \dots u_m)$$

be a word over the alphabet $\mathcal{A}^i$. Note that $|u^{(i)}| = m - i + 1$ where length $|\cdot|$ denotes length over the alphabet $\mathcal{A}^i$. We also denote $u_i^j = u_i u_{i+1} \dots u_j$. Now, we show

$$\begin{aligned} &p_u(n+1) - p_u(n) = -1 \text{ if and only if} \\ &\forall j = 1, \dots m - n - 1 \ \exists! a \in \mathcal{A} \text{ s.t. } |u|_{u_j^{j+n} a} > 0 \end{aligned} \tag{2.25}$$

Such a are usually called "right extensions" in the literature [37]. The converse direction of (2.25) is trivial, and in the forward direction, note that $p_u(n) = \{w \in \mathcal{A}^n, |u^{(n)}|_w \geq 0\}$, that is, the number of distinct letters in $u^{(n)}$. Also, $p_u(n+1) = \{w \in \mathcal{A}^{n+1}, |u^{(n+1)}|_w \geq 0\}$. Combining this with $|u^{(n+1)}| - |u^{(n)}| = m-n-1+1-m+n-1 = -1$ the forward direction is complete. Now let n be as in the assumption of the proposition and let i be such that $1 \leq i \leq m-n-1$, then if there were $a \neq b \in \mathcal{A}$ such that $|u|_{u_j^{j+n+i} a} > 0$ and $|u|_{u_j^{j+n+i} b} > 0$ then this implies that $|u|_{u_j^{j+n} a} > 0$ and $|u|_{u_j^{j+n} b} > 0$ contradicting 2.25. Thus it must be that $p_(n + i + 1) - p_u(n + i) = -1$. □

The convergence of the complexity function on a substitution Markov chain Markov chain is now investigated.

Theorem 2.5.4. *For a primitive substitution Markov chain* Σ*, for each* $n \in$ *and* $a \in \mathcal{A}$ *the quantity*

$$p_{\Sigma_m(a)}(n)$$

converges $_a$ *almost surely as* $m \to \infty$ *to a quantity independent of* a.

Proof. Since Σ is primitive, let Λ be the dominant eigenvalue of M_Σ. Recalling

that $p_u(n) = |\{v \in \mathcal{A}^n, |u|_v > 0\}|$ we have that

$$\begin{aligned} p_{\Sigma_m(a)}(n) &= |\{v \in \mathcal{A}^n, |\Sigma_m(a)|_v > 0\}| \\ &= |\{v \in \mathcal{A}^n, \frac{|\Sigma_m(a)|_v}{\Lambda^m} > 0\}| \end{aligned}$$

Now proposition 2.4.10 implies that $\frac{|\Sigma_m(a)|_v}{\Lambda^m}$ converges $_a$ almost surely as $m \to \infty$ to a vector whose value is independent of a. Since $|\{v \in \mathcal{A}^n, \frac{|\Sigma_n(a)|_v}{\Lambda^m} > 0\}|$ is simply the number of non-zero entries of the above mentioned vector, the result is obtained. □

We can now associate a notion of topological entropy to a substitution Markov chain. The above theorem ensures that the following definition makes sense.

Definition 2.5.5 (Topological entropy)**.** *For Σ a primitive SMC chain on an alphabet of length $|\mathcal{A}| = t$, we define the* ***topological entropy*** *$H_{top}(\sigma)$ as the limit*

$$H_{top}(\Sigma) = \lim_{n\to\infty} \lim_{m\to\infty} \frac{\log_t p_{\Sigma_m(a)}(n)}{n}$$

where the limit with respect to m is $_a$ almost surely and the limit with respect to n is taken in the usual sense.

We now compare this definition to the definition with the definition of topological entropy of a primitive deterministic substitution. Details are few in the following paragraph, but we direct the interested reader to [84]. For σ a deterministic substitution over $|\mathcal{A}| = t$ (that is, an application from $\mathcal{A}$ to $\mathcal{A}^*$ whose domain is extended to $\mathcal{A}^*$ via concatenation), say there is an $a \in \mathcal{A}$ such that $\sigma(a) = a\ldots$, and $\lim_{n\to\infty} |\sigma^n(b)| = \infty$ for all $b \in \mathcal{A}$. Then for $u \in \mathcal{A}^\infty$ an infinite word such that $\sigma(u) = u$, the topological entropy of σ is $H_{top}(\sigma) = \lim_{n\to\infty} \frac{\log_t p_u(n)}{n}$. In [84] it is shown that for such σ, $H_{top}(\sigma) = 0$ always. This is *distinctly* different from the primitive substitution Markov chain case, for example

$$\Sigma : \begin{cases} a \to \begin{cases} a \text{ with prob. } 1/3 \\ b \text{ with prob. } 1/3 \\ ab \text{ with prob. } 1/3 \end{cases} \\ b \to \begin{cases} a \text{ with prob. } 1/3 \\ b \text{ with prob. } 1/3 \\ ba \text{ with prob. } 1/3 \end{cases} \end{cases}$$

has topological entropy 1 since it can easily be seen that $P^{(n)}(a, w) > 0$ for any $w \in \mathcal{A}^n$.

2.5.6 Expected value of topological entropy

We now calculate the expected value of $H_{top}(\Sigma)$ for a particular class of substitution Markov chains Σ. On an alphabet of length $|\mathcal{A}| = t$, we restrict our attention to those substitution Markov chains Σ such that for each $a \in \mathcal{A}$, $g_a(\Sigma_a) = \mathcal{A} \cup \mathcal{A}^2$ and $P_a \equiv \frac{1}{t(1+t)}$. Due to the algorithm described after proposition 2.4.15, it is not too hard to see that such substitution Markov chains have the property that the frequencies of subwords in $L^n(P)$ are uniformily distributed. Thus, we can restrict to taking the expectation with respect to the uniform distribution directly.

Note that for an alphabet $\mathcal{A}$ of length t, there are exactly t^n subwords (counting multiplicity) of length n in a word of length $t^n + n - 1$. Given c distinct subwords (say such subwords are $\{w_1, \ldots, w_c\}$) then for m_i denoting the number of appearances of the subword w_i, there are $\binom{t^n}{m_1,\ldots,m_c}$ words of length t^n+n-1 with those c distinct subwords appearing m_i times respectively. Here $\binom{t^n}{m_1,\ldots,m_c} = \frac{(t^n)!}{m_1!\ldots m_c!}$ is a multinomial coefficient. Note that there are $\binom{t^n}{c}$ different ways to choose c distinct words of length n from $\mathcal{A}$. Summing over all possible appearances of the subwords, we obtain that the number of words over the alphabet $\mathcal{A}$, $|\mathcal{A}| = t$ of length $t^n + n - 1$ that have exactly c subwords

of length n is equal to

$$\binom{t^n}{c} \sum_{\substack{m_1,\ldots,m_c \geq 1 \\ \sum_{i=1}^{c} m_i = t^n}} \binom{t^n}{m_1,\ldots,m_c}$$

Thus the expected value of $H_{top}(u)$ over all words u over the alphabet $\mathcal{A}$ with $|u| = t^n + n - 1$ given the uniform distribution is given by

$$(H_{\text{top}}(\Sigma, n)) = \frac{1}{(t^n)^{t^n}} \sum_{c=1}^{t^n} \left(\frac{\log_t(c)}{n} \binom{t^n}{c} \sum_{\substack{m_1,\ldots,m_c \geq 1 \\ \sum_{i=1}^{c} m_i = t^n}} \binom{t^n}{m_1,\ldots,m_c} \right)$$

To better understand this expected value, we present here a table of values of the above function when the alphabet consists of four letters $|\mathcal{A}| = 4$.

n	$4^n + n - 1$	$(H_{\text{top}}(\Sigma, n))$
1	4	.703583
2	17	.838496
3	66	.890353

Table 2.1. Expected Value of Topological Entropy

While of theoretical interest, the above formula for the expected value is of little computational value due to the fact that the inner summation set grows exponentially in n. When considering the applications of topological entropy to DNA sequences, it will be important to consider the expected value of H_{top} for $n \geq 3$. We thus delay the asymptotic approximation of the expectation until we investigate the biological applications of topological entropy later.

Lastly, we note that given a substitution Markov chain that is particularly analytically tractable, one can use the algorithm given after proposition 2.4.15 to explicitly calculate the associated topological entropy.

Chapter 3

Boundaries associated to an SMC

The boundary theory for Markov chains originated in the work of Doob [23] and has recently found new interest in random walks on trees and abstract hyperbolic graphs (see [113]). The Martin boundary is an important topological boundary that describes all positive harmonic functions by integrals over this boundary; hence the Martin boundary is the most important boundary from a probabilistic and potential theoretic viewpoint. The first thorough treatment of the Martin boundary for Markov chains was by Dynkin [28] who wrote a well-considered and polished account of Hunt's paper [48]. In the context of Markov chains, the standard sources are the books by Kemeny, Snell and Knapp [53], Revuz [87], and Woess [113], [114].

The key difference between the classical literature just mentioned and the class of countable state Markov chains we consider here is that a substitution Markov chain is never irreducible on *any* non-trivial subset of the state-space. Besides Dynkin, the classical literature ([53], [87], [113], [114]) all consider Markov chains (X, P) that are irreducible. Dynkin [28] considered Markov chains (X, P) and initial distributions γ referred to by him as *standard measures*: a measure $\gamma : X \to$ is called *standard* if for all $y \in X$, $\sum_{x \in X} \gamma(x) \sum_{n=0}^{\infty} P^n(x, y) > 0$.

We wish to consider initial distributions that are point masses: $x \in X$, $\gamma = \delta_x$. For substitution Markov chains, such initial distributions are not *standard* according to the definition of Dynkin. This important distinction requires us to investigate the validity of classic convergence theorems. We closely

following the development of the Martin boundary given in [28] but with the necessary changes when considering non-standard measures. After developing the proper definitions, we prove analogs to the classical convergence theorems (convergence to the boundary in theorem 3.2.6 and the integral representation theorem in theorem 3.2.9) and compare these to the classical theorems (when the Markov chain can be divided into irreducible classes or when standard measures are being considered).

It is a natural question to identify the Martin boundary. We identify the Martin boundary in two classes of examples. We also show that the Poisson boundary is always trivial for a substitution Markov chain.

Since substitution Markov chains are not irreducible and the initial distributions we choose are non-standard, it can happen that the Martin kernel $K(\cdot,\cdot)$ (and hence the Martin boundary) depends upon the choice of the root. It is a long-term goal of the author to show how precisely the Martin boundary depends upon the choice of the root. In chapter 4, we present an alternative approach where we define a (reversible) Markov chain associated to an SMC that is in fact irreducible. This will allow us to circumvent non-standard measures for the applications considered in chapter 5.

In this chapter, unless otherwise stated, we assume that the substitution Markov chains are primitive and that for all $a \in \mathcal{A}^*$ and for all $\omega \in \Omega_a$, $|g_a(\omega)| > 1$.

3.1 Definitions

3.1.1 Reducibility

Of primary importance is the recognition that substitution Markov chains are never irreducible as observed in lemma 2.3.7. Traditionally, [114] if a Markov chain is not irreducible, then the state space is simply divided into disjoint sets, each set being irreducible (these are sometimes called *irreducible classes*). This strategy is not applicable here, as lemma 2.3.7 in fact shows that a substitution Markov chain is not irreducible on any non-trivial subset $X \subseteq \mathcal{A}^*$.

Previously, ([20], [21], [22]), a fixed root o has been chosen and the point-

mass on this root δ_o is observed to be a *standard measure* following the definition of Dynkin [28].

In general for substitution Markov chains, no point-mass is ever a standard measure and furthermore the SMC's are not irreducible. We propose to name this particular class of Markov chain as *completely reducible* .

Definition 3.1.2. *A Markov chain (X, P) is said to be* ***completely reducible*** *if for any subset of the state space $Y \subset X$ with $|Y| > 1$, $(Y, P|_Y)$ is not irreducible as per definition 2.3.5.*

Note that lemma 2.3.7 implies in particular that for an SMC$(\mathcal{A}^*, P)$ if for all $a \in \mathcal{A}^*$ and for all $\omega \in \Omega_a$, $|g_a(\omega)| > 1$ then $(\mathcal{A}^*, P)$ is completely reducible .

Definition 3.1.3 (Relatedness)**.** *For a substitution Markov chain $(\mathcal{A}^*, P)$ with $w, v \in \mathcal{A}^*$ if $n \in$ is the smallest such n such that $P^n(w, v) > 0$, we then write $w \ll_n v$. Also, if there is some m such that $P^m(w, v) > 0$ then we write $w \ll v$.*

Definition 3.1.4 (Descendants)**.** *For a substitution Markov chain $(\mathcal{A}^*, P)$ and $w \in \mathcal{A}^*$, let*

$$D(w) = \{v : \exists n \text{ s.t. } P^n(w, v) > 0\}$$

Definition 3.1.5 (Ancestors)**.** *For a substitution Markov chain $(\mathcal{A}^*, P)$ and $w \in \mathcal{A}^*$, let*

$$A(w) = \{v : \exists n \text{ s.t. } P^n(v, w) > 0\}$$

3.1.6 Potential theory

From the perspective of boundary theory, the key difference between completely reducible Markov chains and Markov chains that can be divided into irreducible classes is the failure of the *minimum principle*. This potential-theoretic theorem changes in a key fashion when attention is restricted to completely reducible Markov chains. Before stating the theorem, we recall a number of well known definitions here.

Definition 3.1.7 (Green's function)**.** *For a substitution Markov chain* $(\mathcal{A}^*, P)$*, let*

$$G = \sum_{n=0}^{\infty} P^n$$

Thus $G(x,y) = \sum_{n=0}^{\infty} P^n(x,y)$ with $P^0 = I$ the identity. Note that $x \ll y$ if and only if $G(x,y) > 0$.

Thinking of P and G as (infinite) matrices on l^∞ and functions $f : \mathcal{A}^* \to$ as column vectors and measure $\nu : \mathcal{A}^* \to$ as row vectors, the following is simply matrix multiplication (assuming of course that the following sums exist):

$$Pf(x) = \sum_{y \in \mathcal{A}^*} P(x,y) f(y)$$

$$Gf(x) = \sum_{y \in \mathcal{A}^*} G(x,y) f(y)$$

$$\nu P(x) = \sum_{y \in \mathcal{A}^*} \nu(x) P(x,y)$$

$$\nu G(x) = \sum_{y \in \mathcal{A}^*} \nu(x) G(x,y)$$

Definition 3.1.8 (Harmonic functions)**.** *A function* $f : \mathcal{A}^* \to$ *is said to be harmonic if for all* $x \in \mathcal{A}^*$*,*

$$Pf(x) = f(x)$$

Also, f *is said to be superharmonic if for all* x*,*

$$Pf(x) \leq f(x)$$

Finally, f *is said to be subharmonic if for all* x*,*

$$Pf(x) \geq f(x)$$

Note that the property of being harmonic in essence says that the value of f at a point x is equal to the weighted average (weighted via P) of the value of f on the y such that $x \ll_1 y$.

Definition 3.1.9 (Potential)**.** *For a function $f : \mathcal{A}^* \to$, its potential is the function $g = Gf$.*

Proposition 3.1.10. *If g is a potential of f, then $f = (I - P)g$ and for each $x \in X$, $P^n g(x) \xrightarrow[n\to\infty]{} 0$.*

The following minimum principle is quantitatively different when applied to completely reducible Markov chains.

Proposition 3.1.11 (Minimum principle for completely reducible Markov chains)**.** *If f is a superharmonic function on completely reducible Markov chain (X, P) such that there is an $x \in X$ with $f(x) = \min_X f$, then f is constant on $D(x)$.*

Proof. Let f, x be as in the hypothesis and set $M = \min_X f$, then for $x' \in X$ with $x \ll x'$, due to f being superharmonic, we have

$$\begin{aligned} M = f(x) &\geq P^n f(x) \\ &= P^n(x, x')f(x') + \sum_{y \neq x'} P^n(x, y)f(y) \\ &\geq P^n(x, x')f(x') + \sum_{y \neq x'} P^n(x, y)M \\ &= P^n(x, x')f(x') + (1 - P^n(x, x'))M \\ &= P^n(x, x')f(x') + M - MP^n(x, x') \end{aligned}$$

Hence $f(x') \leq M$, but by definition of M we have $f(x') \geq M$ as well, hence $f(x') = M$. Since this holds for each $x' \in D(x)$, f is constant on $D(x)$. □

In the above proposition, if f is subharmonic, then there exists a maximum principle: simply apply the minimum principle to $-f$.

We recall now the classical minimum principle to illustrate the difference complete reducibility imposes.

Proposition 3.1.12 (Minimum principle for irreducible Markov chains)**.** *If f is a superharmonic function on an irreducible Markov chain (X, P) such that there is an $x \in X$ with $f(x) = \min_X f$, then f is constant all of X.*

Proof. See [114, pg. 5]. □

The minimum (or maximum) principle is a key lemma in proving the classic theorems regarding convergence to the boundary so we devote the next few pages to investigating how these theorems change for completely reducible Markov chains.

3.1.13 Transience

We will be mainly concerned with transient Markov chains, and take the following as our definition of transience.

Definition 3.1.14 (Transient)**.** *A Markov chain (X, P) is said to be transient if for all $x, y \in L(P)$, $G(x, y) < \infty$.*

We will mainly be interested in two classes of substitution Markov chains. A substitution Markov chain from the first class is called a *expansive substitution Markov chain* and one from the second class is called a *point/indel substitution Markov chain*. The terminology of the second class is based on the applications to biological systems. In short, a point mutation is a single letter being replaced with a single letter. Indel stands for insertion/deletion: an insertion is a single letter being replaced with a finite length (greater than 1) word. The deletions (word being replaced with a single letter) will be addressed in the chapter on reversible substitution Markov chains.

Definition 3.1.15 (Expansive substitution Markov chain)**.** *A substitution Markov chain $(\mathcal{A}^*, P)$ is said to be expansive if for every $a \in \mathcal{A}$ and for every $\omega \in \Omega_a$, $|g_a(\omega)| > 1$.*

Note that any constant length $k > 1$ substitution is necessarily expansive.

Definition 3.1.16 (Point/indel substitution Markov chain)**.** *A point/indel substitution Markov chain is a substitution Markov chain $(\mathcal{A}^*, P)$ such that for each $a \in \mathcal{A}$, there exists $\omega_1, \omega_2 \in \Omega_a$ such that $|g_a(\omega_1)| = 1$ and $|g_a(\omega_2)| > 1$.*

An example of a point/indel substitution Markov chain would be the following:

$$\Sigma : \begin{cases} a \to \begin{cases} a \text{ with prob. } 1/4 \\ ab \text{ with prob. } 3/4 \end{cases} \\ b \to \begin{cases} b \text{ with prob. } 1/4 \\ ba \text{ with prob. } 3/4 \end{cases} \end{cases}$$

Theorem 3.1.17 (Transience of expansive substitution Markov chains)**.** *All expansive SMCs are transient.*

Proof. Let $(\mathcal{A}^*, P)$ be an expansive substitution Markov chain, and let

$$M = \min_{a \in \mathcal{A},\ \omega \in \Omega_a} |g_a(\omega)|$$

Let $x, y \in L(P)$ be arbitrary and recall that for any $v, u = u_1 u_2 \cdots \in \mathcal{A}^*$ and for any $\omega = (\omega_1, \ldots, \omega_{|u|}) \in \Omega_u$, $|g_u(\omega)| = \sum_i |g_{u_i}(\omega_i)|$. Thus if $P^{(n)}(x, y) > 0$, then $|y| \geq |x| M^n$. Thus if we let $N \geq n > \log_M \left(\frac{|y|}{|x|} \right)$, then we have that $P^{(N)}(x, y) = 0$. Thus $G(x, y) = \sum_{i=0}^{N} P^{(i)}(x, y)$ is a finite sum and so certainly finite. □

Note that theorem 3.1.17 implies that any constant length $k > 1$ substitution Markov chain is transient.

Theorem 3.1.18 (Transience of point/indel substitution Markov chains)**.** *For a point/indel substitution Markov chain $(\mathcal{A}^*, P)$ over an alphabet $\mathcal{A}$ with $|\mathcal{A}| = t$, let*

$$M = \max_{\substack{a \in \mathcal{A},\ \omega \in \Omega_a \\ |g_a(\omega)| = 1}} P_a(\omega)$$

If $M < \frac{1}{t}$, then $(\mathcal{A}^*, P)$ is transient

Proof. Under the assumptions of the theorem, first let $x, y \in L(P)$ be such that $|x| = |y|$. By a path from x to y of length n we mean a sequence of vertices $w_1, \ldots, w_n$ such that $P(x, w_1) P(w_1, w_2) \ldots P(w_n, y) > 0$. Now since $|x| = |y|$ then there are certainly $t^{|x|(n-1)}$ or less paths from x to y of length n. Note that there are exactly $t^{|x|(n-1)}$ paths precisely when for each $a, b \in \mathcal{A}$,

there exists $\omega \in \Omega_a$ such that $g_a(\omega) = b$. Now by the above definition of M, for a path $w_1, \ldots, w_n$, $P(x, w_1) \ldots P(w_n, y) \leq M^{n|x|}$. Thus

$$P^{(n)}(x, y) \leq (t^{n-1} M^n)^{|x|}$$

Therefore we have that

$$G(x, y) \leq \frac{1}{t^{|x|}} \sum_{n=0}^{\infty} (tM)^{n|x|} < \infty$$

which is convergent since by assumption $M < \frac{1}{t}$ and so the sum is bounded by a convergent p-series.

For the case when $|x| \neq |y|$, recall a property of the Green's function $G(\cdot, \cdot)$. For $W \subset \mathcal{A}^*$, recalling that Σ_n is the n-th coordinate process of the Markov chain $(\mathcal{A}^*, P)$, define the stopping time

$$\mathbf{v}^W = \inf\{n \geq 0 : \Sigma_n \in W\}$$

Then

$$F(x, y) = \quad {}_x[\mathbf{v}^y < \infty]$$

stands for the probability of ever visiting y, starting from x.

Now a well known theorem (see [114], [113]) is that

$$G(x, y) = F(x, y) G(y, y)$$

Now say $|x| \neq |y|$, so without loss of generality say $|x| < |y|$. By the first part of the proof, we have $G(y, y) < \infty$ since $|y| = |y|$ and so

$$G(x, y) = F(x, y) G(y, y) < G(y, y) < \infty.$$

as desired.

□

3.1.19 Martin boundary

The following exposition of basic Martin boundary notions closely follows that of [114] with special attention to the fact that our SMC's are not irreducible. While discarding the assumption of irreducibility leads to key changes in some classical convergence theorems, it does however greatly simplify the actual calculation of the Martin boundary as paths are much more easily described. It is a long-term goal of the author to show that under suitable conditions, the Martin boundary of a non-irreducible Markov chain and that of a certain associated irreducible Markov chain actually coincide.

Now we turn our attention to the Martin Boundary proper. First we want to understand the compact convex set $B = \{u \text{ non} - \text{negative superharmonic} : u(o) = 1\}$ where o is a chosen root of our Markov Chain. As mentioned in the beginning of the chapter, we consider non-standard measures. In particular, we fix a measure $\gamma = \delta_a$ where $\mathcal{A} = \{a, \dots\}$. It is then natural to choose $o = a$ as the root of our SMC under consideration. We develop the relevant definitions and convergence theorems in this case.

Definition 3.1.20. *We denote by K the Martin kernel, given by*

$$K(x, y) = \frac{G(x, y)}{G(a, y)}$$

Note that $PG = G - I$ so for fixed y, $K(\cdot, y) \in B$.

Definition 3.1.21 (Minimal Harmonic Function)**.** *A minimal harmonic function is a function on X with $h \geq 0$ s.t. $Ph = h$, $h(a) = 1$ and whenever $h \geq h_1$ on X for another $h_1 \geq 0$ (h_1 harmonic), $h_1 = ch$ with c a constant.*

We attempt to determine the extremal elements of B: an element $u \in B$ is extremal if it cannot be written as a convex combination of two other elements of B.

It will turn out that the extremal elements of B are precisely the Martin kernels and the minimal harmonic functions. So we could define the Martin Boundary as the closure of X in B under the pointwise topology (note that since X is discrete, this topology coincides with the topology of convergence

on compact subsets). However, this definition is difficult to work with in practice as the topologies involved are particularly obfuscated. We thus make the following definition:

Definition 3.1.22 (Martin Compactification)**.** *The* ***Martin Compactification*** $\hat{X}(P)$ *is the completion of* (X, θ) *where for* w_x *positive weights such that* $\sum_{x\in X} w_x/G(a,x) < \infty$

$$\theta(y_1, y_2) = \sum_{x\in X} w_x \left(|K(x, y_1) - K(x, y_2)| + |\delta_x(y_1) - \delta_x(y_2)|\right)$$

We make some observations regarding this metric.

Proposition 3.1.23. *A sequence* $(y_n)_{n\geq 0}$ *is* θ*-Cauchy iff*

1. *For each* $x \in X$, $K(x, y_n)$ *converges as* $n \to \infty$ *and*
2. *Either* $|y_n| \to \infty$ *(leaves each finite subset of* X*) or else* $(y_n)_{n\geq 0}$ *is eventually constant.*

Proof. In the forward direction, say $(y_n)_{n\geq 0}$ is a θ-Cauchy sequence. Let $x \in X$, then since $\theta(y_n, y_m) \geq w_x|K(x, y_n) - K(x, y_m)|$ we must have that the sequence of real numbers $(K(x, y_n))_{n\geq 0}$ is Cauchy and so convergent. Now we also have for $y_n \neq y_m$ that $\theta(y_n, y_m) \geq w_{y_n} + w_{y_m}$ and so either $|y_n| \to \infty$ or else (y_n) is eventually constant.

In the converse directions, first say that (y_n) is eventually constant, then for n large enough, $y_n = y_m$ so $\theta(y_n, y_m) = 0$ and so (y_n) is θ-Cauchy. Now say that $|y_n| \to \infty$ and $K(x, y_n)$ converges in n for each x. Since $K(x, y_n)$ converges, then $K(x, y_n)$ is a Cauchy sequence of real numbers and hence

$$\lim_{m,n\to\infty} |K(x, y_n) - K(x, y_m)| = 0$$

Now $K(x, y) \leq \frac{1}{G(a,x)}$ so $\sum_x w_x|K(x, y_n) - K(x, y_m)| \leq 2\sum_x \frac{w_x}{G(a,x)} < \infty$ by assumption. Now $|y_n| \to \infty$ as $n \to \infty$ so $\lim_{n,m\to\infty} |\delta_x(y_n) - \delta_x(y_m)| = 0$ for any fixed x. Finally, utilizing that θ is a bounded metric (and so the interchange of summation and limit in the following is justified) we summarize the above to observe that

$$\begin{aligned}
\lim_{n,m\to\infty} \theta(y_n, y_m) &= \lim_{n,m\to\infty} \sum_x w_x \left(|K(x, y_n) - K(x, y_m)| + |\delta_x(y_n) - \delta_x(y_m)|\right) \\
&= \sum_x \lim_{n,m\to\infty} w_x \left(|K(x, y_n) - K(x, y_m)| + |\delta_x(y_n) - \delta_x(y_m)|\right) \\
&= \sum_x \lim_{n,m\to\infty} w_x |K(x, y_n) - K(x, y_m)| \\
&\leq \sum_x \lim_{n,m\to\infty} |K(x, y_n) - K(x, y_m)| \\
&= 0
\end{aligned}$$

Hence $(y_n)_{n\geq 0}$ is a θ-Cauchy sequence as desired.

□

Due to the above proposition, we note that $\hat{X}(P)$ is in fact homeomorphic to the closure of X in B as desired. Thus the Martin compactification is the smallest compactification such that all the Martin kernels $K(x, \cdot)$ extend continuously. Now by construction, X is dense in $\hat{X}(P)$, and since the set of θ-Cauchy sequences for which $|y_n| \to \infty$ is clearly closed, X is also open in $\hat{X}(P)$. We then make the following definition:

Definition 3.1.24 (Martin Boundary)**.** *The* ***Martin boundary*** $\mathcal{M}(P) = \mathcal{M}$ *is defined to be* $\hat{X}(P)/X$*.*

Since we formed $\hat{X}(P)$ via the completion of (X, θ) (equivalence classes of θ-Cauchy sequences), due to the proposition 3.1.23, an equally valid definition of $\hat{X}(P)$ is as equivalence classes of sequences (y_n) where two sequences $(x_n), (y_n)$ are equivalent if for each $z \in X$, $\lim_n K(z, x_n) = \lim_n K(z, y_n)$. When actually finding a tractable description of the Martin boundary, we will often utilize proposition 3.1.23 and determine the elements of $\mathcal{M}$ by finding sequences (y_n) that satisfy

1. For each $x \in X$, $K(x, y_n)$ converges as $n \to \infty$ and
2. $|y_n| \to \infty$

For $\xi \in \mathcal{M}$, we define $K(\cdot, \xi)$ as $\lim_{n\to\infty} K(\cdot, y_n)$ where $(y_n)_{n\geq 0}$ belongs to the equivalence class of sequences defined by ξ. Under this extension, the Martin kernels separate boundary points: for $\xi \neq \psi \in \mathcal{M}$, then for some $x \in X$, $K(x, \xi) \neq K(x, \psi)$.

Now recall that every substitution Markov chain $(\mathcal{A}^*, P)$ has finite range: $|\{y : P(x, y) > 0\}| < \infty$. Thus for $(y_n)_{n\geq 0}$ a θ-Cauchy sequence belonging to the equivalence class ξ, $K(\cdot, \xi)$ is superharmonic as it is the limit of the superharmonic functions $K(\cdot, y_n)$. Thus

$$\begin{aligned}
\lim_{n\to\infty} PK(x, y_n) &= \lim_{n\to\infty} \sum_{y:P(x,y)>0} P(x, y)K(y, y_n) \\
&= \sum_{y:P(x,y)>0} P(x, y) \lim_{n\to\infty} K(y, y_n) \\
&= \lim_{n\to\infty} K(x, y_n) - \frac{\delta_x(y_n)}{K(a, y_n)} \\
&= \lim_{n\to\infty} K(x, y_n)
\end{aligned}$$

Thus $K(\cdot, \xi)$ is a non-negative harmonic function with $K(a, \xi) = 1$ and so $K(\cdot, \xi) \in B$.

The modification to the minimum principle also allows us to determine necessary conditions for the convergence of $\lim_n K(\cdot, x_n)$.

Proposition 3.1.25. *For $K(\cdot, \cdot)$ the Martin kernel, for each $y \in X$, $\lim_{n\to\infty} K(y, x_n)$ exists and is not identically zero in the topology of the Martin compactification only if for all but finitely many n, $x_n \in D(y)$ or else $x_n \notin D(y)$.*

Proof. Arguing by contradiction, say there is a state $y \in X$ such that $x_n \in D(y)$ infinitely often and $x_n \notin D(y)$ infinitely often. Then $K(y, x_n) \neq 0$ for $x_n \in D(y)$ and $K(y, x_n) = 0$ for $x_n \notin D(y)$, hence $K(y, x_n)$ cannot converge, a contradiction to our assumption. □

Since the Martin boundary can be viewed as equivalence classes of θ-Cauchy sequences $(x_n)_{n\geq 0}$, proposition 3.1.25 indicates in which direction to look for such sequences.

3.2 Convergence to the boundary and integral representation

Now that the Martin boundary has been defined, we can investigate two fundamental theorems (convergence to the boundary and integral representation) and see how they change when the Markov chains under consideration are completely reducible .

In the following, let (X, P) be a completely reducible Markov chain, not just a substitution Markov chain. We denote by $(X^{\mathbb{N}_0}, \mathcal{B})$ the trajectory space with Z_n the n-th coordinate process. Before we state the convergence and integral representation theorems, we will need to do some preparatory work regarding downward crossings.

Definition 3.2.1 (Downward Crossings)**.** *Let $(r_n)_{n\geq 0} \in \mathbb{R}^{\mathbb{N}_0}$ and $[a, b] \subset \mathbb{R}$, then we denote the number of downward crossings as:*

$$D_{\downarrow}\left((r_n)|[a,b]\right) = \sup\left\{ k \geq 0 : \begin{array}{l} \text{there are } n_1 \leq n_2 \leq \cdots \leq n_{2k} \\ \text{with } r_{n_i} \geq b \text{ for } i = 1, 3, \ldots, 2k-1 \\ \text{and } r_{n_j} \leq a \text{ for } j = 2, 4, \ldots, 2k \end{array} \right\}$$

The following is a well known lemma.

Lemma 3.2.2 (Doob's Downward Crossings Lemma)**.** *Let (W_n) be a non-negative supermartingale. Then for every interval $[a, b] \subset \mathbb{R}^{+}$,*

$$\left(D_{\downarrow}\left((W_n)|[a,b]\right)\right) \leq \frac{1}{b-a}\ (W_0)$$

Proof. See [114, Lemma 7.29] □

A corollary to lemma 3.2.2 is the following

Corollary 3.2.3. *If $(W_0) < \infty$ then $\lim_{n\to\infty} W_n$ exists almost surely.*

We will apply this lemma to superharmonic functions of the n-th coordinate process Z_n. Consider the Markov chain $(Z_m)_{n\geq 0}$ with probability $\mathbb{P}_x$, let $f : X \to \mathbb{R}$ be a non-negative function and consider $(f(Z_n))_{n\geq 0}$. Now $(f(Z_n))_{n\geq 0}$

is a supermartingale with respect to $(z_n)_{n\geq 0}$ if and only if for each n and $x_0, \ldots, x_{n-1}$ we have

$$\begin{aligned}(f(Z_n)) &= \sum_{y\in X} \delta_x(x_0)p(x_0,x_1)\ldots p(x_{n-1},y)f(y)\\ &\leq \delta_x(x_0)p(x_0,x_1)\ldots p(x_{n-2},x_{n-1})f(x_{n-1})\end{aligned}$$

which is true if and only if f is superharmonic.

We now consider the case when $f(x) = K(y,x)$.

Lemma 3.2.4. *For (X,P) a Markov chain with n-th coordinate process Z_n, for any $x \in X$ and K the Martin kernel, the limit*

$$\lim_{n\to\infty} K(x, Z_n)$$

exists ${}_a$ *almost surely.*

Proof. Let $V \subset X$ be a finite subset of X containing the root, $a \in V$. Define the stopping time

$$e_V = \sup\{n : Z_n \in V\} \tag{3.1}$$

Now recalling that

$$K(x,y) = \frac{G(x,y)}{G(a,y)}$$

we calculate that

$$\begin{aligned}{}_a(K(x,Z_{e_V})) &= \sum_{y\in V} K(x,y)\ {}_a[Z_{e_V} = y]\\ &= \sum_{y\in V} K(x,y)G(a,y)\ {}_y[e_V = 0]\\ &= \sum_{y\in V} G(x,y)\ {}_y[e_V = 0]\\ &= \sum_{y\in V}\ {}_x[Z_{e_V} = y] \leq 1\end{aligned}$$

Then taking a sequence of finite exhausting subsets $V_k \nearrow X$, $(V_k \subset V_{k+1}, \cup_k V_k = X)$ and seeing that $\lim_{k\to\infty} e_{V_k} = \infty$, we can apply corol-

lary 3.2.3 to obtain that $\lim_{n\to\infty} K(x, Z_n)$ converges $_a$ almost surely for each $x \in X$. □

Using lemma 3.2.4, we can prove the first important theorem of convergence to the Martin boundary. Here again, Z_n is the n-th coordinate process with trajectory space $X^{\,0}$ and associated sigma algebra $\mathcal{B}$. We first recall the traditional boundary convergence theorem to emphasize the change when one considers completely reducible Markov chains.

Theorem 3.2.5 (Boundary convergence theorem, irreducible case)**.** *If (X, P) is an irreducible, transient Markov chain with n-th coordinate process Z_n, then there is a random variable Z_∞ taking its values in the Martin boundary $\mathcal{M}$ such that for each $x \in X$,*

$$\lim_{n\to\infty} Z_n = Z_\infty$$

$_x$ *almost surely in the topology of the Martin compactification $\hat{X}(P)$.*

Proof. See [114, Theorem 7.19] □

Theorem 3.2.6 (Boundary convergence theorem, completely reducible case)**.** *If (X, P) is a completely reducible , transient Markov chain with n-th coordinate process Z_n, then there is a random variable Z_∞ taking its values in the Martin boundary $\mathcal{M}$ such that for each $x \in X$ satisfying $a \ll x$,*

$$\lim_{n\to\infty} Z_n = Z_\infty$$

$_x$ *almost surely in the topology of the Martin compactification $\hat{X}(P)$.*

Proof. The proof consists of three parts, we must show

- a) For

  $$\Omega_\infty = \left\{\omega = (x_n) \in X^{\,0} : \exists x_\infty \in \mathcal{M} \; s.t. \; x_n \to x_\infty \text{ in the topologyof } \hat{X}(P)\right\},$$

  Ω_∞ belongs to the sigma algebra $\mathcal{B}$.

- b) For each $x \in X$, $_x(\Omega_\infty) = 1$

- c) $Z_\infty : \Omega_\infty \to \hat{X}(P)$ is measurable with respect to the Borel sigma algebra $\mathcal{B}$.

The proof of part a) consists of a standard measure-theoretic argument; we direct the interested reader to [114] for the detail. Briefly, a) is shown by writing Ω_∞ as an intersection of Borel sets, thus Ω_∞ belongs to the Borel sigma algebra $\mathcal{B}$.

Part b) proceeds as follows: using lemma 3.2.4, we have that

$$\lim_{n\to\infty} K(x, Z_n)$$

exists $\quad_a$ almost surely for each $x \in X$. Thus $\quad_a(\Omega_\infty) = 1$. Usually at this point irreducibility would be used to show that this implies $\quad_x(\Omega_\infty) = 1$ for *any* $x \in X$. Here (X, P) is not assumed to be irreducible, so the choice of the root a is of utmost importance. Let $x \in X$ be arbitrary. Since for each $x \in X$, $a \ll x$, we have that there exists some $k \in$ such that $a \ll_k x$. So there exist states $y_1, \ldots, y_{k-1} \in X$ such that

$$P(a, y_1)P(y_1, y_2) \ldots P(y_{k-1}, x) > 0$$

Thus

$$\begin{aligned} P(a, y_1)P(y_1, y_2) \ldots P(y_{k-1}, x) \quad_x(\Omega/\Omega_\infty) & \\ &= P(a, y_1)P(y_1, y_2) \ldots P(y_{k-1}, x) \quad_x(C(x) \cap (\Omega/\Omega_\infty)) \\ &\leq \quad_a(\Omega/\Omega_\infty) \\ &= 0 \end{aligned}$$

Thus it must be that $\quad_x(\Omega/\Omega_\infty) = 0$ hence $\quad_x(\Omega_\infty) = 1$.

For part c), let Z_∞ be the $\quad_x$ a.s. random variable guaranteed by part b) above combined with lemma 3.2.4. We now show that $Z_\infty : \Omega_\infty \to \hat{X}(P)$ is measurable with respect to the Borel sigma algebra of $\hat{X}(P)$. The argument is similar to part a) above but we provide a few more details here. First, it is easy to see that by the construction of the Martin boundary, a subbasis for

the sigma algebra is given by the sets

$$B_{x,\chi,\epsilon} := \{\nu \in \hat{X}(P) : \ |K(x,\nu) - K(x,\chi)| < \epsilon\}$$

So part c) will be finished if we show that the set $[Z_\infty \in B_{x,\chi,\epsilon}]$ is measurable for each x, χ, ϵ. In the following calculation $[y_1, \ldots, y_n] = \{(x_n) \in \Omega_\infty |\ x_1 = y_1, \ldots, x_n = y_n\}$. Thus

$$\begin{aligned}
&[Z_\infty \in B_{x,\chi,\epsilon}] \\
&= \{\omega = (x_n) \in \Omega_\infty : \ |\lim_{n\to\infty} K(x,x_n) - c| < \epsilon\} \\
&= \{(x_n) \in \Omega_\infty : D_\downarrow(K(x,x_n)|[-\epsilon + c, \epsilon + c]) < \infty\} \\
&= \bigcup_k \bigcap_{l,m\geq k} \{(x_n) : K(x,x_l) \geq b\} \cap \{(x_n) : K(x,x_m) \leq a\} \\
&= \bigcup_k \bigcap_{l,m\geq k} \left(\bigcup_{y_1,\ldots,y_l : K(x,y_l)\geq b} [y_1, \ldots, y_l] \ \cap \bigcup_{y_1,\ldots,y_m : K(x,y_m)\leq a} [y_1, \ldots, y_m] \right) \\
&\in \mathcal{B}
\end{aligned}$$

□

Now that we have verified theorem 3.2.6 in the completely reducible case, the proof of the following classical Martin boundary theorem holds just as in the irreducible case. In the following, we let $\nu_x(B) := \ _x[Z_\infty \in B]$.

Theorem 3.2.7. *For $x \in X$ satisfying $a \ll x$, the measure ν_x is absolutely continuous with respect to ν_a and a realization of the Radon-Nikodym derivative is given by*

$$\frac{d\nu_x}{d\nu_a} = K(x, \cdot)$$

Proof. See [114, Theorem 7.42]. The proof relies on the boundary convergence theorem and we have shown that in the completely reducible case, this theorem remains intact as long as $a \ll x$. □

We now present the integral representation theorem. Note the difference between the following theorem (which holds in the completely reducible case) and the classical theorem which holds in the irreducible case. In particular,

since the minimum principle (lemma 3.1.11) for a reducible Markov chain does not guarantee that a superharmonic function is strictly positive, not all superharmonic functions can be represented as an integral over the Martin boundary in the completely reducible case.

Theorem 3.2.8 (Integral representation, irreducible case)**.** *Let (X, P) be an irreducible, transient Markov chain with Martin compactification $\hat{X}$ and Martin boundary $\mathcal{M}$. Then for every non-negative superharmonic function h, there is a Borel measure ν^h on $\hat{X}$ such that for every $x \in X$,*

$$h(x) = \int_{\hat{X}} K(x, \cdot) d\nu^h$$

If h is harmonic, then $\text{supp}(\nu^h) \subset \mathcal{M}$.

Proof. See [114, Theorem 7.45] □

The proof involves the introduction of an h-process where the Martin kernel is modified to be $K_h(x, y) = K(x, y)\frac{h(a)}{h(x)}$. The minimum principle for the superharmonic function h on an irreducible Markov chain implies that if h is non-constant (the only non-trivial case), then non-negativity actually implies strict positivity. Thus $h(x) > 0$ for all $x \in X$ if h is non-constant. The minimum principle for a *completely reducible* Markov chain only guarantees the strict positivity of h on sets $D(x)$ where h does not achieve its minimum on x. Thus the proper statement of the theorem is as follows.

Theorem 3.2.9 (Integral representation, completely reducible case)**.** *Let (X, P) be a completely reducible transient Markov chain with Martin compactification $\hat{X}$ and Martin boundary $\mathcal{M}$. Then for every strictly positive, superharmonic function h, there is a Borel measure ν^h on $\hat{X}$ such that for every $x \in X$,*

$$h(x) = \int_{\hat{X}} K(x, \cdot) d\nu^h$$

If h is harmonic, then $\text{supp}(\nu^h) \subset \mathcal{M}$.

Proof. The following proof follows the exposition of [114] regarding h-processes. Let

$$P_h(x, y) = \frac{P(x, y)h(y)}{h(x)}$$

hence

$$K_h(x,y) = \frac{K(x,y)h(y)}{h(x)}$$

Now clearly $\lim_{n\to\infty} K(x,x_n)$ converges if and only if $\lim_{n\to\infty} K_h(x,x_n)$ (here we use strict positivity of h). Now let $\tilde{\nu_x}(B) := {}^h_x[Z_\infty \in B]$. Then by theorem 3.2.7, setting $B = \hat{X}(P)$, we have $K_h(x,\cdot) = d\tilde{\nu}_x/d\tilde{\nu}_a$ so

$$1 = \tilde{\nu}_x(\hat{X}(P) = \int_{\hat{X}(P)} K_h(x,\xi)d\tilde{\nu}_a(\xi)$$

Then set $\nu^h = h(a)\tilde{\nu}_a$ and note that lemma 3.1.11 implies that if h is not identically zero, then $h(a) \neq 0$. Calculating

$$\begin{aligned} h(x) = h(x)\tilde{\nu}_x(\hat{X}(P)) &= \int_{\hat{X}(P)} h(x)K_h(x,\xi)d\tilde{\nu}_a(\xi) && (3.2)\\ &= \int_{\hat{X}(P)} K(x,\xi)h(a)d\tilde{\nu}_a(\xi) && (3.3)\\ &= \int_{\hat{X}(P)} K(x,\xi)d\nu^h(\xi) && (3.4) \end{aligned}$$

As for the second part of the theorem, assuming that h is harmonic suppose that $y \in \text{supp}(\nu^h)$ such that $y \in X$. Now let $n = \nu^h(y)$ and note that since X is discrete in $\hat{X}(P)$, $n > 0$. Thus we can decompose $\nu^h = n\delta_y + \nu'$ with $\text{supp}(\nu') \subset \hat{X}(P)/y$. So by 3.2 above, $h(x) = nK(x,y) + h'(x)$ where $h'(x) = \int_{\hat{X}(P)} K(x,\xi)d\nu'(\xi)$. But $K(\cdot,y)$ is strictly superharmonic and $h(x)$ is superharmonic, thus $h(x)$ is strictly superharmonic, a contradiction. Thus $\text{supp}(\nu^h) \subset \hat{X}(P)/X$.

□

3.3 Poisson boundary

By theorem 3.2.9, there exists a Borel measure ν^1 associated to the strictly positive harmonic constant function 1.

Definition 3.3.1 (Poisson boundary)**.** *For the transient Markov chain (X,P) with Martin boundary $\mathcal{M}$, the Poisson boundary is defined to be the measure*

space $(\mathcal{M}, \nu^1)$.

It should be noted that the Poisson boundary is a measure space, and so two models of the Poisson boundary are identified when they are isomorphic as measure spaces. In contrast, even though the Martin boundary is always metrizable (via Urysohn's metrization theorem), it is primarily a topological space and so two representations of the Martin boundary are identified when they are homeomorphic as topological spaces.

The Poisson boundary is considered to be trivial if the measure $\nu^1 = \delta_x$ is a point mass.

Corollary 3.3.2. *For a transient Markov chain* (X, P)*, the following statements are equivalent*

1. *The Poisson boundary is trivial.*
2. *All bounded harmonic functions are constant (the weak Liouville property)*
3. *The constant function* $\mathbf{1}$ *is minimal harmonic.*
4. *There is a* $\xi \in \mathcal{M}$ *such that* $X_n \to \xi \quad {}_x$*-almost surely for every* x*.*

The Poisson boundary is easy to determine for many SMC's due to the special structure of the transition operator P.

Theorem 3.3.3. *The Poisson boundary of an expansive substitution Markov chain is trivial.*

Proof. We utilize theorem 2.8 from [50] which states the following: For a Markov operator P with initial distribution m, the Poisson boundary is trivial if and only if for any two probabilities $\nu_1, \nu_2 \prec m$,

$$\lim_{n\to\infty} \frac{1}{n} || \sum_{i=1}^{n} (\nu_1 P^n - \nu_2 P^n) || = 0 \tag{3.5}$$

Here, the norm $||\cdot||$ is the total variation norm.

Endowing the basis $X \times X$ with the order inherited by the relation "$\ll$" given in definition 3.1.3, the transition operator P is strictly upper-triangular.

Since the initial distributions m we are considering are point-masses $m = \delta_a$, then for sufficiently large n, $\nu_1 P^n = \nu_2 P^n = 0$ for $\nu_1, \nu_2 \prec m$. So $\lim_{n\to\infty} || \sum_{i=1}^{n} (\nu_1 P^n - \nu_2 P^n)||$ is a constant, and hence equation (3.5) is satisfied, and so the Poisson boundary is trivial. □

Thus in a reducible Markov chain, if a superharmonic function attains its minimum at some state x, then it's constant on $D(x)$. This allows us to determine necessary conditions for the convergence of $\lim_n K(\cdot, x_n)$.

Proposition 3.3.4. *For $K(\cdot,\cdot)$ the Martin kernel, for each $y \in X$, $\lim_{n\to\infty} K(y, x_n)$ exists and is not identically zero in the topology of the Martin compactification only if for all but finitely many n, $x_n \in D(y)$ or else $x_n \notin D(y)$.*

Proof. Arguing by contradiction, say there is a state $y \in X$ such that $x_n \in D(y)$ infinitely often and $x_n \notin D(y)$ infinitely often. Then $K(y, x_n) \neq 0$ for $x_n \in D(y)$ and $K(y, x_n) = 0$ for $x_n \notin D(y)$, hence $K(y, x_n)$ cannot converge, a contradiction to our assumption. □

Since the Martin boundary can be viewed as equivalence classes of θ-Cauchy sequences $(x_n)_{n\geq 0}$, proposition 3.3.4 indicates in which direction to look for such sequences.

3.4 Identifying the Martin boundary

We take the following strategy in identifying the Martin boundary: identify a particularly "nice" sequence (y_n) in each of the equivalence classes mentioned at the end of the previous section. We then identify to what space this collection (equipped with the extended metric also denoted as θ) is homeomorphic to.

3.4.1 Martin boundary of the SMC Σ_{eg_1}

Throughout this subsection, let Σ_{eg_1} be the substitution Markov chain defined by

$$\Sigma_{eg_1} : \begin{cases} a \to \begin{cases} ab \text{ with prob. } 1/2 \\ ba \text{ with prob. } 1/2 \end{cases} \\ \\ b \to b \end{cases}$$

We first note that each word generable by Σ_{eg_1} is of the form $\overbrace{b\cdots b}^{n} a \overbrace{b\cdots b}^{m}$ for $n, m \geq 0$ and both words ba and ab connect to the word bab in one application of Σ_{eg_1}. Thus the Markov chain (X, P) to which Σ_{eg_1} is associated is simply the nearest neighbor random walk on the integer lattice $X \cong \overset{2}{+}$ with the restriction that one can only move north or east: for a, b, c, d natural numbers,

$$P((a,b),(c,d)) = \begin{cases} 1/2 & c = a+1 \\ 1/2 & d = b+1 \\ 0 & \text{otherwise} \end{cases}$$

We are now in a position to compute the Martin kernel. First, let $x \in X$ and $(y_n)_{n\geq 0} \in X$, we use the identification that $x = (x_1, x_2)$ and $y_n = (y_1^{(n)}, y_2^{(n)})$. Then

$$K(x, y_n) = \begin{cases} \frac{G(x,y_n)}{G(a,y_n)} & x \ll y_n \\ 0 & x \not\ll y_n \end{cases}$$

So in the case that $x \ll y_n$, suppressing the dependence of $y_1^{(n)}$ and $y_2^{(n)}$ on n, and considering the case where $y_1 y_2 = 0$ later, we have that

$$\frac{G(x,y_n)}{G(a,y_n)} = \frac{2^{-y_1-y_2+x_1+x_2}\binom{y_1+y_2-x_1-x_2}{y_1-x_1}}{2^{-y_1-y_2}\binom{y_1+y_2}{y_1}} \tag{3.6}$$

$$= \frac{2^{x_1+x_2}(y_1+y_2-x_1-x_2)!y_1!y_2!}{(y_1-x_1)!(y_2-x_2)!} \tag{3.7}$$

$$= \frac{2^{x_1+x_2}y_1^{x_1}y_2^{x_2}(1-\frac{1}{y_1})\cdots(1-\frac{x_1-1}{y_1})(1-\frac{1}{y_2})\cdots(1-\frac{x_2-1}{y_2})}{(y_1+y_2)^{x_1+x_2}(1-\frac{1}{y_1+y_2})\cdots(1-\frac{x_1+x_2-1}{y_1+y_2})} \tag{3.8}$$

Examining (3), if one of y_1, y_2 is bounded and the other tends to infinity as

$n \to \infty$, then for c a constant depending on x, if we take a smooth interpolation of the functions $y_1^{(n)}$ and $y_2^{(n)}$, upon applying L'Hospital's rule $\max(x_1, x_2)$ times, (3) becomes

$$\frac{G(x, y_n)}{G(a, y_n)} = O(c)\frac{1}{(y_1 + y_2)^{x_2}} \to 0$$

We can thus conclude by the characterization of θ-Cauchy sequences

Proposition If $(y_n)_{n\geq 0} = (y_1^{(n)}, y_2^{(n)})_{n\geq 0}$, $|y_n| \to \infty$ is to be θ-Cauchy and $\lim_n K(\cdot, y_n)$ is not identically zero, then both $|y_1^{(n)}| \to \infty$ and $|y_2^{(n)}| \to \infty$ as $n \to \infty$.

By this proposition, we can then finish the calculation of $K(x, y_n)$. Assuming that both $|y_1|, |y_2| \to \infty$, we have by (3) that

$$\lim_n K(x, y_n) = \lim_n \frac{2^{x_1+x_2} y_1^{x_1} y_2^{x_2}}{(y_1 + y_2)^{x_1+x_2}} \tag{3.9}$$

$$= \lim_n 2^{x_1+x_2} \left(\frac{y_1}{y_1 + y_2}\right)^{x_1} \left(1 - \frac{y_1}{y_1 + y_2}\right)^{x_2} \tag{3.10}$$

We now turn our attention to determining the Martin Boundary. Thus we attempt to characterize all θ-Cauchy sequences (y_n) such that $|y_n| \to \infty$. By the above proposition, we need only consider those $y_n = (y_1^{(n)}, y_2^{(n)})$ such that both $|y_1^{(n)}|, |y_2^{(n)}| \to \infty$. Given such a sequence, we now construct "nicer" and yet equivalent sequence: we want to show

Proposition 3.4.2. *If (y_n) is θ-Cauchy, we can construct a sequence (y'_n) equivalent to (y_n) (that is, $\lim_{n,m} \theta(y_n, y'_m) = 0$) such that $y'_n \ll_1 y'_{n+1}$ for all n.*

Proof. By proposition 3.1.25, we have that for each $i > 0$, $y_n \ll y_{n+i}$. Let $y'_1 = y_1$. Since $y_1 \ll y_2$, there are x_i's such that

$$y_1 \ll_1 x_1 \ll_1 x_2 \cdots \ll_1 x_t \ll_1 y_2$$

Letting $y'_i = x_i$ for $i = 1, \cdots, t$ and proceeding in this fashion, we obtain a sequence (y'_n) that is clearly θ-Cauchy for which $y'_n \ll_1 y'_{n+1}$ and such that

$\lim_{n,m} \theta(y_n, y'_m) = 0$ and the proposition is proved. □

Using this proposition, we can now rewrite (3.10) utilizing a sequence (y'_n) such that $y'_n \ll_1 y'_{n+1}$. Note that then for $y'_n = (y'_1, y'_2)$, we have $y'_1 + y'_2 = |y'_n|$ where by $|x|$ we mean the integer t such that $a \ll_t x$. By adding on finitely many terms in the beginning of the sequence (y'_n), we can assume that $y'_0 = a = (0,0)$ and so $|y'_n| = n$. Thus

$$\lim_n K(x, y_n) = \lim_n K(x, y'_n) \tag{3.11}$$

$$= 2^{x_1+x_2} \lim_n \left(\frac{y'_1}{n}\right)^{x_1} \left(1 - \frac{y'_1}{n}\right)^{x_2} \tag{3.12}$$

We can now easily take care of the case when for $(y_n)_{n\geq 0}$, $y_n = (y_1, y_2)$ (suppressing the dependence on n), $y_1 y_2$=0. Since then, by the above proposition, we construct an equivalent sequence y'_n all related in one step. Thus,

$$\lim_n K(x, y_n) = \begin{cases} \lim_n K(x, y'_n) & x \ll y'_n \\ 0 & x \not\ll y'_n \end{cases} \tag{3.13}$$

$$= \begin{cases} \lim_n \frac{\binom{n-x_1-x_2}{n-x_1}}{\binom{n}{n}} & x \ll y'_n \\ 0 & x \not\ll y'_n \end{cases} \tag{3.14}$$

$$= \begin{cases} \lim_n \frac{(n-x_1-x_2)\cdots(n-x_1+1)}{x_2!} & x \ll y'_n \\ 0 & x \not\ll y'_n \end{cases} \tag{3.15}$$

$$= \begin{cases} \infty & x \ll y'_n \\ 0 & x \not\ll y'_n \end{cases} \tag{3.16}$$

And so if (y_n) is to be θ-Cauchy, then $\lim_n K(\cdot, y_n)$ must be identically zero.

Summarizing the last few pages, we have successfully identified a representative $(y^c_n)_{n\geq 0}$ θ-Cauchy sequence in each equivalence class of Cauchy sequences: for $c \in [0,1]$ an independent constant let $y^c_n = (\lfloor cn \rfloor, \lfloor (1-c)n \rfloor)$, then by the above calculations and the squeeze theorem

$$\lim_n K(x, y^c_n) = 2^{x_1+x_2} c^{x_1} (1-c)^{x_2} \tag{3.17}$$

We are now in a position to prove the following

Theorem 3.4.3. *The Martin Boundary of the substitution Markov chain Σ_{eg_1} is homeomorphic to the unit interval $[0,1]$ equipped with the standard metric.*

Proof. Let $F : [0,1] \to M(P)$ be the map from the unit interval to the Martin boundary (equipped with the extended θ metric, also denoted by θ) defined by $F(c) = [(y_n)_{n\geq 0}]$ where c is mapped to the equivalence class of θ-Cauchy sequences of elements of X such that $(y_n) \sim (y_n^c)$ where $y_n^c = (cn, (1-c)n)$. Such a map is well defined by all of our previous propositions (see (12) and the comments regarding it). Now to show that such a map is continuous, observe that for $(y_n)_{n\geq 0}$ and $(y'_n)_{n\geq 0}$ in the same equivalence class of θ-Cauchy sequences as $(y_n^{c_1})$ and $(y_n^{c_2})$ respectively,

$$\theta((y_n),(y'_n)) = \sum_{x\in X} w_x \left(|K(x,(y_n)) - K(x,(y'_n))| + |\delta_x((y_n)) - \delta_x((y'_n))|\right) \tag{3.18}$$

$$= \sum_{x\in X} w_x \left(|2^{x_1+x_2} c_1^{x_1}(1-c_1)^{x_2} - 2^{x_1+x_2} c_2^{x_1}(1-c_2)^{x_2}|\right) \tag{3.19}$$

$$\leq |c_1(1-c_1) - c_2(1-c_2)| \tag{3.20}$$

$$\leq |c_1 - c_2|(1 + c_1 + c_2) \tag{3.21}$$

$$< 3|c_1 - c_2| \tag{3.22}$$

Thus for any given ϵ, for $\delta \leq \epsilon/3$, we have that $|c_1 - c_2| < \delta$, then $\theta(F(c_1), F(c_2)) < \epsilon$. Thus F is continuous. Now by 3.17 bijectivity is clear. Also, since $[0,1]$ is compact, and $M(P)$ is Hausdorff, elementary topology arguments give us that F is a homeomorphism. □

3.4.3.1 Harmonic measure and Poisson boundary

In our case, for $x = (x_1, x_2)$ and $\xi \in M$ represented by the Cauchy sequence $y_n = (cn, (1-c)n)$, we have

$$\begin{aligned} 1 &= \int_M K(x,\xi) d\nu^1(\xi) \\ &= \int_0^1 2^{x_1+x_2} c^{x_1}(1-c)^{x_2} d\nu^1(c) \end{aligned}$$

$$= 2^{x_1+x_2}\frac{x_1!x_2!}{(x_1+x_2+1)!}$$

So we see that $\nu^1 = \delta_{1/2}$ the point mass at $1/2$ in the representation of the Martin boundary as the unit interval. Thus the Poisson boundary is trivial.

3.4.4 Martin boundary when the substitution Markov chain is a tree, Σ_{eg_2}

We calculate the Martin boundary of another class of SMC's: those having the structure of a tree. It is tempting to apply the work of [114] and [113], however lack of irreducibility prevents this. Furthermore, the following examples emphasize how complete reducibility makes the calculation of the Martin boundary easier in general.

We aim to show that the Martin Boundary associated to the substitution Markov chain

$$\Sigma_{eg_2}: \begin{cases} a \to \begin{cases} aaa \text{ with prob. } 1/2 \\ aba \text{ with prob. } 1/2 \end{cases} \\ \\ b \to \begin{cases} bbb \text{ with prob. } 1/2 \\ bab \text{ with prob. } 1/2 \end{cases} \end{cases}$$

is homeomorphic to a specific Cantor set.

To that end, let (X, P) be the Markov chain associated to Σ_{eg_2} with unit mass on the letter a. Note that the state space will then be $X = \{\Sigma^n_{eg_2}(a) : n \in \ \}$, all words that can be reaching via applying the substitution Markov chain to the letter a. For $x, y \in X$, we write $x \ll_n y$ if $\Sigma^n_{eg_2}(x) = y$. Also, $x \ll y$ if $\exists n \geq 0$ such that $x \ll_n y$. For $x \in X$, we use $|x|$ to denote the integer n such that $a \ll_n x$. An important fact that we will refer to frequently is that due to the combinatorial structure of Σ_{eg_2}

Proposition 3.4.5. *The Markov chain (X, P) associated to Σ_{eg_2} is a **tree**: For each $y \in X$ there exists a unique $x \in X$ such that $\Sigma_{eg_2}(x) = y$.*

Proof. The proof relies on the fact that $\{aaa, aba\} \cap \{bbb, bab\} = \emptyset$ and the fact that the substitution is constant length, and so we may "desubstitute"

by canceling on the left. Let $y = y_1y_2y_3\cdots \in X$, thus by definition of X, there exists some n such that $\Sigma^n_{eg_2}(a) = y$. We inductively construct the desired unique $x = x_1x_2x_3\ldots$ such that $\Sigma_{eg_2}(x) = y$. For the base case, since $\Sigma^n_{eg_2}(a) = y$ and $\Sigma_{eg_2}(a) = aaa$ or else $\Sigma_{eg_2}(a) = aba$, then $y = aaay_3y_4\ldots$ or else $y = abay_3y_4\ldots$. Accordingly, let $x_1 = a$. Before the induction step, we observe that since Σ_{eg_2} is a constant length substitution (constant of length 3), then for each triplet $y_{3k+1}y_{3k+2}y_{3k+3}$, $k \in \ ^+$, it must be that $\Sigma_{eg_2}(a) = y_{3k+1}y_{3k+2}y_{3k+3}$ or else $\Sigma_{eg_2}(b) = y_{3k+1}y_{3k+2}y_{3k+3}$. Now for the induction step, say we have determined $x_1, \ldots, x_n$. Considering $y_{3n+1}y_{3n+2}y_{3n+3}$, if $y_{3n+1}y_{3n+2}y_{3n+3} \in \{aaa, aba\}$ then it must be that $\Sigma_{eg_2}(a) = y_{3n+1}y_{3n+2}y_{3n+3}$ and so $x_{n+1} = a$ uniquely. If $y_{3n+1}y_{3n+2}y_{3n+3} \in \{bbb, bab\}$ then it must be that $\Sigma_{eg_2}(b) = y_{3n+1}y_{3n+2}y_{3n+3}$ and so $x_{n+1} = b$ uniquely. Having uniquely determined x_{n+1} the proof is complete by induction. □

Corollary 3.4.6. *If $x \ll y$, then there exists a unique $n \in$ such that $x \ll_n y$.*

Proof. Let $x, y \in X$ such that $x \ll y$. By definition, there exists an n such that $\Sigma^n_{eg_2}(x) = y$. By applying proposition 3.4.5 n times to y, we see that such an n is unique: by proposition 3.4.5, $\exists! z_1, z_2, \ldots z_n$ such that $\Sigma_{eg_2}(z_1) = \Sigma^2_{eg_2}(z_2) = \cdots = \Sigma^n_{eg_2}(z_n) = y$. Since the z_i are unique, then it must be that $z_n = x$ and n is unique too. □

Corollary 3.4.7. *For $x \ll_n y$, there exist $n-1$ unique states $x_1, \ldots x_{n-1}$ such that $x \ll_1 x_1 \ll_1 \cdots \ll_1 x_{n-1} \ll_1 y$.*

Proof. This is simply applying proposition 3.4.5 $n-1$ times to y. □

Using the n-step transition probabilities $P^n(x, y)$ and taking advantage of corollary 3.4.6, we calculate the Green's function.

Proposition 3.4.8.

$$G(x,y) = \sum_{i=0}^{\infty} P^i(x,y) = \begin{cases} P^n(x,y) & \text{if } x \ll_n y \\ 0 & \text{o.w.} \end{cases}$$

Proof. Let $x, y \in X, x \ll y$, then by corollary 3.4.6, we have that there exists a unique n such that $\Sigma^n_{eg_2}(x) = y$. Thus $G(x, y) = \sum_{i=0}^{\infty} P^i(x, y) = P^n(x, y)$. If $x \not\ll y$, then $G(x, y) = 0$ clearly. □

The transition probabilities enjoy nice properties due to the tree structure of the Markov chain.

Proposition 3.4.9. *For $x \ll_m y \ll_n z$, $P^{m+n}(x,z) = P^m(x,y)P^n(y,z)$*

Proof. By corollary 3.4.7, there exist unique states $x_1, \ldots, x_{m-1}, y_1, \ldots, y_{n-1}$ such that $x \ll_1 x_1 \ll_1 \cdots \ll_1 x_{m-1} \ll_1 y =: x_m \ll_1 \cdots \ll_1 x_{m+n-1} \ll_1 z$. Thus

$$\begin{aligned} P^{m+n}(x,z) &= \sum_{w_1,\ldots w_{m+n-1}} P(x,w_1)P(w_1,w_2)\ldots P(w_{m+n-1},z) \\ &= P(x,x_1)P(x_1,x_2)\ldots P(x_{m+n-1},z) \\ &= \sum_{w_1,\ldots,w_{m-1}} P(x,w_1)\ldots P(w_{m-1},y) \\ &\qquad \times \sum_{w_{m+1},\ldots,w_{n+m-1}} P(y,w_{m+1})\ldots P(w_{m+n-1},z) \\ &= P^m(x,y)P^n(y,z) \end{aligned}$$

□

We also compute the Martin kernel:

Proposition 3.4.10.

$$k(x,y) = \begin{cases} \frac{1}{G(a,x)} & \text{if } x \ll y \\ 0 & o.w. \end{cases}$$

Proof. By definition, $k(x,y) = \frac{G(x,y)}{G(a,y)}$. Thus using propositions 3.4.8 and 3.4.9, we calculate that if $a \ll_m x \ll_n y$,

$$k(x,y) = \frac{P^n(x,y)}{P^{n+m}(a,y)} = \frac{P^n(x,y)}{P^m(a,x)P^n(x,y)} = \frac{1}{P^m(a,x)} = \frac{1}{G(a,x)}$$

And $k(x,y) = 0$ otherwise. □

Note that due to the structure of the Markov Chain in question (each word has a single ancestor), for any $x \ll_n y$, there exists a unique *geodesic* $\pi(x,y)$ connecting x to y.

Definition 3.4.11. *Let $\pi(x,y) = [x = x_0, x_1, \ldots, x_{n-1}, x_n = y]$ be the unique sequence of vertices guaranteed by corollary 3.4.7 such that $x \ll_1 x_1 \ll_1 \cdots \ll_1 y$.*

We now characterize the θ-Cauchy sequences, or equivalently by proposition 3.1.23, all sequences $(x_n)_{n\geq 0}$ such that for all y, $\lim_n k(y, x_n)$ converges.

Proposition 3.4.12. *A sequence $(x_n)_{n\geq 0}$ has the property that for all y, $\lim_n k(y, x_n)$ converges to a finite real number if and only if for all but finitely many indices i, $x_i \ll x_{i+1}$.*

Proof. Fix a sequence $(x_n)_{n\geq 0}$ and $y \in X$ and say $\lim_n k(y, x_n)$ converges to a finite real number. Then for the indices such that $y \ll x_n$, by proposition 3.4.10, $k(y, x_n) = 1/G(a, y) \neq 0$. Thus by proposition 3.3.4, it must be that for all but finitely many indices i, $x_i \ll x_{i+1}$. Now for the other direction say that for all but finitely many indices i, $x_i \ll x_{i+1}$ (note that this implies that $|x_n| \to \infty$). Thus there exists an N such that for all $n \geq N$, $x_n \ll x_{n+1}$. Then using proposition 3.4.10, for all $n \geq N$,

$$k(y, x_n) = \begin{cases} \frac{1}{G(a,y)} & y \ll x_N \\ 0 & \text{o.w.} \end{cases}$$

and so $\lim_n k(y, x_n)$ converges to a finite real number.

□

Now that a nice characterization of the θ-Cauchy sequences have been found, we exhibit a much simpler metric $\bar{\rho}$ that is uniformly equivalent to θ (and so the completion of X with respect to $\bar{\rho}$ is homeomorphic to its completion with respect to θ). First, we introduce some notation:

Definition 3.4.13. *For any $x, y \in X$, we define the confluent or most recent common ancestor $x \wedge y$ as the last common vertex in $\pi(a, x)$ and $\pi(a, y)$.*

The confluent exists and is unique by corollary 3.4.7. Now recall that for $x \in X$, we use $|x|$ to denote the integer n such that $a \ll_n x$, and let

$$\bar{\rho}(x, y) = \begin{cases} e^{-|x \wedge y|} & \text{if } x \neq y \\ 0 & \text{o.w.} \end{cases}$$

Proposition 3.4.14. *The function $\bar{\rho}$ is an ultrametric.*

Proof. The proof is classical, but we recall it here for completeness. For any $x, y \in X$, clearly $\bar{\rho}(x, y) \geq 0$ and $\bar{\rho}(x, y) = \bar{\rho}(y, x)$. By definition, $\bar{\rho}(x, y) = 0$ iff $x = y$. As for the strong triangle inequality: $\bar{\rho}(x, z) \leq \max\{\bar{\rho}(x, y), \bar{\rho}(y, z)$ we prove the classically equivalent 3-point condition: For any points a, b, c there exists a renaming such that $\bar{\rho}(a, b) \leq \bar{\rho}(b, c) = \bar{\rho}(a, c)$. To that end, without loss of generality assume that $\bar{\rho}(a, b) \leq \bar{\rho}(a, c)$. Now $\bar{\rho}(a, c) = \bar{\rho}(a, a \wedge c)$ and $\bar{\rho}(a, b) = \bar{\rho}(a, a \wedge b)$. But $a \wedge c \ll a$ and $a \wedge b \ll a$, thus it must be that $a \wedge c \ll a \wedge b$. Now a path from b to c starting a b is to go from b to $a \wedge b$ to $a \wedge c$ to c. Since X is a tree, this must be the only path. Thus $\bar{\rho}(b, c) = \bar{\rho}(b, a \wedge c) = \bar{\rho}(a, a \wedge c) = \bar{\rho}(a, c)$ and the result holds. □

Lemma 3.4.15. *The metrics θ and $\bar{\rho}$ are uniformly equivalent.*

Proof. We proceed by showing that the identity map $i : (X, \bar{\rho}) \to (X, \theta)$ and its inverse are uniformly continuous. First we show that the identity map $i : (X, \bar{\rho}) \to (X, \theta)$ is uniformly continuous. Let $x \neq y \in X$ be arbitrary. If for $x \neq y \neq z \in X$, $z \ll x$ and $z \ll y$, then by proposition 3.4.10, for w_z the weights defined in definition 3.1.19,

$$w_z(|k(z, x) - k(z, y)| + |\delta_z(x) - \delta_z(y)|) = w_z(|G(a, z) - G(a, z)| + |0 - 0|) = 0$$

Also by the same proposition, if $x \neq y \neq z \in X$, $z \not\ll x$ and $z \not\ll y$, then

$$w_z(|k(z, x) - k(z, y)| + |\delta_z(x) - \delta_z(y)|) = w_z(|0 - 0| + |0 - 0|) = 0$$

Thus the following equality is true

$$\begin{aligned}
&\sum_{z \in X} w_z(|k(z, x) - k(z, y)| + |\delta_z(x) - \delta_z(y)|) \\
&= \sum_{z \not\ll x, z \ll y} w_z(|k(z, y)| + |\delta_z(x) - \delta_z(y)|) \\
&\quad + \sum_{z \ll x, z \not\ll y} w_z(|k(z, x)| + |\delta_z(x) - \delta_z(y)|)
\end{aligned}$$

But those z such that $z \not\ll x$ and $z \ll y$ are precisely those $z \in \pi(x \wedge y, y)$

such that $z \neq x \wedge y$ and similarly for x, thus

$$\sum_{z\in X} w_z(|k(z,x)-k(z,y)|+|\delta_z(x)-\delta_z(y)|) \tag{3.23}$$

$$= \sum_{z\in\pi(x\wedge y,y),z\neq x\wedge y} w_z(|k(z,y)|+|\delta_z(y)|) + \sum_{z\pi(x\wedge y,x),z\neq x\wedge y} w_z(|k(z,x)|+|\delta_z(x)|) \tag{3.24}$$

$$= \sum_{z\in\pi(x\wedge y,y)} \frac{w_z}{G(a,z)} + \sum_{z\in\pi(x\wedge y,x)} \frac{w_z}{G(a,z)} + w_x + w_y \tag{3.25}$$

$$\leq 4 \sum_{|z|\geq|x\wedge y|} \frac{w_z}{G(a,z)} \tag{3.26}$$

Now let $\epsilon > 0$. Recall from definition 3.1.19 that $\sum_{z\in X} \frac{w_z}{G(a,z)} < \infty$ and so

$$\lim_{|x|\to\infty} \sum_{|z|>|x|} \frac{w_z}{G(a,z)} = 0$$

since this is a tail of a convergent series. So there exists an $N > 0$ such that for all $|x| > N$,

$$\lim_{|x|\to\infty} \sum_{|z|>|x|} \frac{w_z}{G(a,z)} < \epsilon/4.$$

Let $\delta < e^{-N}$. Thus for $\bar{\rho}(x,y) < \delta$, $|x \wedge y| > N$ and by (3.26),

$$\theta(x,y) \leq 4 \sum_{|z|\geq|x\wedge y|} \frac{w_z}{G(a,z)} < 4\epsilon/4 = \epsilon$$

and we have shown uniform continuity.

Showing that the inverse $i^{-1} : (X,\theta) \to (X,\bar{\rho})$ is uniformly continuous proceeds similarly: by equation 3.26 above, $\frac{4w_{|x\wedge y|}}{G(a,x\wedge y)} \leq \theta(x,y)$. Now since the w_z were arbitrarily chosen and $\sum_{z\in X} \frac{w_z}{G(a,z)} < \infty$, we can chose the w_z such that $\frac{4w_z}{G(a,z)}$ decreases to zero monotonically in $|z|$. Thus $\frac{4w_z}{G(a,z)} < \delta$ implies that $|z| \geq N$. Now let $\epsilon > 0$ be chosen. Then pick δ such that $N > -\ln\epsilon$. Thus for all x, y such that $\frac{4w_{|x\wedge y|}}{G(a,x\wedge y)} \leq \theta(x,y) < \delta$, we have that $|x \wedge y| \geq -\ln\epsilon$ and so $\bar{\rho}(x,y) < \epsilon$.

□

We now elaborate on proposition 3.4.12 by showing the following proposition

Proposition 3.4.16. *For a sequence $(x_n)_{n\geq 0}$ such that for every $y \in X$, $\lim_n k(y, x_n)$ converges, there exists a sequence $(x'_n)_{n\geq 0}$ such that $x'_n \ll_1 x_{n+1}$ and $\lim_n \theta(x_n, x'_n) = 0$.*

Proof. We construct $(x'_n)_{n\geq 0}$ inductively: By proposition 3.4.12, there exists an N such that for all $n \geq N$, $x_n \ll x_{n+1}$. Let $x'_N := x_N$. Then for $i > N$,let

$$x'_{|\pi(x_{i-1},x_i)|+1}, x'_{|\pi(x_{i-1},x_i)|+2}, \ldots, x'_{|\pi(x_i,x_{i+1})|} \equiv x_{i+1}$$

be all but the first vertices listed in $\pi(x_i, x_{i+1})$. Then define $x'_0, \ldots, x'_{N-1}$ to be all but the first vertices listed in $\pi(a, x_N)$. By construction we have that $x'_n \ll_1 x_{n+1}$ as desired. Now by construction, we have that for all but finitely many $z \in X$, there exists an M such that for all $n \geq M$, both $x_n, x'_n \in D(z)$. Thus $z \ll x_n \wedge x'_n$ and so $|x_n \wedge x'_n| > |z|$. Since this is true for all but finitely many z, taking $|z| \to \infty$ it must be that $|x_n \wedge x'_n| \to \infty$. Then just as in the proof of lemma 3.4.15, we have

$$\begin{aligned}\theta(x_n, x'_n) &= \sum_{z\in\pi(x_n\wedge x'_n, x'_n)} \frac{w_z}{G(a,z)} + \sum_{z\in\pi(x_n\wedge x'_n, x_n)} \frac{w_z}{G(a,z)} + w_{x'_n} + w_{x_n} \\ &\leq 4 \sum_{|z|\geq|x_n\wedge x'_n|} \frac{w_z}{G(a,z)} \xrightarrow[|x_n\wedge x'_n|\to\infty]{} 0\end{aligned}$$

Where the limit goes to zero in the last step since it is the tail of the convergent series $\sum_z w_z/G(a,z)$. □

As an intermediary step to showing that the Martin boundary is homeomorphic to a Cantor set, we first show that it is isomorphic to a particular sequence space.

Proposition 3.4.17. *The Martin boundary $\hat{X}(P)/X$ is isomorphic to the space (S, ρ) where $S = \{(a = x_0, x_1, \ldots) \text{ s.t. } \forall i, x_i \ll_1 x_{i+1}\}$ and ρ is the metric defined for $x, y \in S$ by $\rho(x, y) = \lim_{n\to\infty} \bar{\rho}(x_n, y_n)$.*

Proof. This is simply by definition of the Martin boundary: by definition 3.1.24 and proposition 3.1.23 we have that $\mathcal{M}$ is defined to be the space of

all θ-equivalence classes of Cauchy sequences $(x_n)_{n\geq 0}$ such that $|x_n| \xrightarrow[n\to\infty]{} \infty$. Proposition 3.4.15 proved that this space is isomorphic to all $\bar{\rho}$-equivalence classes of Cauchy sequences $(x_n)_{n\geq 0}$ such that $|x_n| \xrightarrow[n\to\infty]{} \infty$. Propositions 3.4.16 and 3.4.12 have shown that we can choose as representatives of equivalence classes sequences $(x_n)_{n\geq 0}$ such that for all n, $x_n \ll_1 x_{n+1}$ (these are automatically $\bar{\rho}$-Cauchy as seen in the proof of proposition 3.4.16). Thus, by definition of the completion as found in any elementary real analysis text, for ρ the extended metric, $\mathcal{M}$ is isomorphic to the space (S, ρ).

□

In preparation for showing that the Martin boundary of our substitution Markov chain Σ_{eg_2} is homeomorphic to a Cantor set, we first construct the Cantor set $\mathcal{C}$. Let $N(t) := 2^{\frac{3^t-1}{2}} = |\{x \text{ s.t. } \Sigma^n_{eg_2}(a) = x\}|$ be the number of words at step n. Let $Len(n)$ denote the "length of a subinterval" given inductively by $Len(0) = 1$, $Len(n) = \frac{Len(n-1)}{2N(n)-1}$. Let $I_0 = [0, 1]$, then inductively define

$$I_n = Len(n) * \left(\bigcup_{k=0}^{\infty} [2k, 2k+1] \right) \bigcap I_{n-1}$$

As constructed, each set I_n consists of subintervals in bijective correspondence to the number of distinct words $w = \Sigma^n_{eg_2}(a)$. Reference figure 3.4.4 for a visual representation of the sets I_n.

Then finally we define the Cantor set $\mathcal{C} : \bigcap_n I_n$. Note that the construction of $\mathcal{C}$ can be described as follows: Each set I_n consists of subintervals in bijective correspondence to the number of distinct words $w = \Sigma^n_{eg_2}(a)$. If two words are related, then the intervals corresponding to these words have non-empty intersection. Note that in the above picture, $E_{\{i,j,l\}}$ denotes the left end point of the j^{th} subinterval of the set I_i.

Theorem 3.4.18. *The Martin boundary of the Markov chain X associated to the substitution Markov chain Σ_{eg_2} is homeomorphic to the Cantor-like set $\mathcal{C}$.*

Proof. Since we know that the Martin boundary $\hat{X}/X$ is isomorphic to the sequence space S, we need only show that S is homeomorphic to $\mathcal{C}$. Let $x = (x_0 = a, x_1, \dots) \in S$. As mentioned above, by construction the subintervals of I_n are in bijective correspondence to the words in $L(n)$ so that the geodesic ray

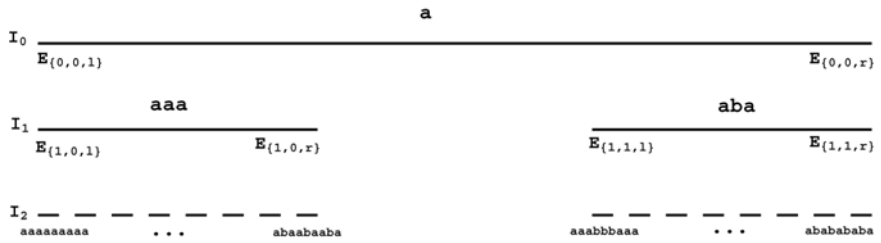


Figure 3.1. Constructing the Martin boundary

$\pi(a, x)$ corresponds to a decreasing sequence of intervals whose intersection is a point $y \in \mathcal{C}$. We may then define the mapping $F : Y \to \mathcal{C}$ by $F(x) = y$. This mapping is continuous since for $x, x' \in Y$ and n such that $a \ll_n |x \wedge x'|$,

$$|F(x) - F(x')| \leq \prod_{j=0}^{n} \frac{1}{M(j)} \leq 2^{-|x \wedge x'|} = \rho(x, x')^{\log(2)}$$

Also, F is injective since if $F(x) = F(x')$ then the geodesics $\pi(a, x)$ and $\pi(a, x')$ coincide for all but finitely many vertices, but since Σ_{eg_2} is injective, this implies that $x = x'$. Surjectivity of F is also clear since for any $y \in \mathcal{C}$, $y = \cap_n [E_{\{n,i_n,l\}}, E_{\{n,i_n,r\}}]$ and so for $x = (a, x_1, \dots)$ with x_n corresponding to the interval $[E_{\{n,i_n,l\}}, E_{\{n,i_n,r\}}]$, $F(x) = y$.

Finally, since S is a compact space and $\mathcal{C}$ is Hausdorff, an elementary point-set topology argument shows that F is a homeomorphism. □

As a final note, Theorem 3.4.18 only relies on the injectivity of the substitution Markov chain

$$\Sigma_{eg_2} : \begin{cases} a \to \begin{cases} aaa \text{ with prob. } 1/2 \\ aba \text{ with prob. } 1/2 \end{cases} \\ \\ b \to \begin{cases} bbb \text{ with prob. } 1/2 \\ bab \text{ with prob. } 1/2 \end{cases} \end{cases}$$

and not even on the weights p_i (with some slight and obvious modifications to the construction of $\mathcal{C}$). Thus Theorem 3.4.18 can be restated as follows:

$|g_a(\omega_a)| = |g_b(\omega_b)|$

Definition 3.4.19. *We call a substitution Markov chain Σ injective if for each $a \neq b \in \mathcal{A}$, $g_a(\Omega_a) \cap g_b(\Omega_b) = \emptyset$.*

Theorem 3.4.20. *The Martin boundary of the Markov chain X associated to an injective substitution Markov chain is homeomorphic to a Stone space (compact, Hausdorff, totally disconnected) and in particular, a Cantor-like set.*

Chapter 4

Reversible substitution Markov chains

As noted in Chapter 3, an SMC is not irreducible and point-mass initial distributions δ_a are not *standard measures* (see [28]). The Martin boundary is thus dependent on the choice of root. We present a strategy for circumventing this problem by associating an irreducible, reversible Markov chain associated to any given SMC. This frees us from needing to consider the root of the substitution Markov chain (or the initial distribution). Furthermore, by relating the substitution matrices of the irreducible and the completely reducible Markov chains, we are able to show in theorem 4.3.12 that subword frequencies converge almost surely in the irreducible case as well.

Secondly, this chapter also forms the basis for the model of molecular evolution presented in chapter 5.

4.1 Introduction

As noted before in lemma 2.3.7, a substitution Markov chain is not irreducible. We plan to develop in this section a method to obtain a structurally similar and yet irreducible Markov chain associated to an SMC. This Markov chain will be called a *substitution Markov chain with insertions and deletions* or SMC(R) for short. Furthermore, for a substitution Markov chain $(\mathcal{A}^*, P)$ if $P(u, v) > 0$ then for a particular sample path, v is obtained from u by replacing

letters of u with single letters or words $g_a(\omega)$. A reversed substitution Markov chain will allow v to be obtained from u by not only *replacing* letters of u with words, but by also *removing* subwords from u. As mentioned in chapter 3, forming an irreducible Markov chain associated to an SMC may present a general approach to determining Martin boundaries. In later work, we hope to relate the boundary of an SMC with the boundary of the associated SMC(R).

4.2 Reversible SMC

We now define a reversible substitution Markov chain.

Definition 4.2.1 (Reversible Markov chain)**.** *A Markov chain (X, P) with state space X and transition operator P, is said to be reversible if there exists a function $m : X \to (0, \infty)$ such that for all $x, y \in X$,*

$$m(x)P(x, y) = m(y)P(y, x)$$

We now define the Markov chain which will serve as our comprehensive model of molecular evolution.

Definition 4.2.2 (SMC(R))**.** *We define the substitution Markov chain with insertions and deletions, SMC(R), $(\mathcal{A}^*, R)$ to be the Markov chain on the state space $\mathcal{A}^*$ with the transition operator (matrix) given for $x, y \in \mathcal{A}^*$ as*

$$R(x, y) = \frac{P(x, y) + P(y, x)}{1 + \sum_{z \in \mathcal{A}^*} P(z, x)} \tag{4.1}$$

This Markov chain is reversible with $m(x)$ given by

$$m(x) = 1 + \sum_{z \in \mathcal{A}^*} P(z, x).$$

As defined, every $(\mathcal{A}^*, P)$ has finite range $|\{y : P(x, y) > 0\}| < \infty$, so equation (4.1) is well defined. We use the notation $R(\cdot, \cdot)$ for the transition operator associated to an SMC(R) to emphasize the property of reversibility that SMC(R).

For notational simplicity, we will use the same notation Σ_n to refer to the n-th coordinate random variable associated to both $P(\cdot,\cdot)$ and $R(\cdot,\cdot)$. Similarly for expectation ; we will attempt to make clear the distinction via context.

4.3 Frequencies of SMC(R)

To determine the frequencies of n-mer subwords of $(\mathcal{A}^*, R)$, we present another perspective on the substitution matrices.

4.3.1 Substitution matrices of SMC(R)

We first give the notation for the substitution matrix associated to the SMC(R) $(\mathcal{A}^*, R)$ to distinguish it from M_Σ. Note that $L^{(m)}(P) = L^{(m)}(R)$.

Definition 4.3.2. *For $i, j \in L^{(m)}(P)$, $m \in$, let*

$$\left({}_R M_\Sigma^{(m)}\right)_{ij} = \ {}_j|\Sigma^{(m)}|_i$$

where the expectation is taken with respect to the measure induced by $R(\cdot,\cdot)$. For simplicity, we use ${}_R M_\Sigma$ for ${}_R M_\Sigma^{(1)}$.

4.3.3 Relationship between ${}_R M_\Sigma$ and M_Σ

The following matrices will be used to elucidate the relationship between ${}_R M_\Sigma$ and M_Σ.

Definition 4.3.4. *By $H^{(m)}$ we denote the following matrix, for $i \in \mathcal{A}^*$, $j \in L^{(m)}(P)$,*

$$(H^{(m)})_{ij} = |i|_j$$

Definition 4.3.5. *By $I^{(m)}$ we denote the following matrix, for $i \in L^{(m)}$ and $j \in \mathcal{A}^*$*

$$(I^{(m)})_{ij} = \begin{cases} 1 & i = j \in L^{(m)}(P) \\ 0 & \text{otherwise} \end{cases}$$

For simplicity, let $H = H^{(1)}$ and $I = I^{(1)}$ (note this is not the usual identity).

Proposition 4.3.6. *For a substitution Markov chain $(\mathcal{A}, P)$, the Markov transition matrix P and the substitution matrix M_Σ are related by the following*

$$M_\Sigma = IPH$$

Similarly,

$$_R M_\Sigma = IRH$$

Proof. Calculating:

$$(M_\Sigma)_{ij} = \ _j|\Sigma|_i = \sum_{z \in \mathcal{A}^*} P(j,z)|z|_i = IPH$$

The $_R M_\Sigma$ case is completely analogous. □

Definition 4.3.7. *We denote by D the diagonal matrix with basis $L(P)$ whose diagonal entries are given by:*

$$D = \text{diag}_{x \in L(P)} \left(\frac{1}{1 + \sum_{z \in \mathcal{A}^*} P(z,x)} \right)$$

By $D(x)$ we mean $\frac{1}{1+\sum_{z \in \mathcal{A}^*} P(z,x)}$. Also, by $D|_\mathcal{A}$ we mean the matrix D restricted to the basis elements in $\mathcal{A}$.

Using the above definition, we have the matrix equation $R = D(P + P^T)$ where P^T is the transpose of the matrix P (compare with definition 5.3.3). Also, by proposition 4.3.6, we have $IDPH = D|_\mathcal{A} M_\Sigma$.

Proposition 4.3.8. *For $i, j \in \mathcal{A}$,*

$$\left(IP^T H\right)_{i,j} = P(i,j)$$

Hence, $IP^T H = P|_\mathcal{A}$, the matrix P restricted to the basis elements in $\mathcal{A}$.

Proof. For arbitrary $i, j \in \mathcal{A}$,

$$\left(IP^T H\right)_{i,j} = \sum_{z \in \mathcal{A}^*} P(z,j)|z|_i \tag{4.2}$$

$$= \sum_{z \in \mathcal{A}} P(z,j)|z|_i \tag{4.3}$$

$$= P(i,j) \tag{4.4}$$

The restriction of the summation set in line 4.3 is due to definition 2.3.1 implying that for $j \in \mathcal{A}$ and $P(z,j) > 0$, then $|z| = 1$. □

We are now ready to prove the following:

Proposition 4.3.9. *The substitution matrices associated to R and P are related by:*

$$_R M_\Sigma = D|_{\mathcal{A}} \left(M_\Sigma + P|_{\mathcal{A}} \right)$$

Proof. Combining proposition 4.3.6 and the following propositions, we have:

$$_R M_\Sigma = IRH \tag{4.5}$$

$$= ID(P + P^T)H \tag{4.6}$$

$$= IDPH + IDP^T H \tag{4.7}$$

$$= D|_{\mathcal{A}} \left(M_\Sigma + P|_{\mathcal{A}} \right) \tag{4.8}$$

□

Proposition 4.3.10. *The matrix $_R M_\Sigma$ is primitive.*

Proof. Using proposition 4.3.9, we have

$$_R M_\Sigma = D|_{\mathcal{A}} \left(M_\Sigma + P|_{\mathcal{A}} \right) \geq \min_{i \in \mathcal{A}} D(i) M_\Sigma \tag{4.9}$$

where the inequality is meant entry-wise. Hence

$$_R M_\Sigma^n \geq (\min_{i \in \mathcal{A}} D(i))^n M_\Sigma^n \tag{4.10}$$

Since M_Σ is primitive, n can be made large enough that

$$(\min_{i \in \mathcal{A}} D(i))^n M_\Sigma^n > 0.$$

□

Proposition 4.3.11. *If*

$$\sum_{\substack{z\in\mathcal{A}^{>1}\\ j\in\mathcal{A}}} P(j,z)|z| > 1 \tag{4.11}$$

Then

$$|\Sigma_n| \xrightarrow[n\to\infty]{R-a.s.} \infty$$

Note that condition (4.11) can be interpreted as saying that the P-weighted average length (taken over all $a \in \mathcal{A}$) of the words in $g_a(\Omega_a) \cap \mathcal{A}^{>1}$ is greater than 1 (recall definition 2.3.1). Hence we need "some long words w to have enough weight $P(a,w)$".

Proof. By proposition 4.3.10, ${}_RM_\Sigma \geq O\left((\Lambda \times \min_{i\in\mathcal{A}} D(i))^n\right)$ entry wise for Λ the Perron-Frobenius eigenvalue for M_Σ. Applying corollary 2.4.6 to $(\mathcal{A}^*, R)$ and using the remarks following lemma 2.4.5, if we can show $\Lambda \times \min_{i\in\mathcal{A}} D(i) > 1$ then the Perron-Frobenius eigenvalue associated to ${}_RM_\Sigma$ would be greater than one. Accordingly, this would show that $|\Sigma_n| \xrightarrow[n\to\infty]{R-a.s.} \infty$. Now the Perron-Frobenius theorem applied to M_Σ implies that

$$\Lambda \geq \min_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} (M_\Sigma)_{ij}$$

So now

$$\begin{aligned}
\Lambda \geq \min_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} (M_\Sigma)_{ij} &= \min_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} {}_j|\Sigma|_i && (4.12)\\
&= \min_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} \sum_{i\in\mathcal{A}} \sum_{z\in\mathcal{A}^*} P(j,z)|z|_i && (4.13)\\
&= \min_{i\in\mathcal{A}} \sum_{j\in\mathcal{A}} \sum_{z\in\mathcal{A}^*} P(j,z) \left(\sum_{i\in\mathcal{A}} |z|_i\right) && (4.14)\\
&= \sum_{j\in\mathcal{A}} \sum_{z\in\mathcal{A}^*} P(j,z)|z| && (4.15)\\
&= \sum_{\substack{z\in\mathcal{A}^{>1}\\ j\in\mathcal{A}}} P(j,z)|z| + ||P|_{\mathcal{A}}|| && (4.16)\\
&> 1 + ||P|_{\mathcal{A}}|| && (4.17)
\end{aligned}$$

$$\geq \frac{1}{\min_{i \in \mathcal{A}} D(i)} \tag{4.18}$$

□

Theorem 4.3.12 (Frequencies of SMC(R))**.** *For a primitive SMC(R), $\alpha \in \mathcal{A}$, and $w \in L^{(m)}(P)$ the sequence of real numbers, if*

$$\sum_{\substack{z \in \mathcal{A}^{>1} \\ j \in \mathcal{A}}} P(j,z)|z| > 1 \tag{4.19}$$

then

$$\frac{|{}_R\Sigma_n(\alpha)|_w}{|{}_R\Sigma_n(\alpha)|}$$

converges $_\alpha$ almost surely to a limit whose value is independent of α.

Proof. It is easy to check that the results contained in section 2.4.12 hold for the Markov chain $(\mathcal{A}^*, R)$ as well as for $(\mathcal{A}^*, P)$ as long as ${}_RM_\Sigma$ is primitive and that $|\Sigma_n|$ diverges almost surely with respect to the measure induced by R as $n \to \infty$. We have demonstrated that the first condition is true in general (proposition 4.3.10), and the second under the given assumption (proposition 4.3.11). Note that the second condition implies the transience of the Markov chain $(\mathcal{A}^*, R)$. As always, we are assuming that M_Σ is a primitive matrix. □

We have thus shown that under the assumption $\sum_{\substack{z \in \mathcal{A}^{>1} \\ j \in \mathcal{A}}} P(j,z)|z| > 1$, the frequencies of letters $a \in \mathcal{A}$ in the reversed substitution Markov chain $(\mathcal{A}^*, R)$ can be calculated using the Perron-Frobenius eigenvector associated to ${}_RM_\Sigma = D|_{\mathcal{A}}\,(M_\Sigma + P|_{\mathcal{A}})$. There exist similar equations for the induced matrices ${}_RM_\Sigma^{(m)}$, and there exists an algorithm for calculating n-mer frequencies for $(\mathcal{A}^*, R)$ (analogous to the end of section 2.4.12). We do not include these equations and algorithms due to their notational (but not computational) intractability.

A corollary to proposition 4.3.12 is that the notion of topological entropy introduced in 6.1.5 exists on a reversed substitution Markov chain as well. Since existence of topological entropy depended only on convergence of the

complexity function which in turn depended on convergence of frequencies, we are justified in making the following definition:

Definition 4.3.13 (Topological entropy for an SMC(R))**.** *For $(\mathcal{A}^*, P)$ satisfying the conditions of proposition 4.3.12, we define the* ***topological entropy*** *$H_{top}(\sigma^R)$ of SMC(R) as*

$$H_{top}(\Sigma^R) := \lim_{n\to\infty} \lim_{m\to\infty} \frac{\log_t \left(P_{\Sigma^R_m(a)}(n)\right)}{n}$$

4.3.14 Martin boundary of SMC(R) associated to Σ_{eg_1} and Σ_{eg_2}

We now consider the reversible Markov chains SMC(R) associated to the examples Σ_{eg_1} and Σ_{eg_2} given in sections 3.4.1 and 3.4.4. Using proposition 4.3.11 and definition 4.2.2, one can observe that the SMC(R) associated to the examples Σ_{eg_1} and Σ_{eg_2} are traditional transient, irreducible, countable-state Markov chains, the boundary theory of which is well studied. We can then observe that the Martin boundaries associated to these SMC(R)'s are actually the same is the the non-reversible, non-irreducible cases.

Proposition 4.3.15 (Equivalence of Martin boundaries for SMC and SMC(R) for Σ_{eg_1})**.** *The Martin boundary of the SMC associated to Σ_{eg_1} is homeomorphic to the Martin boundary of the SMC(R) associate to Σ_{eg_1}.*

Proof. It has already been seen in theorem 3.4.3 that the Martin boundary of the SMC associated to Σ_{eg_1} is homeomorphic to the unit interval. Using theorem 1.2 from [74], it can be seen that the Martin boundary of the SMC(R) associated to Σ_{eg_1} is *also* homeomorphic to the unit interval. □

Proposition 4.3.16 (Equivalence of Martin boundaries for SMC and SMC(R) for Σ_{eg_2})**.** *The Martin boundary of the SMC associated to Σ_{eg_2} is homeomorphic to the Martin boundary of the SMC(R) associate to Σ_{eg_2}.*

Proof. It has already been seen in theorem 3.4.18 that the Martin boundary of the SMC associated to Σ_{eg_2} is homeomorphic to the Cantor set. Using theorem 3 from [27], it can be seen that the Martin boundary of the SMC(R) associated to Σ_{eg_2} is *also* homeomorphic to the Cantor set. □

We believe that these examples hint at a deeper connection between the Martin boundaries of an SMC and the corresponding SMC(R). The calculation of the Martin boundary is significantly easier for an SMC than for an SMC(R) (compare the proof of theorem 3.4.3 to that of theorem 1.2 in [74]), so it would be very advantageous to find a broad class of substitution Markov chains in which the Martin boundaries of the SMC and SMC(R) are homeomorphic. The set of injective SMC's (see definition 3.4.19) are the only such class the author has found thus far.

Chapter 5

Comprehensive model of molecular evolution with alignment free parameter estimation via SMC's

The most prominent application of substitution Markov chains to date is in the modeling of molecular evolution. As noted in the very nice review [65], most models of molecular evolution do not incorporate the very important phenomena of insertions and deletions. In fact the three most commonly utilized models are the HKY [46], Jukes-Cantor [49], and REV [100] models. These models describe the probabilities of a single nucleotide mutating into another single nucleotide as a four state Markov chain; each simply has different transition probabilities. It has been observed that insertions and deletions play a vital role in molecular evolution.

In this chapter, we present a comprehensive new framework (based on substitution Markov chains) for handling biologically accurate models of molecular evolution. This model provides a systematic framework for studying models of molecular evolution that implement heterogeneous rates, conservation of reading frame, differing rates of insertion and deletion, customizable parametrization of the probabilities and types of substitutions, insertions, and deletions, as well as neighboring dependencies.

Most importantly, we utilize the convergence of subword frequencies (theorems 2.4.10 and 4.3.12) from chapters 2 and 4 to develop an alignment-free

parameter estimation technique. This alignment-free technique circumvents many of the nuanced issues related to alignment-dependent parameter estimation.

This model is of particular importance due to the lack of a robust and comprehensive probabilistic model of molecular evolution that includes insertions and deletions (as remarked in [10]).

The construction of the model is based upon a reversible SMC with the only additional requirement being that for each letter $a \in \mathcal{A}$, the function $g_a : \Omega_a \to \mathcal{A}^*$ has the property that if $|g_a(\omega)| > 1$, then the word $g_a(\omega)$ begins with a (see section 5.3).

5.1 Basics of molecular evolution

Before defining these Markov chains, it will be beneficial to review the (oversimplified) basics of molecular evolution. A DNA sequence is a finite string of characters over the alphabet $\{A, C, T, G\}$ with the letters referred to as *nucleotides*. When a coding DNA sequence is to be used in the production of a protein, nucleotides are read in non-overlapping groups of three. This grouping of three is called a *codon*, and codons become amino acids according to the Genetic Code. When a DNA strand is replicated, some errors may occur that lead to mutations. These mutations are classified as either *small scale* (substitutions, insertions, deletions) or *large scale* (amplification, inversion, etc). *When modeling molecular evolution, one usually considers only small scale mutations* ([61], [103]).

We now explain what is meant by substitutions, insertions, and deletions in the context of DNA replication. In DNA replication the double helix is unwound and each strand acts as a template for a new strand. Considering one of these templates, the sequence is "read" one nucleotide at a time from (say) left to right and a new DNA sequence is created. The goal is to make an exact copy of the template strand. However, it might happen that a single nucleotide in the template strand (say A) is replicated to a different nucleotide (say C) in the new strand. This is called a *substitution*. It might happen that between two adjacent nucleotides (say AT) in the template strand, a finite

word is inserted between these nucleotides in the new strand (so AT in the template would become, say $ACCCCT$ in the new strand). This is called *insertion*. It might also happen that between two non-adjacent nucleotides in the template strand (say $AGGGGGC$), the nucleotides between them are deleted (so say $AGGGGGC$ in the template becomes AC in the new strand). This is called *deletion*. Insertions and deletions are collectively called *indels* and (non-overlapping) insertion and deletion events can both happen in the replication of a single DNA sequence.

Molecular evolution (in our context) is the aggregation over a long period of time (and so many instances of replication) of such mutations (substitutions and indels) described above. Other mutational evens can also be considered, but this will be detailed further on.

5.2 SMC model of molecular evolution

While substitution models of molecular evolution have been well studies and developed, the inclusion of insertions and deletions (indels) into biologically accurate models has enjoyed less success. As remarked in [10], a robust and comprehensive understanding of probabilistic indel analysis (and its relationship to sequence alignment) is still lacking. A number of models of molecular evolution that include substitutions, insertions and deletions have been proposed ([11], [71], [72], [102], [101]), but there has yet to be developed a comprehensive mathematical structure in which biologically accurate models of molecular evolution can be developed. In fact, it is this lack of a well-studied mathematical structure that leads to the analytic intractability of some proposed indel models (as mentioned in [71]). This lack of a unifying structure not only gives rise to a variety of non-biologically motivated constructs (such as "fragments" [102], [101] and the embedding of a given sequence into an infinite sequence [71]), but also leads to difficulties in comparing models, their assumptions, and their applicability. For example, the relationships between various substitution models of molecular evolution are well understood and these relationships can be easily compared and contrasted (as in [110] and [65]). This is due to most traditional substitution models being stated in

terms of instantaneous rate matrices of a finite state Markov chain. Due to the variety of mathematical tools that have been used to implement indels in various applications (for example: HMM's [11], rate grammars [71], transducers [54], birth-death processes [102] [101]) no such immediate comparison of models is possible.

A few structures have been suggested, most notably the framework of a finite state transducer ([10]). While finite state transducers indeed seem promising, we take a route that is more probabilistically motivated and takes advantage of the well-developed theory for countable-state Markov chains.

Our model allows for the incorporation of substitutions, insertions, deletions, inversions, and duplications of any length less than or equal to a specified length N. We also incorporate rate heterogeneity, and neighboring dependencies (context dependency) up to a specified distance. The model is discrete time (it can be viewed as a generalization of a stochastic grammar) and the biological assumptions are clearly stated: First, we assume that in one time unit, besides substitutions, only one mutation event (deletion, insertion, inversion, etc.) is allowed. Secondly, we assume reversibility, though as we will see this assumption can easily be relaxed. These are the only inherent assumptions in this model. Our model contains a high degree of flexibility further assumptions can be made if, for example, one insists on using the HKY ([46]) model as the underlying substitution model.

The mathematical language in which we cast this model is that of a discrete, time-homogeneous Markov chain on infinitely (countably) many states. This approach allows for immediate application of probabilistic methods. This approach also allows for the possibility of recasting the model in terms of a time-*inhomogeneous* Markov chain so as to allow for the evolution of the rate of evolution, but here we remain in the time-homogeneous case for notational simplicity.

5.3 Definition of the model

We present now the rigorous definition of the model. We proceeded in two steps: First we define an infinite state Markov chain to incorporate insertions

and substitutions. Second we construct the induced reversible Markov chain which will incorporate deletions. To construct the first Markov chain, we need:

1. $\mathcal{A} = \{a_1, \ldots, a_t\}$ a finite set of ordered symbols referred to as an *alphabet*.
2. For each letter $a \in \mathcal{A}$, (Ω_a, P_a) a finite (non-empty) probability space.
3. For each letter $a \in \mathcal{A}$, a function $g_a : \Omega_a \to \mathcal{A}^*$ with the property that if $|g_a(\omega)| > 1$, then the word $g_a(\omega)$ begins with a.

Let $\mathcal{A}^n$ denote all words of length n formed from letters of $\mathcal{A}$. Then $\mathcal{A}^* = \cup_{n\geq 1}\mathcal{A}^n$ is the set of finite length words formed from $\mathcal{A}$. The alphabet $\mathcal{A}$ is usually equal to either $\{A, C, T, G\}$ for DNA models, or $\{R, H, K, \ldots, V\}$ for amino acid models. The probability spaces (Ω_a, P_a) encapsulate the probabilities of insertion, substitution, and other desired mutational events. In particular, the cardinality of Ω_a gives the number of different substitution, insertion, and deletion types that are allowed to occur at the letter $a \in \mathcal{A}$. The functions $g_a : \Omega_a \to \mathcal{A}^*$ (and particularly, the ranges of functions g_a) specify the set of allowable substitutions and insertions. In particular, if one wishes to allow the substitution of the letter $b \in \mathcal{A}$ to occur at the letter $a \in \mathcal{A}$, then the function g_a should evaluate to b on some P_a-non-zero element ω_1 of Ω_a: $g_a(\omega_1) = b$. If one wishes to allow the insertion of the n-length word $v_1 \ldots v_n \in \mathcal{A}^n$ to occur after the letter $a \in \mathcal{A}$, then the function g_a should evaluate to $av_1 \ldots v_n$ on some P_a-non-zero element ω_2 of Ω_a: $g_a(\omega_2) = av_1 \ldots v_n$. Notice that the word $v_1 \ldots v_n$ is preceded by a in $g_a(\omega_2) = av_1 \ldots v_n$, this is to assure that $v_1 \ldots v_n$ has been genuinely inserted into the sequence. Lack of the initial a would cause the net effect of a being deleted, then $v_1 \ldots v_n$ being inserted into the created gap.

We now define the Markov chain representing substitution Markov chains and insertions.

Definition 5.3.1 (SMC(I))**.** *A substitution Markov chain for insertions SMC(I) (with fixed $\mathcal{A}$, $\{(\Omega_a, P_a)\}_{a\in\mathcal{A}}$, and $\{g_a\}_{a\in\mathcal{A}}$) is an infinite state Markov chain $(\mathcal{A}^*, P)$ with transition operator P defined in the following way. For $u = b_1 \ldots b_n \in \mathcal{A}^*$ a word, we let $\Omega_u = \Omega_{b_1} \times \cdots \times \Omega_{b_n}$ and $P_u = P_{b_1} \times \cdots \times P_{b_n}$. We*

define $g_u : \Omega_u \to \mathcal{A}^*$ *via concatenation of words: for* $\omega = (\omega_1, \ldots, \omega_n) \in \Omega_u$, $g_u(\omega) = g_{b_1}(\omega_1) \ldots g_{b_n}(\omega_n)$. *Now define* P *by*

$$P(u, v) = \sum_{\omega \in g_u^{-1}(v)} P_u(\omega) \tag{5.1}$$

So one can think of the given model in the following way: instead of modeling the evolution of individual nucleotides (or fragments) the Markov transition operator P operates on entire sequences. Indeed, the state space of this model is $\mathcal{A}^*$: the set of all sequences. The transition operator $P(u, v)$ (probability of transition from the sequence u to the sequence v) takes into consideration *every* combination of insertions and substitutions possible in one time unit to compute the appropriate probability.

We now construct the induced reversible Markov chain, which will serve to incorporate deletions. Recall first the definition of a reversible Markov chain:

Definition 5.3.2 (Reversible Markov chain)**.** *A Markov chain* (X, P) *with state space* X *and transition operator* P*, is said to be reversible if there exists a function* $m : X \to (0, \infty)$ *such that for all* $x, y \in X$*,*

$$m(x)P(x, y) = m(y)P(y, x)$$

We now define the Markov chain which will serve as our comprehensive model of molecular evolution.

Definition 5.3.3 (Reversible Substitution Markov Chain for Insertions and Deletions, SMC(R|I|D))**.** *we define the reversible SMC for insertions and deletions, SMC(R|I|D), model* $(\mathcal{A}^*, R)$ *to be the Markov chain on the state space* $\mathcal{A}^*$ *with the transition operator (matrix) given for* $x, y \in \mathcal{A}^*$ *as*

$$R(x, y) = \frac{P(x, y) + P(y, x)}{1 + \sum_{z \in \mathcal{A}^*} P(z, x)} \tag{5.2}$$

This Markov chain is reversible with $m(x)$ given by
$m(x) = 1 + \sum_{z \in \mathcal{A}^*} P(z, x)$. As defined, every $(\mathcal{A}^*, P)$ has finite range $|\{y : P(x, y) > 0\}| < \infty$, so equation (5.2) is well defined. Note that as mentioned

in the introduction, we can completely circumvent the reversibility criterion (and simultaneously allow for different rates for insertions and deletions) by modifying the above definition in the following way. If we wish insertions to have a rate π_i and deletions to have a rate of π_d (one can easily make these rates depend on time, or location in a given sequence), then we can use the following SMC(I|D) model that drops the reversiblity requirement: $(\mathcal{A}, R)$ with R given by

$$R(x,y) = \frac{\pi_i P(x,y) + \pi_d P(y,x)}{1 + \sum_{z \in \mathcal{A}^*} P(z,x)} \tag{5.3}$$

We wish to model mutations that occur due to inherent DNA replication infidelities, not mutations due to environmental factors. Such mutations can be accurately modeled in a discrete time fashion. The transition probability $R(u,v)$ between two sequences u and v takes into account every possible substitution that could have happened when evolving the sequence u into v in one evolutionary step (the time unit can be taken to be a single replication). The transition probability $R(u,v)$ also takes into account all possible substitution and insertion paths leading from u to v, as well as all possible substitution and deletion paths leading from u to v. Again, the first assumption is that both insertions and deletions do not simultaneously happen in one step. Rather, to allow insertions *and* deletions, one needs to consider the n-th step transition matrix: $R^{(n)}(u,v)$ which is the $(u,v)^{\text{th}}$ entry in the n^{th} matrix product of R with itself. This n-step transition probability represents the probability (summed over all possible paths) of evolving the sequence u into the sequence v in n time units. So for the purpose of measuring the total evolutionary distance from u to v, one considers the so called *Green's function*:

$$G(u,v) = \sum_{n=0}^{\infty} R^{(n)}(u,v) \tag{5.4}$$

The Green's function represents the sum of probabilities of ever evolving the sequence u into the sequence v.

5.4 Example

We present here a toy example to elucidate the above definitions. We also present notation that succinctly summarizes given model. In this example, say we wish to have a mathematical object that represents a model where substitutions are described by Kimura's two parameter model [55] with transition probability equal to 0.2 and transversion probability equal to 0.1. Thus the instantaneous rate matrix for substitutions has the form:

$$Q = \begin{array}{c} \\ A \\ C \\ T \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & C & T & G \end{array} \\ \left[\begin{array}{cccc} 0.6 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.6 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.6 & 0.1 \\ 0.2 & 0.1 & 0.1 & 0.6 \end{array}\right] \end{array}$$

Say we also desired the model to allow only one letter insertions or deletions, with the probability of an indel occurring being 0.01, and the choice of the particular indel type being given by the uniform distribution. Thus, transitions occur with probability $0.2 * (1 - 0.01) = 0.198$, transversions occur with probability $0.1 * (1 - 0.01) = 0.099$ and insertions and deletions both occur with probability $0.01/4 = 0.0025$.

Using the notation introduced in section 5.3, the alphabet is given by $\mathcal{A} = \{A, C, T, G\}$. Each of the probability spaces Ω_a consist of eight elements. We provide the details regarding Ω_A, P_A, and g_A, the other definitions are completely analogous. Now as stated $\Omega_A = \{1, 2, 3, 4, 5, 6, 7, 8\}$ since there are eight allowable events that can happen at the base A: substitution to one of four other bases, plus four possible indels. Also, P_A and g_A are defined by

Ω_A	P_A	g_A
1	0.594	A
2	0.099	C
3	0.099	T
4	0.198	G
5	0.0025	AA
6	0.0025	AC
7	0.0025	AT
8	0.0025	AG

So, for example, we have that $P_A(1) = 0.594$ and $g_A(1) = A$. Note that the last four rows of the above table represent the insertion or deletion of A, C, T, or G respectively. The notation of Table 1. can be used to summarize the probability spaces and functions of this example.

Definition 5.4.1 (SMC as a Random Variable)**.** *For $v \in \mathcal{A}$, by $\Sigma_n(v)$ we denote the n-th coordinate random variable associated to the Markov chain $(\mathcal{A}^*, P)$ with initial distribution unit mass on v.*

Note that for Ω^∞ the trajectory space of $(\mathcal{A}^*, P)$, $\Sigma_n(v) : \Omega^\infty \to \mathcal{A}^*$. Thus for particular trajectory $\omega \in \Omega^\infty$, step n and initial word v, $\Sigma_n(v)(\omega) \in \mathcal{A}^*$. We will usually suppress the dependence of Σ_n on both $v \in \mathcal{A}^*$ and especially $\omega \in \Omega^\infty$. We will refer to both $(\mathcal{A}, P)$ and Σ as *substitution Markov chains* for brevity.

5.5 Flexibility of the model

In this section we present the flexibility of the SMC(R|I|D) model given in definition 5.3.3.

5.5.1 Implementable biological phenomena

Due to the flexibility of this model, we summarize here the various biological phenomena that can be implemented by using the SMC(R|I|D) model and its variants mentioned above. The SMC(R|I|D) model provides a systematic

Table 5.1. Notation describing the Markov chain example from section 5.4

$$
\Sigma : \begin{cases}
A \to \begin{cases}
A \text{ with probability } 0.594 \\
C \text{ with probability } 0.099 \\
T \text{ with probability } 0.099 \\
G \text{ with probability } 0.198 \\
AA \text{ with probability } 0.0025 \\
AC \text{ with probability } 0.0025 \\
AT \text{ with probability } 0.0025 \\
AG \text{ with probability } 0.0025
\end{cases} \\
C \to \begin{cases}
A \text{ with probability } 0.099 \\
C \text{ with probability } 0.594 \\
T \text{ with probability } 0.198 \\
G \text{ with probability } 0.099 \\
CA \text{ with probability } 0.0025 \\
CC \text{ with probability } 0.0025 \\
CT \text{ with probability } 0.0025 \\
CG \text{ with probability } 0.0025
\end{cases} \\
T \to \begin{cases}
A \text{ with probability } 0.099 \\
C \text{ with probability } 0.198 \\
T \text{ with probability } 0.594 \\
G \text{ with probability } 0.099 \\
TA \text{ with probability } 0.0025 \\
TC \text{ with probability } 0.0025 \\
TT \text{ with probability } 0.0025 \\
TG \text{ with probability } 0.0025
\end{cases} \\
G \to \begin{cases}
A \text{ with probability } 0.198 \\
C \text{ with probability } 0.099 \\
T \text{ with probability } 0.099 \\
G \text{ with probability } 0.594 \\
GA \text{ with probability } 0.0025 \\
GC \text{ with probability } 0.0025 \\
GT \text{ with probability } 0.0025 \\
GG \text{ with probability } 0.0025
\end{cases}
\end{cases}
$$

framework for studying models of molecular evolution that implement heterogeneous rates, conservation of reading frame (through careful selection of the functions g_a), variation in conservation, differing rates of insertion and deletion, customizable parameterization of the probabilities and types of substitutions, insertions, and deletions available (through the specification of the probabilities (Ω_a, P_a)), as well as neighboring dependencies.

5.5.1.1 Traditional substitution models of molecular evolution

The SMC(R|I|D) model allows for the implementation of most previous substitution models of molecular evolution. For example, by using the alphabet $\mathcal{A} = \{A, C, T, G\}$, and for each $a \in \mathcal{A}$, letting $\Omega_a = \{1, 2, 3, 4\}$, and $g_a(\Omega_a) = \{A, C, T, G\}$, and choosing the probabilities P_a appropriately, the SMC(R|I|D) model given in definition 5.3.3 completely encompasses the JC [49], HKY [46], FEL81 [32], K2P [55], and REV [100] models. In fact, in this particular case the SMC(R|I|D) model is a generalization of *all* possible homogeneous rate Markov models of DNA or amino acid evolution. This is due to the fact that if $g_a(\Omega_a) = \mathcal{A}$ for each $a \in \mathcal{A}$, then the SMC(R|I|D) model simply becomes a traditional finite state Markov chain (with as many states as the cardinality of the alphabet $\mathcal{A}$).

5.5.1.2 General mutational events

Other mutational events such as inversions, duplications, and translocations can also be implemented in this model by considering *induced substitution(s)*. The definition of an induced substitution is given in definition 2.3.12. Briefly, an induced substitution enlarges the alphabet on which the SMC(R|I|D) is defined to include n-length concatenations of elements of $\mathcal{A}$. We do not consider such mutational events in our current exposition, but note the possibility of their inclusion.

5.5.1.3 Heterogeneous rates

Models utilizing heterogeneous rates of evolution can be introduced by slightly modifying definition 2.3.1 and consequently 5.3.3. Instead of fixed probability

spaces (Ω_a, P_a), we allow the probability space to change. Let $\mathcal{P}(g_a(\Omega_A))$ denote the set of probability measures on $g_a(\Omega_a)$. Then the desired heterogeneity can be introduced with a *random probability* (also known as a random element [5]) i.e. a probability-valued random variable: $X_a : (\Omega_u, P_u) \to \mathcal{P}(g_a(\Omega_a))$. Then definitions 2.3.1 and 5.3.3 work just as before, but instead by utilizing the spaces (Ω_a, X_a). Hence, the SMC(R|I|D) model can also incorporate heterogeneous evolution rates. This is similar in spirit to the idea of the "variety of fragments" utilized in [101].

5.5.1.4 Neighboring dependencies

We can introduce neighboring dependencies by again slightly modifying definition 2.3.1. For $u = b_1 \dots b_n \in \mathcal{A}^*$, instead of using the probability $P_u = P_{b_1} \times \cdots \times P_{b_n}$, we can use a coupling P_u (i.e. whose marginal distributions correspond to the $P_{b_1}, \dots, P_{b_n}$). Hence the original definition 2.3.1, which assumes that what happens at a specific nucleotide (be it substitution, insertion, or deletion) is independent of its neighbors, simply uses the null coupling. Of course, the specific coupling to be used depends on the situation at hand, we are simply enumerating the various mathematical constructs that may be used to implement the desired biological properties.

5.5.1.5 Parameterization

Now the SMC(R|I|D) model is not meant to be implemented in its most general form, but rather parameterized to a certain degree taking into consideration the problem at hand. For example, we sketch here a possibility of parameterizing the indel appearance rate to depend only on a two parameter α, β power law $\alpha L^{-\beta}$ on the length L of the particular indel. To accomplish this, all one needs to do is define the P_a in the following way: for $\omega \in \Omega_a$, let $P_a(\omega) = \alpha |g_a(\omega)|^{-\beta}$. Further nuanced parameterizations are possible and easily implemented into the SMC(R|I|D) model. For example, one can easily use a distribution on the possible indels that not only takes length into consideration, but also GC content.

5.5.1.6 Algorithms and implementation

The transition probabilities

$$P(u,v) = \sum_{\omega \in g_u^{-1}(v)} P_{u_1}(\omega_1) P_{u_2}(\omega_2) \dots P_{u_n}(\omega_n) \tag{5.5}$$

can be calculated using standard dynamical programming techniques. If one wishes to measure the total (evolutionary) distance between the sequences u and v by using the Green's function $G(u,v)$ found in equation 5.4, it is typically intractable to attempt to compute the entire infinite sum $\sum_{i=1}^{\infty} R^{(i)}(u,v)$, but in [58], it is proved that the approximation

$$G_n(u,v) = \sum_{i=0}^{n} R^{(i)}(u,v) \tag{5.6}$$

converges to the full summation $G(u,v)$ in geometric speed. So in practice one needs only calculate $G_n(u,v)$ for some adequately large value of n (representing one to n generations of mutations from u to v).

5.6 Frequencies and parameter estimation

Normalizing the Perron-Frobenius eigenvector of ${}_R M_\Sigma$ gives the expected frequency of appearance of the letter $\mathcal{A}$. Furthermore, the observed frequencies converge at geometric speed to the expected frequencies. Similarly, couplet frequencies are given by the dominant eigenvector of ${}_R M_\Sigma^{(2)}$. These frequencies are given in terms of the various parameters that were chosen for the probabilities P_a in the definition of 2.3.1. For single letters frequencies in the DNA case, this results in four constraints (linear in the entires of ${}_R M_\Sigma$), and hence can estimate up to four parameters. Counting couplets of nucleotides gives 16 quadratic constraints. As we demonstrated, n-mer frequencies for $n > 2$ depend linearly on couplet frequencies. Hence this model can have at most 20 parameters to describe the mutational events. Such a high number of possible parameters gives great flexibility in describing the desired mutational events.

To estimate the parameters, set these equations describing expected fre-

quencies equal to the observed frequencies and then utilize standard numerical optimization techniques. Complications may arise due to multiple optima and over-determined systems, but these issues were not observed in our application. Furthermore, care must be taken when using large data sets as it might be more appropriate to utilize the heterogeneous rates approach mentioned in section 5.5.

As an example, if one was given a data set and wished to model it using $(\mathcal{A}^*, R)$ that used the Kimura two parameter γ, δ model ([55]) to describe the one letter substitutions and a two-parameter power law ($P_a(\omega) = \alpha|g_a(\omega)|^{-\beta}$) to describe the indel distribution, one would only need to count single letter frequencies to obtain four equations in the four parameters $\alpha, \beta, \gamma, \delta$.

5.6.1 Alignment-free nature of parameter estimation

It is of utmost important to note that this method of parameter estimation is completely alignment free. This circumvents the myriad issues involved when, for example, estimating parameters in a classical substitution model of molecular evolution: choosing a particular alignment algorithm, a particular alignment parameterization (linear, log-linear, affine), particular mismatch, match, gap opening, and gap extensions penalties. It is hard to overstate the advantages of having an alignment-free parameterization technique. Choosing a particular alignment scheme is a nuanced endeavor where slight changes in implementation can lead to large changes in alignment outcome. Furthermore, it has been observed that various algorithms have potential to introduce hidden bias (see [67], [66], [69], [70], [36], [106], [35]).

5.6.2 Alternative parameter estimation

Alternatively, the parameters in this model can be estimated using standard maximum likelihood methods: counting indel and substitution types and frequencies over a number of optimal and sub-optimal alignments (or all alignments) and parameterizing accordingly.

5.7 Conclusion

We have presented a comprehensive new framework for handling biologically accurate models of molecular evolution. As we have demonstrated, the number of implementable biological phenomenon is vast. One profound advantage of stating the SMC(R|I|D) in the language of an infinite state Markov chain is that one can utilize the vast mathematical literature to rigorously analyze a given implementation. We used such theorems to develop an alignment-free parameter estimation technique. This alignment-free parameter estimation technique circumvents many nuanced issues related to alignment-dependent estimation.

Chapter 6

Topological entropy of finite sequences with applications to DNA

We have already introduced topological entropy in section 6.1.5 and seen that it converges in expectation under the model of molecular evolution introduced in section 5.3.3. Due to the compatibility of topological entropy and our model of molecular evolution, we expect topological entropy to be a useful tool in genomic analysis. We pursue this line of research in this chapter by investigating the appropriate finite implementation of topological entropy for genomic analysis.

Entropy, as a measure of information content and complexity, was first introduced by Shannon [94]. Since then entropy has taken on many forms, namely topological, metric (due to Shannon), Kolmogorov-Sinai, and Rènyi entropy. These entropies were defined for the purpose of classifying a system via some measure of complexity or simplicity. These definitions of entropy have have been applied to DNA sequences with varying levels of success. Topological entropy in particular is infrequently used due to high-dimensionality problems and finite sample effects. These issues stem from the fact that the mathematical concept of topological entropy was introduced to study *infinite* length sequences. It is universally recognized that the most difficult issue in implementing entropy techniques is the convergence problems due to finite sample

effects ([105], [57]). A few different approaches to circumvent these problems with topological entropy and adapt it to *finite* length sequences have been attempted before. For example, in [104], linguistic complexity (the fraction of total subwords to total possible subwords) is utilized to circumvent finite sample problems. This though leads to the observation that the complexity/randomness of intron regions is *lower* than the complexity/randomness of exon regions. However, in [16] it is found that the complexity of randomly produced sequences is *higher* than that of DNA sequences, a result one would expect given the commonly held notion that intron regions of DNA are free from selective pressure and so evolve more randomly than do exon regions. Also, little has been done in the way of mathematically analyzing other finitary implementations of entropy due to most previous implementations using an entire function instead of a single value to represent entropy (thus the expected value would be very difficult to calculate)

In this chapter we focus on topological entropy, introducing a new definition that has all the desired properties of an entropy and still retains connections to information theory. This approximation, as opposed to previous implementations, is a *single* number as opposed to an entire function, thus greatly speeding up the calculation time and removing high-dimensionality problems while allowing more mathematical analysis. This definition will allow the comparison of entropies of sequences of differing length, a property no other implementation of topological entropy has been able to incorporate. We will also calculate the expected value of the topological entropy to precisely draw out the connections between topological entropy and information content. We will then apply this definition to the human genome to observe that the entropy of intron regions is in fact lower than that of exon regions in the human genome as one would expect. We then provide evidence indicating that this definition of topological entropy can be used to detect sequences that are under selective pressure.

6.1 Methods

6.1.1 Definitions and preliminaries

We restrict our attention to the alphabet $\mathcal{A} = \{A, C, T, G\}$. For a finite sequence w over the alphabet $\mathcal{A}$, we use $|w|$ to denote the length of w. Of primary importance in the study of topological entropy is the complexity function of a sequence w (finite or infinite) formed over the alphabet $\mathcal{A}$.

Definition 6.1.2 (Complexity function)**.** *For a given sequence w, the complexity function $p_w :$ $\rightarrow$ is defined as*

$$p_w(n) = |\{u : |u| = n \text{ and u appears as a subword of } w\}|$$

That is, $p_w(n)$ represents the number of different n-length subwords (overlaps allowed) that appear in w.

Now the traditional definition of topological entropy of an *infinite* word w is the asymptotic exponential growth rate of the number of different subwords:

Definition 6.1.3. *For an infinite sequence w formed over the alphabet $\mathcal{A}$, the topological entropy is defined as*

$$\lim_{n\to\infty} \frac{\log_4 p_w(n)}{n}$$

Due to the limit in the above definition, it is easily observed that this definition will always lead to an answer of zero if applied directly to finite length sequences. This is due to the fact that the complexity function of infinite length sequences is non-decreasing, while of finite length sequences it is eventually zero. We include in figures 6.1 and 6.2 a log-linear plot of the complexity functions for the gene ACSL4 found on ChrX:108906440-108976621 (hg19) as well as for an infinite string generated by a Markov chain on four states with equal transition probabilities.

The graph of the complexity function of the gene found in figure 6.1 is entirely typical of the graph of a complexity function for a finite sequence as can be seen by the following proposition. The proof can be found in the nice

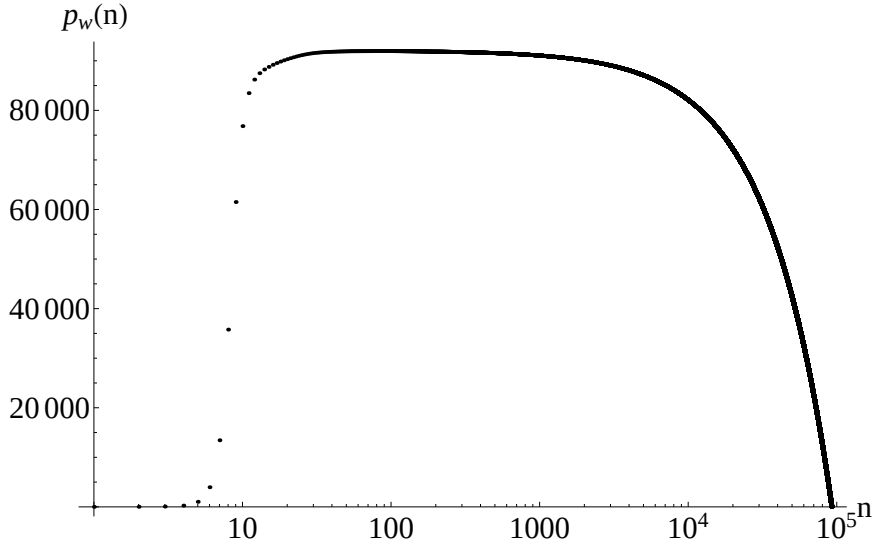


Figure 6.1. Log-Linear Plot of the Complexity Function of the Gene ACSL4

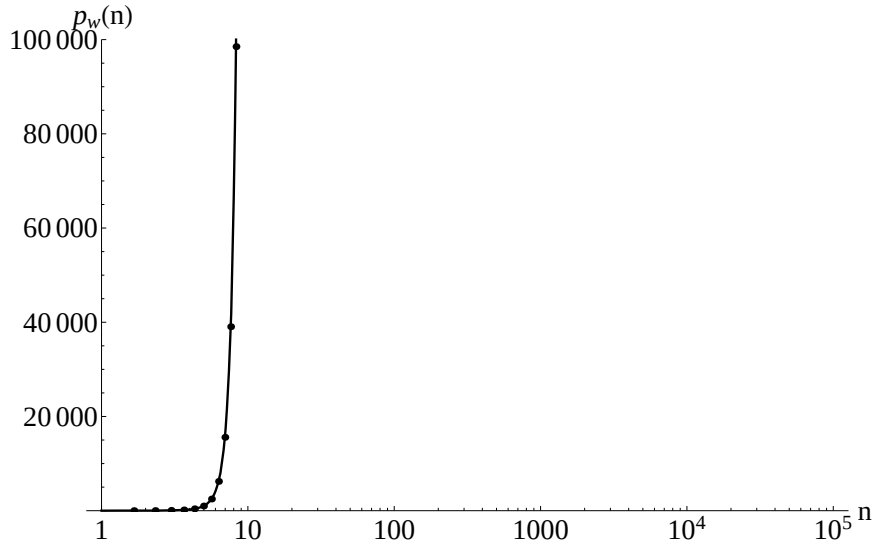


Figure 6.2. Log-Linear Plot of the Complexity Function of a Random Infinite Sequence.

summary by [16]. Note that in the following m and M are numbers whose calculation is straightforward.

Proposition 6.1.4 (Shape of Complexity Function)**.** *For a finite sequence w, there are integers m, M, and $N = |w|$, such that the complexity function*

$p_w(n)$ is strictly increasing in the interval $[0, m]$, non-decreasing in the interval $[m, M]$ and strictly decreasing in the interval $[M, N]$. In fact, for n in the interval $[M, N]$ we have $p_w(n+1) - p_w(n) = -1$.

Now for a finite sequence w we desire that an approximation of topological entropy $H_{top}(w)$ should have the following properties:

1. $0 \leq H_{top}(w) \leq 1$
2. $H_{top}(w) \approx 0$ if and only if w is highly repetitive (contains few subwords)
3. $H_{top}(w) \approx 1$ if and only if w is highly complex (contains many subwords)
4. For different length sequences v, w, $H_{top}(w)$ and $H_{top}(v)$ should be comparable

It should be noted that item 4 on this list is of utmost importance when implementing topological entropy. It is very important to normalize with respect to length since otherwise when counting the number of subwords, longer sequences will appear artificially more complex simply due to the fact that since the sequence is longer, there are more chances for subwords to show up. This explains the "linear correlation" between sequence length and the implementations of topological entropy used in [51] and [57]. This also hints at the incomparability of the notions of entropy contained in [51], [16], [57], and [91].

Recall that an approximation of topological entropy should give an approximate asymptotic exponential growth rate of the number of subwords. With this and the above properties in mind, it is immediately concluded that we can disregard the values of $p_w(n)$ for n in the interval $[m, N]$ mentioned in proposition 6.1.4. In fact, as in [16] the only information gained by considering $p_w(n)$ for n in the interval $[m, N]$ has to do with the specific combinatorial arrangement of "special factors" and has little to do with the complexity of a sequence.

We define the approximation to topological entropy as follows

Definition 6.1.5 (Topological Entropy)**.** *Let w be a finite sequence of length $|w|$, let n be the unique integer such that*

$$4^n + n - 1 \leq |w| < 4^{n+1} + (n+1) - 1$$

Then for $w_1^{4^n+n-1}$ *the first* $4^n + n - 1$ *letters of* w*,*

$$H_{top}(w) := \frac{\log_4(p_{w_1^{4^n+n-1}}(n))}{n}$$

The reason for truncating w to the first $4^n + n - 1$ letters is due to the following two facts whose proofs are omitted.

Lemma 6.1.6. *A sequence* w *over the alphabet* $\{A, C, T, G\}$ *of length* 4^n+n-1 *can contain at most* 4^n *subwords of length* n*. Conversely, if a word* w *is to have* 4^n *subwords, it must have length at least* $4^n + n - 1$*.*

Thus if we had taken an integer $m > n$ in the above definitions and instead utilized $\frac{\log_4(p_w(m))}{m}$, w would not be long enough to contain all different possible subwords.

Lemma 6.1.7. *Say a sequence* w *has length* $4^n + n - 1$ *for some integer* n*, then if* w *contains all possible subwords of length* n *formed on the alphabet* $\{A, C, T, G\}$*, then* $H_{top}(w) = 1$

Thus if a sequence of length $4^n + n - 1$ is "as random as possible" (i.e. contains every possible subword), its topological entropy is 1, just as we would expect in the infinite sequence case. Similarly, if w is "as nonrandom as possible", that is, if w is simply the repetition of a single letter $4^n + n - 1$ times, then $H_{top}(w) = 0$.

Furthermore, if we had not used truncation in definition 6.1.5, then for a sequence v such that $|v| > |w|$, the topological entropy of v would on average be artificially higher due to v being a longer sequence and thus has more opportunity for the appearance of subwords. Thus, by truncating we have allowed sequences of different lengths to have comparable topological entropies.

This definition of topological entropy serves as a measure of the randomness of a sequence: the higher the entropy, the more random the sequence. The justification for this finite implementation giving an approximate characterization of randomness is given in [76] in which it is shown that functions of entropy are the only finitely observable invariants of a process.

6.1.8 Expected value

While topological entropy has been well studied for infinite sequences, very little has been done by way of mathematically analyzing topological entropy for finite sequences. This lack of analysis is most likely due to topological entropy as in the literature ([57], [19], [91]) being considered not as a single number to be associated to a DNA sequence, but rather the entire function $\frac{\log_4 p_w(n)}{n}$ is considered for *every* n. This approach turns topological entropy (which should be just a single number associated to a DNA sequences) into a very high dimensional problem. In fact, as many dimensions as is the length of the DNA sequence under consideration. Our definition given above (definition 6.1.5) does in fact associate just a single number (instead of an entire function) to a sequence, and so is much more analytically tractable.

We now utilize the results of [41] to compute the expected value of the above topological entropy. This will assist us in determining what constitutes "high" or "low" entropy. First, we calculate the expected value of the complexity function $p_w(n)$. As is commonly assumed ([65], [46], [49]), we now assume that DNA sequences evolve in the following way: each state in a Markov fashion independent of neighboring states. We do not assume a single model of molecular evolution, but rather just assume that there is some set of probabilities $\{\pi_A, \pi_C, \pi_T, \pi_G\}$ such that the probability of appearance of a sequence w is given by the following: for n_A the number of occurrences of the letter A in w, n_C the number of occurrences of the letter C in w, etc., the probability of the sequence w appearing is given by:

$$(w) = \pi_A^{n_A}\pi_C^{n_C}\pi_T^{n_T}\pi_G^{n_G}$$

This assumption regarding the probability of appearance of a DNA sequence is used only to procure a distribution against which we may calculate the expected number of subwords. The actual calculation of topological entropy as in definition 6.1.5 does not make any such assumption about the probability of appearance.

Theorem 6.1.9 (Expected Value of the Complexity Function)**.** *The expected value of the complexity function* $p_w(n)$ *taken over sequences of length* $|w| =$

$n+k-1$ is given by

$$[p_w(n)] = 4^k - \sum_w (1 - \quad (w))^n + \mathcal{O}(n^{-\epsilon}\mu^n) \tag{6.1}$$

where the summation is over all sequences w of length n, and $0 < \epsilon < 1$, $\mu < 1$ (these are explicitly computed constants based on the π_i defined above, see [41]).

Proof. See [41]. □

This theorem has a particularly nice reduction when one assumes that the probability of appearance of each subletter is the same (equivalent to the the expected value being computed with a uniform distribution on the set of all sequences of a certain length).

Corollary 6.1.10. *Assuming that $\pi_A = \pi_C = \pi_T = \pi_G = 1/4$, the expected value of complexity function taken over sequences of length $|w| = n+k-1$ is given by*

$$[p_w(n)] = 4^k - 4^k(1 - (\frac{1}{q})^k)^n + \mathcal{O}(n^{-\epsilon}\mu^k) \tag{6.2}$$

While clearly there *is* a mononucleotide bias for different genomic regions and DNA sequences do not occur uniformly randomly, we do assume equal probability of appearance of each nucleotide as then the calculation of the expected number of subwords reduces in computational complexity from exponential to linear in the length of the sequence.

It is a straightforward calculation to combine formula 6.2 with definition 6.1.5 and compute the constants ϵ and μ as set forth in [41]. Doing so, we obtain the following expected value for the topological entropy.

Theorem 6.1.11 (Expected Value of Topological Entropy)**.** *The expected value of topological entropy taken over sequences of length $|w| = 4^n + n - 1$ is given by*

$$[H_{top}] = \frac{\log_4(4^n - 4^n(1 - 1/4^n)^{4^n} + \mathcal{O}((\frac{1}{\sqrt{2}}))^n)}{n} \tag{6.3}$$

We now present in table 6.1 the calculated estimation of the expected value of H_{top} using the above formula. Keep in mind that the convergence of this calculation to the actual expected value is exponentially quick (the term $\mathcal{O}((\frac{1}{\sqrt{2}}))^n)$ as n increases (and so also the length of the sequence). We thus ignore the $\mathcal{O}((\frac{1}{\sqrt{2}}))^n$ term in the following calculation.

Table 6.1. Calculated Expected Value of Topological Entropy

n	$4^n + n - 1$	Calculated Value of H_{top}	Expected
1	4	.725606	
2	17	.841242	
3	66	.890810	
4	249	.917489	
5	1028	.933868	
6	4101	.944865	
7	16390	.952736	
8	65543	.958642	
9	262152	.963237	
10	1048585	.966914	
11	4194315	.969921	
12	16777227	.972428	

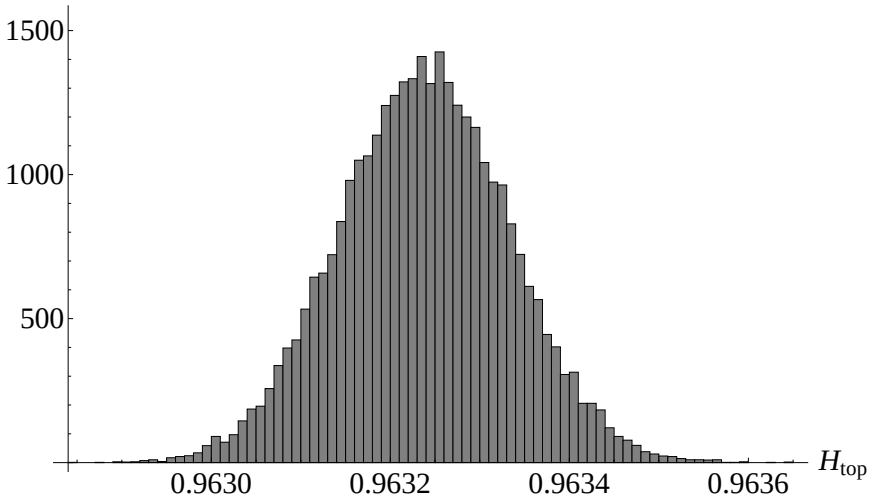
For comparison's sake, we present in table 6.2 the sampled expected values for $n = 1, \ldots, 9$ along with sampled standard deviations (the calculation where made by explicitly computing the topological entropy of uniformly randomly selected sequences).

Summarizing this table, the topological entropy of randomly selected sequences is tightly centered around the expected value which itself is close to one. Furthermore, the distribution of topological entropy is very close to a normal distribution as can be observed from the histogram of topological entropy for sequences of length $4^9 + 9 - 1$ included in figure 6.3. The skewness and kurtosis are .0001996 and 2.99642 respectively.

Table 6.2. Sampled Expected Value and Standard Deviation of Topological Entropy

n	$4^n + n - 1$	Sampled Expected Value of H_{top}	Sampled Standard Deviation	Sample Size
1	4	.703583	.184798	256
2	17	.838956	.0508640	300000
3	66	.890576	.0176785	300000
4	249	.917457	.00674325	300000
5	1028	.933869	.0027160	300000
6	4101	.944861	.00113176	300000
7	16390	.952733	.000486368	300000
8	65543	.958642	.000212283	300000
9	262152	.963237	.000094481	300000

Figure 6.3. Histogram of Topological Entropy of Randomly Selected Sequences of Length $4^9 + 9 - 1 = 262152$



6.2 Algorithm

An implementation of this approximation to topological entropy is available at:

http://www.math.psu.edu/koslicki/entropy.nb

We mention a few notes regarding this estimation of topological entropy. First,

if a sequence w in consideration has a length such that for some n, $4^n+n-1 < |w| < 4^{n+1}+n$ it will be more accurate to use a sliding window to compute the topological entropy. For example, if $|w| = 16000$, we would normally truncate this sequence to the first 4101 letters. This might misrepresent the actually topological entropy of the sequence. Accordingly, we could instead compute the average of the topological entropy of the following sequences (where w_n^m means the subsequence of w consisting of the n^{th} to m^{th} letters of w):

$$w_1^{4101}, w_2^{4102}, w_3^{4103}, \ldots, w_{11899}^{16000}$$

This is computationally intensive, so for longer sequences, one might instead choose to take non-overlapping windows, so finding the average of the topological entropy of the sequences

$$w_1^{4101}, w_{4102}^{8203}, w_{8204}^{12305}, \ldots$$

The above website includes serial and parallel versions of the algorithm. The fastest version utilizes Nvidia CUDA GPU computing, has complexity $\mathcal{O}(n)$ for a sequence of length n, and takes an average of 5.2 seconds to evaluate on a DNA sequence of length 16,777,227 when using an Intel i7-950 3.6 GHz CPU and an Nvidia GTX 460 GPU.

6.2.1 Comparison to traditional measures of complexity

Other measures of DNA sequence complexity similar to this approximation of topological entropy include: previous implementations of topological entropy [57], special factors [16], Shannon's metric entropy ([57], [30]), Rènyi continuous entropy ([105], [85]), and linguistic complexity (LC) ([104], [39]).

The implementation of topological entropy in [57] does not produce a single number representing entropy, but rather an entire sequence of values. Thus while the implementation of [57] does distinguish between artificial and actual DNA sequences, Kirillova notes that the implementation is hampered by high-dimensionality and finiteness problems.

In [16], it is noted that the special factors approach does not differentiate

between introns and exons.

Note also that the convergence of our approximation of topological entropy is even faster than that of Shannon's metric entropy. Shannon's metric entropy of the sequence u for the value n is defined as

$$H_{met}(u,n) = \frac{-1}{n}\sum_{w}\mu_u(w)\log(\mu_u(w))$$

where the summation is over all words of length n and $\mu_u(w)$ is the probability (frequency) of the word w appearing in the given sequence u. Thus Shannon's metric entropy requires not only the appearance of subwords, but for the actual frequency of appearance of the subwords to converge as well. As can be seen from definition 6.1.5, our notion of topological entropy does not require the use of the actual subword frequencies. So topological entropy will in general be more accurate than Shannon's metric entropy for shorter sequences. Accordingly, the convergence issues mentioned in [30] (even with the clever Lempel-Ziv estimator) can be circumvented.

Furthermore, it is not difficult to show (as in [6], Proposition 1.2.5) what is known as the *Variational Principle*, that is, topological entropy dominates metric entropy: for any sequence u (finite or not) and integer n

$$H_{met}(u,n) \leq H_{top}(u,n) \tag{6.4}$$

Thus topological entropy retains connections to the information theoretic interpretation of metric entropy as set forth by [94]. Since topological entropy bounds metric entropy from above:

> Low topological entropy of a sequence implies that it is
> "less chaotic" and is "more structured."

This connection to information theory is also an argument for the use of topological entropy over Rènyi continuous entropy of order α (see [105] for more details). [85] showed that for $\alpha \neq 1$, one cannot define conditional and mutual information functions and hence Rènyi continuous entropy does not measure "information content" in the usual sense. So while Rènyi entropy does allow

for the identification of statistically significant motifs ([105], one cannot conclude that higher/lower Rènyi continuous entropy for $\alpha \neq 1$ implies more/less information content or complexity in the usual sense.

Thus LC is the only other similar measurement of sequence complexity that produces a single number representing the complexity of a sequence. Like our implementation of topological entropy, the implementation of LC contained in [104] also runs in linear time. A comparison of our implementation of topological entropy and LC is contained in section 6.3.4.

6.3 Application to exons/introns of the human genome

6.3.1 Method

We now apply our definition of topological entropy to the intron and exon regions of the human genome.

We retrieved the February 2009 GRCh37/ hg19 human genome assembly from the UCSC database and utilized Galaxy ([8], [7]) to extract the nucleotide sequences corresponding to the introns and exons of each chromosome (including ChrX and ChrY). Now even though as argued above topological entropy converges more quickly than metric entropy, one must be careful to not use this definition of topological entropy on sequences that are too short as this would lead to significant noise. For example, the UCSC database contains exons that consist of a single base and it is meaningless to attempt to measure topological entropy of such sequences. Hence we selected the longest 100 different intron and exon sequences from each chromosome.

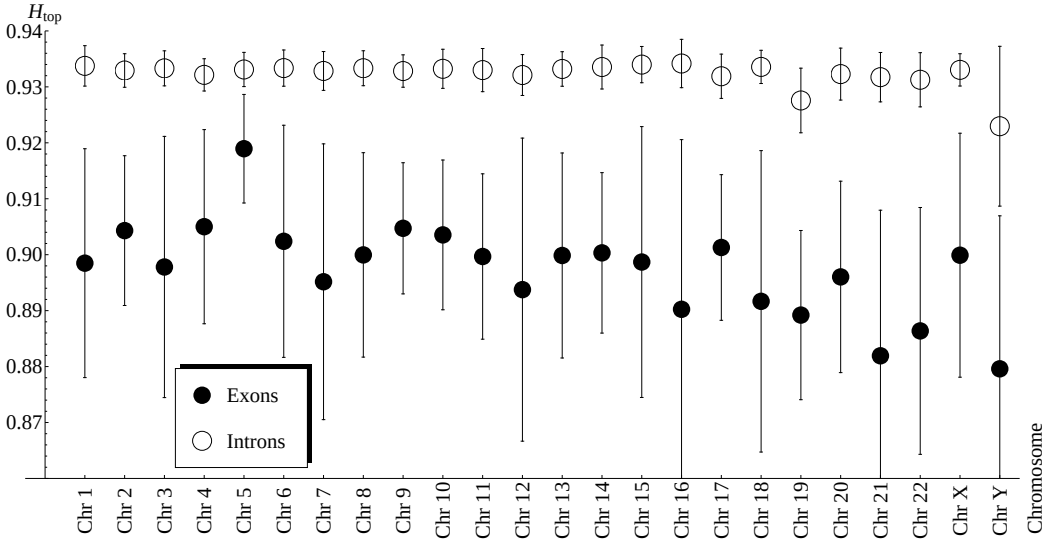
After ensuring that each sequence consisted only of letters from $\{A, C, T, G\}$, we then applied the approximation of topological entropy found in definition 6.1.5 to the resulting sequences. For comparison's sake we also applied the approximation of topological entropy to the longest 50, 200, and 400 sequences, as well as to *all* the intron and exon sequences. The salient observed features persist throughout. Though as expected, when shorter sequences are allowed, the results become noisier.

To investigate in more detail the relationship between regions under selective pressure and the value of topological entropy, we also selected each 5' and 3' UTR on chromosome Y that consisted of more than $4^3 + 3 - 1 = 66$ bp.

6.3.2 Data

Figure 6.4 displays the error bar plot for the longest 100 exons and introns. The error bar plots for the longest 50, 200, and 400 sequences, as well as the plot for all the intron and exon sequences are, for brevity's sake, not shown. Figure 6.5 displays the error bar plot for chromosome Y 5' and 3' UTRs which are longer than 66bp long.

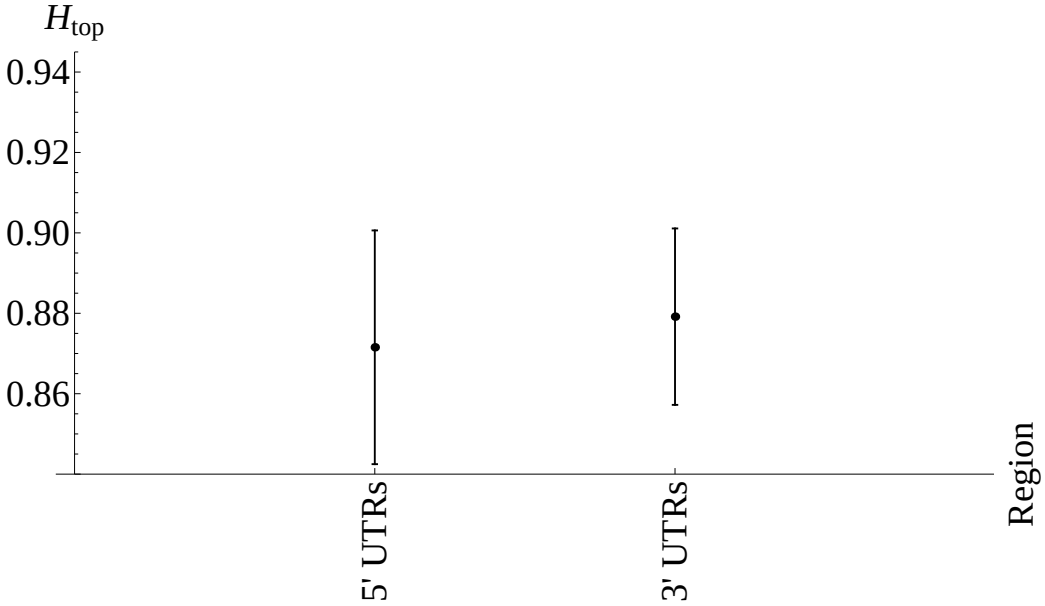
Figure 6.4. Error bar plot of average topological entropy for the longest 100 introns and exons in each chromosome



6.3.3 Analysis and discussion

We first discuss the results regarding intron and exon regions. As figure 6.4 demonstrates, the topological entropies of intron regions of the human genome are larger than the topological entropies of the exon regions. For example, the mean of the entropies of the introns on chromosome 21 is more than 11 standard deviations away from the mean of the entropy of the exons on the same chromosome. This result supports the commonly held notion that intron

Figure 6.5. Error bar plot of chromosome Y 5' and 3' UTRs longer than 66bp long



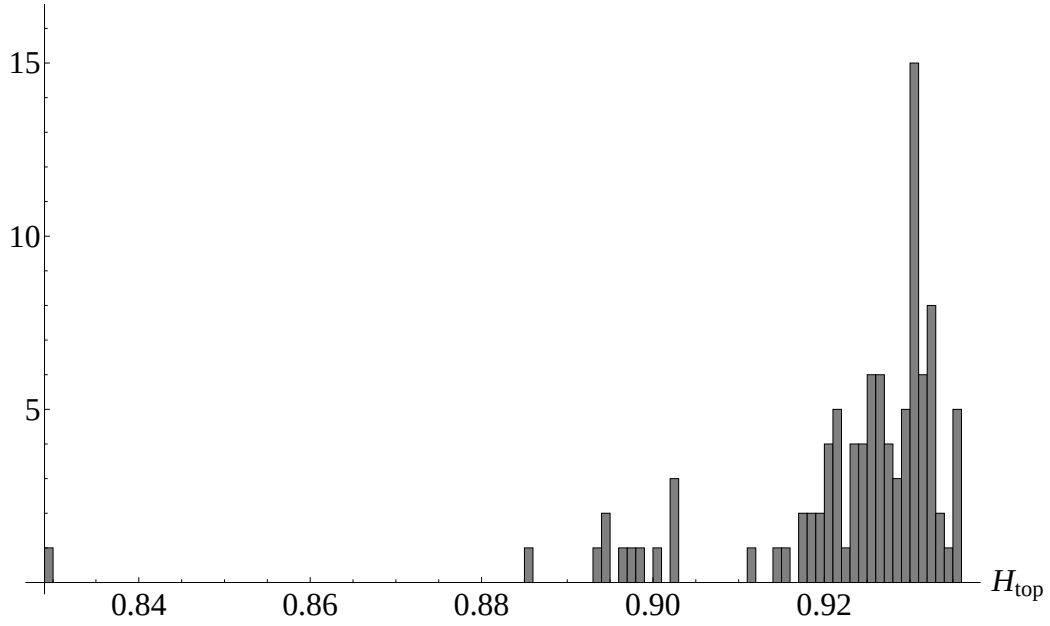
regions of DNA are mostly free from selective pressure and so evolve more randomly than do exon regions. We thus suggest that the observation of [51], [104], [68], and [98] that intron entropy is *smaller* than exon entropy is due to the aforementioned finite sample effects and high-dimensionality problems related to previous implementations of entropy.

Interestingly, even though we observe that intron entropy is larger than exon entropy, the entropies of *both* regions are much lower than expected (here expectation is as calculated in table 6.1). Indeed, of the longest 100 sequences, the average intron length is 180880 and the average exon length is 2059, so according to tables 6.1 and 6.2, we would expect the entropies to be .966914 and .933853 respectively. We find, though, that the average entropy for introns is .9323166 and for exons is .897451. Note that the largest intron sequence entropy (H_{top} = .943627 for an intron of length 1.1Mbp found on chromosome 16) is significantly lower than the expected value of .969921 (at least 60 standard deviations from the expectation). This is not too surprising considering that the expectation as calculated in theorem 6.1.11 uses the uniform distribution. This supports the conclusion that while intron regions do evolve more randomly than exon regions, introns do not evolve uniformly randomly.

Note the disparity between the entropies of the sex chromosomes: The entropy of chromosome X in both intron and exon regions is significantly higher than in chromosome Y. In fact, the mean of chromosome X intron entropies is 3.5 standard deviations higher than the mean of chromosome Y intron entropies; the mean of chromosome X exon entropies is 1 standard deviation higher than the mean of chromosome Y exon entropies. Thus the X chromosome has intron and exon entropy similar to that of the autosomes, but chromosome Y has significantly differing exon and intron entropy. This is a particularly puzzling result considering that chromosome Y is known to have a high mutation rate and a special selection regime ([112], [111], [43]), and so one would expect the entropy of chromosome Y (both intron and exon regions) to be much higher than it is. In fact, the chromosome Y introns have the lowest mean topological entropy of any intron region across the entire genome. This would suggest that the accumulation of "junk" DNA and the massive accumulation of retrotransposable elements mentioned in Graves (2006) have some underlying function or structure. More specifically, it appears that the intron regions in chromosome Y might fall into two categories: the truly "junk" DNA consisting of the introns with topological entropy greater than .910, and the introns that have hidden structure consisting of those sequences with entropy less than .910. We present in figure 6.3.3 a histogram of the topological entropy on chromosome Y demonstrating the distinction between the two categories.

Remaining on chromosome Y, we now present evidence that topological entropy can be used to detect sequences that are under selective pressure. Note that [95] showed that both 5' and 3' UTRs are among the most conserved elements in vertebrate genomes. Thus one would expect that the topological entropy of these regions would be very low (as this is indicative of a high degree of structure). As indicated in figure 6.5, the entropy of both the 5' and 3' region are low in comparison to the entropy of the intron and exon regions across the autosomes. In fact the mean of the topological entropy of the 5' and 3' UTRs ($.871545 \pm .0290619$ and $.879163 \pm .0219371$) are lower than the mean entropy of *any* intron or exon region across every chromosome. The lowest mean topological entropy for an autosome is $.927802 \pm .00539$ on chromosome 19, this is more than nine standard deviations *higher* than the

Figure 6.6. Histogram of topological entropy of introns in chromosome Y

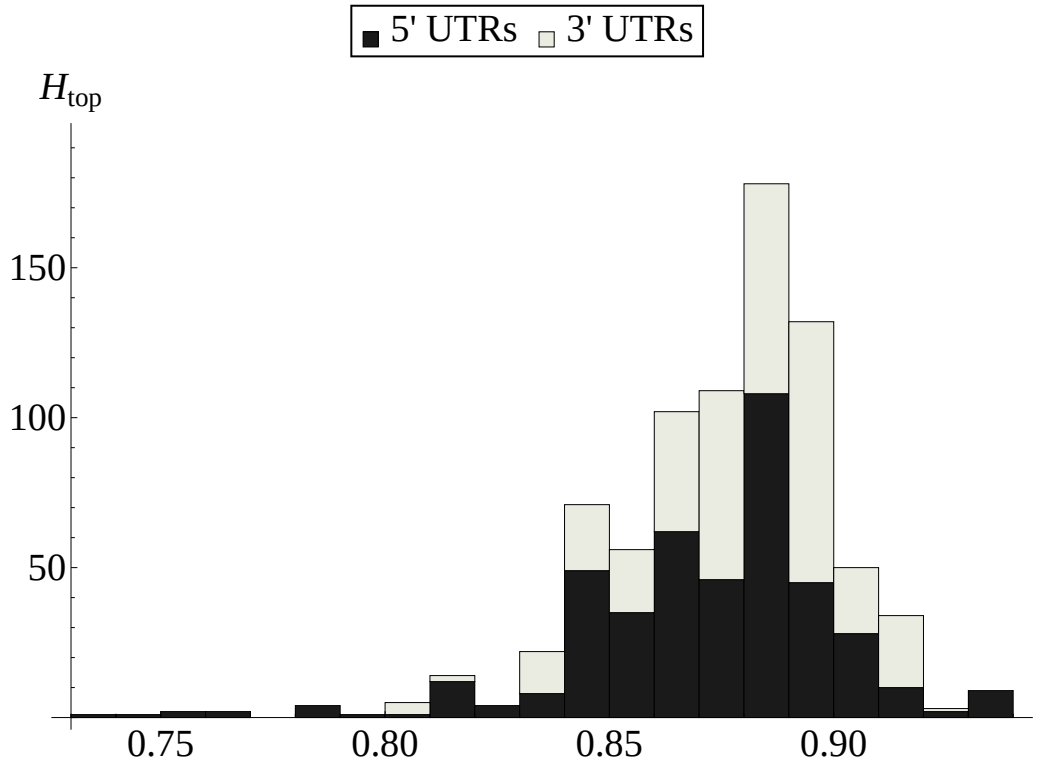


mean of topological entropy for either the 3' or 5' UTRs. This lends support to the assertion that topological entropy can be used to detect functional regions and regions under selective constraint.

6.3.4 Comparison to linguistic complexity

As mentioned in section 6.2.1, LC is the only other similar measurement of sequence complexity that produces a single number to represent the complexity of a sequence. We applied the algorithm described in [104] and written by [59] to the same data set contained in section 6.3.2. To obtain directly comparable results, we used a window size as big as the given sequence is long. As can be seen in figure 6.8, LC does distinguish between introns and exons to an extent, though not to the same quality of resolution as that of topological entropy (compare to figure 6.4). For example, while topological entropy consistently measures introns as more random than exons, LC does not. This discrepancy is most likely due to linguistic complexity being effectively utilized ([104] as a sliding window method to detect repetitive motifs, not as a holistic measure of sequence information content. So we also applied LC using a sliding window of

Figure 6.7. Histogram of topological entropy for 5' and 3' UTRs in chromosome Y



2000bp, taking the average value of LC on a given sequence, and then averaging on a given chromosome (see figure 6.9). Using the sliding window, LC does give a higher value to introns than to exons (except on chromosome 5). While the separation between the LC of introns and exons becomes more pronounced, the resolution is still not nearly as clear as with topological entropy since a large amount of error persisted. The LC values amongst introns and exons are well within one standard deviation of each other across the entire genome.

6.4 Conclusion

This implementation of topological entropy is free from issues that other implementations have encountered. Namely, this definition allows for the comparison of sequences of different length and does not suffer from multi-dimensionality complications. Since this definition supplies a single value to

Figure 6.8. Error bar plot of linguistic complexity on introns and exons using window as long as the sequence.

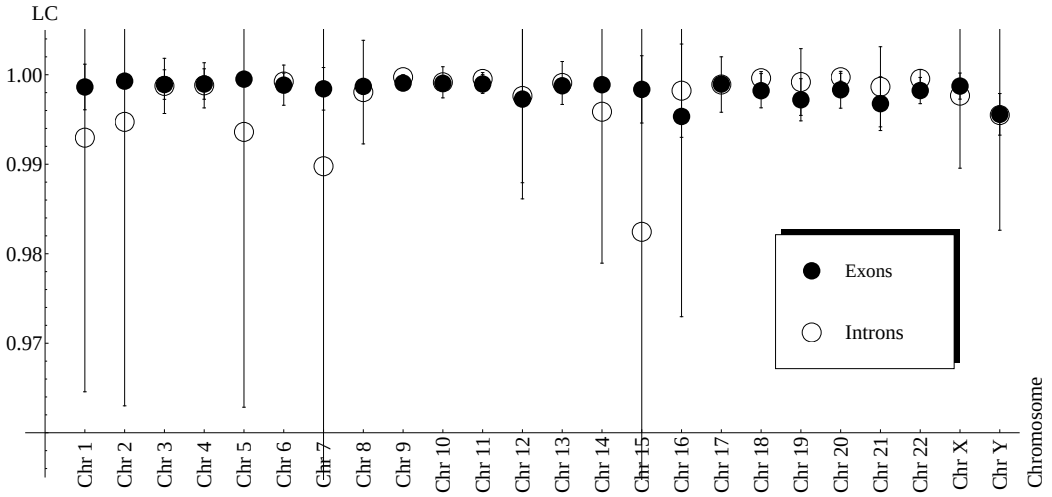
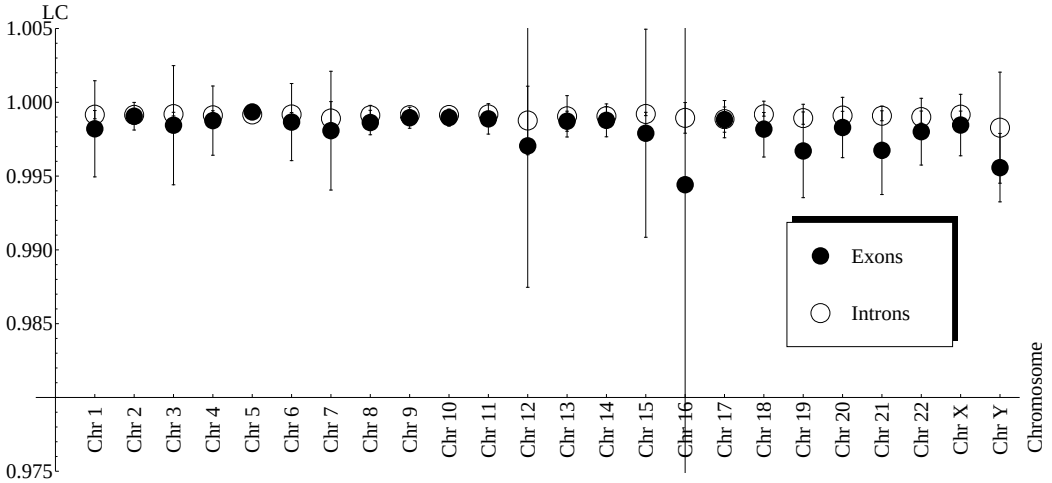


Figure 6.9. Error bar plot of linguistic complexity on introns and exons using 2000bp windows.



characterize the complexity of a sequence, it is much more capable of being mathematically analyzed. Beyond measuring the complexity or simplicity of a sequence, we presented evidence that our approximation to topological entropy might detect functional regions and sequences free from or under selective constraint. The speed and simplicity of this implementation of topological entropy makes it very suitable for utilization in detecting regions of high/low complexity. For example, we observe the novel phenomena that the introns on chromosome Y have atypically low and bi-modal entropy, possibly corre-

sponding to random sequences and sequences that posses hidden structure or function.

Chapter 7

Topological pressure of finite sequences, coding sequence density estimation, and synonymous codon bias

This chapter represents joint work with Daniel J. Thompson.

7.1 Introduction

As we saw in chapter 6, topological entropy is a useful analytic tool in the field of genomic analysis. The model of molecular evolution given in chapter 5 demonstrates the robustness of topological entropy under mutation influences. In this chapter, we continue to utilize the field of symbolic dynamics as a source of techniques for genomic analysis. We introduce a new tool called *topological pressure* (or simply *pressure*), which we apply to the study of the human genome. Pressure can be interpreted as a weighted measure of complexity and is the natural generalization of topological entropy for finite sequences introduced in chapter 6 (see also [58]). The primary goals of our analysis are to demonstrate how pressure can predict the distribution of coding sequences across a genome and to use this to recover quantitative data on codon usage

bias. This could shed light on the issue of mammalian codon bias, where it is recognized that a complete understanding has not yet been achieved [14, 47, 83].

The pressure of a finite sequence is given by counting (with weights) all subwords of an exponentially shorter length that appear in the original word. Each subword is weighted through the use of a function φ, which we call the *potential*. We focus on potentials which depend on only 3 symbols, so φ is essentially a choice of weighting for each codon. Pressure detects a trade-off between complexity in the sequence and frequency of occurrence of 'favored' codons. This intuition is made rigorous by the Variational Principle from ergodic theory [107], which we recall in §7.5.3. If a potential can be found so that the pressure correlates strongly with an observed biological phenomena, this gives evidence for the biological importance of those codons which are weighted strongly by that potential.

Our main focus is the selection of potentials for which the topological pressure correlates strongly with the observed distribution of coding sequences when using windows of size approximately $60,000$bp. We optimize this correlation independently on each chromosome with respect to the parameters of the potential and find that the Pearson's correlation coefficient between the coding sequence density and the topological pressure is above 0.9. The parameters obtained on each chromosome are close together (at least for the autosomes) so we average them to obtain a 'canonical' potential for CDS density estimation which we denote by φ_{hs}. We check that a similar potential is obtained when we optimize the correlation across the whole human genome simultaneously. We give a detailed analysis of the pattern of codon weights given by the potential φ_{hs} and observe a number of striking qualitative features that contribute to the investigation of codon usage bias.

This chapter also contributes to the investigation of codon usage bias. Recent [83, 14, 13, 15, 25, 75, 92] has focused on analyzing the nuanced and oft-debated question of the nature and cause of codon usage bias. While many studies have successfully analyzed a particular influence on synonymous codon usage (context dependency [31], GC content [15], tRNA adaptation [25], etc.), a comprehensive understanding of codon bias (particularly in mammals) re-

mains a challenging open problem. The difficulty in analyzing codon bias can be attributed partly to an over-abundance of plausible statistical and theoretical approaches which are often mutually contradictory [83]. Furthermore, there is general disagreement on how to properly take into consideration features of the sequence such as GC content and context dependencies, as well as the inherent randomness of nucleotide composition. For example, it has been argued that the codon adaptation index [83] should [89] and should not [45] be used to determine the influence of synonymous codon usage on gene expression levels.

The advantage of utilizing topological pressure is its relative simplicity. The definition is entirely combinatorial and implicitly takes account of important considerations such as neighboring dependencies, different choices of reading frame, autocorrelation, background codon frequencies, and GC content. Furthermore, for sequences of suitably large scale, the definition is robust enough to absorb the inherent randomness and noise in nucleotide composition.

These advantages allow us to compare synonymous codon usage between species and its relationship with CDS density estimation. Using the potential φ_{hs}, we show that pressure has good correlation with the CDS distribution of *mus musculus*. In addition, we optimize the correlation of CDS density with pressure over the mouse genome. We observe that the potential obtained this way shares many qualitative features with φ_{hs}. This gives evidence that the parameters in φ_{hs} are biologically meaningful and is a first step in the investigation of interspecies codon usage via topological pressure.

Inspired once more by the techniques of ergodic theory, we demonstrate that any potential φ canonically defines a probability measure on finite sequences via the Variational Principle. This measure, called the *equilibrium measure for* φ, reflects the properties of the potential and can be used to analyze sequences that are orders of magnitude shorter than those on which pressure is utilized. This represents a strategy in which large scale information (pressure) can be utilized to extract information at a much smaller scale (measure of a sequence). The development of robust techniques that detect the coding potential of short sequences is an important area of research [18, 34, 40, 44, 62, 63, 90, 109] with applications to sequence annotation as

well as gene prediction. It has been recognized that measures of coding potential based on single sequence nucleotide composition [63, p.i281] are an important part of the problem of differentiating between short reads of coding and non-coding sequences and are complementary to the very effective comparative techniques developed in, for example, [109]. We contribute to this line of research by showing that the equilibrium measure associated with φ_{hs} can effectively distinguish between randomly selected introns and exons in the human genome.

The layout of the chapter is as follows: In §7.2, we define topological pressure for finite sequences. In §7.3, we investigate the correlation of topological pressure and CDS density in the human genome. In §7.4, we briefly investigate applications of topological pressure to the mouse genome. In §7.5, we demonstrate how topological pressure defines a measure on finite sequences, and show that this measure can distinguish between coding sequences and non-coding sequences.

7.2 Topological pressure for finite sequences

We rigorously develop our implementation of topological pressure for any finite sequence. Let $\mathcal{A}$ be our alphabet, that is, a finite collection of symbols. Since our application is to the study of DNA sequences, we mainly consider the alphabet $\mathcal{A} = \{A, C, T, G\}$. We consider various spaces of sequences on the alphabet $\mathcal{A}$. We denote the space of sequences of length n by $\mathcal{A}^n$, the space of finite sequences (of any length) $\mathcal{A}^<$, the space of finite sequences of length at least n by $\mathcal{A}^{\geq n}$ and the space of infinite sequences by $\Sigma = \mathcal{A}$. For $w = (w_1, w_2, \ldots) \in \Sigma$ or $w = (w_1, w_2, \ldots, w_m) \in \mathcal{A}^{\geq n}$, let w_1^n denote the finite word $(w_1, \ldots, w_n)$. For $w \in \mathcal{A}^n$, let $[w]$ be the set of sequences $v \in \Sigma$ so that $v_1^n = w$. Let σ be the shift map: For $w = (w_1, w_2, w_3, \ldots) \in \mathcal{A}^{\geq 2} \cup \Sigma$, $\sigma((w_1, w_2, w_3, \ldots)) = (w_2, w_3, \ldots)$.

When we consider norms of matrices $M = (m_{ij})$ and vectors $v = (v_i)$, we consistently use the sum norm, so that $\|M\| = \sum_{i,j} |m_{ij}|$ and $\|v\| = \sum_i |v_i|$.

Definition 7.2.1. *We say a function ψ on Σ (or on $\mathcal{A}^{\geq m}$) depends on the first m symbols of a word if*

1. *For all $v \in \mathcal{A}^m$, the restriction of ψ to $[v]$ is a constant function.*

2. *There exists $w \in \mathcal{A}^{m-1}$ for which the restriction of ψ to $[w]$ is not a constant function.*

If ψ depends on m symbols, then for $v \in \mathcal{A}^m$, we write $\psi(v)$ for the common value of ψ on $[v]$.

We define topological pressure for finite sequences $w \in \mathcal{A}^{\leq}$. Define

$$SW_n(w) = \{u : |u| = n \text{ and } u \subset w\}.$$

The definition depends on the cardinality of the alphabet. To keep the presentation close to our applications, we give the definitions under the assumption that $\#\mathcal{A} = 4$. For alphabets of different cardinality, we simply replace the occurrences of 4 with $\#\mathcal{A}$.

Definition 7.2.2. *For a word w such that $|w| = 4^n + n - 1$ and a potential function ψ which depends on m symbols, where $n \geq m$, we define the topological pressure of ψ on w to be*

$$P(w, \psi) = \frac{1}{n} \log_4 p(w, \psi), \tag{7.1}$$

where

$$p(w, \psi) = \sum_{u \in SW_n(w)} \exp \sum_{i=0}^{n-m} \psi(\sigma^i u). \tag{7.2}$$

We denote the greatest topological pressure for such words by

$$P_{\max}(n, \psi) = \max\{P(w, \psi) : |w| = 4^n + n - 1\}. \tag{7.3}$$

Remark. When $\psi = \log \varphi$ (log denotes natural logarithm) for a function $\varphi > 0$,

$$P(w, \log \varphi) = \frac{1}{n} \log_4 \left(\sum_{u \in SW_n(w)} \prod_{i=0}^{n-m} \varphi(\sigma^i u) \right). \tag{7.4}$$

We extend this definition to words of an arbitrary finite length.

Definition 7.2.3. *For a word w with $4^n + n - 1 \leq |w| < 4^{n+1} + n$, we define the topological pressure of ψ on w to be*

$$P(w, \psi) = P(w_1^{4^n+n-1}, \psi). \tag{7.5}$$

That is, the pressure of ψ on w is defined to be the pressure of ψ on the first $4^n + n - 1$ symbols of w.

Remark. An elementary argument given in [58] shows that for each n, there exists a word v^n of length $4^n + n - 1$ which has every word of length n as a subword. It follows that $P_{\max}(\psi, n) = P(v^n, \psi)$ for any function ψ.

Remark. When $\psi = 0$, (7.1) reduces to the definition of topological entropy for finite sequences due to the first named author in [58]. The reason we take the logarithm in base 4 in (7.1) is so that $P_{\max}(n, 0) = 1$.

7.2.4 Convergence of pressure

Utilizing an SMC as the model of molecular evolution (as in section 5.3.3), and propositions 4.3.12 and 2.4.12, it is not difficult to observe that topological pressure converges in expectation.

Proposition 7.2.5 (Convergence of Pressure)**.** *For a primitive SMC Σ, initial word $v \in \mathcal{A}^*$, and potential ψ that depends on finitely many symbols, the quantity*

$$P(\Sigma_n(v), \psi)$$

converges as $n \to \infty$. Here, the expectation is taken with respect to $\quad_v$.

This allows us to be confident that topological pressure, as a technique for genomic analysis, is robust under mutational influences.

7.2.6 Normalization of potentials

An arbitrary potential $\psi = \log \varphi$ can be normalized by the addition of a constant. This is useful for a number of reasons, and does not affect the

quantities associated to pressure that we study in this chapter, such as the equilibrium measures introduced in §7.5 and correlation with the CDS density developed in §7.3. For any $t > 0$, we have the formula

$$P(w, \log t\varphi) = \frac{n-m}{n} \log_4 t + P(w, \log \varphi). \tag{7.6}$$

This allows for a variety of normalizations. For us, the most useful normalization is to let $t = \|\varphi\|^{-1}$ so that $t\varphi$ is described by a probability vector. We use this normalization frequently in §7.2.8.

7.2.7 Interpretation of high pressure sequences

A sequence with high pressure has a good mix of complexity and frequency of 'favored' codons. When using the 0 potential (i.e. entropy), we simply detect high complexity. In [58], it was shown that an intron region of a DNA sequence tends to have higher entropy than an exon region. This is due to the exons having more structure (and hence less randomness). However, for windows of larger size, which may contain numerous intron and exon regions, entropy is a poor indicator of CDS density (see figure 7.1). In §7.3.7, we demonstrate that with an appropriate choice of potential, the high pressure sequences correlate very well with those with high coding sequence density. Further insight into the meaning of pressure is given by the Variational Principle from ergodic theory, which we recall in §7.5.3. The Variational Principle makes precise the intuition that high pressure sequences are those that balance high complexity against high frequency of favored codons.

7.2.8 Selection of the potential

Two perspectives can be taken regarding selection of the potential φ. The first perspective is to obtain a potential via maximizing the correlation of topological pressure with a given set of biological data. We take this approach in section §7.3 to select potentials based on the correlation of pressure with the probability distribution of known coding sequences.

The second perspective is to construct a potential based on known biolog-

ical phenomena and then utilize topological pressure to analyze the desired feature. Next, we give an example of such a potential.

7.2.9 A 1-parameter family of examples

We give a simple family of examples to illustrate the role of pressure. We write down a potential adapted to detecting regions with high GC content. Since we focus on a much broader and more sophisticated class of potentials in the rest of this chapter, this example should be understood as an illustrative toy model. Let $\mathbf{1}_A$ denote the characteristic function of $[A]$, and suppose that $|w| = 4^n + n - 1$. Consider the family of functions

$$\varphi_t = \mathbf{1}_A + \mathbf{1}_T + t(\mathbf{1}_G + \mathbf{1}_C),$$

where $t > 0$. Then

$$p(w, \log \varphi_t) = \sum_{u \in SW_n(w)} \prod_{i=0}^{n-1} \varphi_t(\sigma^i u) = \sum_{u \in SW_n(w)} t^{GC(u)},$$

where $GC(u)$ denotes the total number of occurrences of G and C in the word u. Then

$$P(w, \log \varphi_1) = H_{top}(w)$$

and as t increases from 1, $P(w, \log \varphi_t)$ gives a measure of complexity which assigns increasing importance to sequences with greater GC content. In §7.3.13, we investigate this family of potentials and how best to choose t in the context of CDS density estimation in the the human genome.

7.3 Topological pressure and CDS density estimation

We show that pressure can be used as an effective predictor of coding sequence density for the human genome. The challenge is to make a good choice of potential function. The discovery of a potential function which correlates well

with coding sequence density then yields biologically relevant information on the roles of different codons.

7.3.1 Coding sequence density of the human genome

The *coding sequence density* is the probability density function representing the percentage of coding sequences versus non-coding sequences in non-overlapping windows of a given size. We introduce some notation in order to define the coding sequence density precisely.

Notation 7.3.1. Let $\mathrm{Chr}(i)$ denote the string which represents the i^{th} chromosome of the human genome, and $\mathrm{Chr}(i, [n, m])$ denote the substring which starts at position n and ends at position m. For convenience, we refer to the X and Y chromosomes as the 23^{rd} and 24^{th} chromosomes respectively.

We utilize the NCBI hg18 build 36.3 with coding sequences defined by NCBI RefSeq genes and accessed via Wolfram's Mathematica 8.0 [115]. We choose a chromosome and fix an integer window size m to divide the chromosome into non-overlapping windows of length m. The most suitable window sizes for comparision with topological pressure are those of the form $m = 4^n + n - 1$.

Definition 7.3.2. *For some fixed window size $m = 4^n + n - 1$, we define*

$$\#CS(i, n, x) := \#\{\text{Known coding sequences with initial nucleotide contained in } \mathrm{Chr}(i, [xm + 1, xm + m])\},$$

assuming the chromosome is read in the p to q direction. The coding sequence density is defined to be

$$\mathrm{CDS}(i, n, x) := \#CS(i, n, x)/\#CS(i),$$

where $\#CS(i) := \#\{\text{Known coding sequences in } \mathrm{Chr}(i)\}$.

Thus, the indices i and n tell us to look at the i^{th} chromosome using a window of size $4^n + n - 1$, and the index x describes the starting point of

the window along the given chromosome. For fixed i and n, $\mathrm{CDS}(i, n, x)$ is a probability density function of x.

7.3.3 Topological pressure of the human genome

We now set up notation for our application of topological pressure to the human genome.

Definition 7.3.4. *For a potential $\varphi > 0$ and $m = 4^n + n - 1$, let*

$$P^{\mathrm{hs}}(i, n, x, \varphi) := P(\mathrm{Chr}(i, [mx + 1, mx + m]), \log \varphi),$$

where $P(\cdot, \cdot)$ is the topological pressure defined in equation (7.5)*.*

Thus, $P^{\mathrm{hs}}(i, n, x, \varphi)$ is the topological pressure associated to the x^{th} window of size $4^n + n - 1$ on the human chromosome i, using the potential $\log \varphi$.

7.3.5 Selection of φ via maximum correlation with CDS density

For fixed i and n, we consider $\mathrm{CDS}(i, n, x)$ and $P^{\mathrm{hs}}(i, n, x, \varphi)$ as functions in x. We use the Nelder-Mead [73] method to maximize the correlation between $P^{\mathrm{hs}}(i, n, x, \varphi)$ and $\mathrm{CDS}(i, n, x)$ with respect to potentials φ which depend on 3 symbols. Due to (7.6), we can without loss of generality restrict our attention to the set of potentials whose parameters sum to 1. We thus obtain a set of potentials $\varphi^i_{\max}$ whose associated pressure correlates very well with the coding sequence distribution in the chromosome $\mathrm{Chr}(i)$. We expand on our methodology below, and then present and analyze our results.

7.3.6 Methodology

Considered as functions in x, both $\mathrm{CDS}(i, n, x)$ and $P^{\mathrm{hs}}(i, n, x, \varphi)$ are inherently noisy due to random fluctuations in nucleotide composition in a given chromosome as well as due to incomplete knowledge regarding coding sequences (eg. incorrectly annotated sequences). The noise in both functions is easily suppressed by utilizing a Gaussian filter (convolution with a Gaussian

kernel of radius r). We checked that other standard smoothing techniques lead to similar results, and chose the Gaussian filter for our analysis due to its simplicity and speed of implementation. The filter is applied after removing from both $\mathrm{CDS}(i,n,x)$ and $P^{\mathrm{hs}}(i,n,x)$ those x where $\mathrm{Chr}(i,[xm+1,xm+m])$ contained any symbols besides $\{A,C,T,G\}$. The radius of the Gaussian filter is chosen so that $\mathrm{CDS}(i,n,x)$ coincides at each x with the probability density function obtained from a Gaussian kernel density estimation of $\mathrm{CDS}(i,n,x)$ considered as a function of x: that is, we linearly interpolate the quantity

$$\frac{1}{h*m}\sum_{j=1}^{m} k\left(\frac{x-\mathrm{CDS}(i,n,x_j)}{h}\right), \tag{7.7}$$

where $m = \lfloor \frac{|\mathrm{Chr}(i)|}{4^n+n-2} \rfloor$, $k(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$, and the bandwidth h is selected according to Silverman's rule [96].

The selection of the window size in $P^{\mathrm{hs}}(i,n,x,\varphi)$ exhibits the typical trade-off between sensitivity and specificity: a smaller window size allows for a finer approximation of the CDS distribution, but exhibits a higher sensitivity to fluctuations in nucleotide composition. We focus on a window size of $65,543$ ($n=8$), as this seems to achieve a good balance. This corresponds to dividing $\mathrm{Chr}(1)$ into roughly 3700 non-overlapping windows.

After fixing i and n, we utilize the Nelder-Mead [73] method to maximize the correlation between $P^{\mathrm{hs}}(i,n,x,\varphi)$ and $\mathrm{CDS}(i,n,x)$ with respect to potentials φ which depend on 3 symbols and whose parameters sum to one. The precision threshold for the convergence of this heuristic maximization technique was set to 10^{-6} and convergence was typically achieved in 4000 steps of the algorithm. We denote the potential thus obtained on the i^{th} chromosome by $\varphi^i_{\max}$.

7.3.7 Results

For each chromosome, we obtain a potential $\varphi^i_{\max}$ for which $CDS(i,8,x)$ and $P^{\mathrm{hs}}(i,8,x,\varphi^i_{\max})$ display very strong positive correlation. The value of the Pearson correlation coefficient on each chromosome is shown in figure 7.1, and is above 0.9 in all cases. Figure 7.1 also demonstrates that topological entropy

is not a good estimator of coding sequence density. This is unsurprising since we have no theoretical reason to expect correlation between entropy and coding sequence density since multiple intron and exon regions may be contained in windows of this size. The parameter values for each $\varphi^i_{\max}$ can be found at

http://www.math.psu.edu/koslicki/potentials.xls

We also provide in figure 7.2 a plot of the standardized values of both $\mathrm{CDS}(5,8,x)$ and $P^{\mathrm{hs}}(5,8,x,\varphi^5_{\max})$ to show the goodness of fit, and overlay these plots on the Ensemble Genome Browser [29] histogram of known genes.

7.3.8 Comparison of the potentials $\varphi^i_{\max}$

Let $\mathbf{r}^i = (r^i_1, r^i_2, \ldots, r^i_{64})$ represent the 64 parameters of the potential $\varphi^i_{\max}$. We show that the parameters for $\varphi^i_{\max}$ exhibit a consistent codon bias by demonstrating that the probability vectors $\mathbf{r}^i$ are relatively close in the standard Euclidean metric. Figure 7.3 is a plot of the pairwise Euclidean distances between each of the chromosomes. We have

$$\max_{i,j\in\{1,\ldots,24\}} d(\mathbf{r}^i,\mathbf{r}^j) = .319 \qquad \text{and} \qquad \operatorname*{mean}_{i,j\in\{1,\ldots,24\}} d(\mathbf{r}^i,\mathbf{r}^j) = .203$$

The sex chromosomes X and Y are clear outliers, so focusing on the autosomes, these values improve to

$$\max_{i,j\in\{1,\ldots,22\}} d(\mathbf{r}^i,\mathbf{r}^j) = .284 \qquad \text{and} \qquad \operatorname*{mean}_{i,j\in\{1,\ldots,22\}} d(\mathbf{r}^i,\mathbf{r}^j) = .195$$

This is relatively close against a maximum possible distance of $\sqrt{2}$.

In [58], it was observed that the sex chromosomes exhibit a distinctly different entropy distribution than the autosomes. This observation coincides with the fact that the sex chromosome potentials were furthest from the autosomal potentials. Interestingly, the potential $\varphi^7_{\max}$ corresponding to chromosome 7 was similarly distant from the other chromosomes. This is consistent with the fact that chromosome 7 contains many regions identical to the sex chromosomes (of $30,000$ non-overlapping sequences of length $5,000$ from Chr(7), over 77% matched identically with a sequence in chromosome Y).

Figure 7.1. Correlation between pressure and CDS density

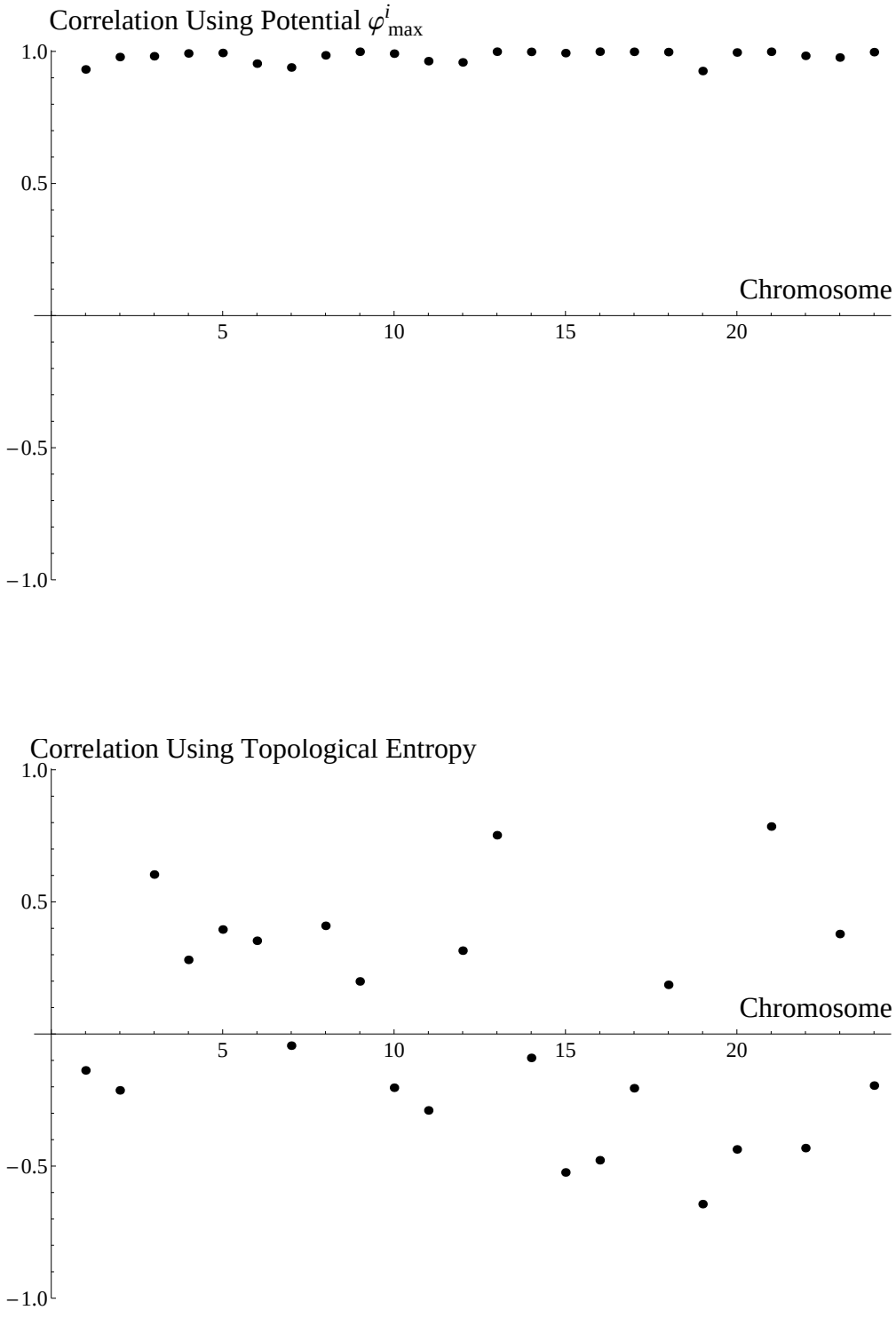


Figure 7.2. Coding sequence density, pressure, and Ensembl known CDS histogram

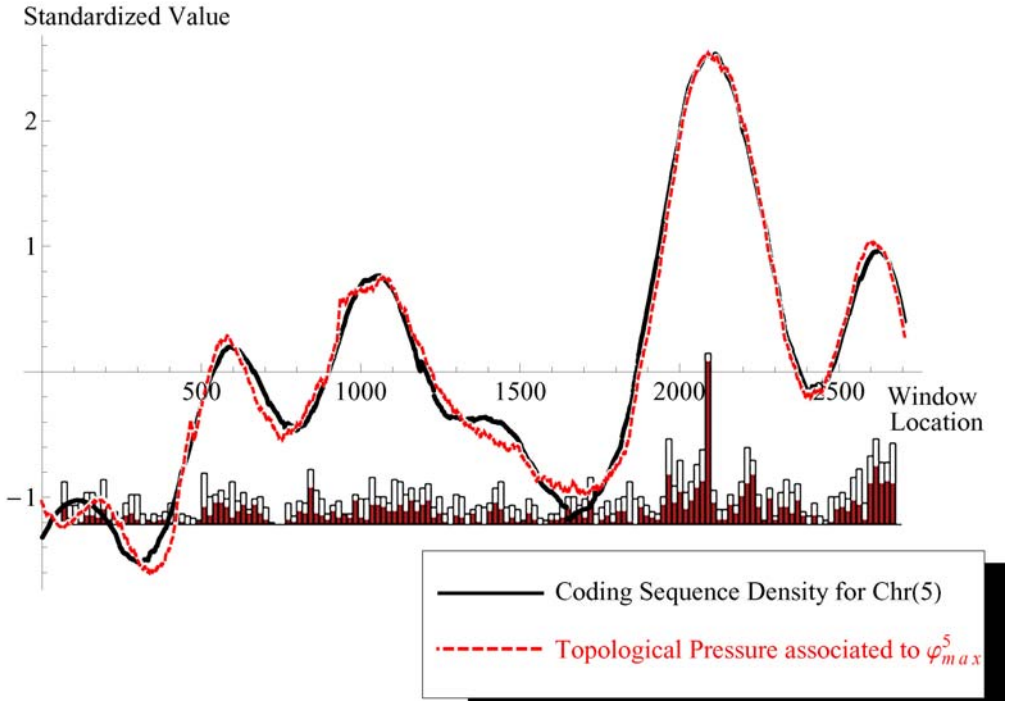
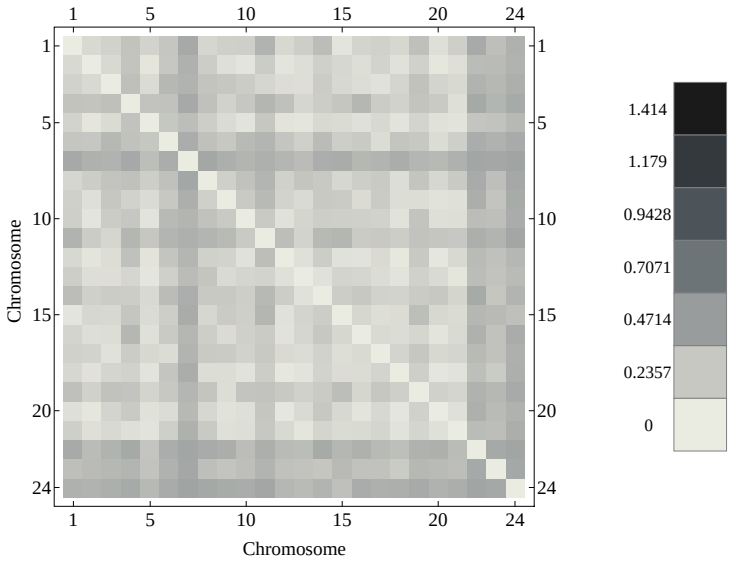


Figure 7.3. Pairwise Euclidean distance of the parameter values for $\varphi^i_{\max}$. The darker the square in position (i, j), the greater the distance between $\mathbf{r}^i$ and $\mathbf{r}^j$.



7.3.9 The best choice of potential for CDS density estimation

We make a 'canonical' choice of potential for estimation of CDS density on the human genome by taking a suitable average of the potentials $\varphi^i_{\max}$. There are

various natural ways to do this, each yielding qualitatively similar results. The resulting potential is meaningful because, as shown in §7.3.8, the individual potentials are close to each other.

Definition 7.3.10. *We define a 'canonical' potential for detecting coding sequence density in the human genome which we denote by* φ_{hs}*. For each codon* w*, we let*

$$\varphi^0_{\text{hs}}(w) := \text{median}\{\varphi^1_{\max}(w), \ldots, \varphi^{24}_{\max}(w)\},$$

and define

$$\varphi_{\text{hs}} := \frac{\varphi^0_{\text{hs}}}{\|\varphi^0_{\text{hs}}\|}$$

Other natural ways to obtain the 'canonical' potential would be to take the mean of the parameter values of each $\varphi^i_{\max}$, or to take the median/mean after omitting the outlying chromosomes X, Y and 7 from the data set. Each of these approaches yields a very similar potential. Alternatively, we can perform the maximization procedure on the sequence formed by concatenating all the autosomes. This approach yields a potential which is close to φ_{hs} (Euclidean distance less than .148) and qualitatively identical.

We include a visualization of the parameters of φ_{hs} in figure 7.4. The results of the correlation between the topological pressure $P(i, 8, x, \varphi_{\text{hs}})$ and $\text{CDS}(i, 8, x)$ for the autosomes are contained in figure 7.5.

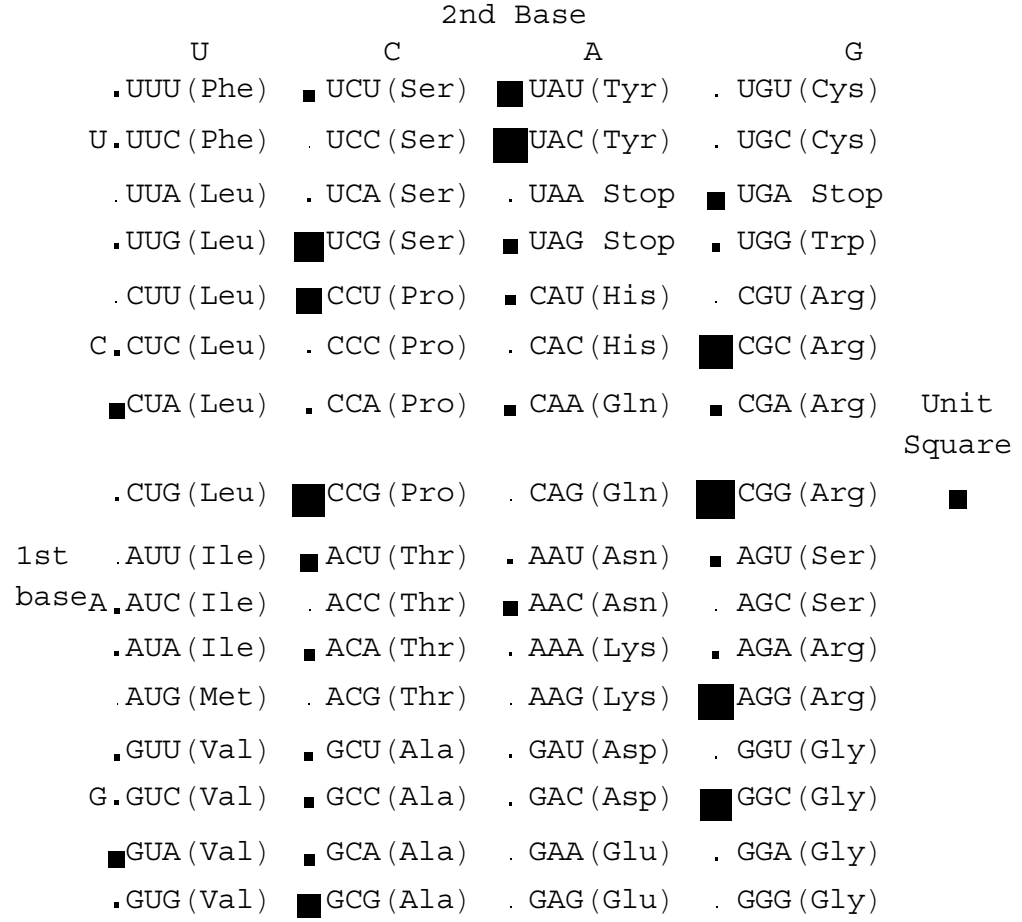
7.3.11 Analysis of parameter values for φ_{hs}

We give an in-depth analysis of our canonical potential. As can be observed from figure 7.4, it is clear that φ_{hs} exhibits a distinct codon bias. We summarize and attempt to explain some of the most distinctive features of φ_{hs}:

- The codons UCG, CCG, UAC, CGC, CGG, AGG and GGC are the most heavily weighted, and the codons CGU, ACU, GCG, UAU, UGA have quite strong weightings.

- Codons which contain the pair CG or GC tend to be highly weighted (for example UCG, CCG, GCG, CGC, CGG, GCC). This is explained by the well known connection between GC content and CDS density.

Figure 7.4. Plot of 50 times the parameter values of φ_{hs}. The area of a square is equal to the corresponding potential value.



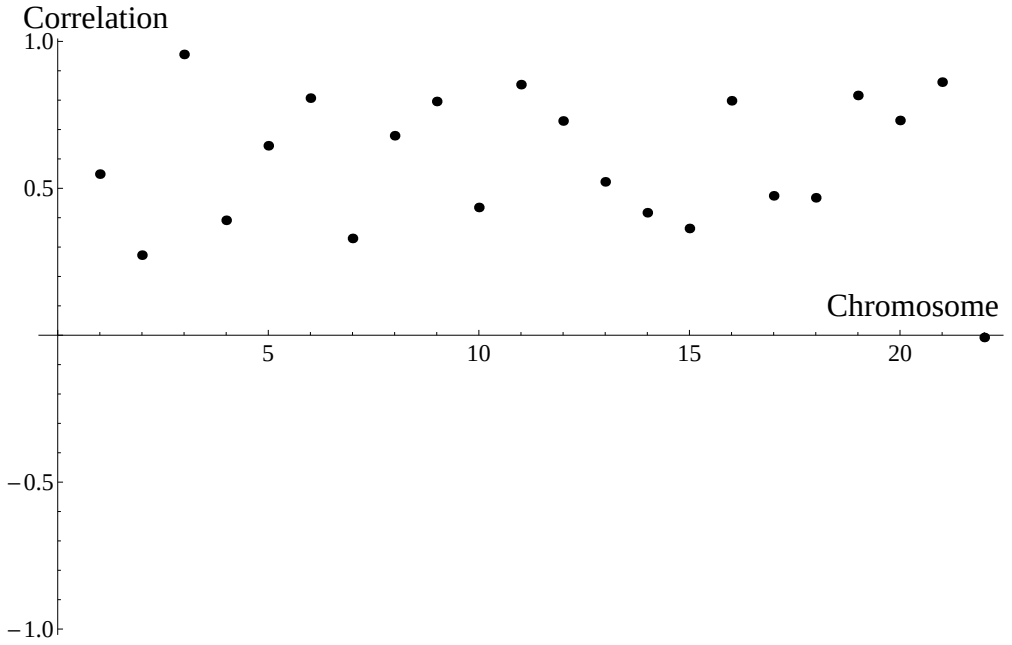
However, we will see in §7.3.13 that basing a potential on GC content alone is not sufficient for accurate estimation of CDS density.

- According to the weights in φ_{hs}, the expected GC content of a sequence is 58.4%, which is moderately high since it was shown in [88] that average GC content for a 100-kb segment of the human genome is between 35% and 60%. We calculate expected GC content by the formula

$$\sum_{w \in \mathcal{A}^3} \varphi_{\text{hs}}(w)(\mathrm{N}_G(w) + \mathrm{N}_C(w))/3,$$

where $\mathrm{N}_G(w)$ denotes the number of times the letter G appears in the word w (similarly for $\mathrm{N}_C(w)$). This supports the commonly held notion

Figure 7.5. Plot of Pearson correlation coefficient between the coding sequence density of the human genome and the topological pressure associated to the potential φ_{hs}.



that high GC content corresponds to high coding sequence density in the human genome [4, 44].

- The start codon (AUG) is weighted near zero. This may indicate that from a large scale perspective, start codons provide too weak a signal to utilize in estimating CDS density.

- The stop codons UGA, UAG, UAA exhibit a decreasing order of significance. This reflects the observation contained in [99] that UGA is utilized most frequently to terminate transcription in the human genome. Furthermore, this pattern of decreasing importance of stop codons reflects the alternate decoding of stop codons (see [24], [97], [116], etc.). In particular, the two stop codons that can be alternately transcribed (UGA and UAG into Selenocysteine and Pyrrolysine respectively) are weighted much more strongly than UAA.

- Codons made up of a single repeating nucleotide receive consistently low

weights. This can be explained by the presence of long repetitive regions in non-coding regions.

- We analyzed a number of physical properties associated to amino acids and codons (e.g. acidity, polarity, hydropathy, etc.), and found a weak (.293) but statistically significant ($p < .025$) correlation between the values of φ_{hs} and heat of combustion of the corresponding codons. We are not aware of any results in the literature which would give a theoretical basis for this observation.

- Synonymous codons may receive very different weightings. For example, among the codons which specify Glycine, GGC is strongly weighted but GGU, GGA and GGG are all weighted near zero.

- Of the five amino acids with four-fold degenerate sites, distinct codon bias is observed: each amino acid with a four-fold degenerate site shows a clear bias towards a nucleotide in the third site (with the exception of Proline where two particular codons are favored). This corresponds with the observations of previous comparative studies, for example [12, 86], where it was observed that there exists selectively driven codon usage at four-fold degenerate sites for mammals (with a weak bias towards C). Our study suggests that any of the four nucleotides may be favored in the third position (A for Val, G for Ala, C for Gly, U for Thr).

- Amino acids with twofold degenerate sites seem typically to carry similar weightings. For example, GAA and GAG both have negligible weightings, while UAU and UAC are both weighted quite strongly. The mean variance of the weighting at twofold degenerate sites was 4.7×10^{-5} while the mean variance over all amino acids was 1.6×10^{-4}.

- For most amino acids, either exactly one codon is weighted strongly (Leu, Val, Ser, Thr, Ala, His, Gly) (or at least more strongly than the others (His, Gln, Asn)), or no codons are weighted strongly (Phe, Ile, Met, Lys, Asp, Glu, Cys, Trp). A notable exception is Arginine where three out of its six synonymous codons are weighted strongly. This may suggest that Arginine has a particularly important role. This could reflect the

evolutionary pressure exerted on Arginine as observed in [56], where it is noted that Arginine has a much lower frequency of appearance than expected. Recently, in [60] it was shown that in yeast, preferential synonymous codon usage for Arginine greatly affects expression levels via influencing translational efficiency. Our results may indicate that a similar phenomenon occurs in the human genome, as this would be another explanation for the strong weighting of Arginine.

7.3.12 Selecting potentials using intron/exon density

Many single sequence techniques for measuring the coding potential of DNA sequences are based upon frequencies of codons or n-mers in known intronic and exonic regions [2, 17, 18, 52]. We can use this principle to write down potentials $\varphi^i_{\text{intron}}$ and φ^i_{exon} which are based simply on the frequency of codons in the intron (or exon) sequences.

More precisely, we let Introns(i) denote the collection of all segments of chromosome i which correspond to known intron regions. For each $w \in$ Introns(i), we let $\mathrm{N}_v(w)$ denote the number of times a given codon v appears in w, and we note that the total number of codons (with overlap) in w is $|w| - 2$. We define a potential $\varphi^i_{\text{intron}}$ by assigning each codon a weight by the formula

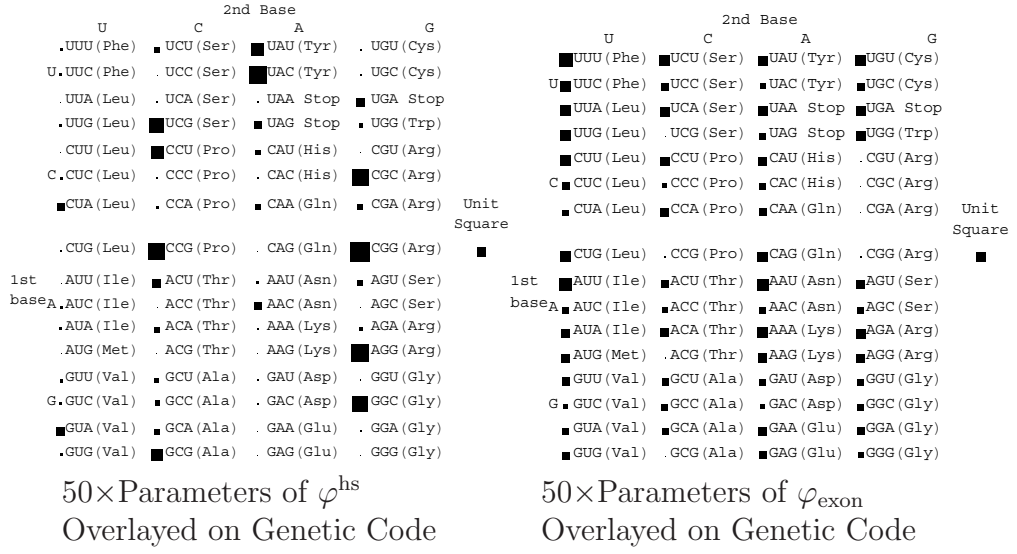
$$\varphi^i_{\text{intron}}(v) := \sum_{w \in \text{Introns}(i)} \frac{\mathrm{N}_v(w)}{|w| - 2}. \tag{7.8}$$

We define potentials φ^i_{exon} analogously, using the frequencies of codons that appear in known exon regions of chromosome i. Finally, as in definition 7.3.10, let $\varphi^0_{\text{exon}}(w) := \text{median}\{\varphi^1_{\text{exon}}(w), \ldots, \varphi^{24}_{\text{exon}}(w)\}$ and define $\varphi_{\text{exon}} := \frac{\varphi^0_{\text{exon}}}{\|\varphi^0_{\text{exon}}\|}$.

As one would expect, the pressure taken with respect to the potentials $\varphi^i_{\text{intron}}$ (resp. φ^i_{exon}) tends to have significant negative (resp. positive) correlation with the coding sequence density: see figure 7.7. The mean correlation between $P(i, 8, x, \varphi^i_{\text{intron}})$ and CDS$(i, 8, x)$ was $-.531$. The mean correlation between $P(i, 8, x, \varphi^i_{\text{exon}})$ and CDS$(i, 8, x)$ was .376. While this clearly shows a correlation, it is significantly weaker than that obtained using the potentials $\varphi^i_{\max}$. We conclude that potentials which are based simply on frequencies of

occurrence of codons in intron/exon regions are useful to an extent, but that more sophisticated potentials, such as φ_{hs}, yield much better results. See figure 7.6 for a comparison of the potentials φ_{hs} and φ_{exon}.

Figure 7.6. Visualization of parameter values of potentials. The area of a square is equal to the corresponding potential value.



7.3.13 Selecting potentials to detect GC content

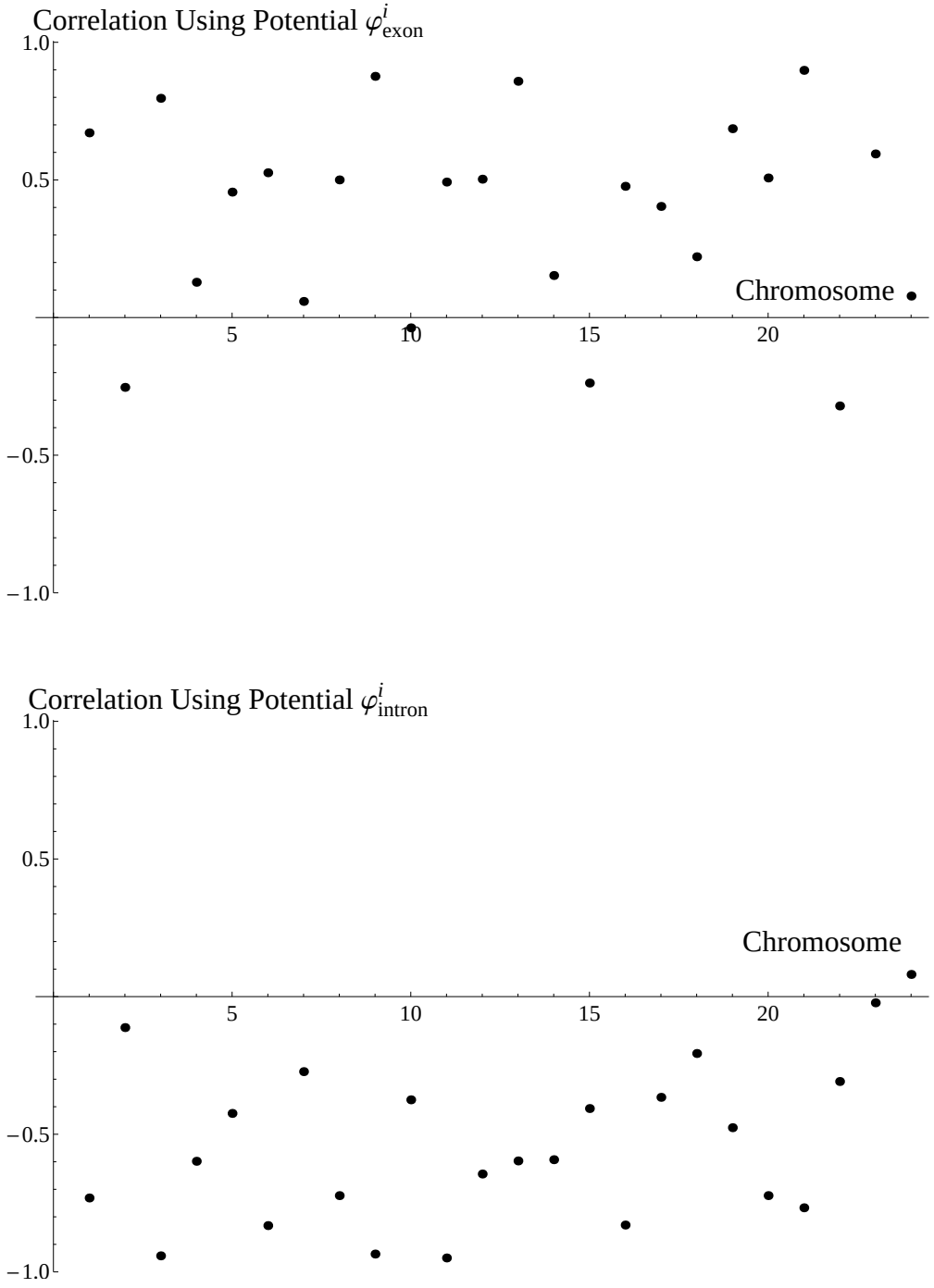
We investigate the pressure of the family of potentials introduced in §7.2.9:

$$\varphi_t = \mathbf{1}_A + \mathbf{1}_T + t(\mathbf{1}_G + \mathbf{1}_C),$$

where $t > 0$. It is a commonly held notion that high GC content corresponds to high coding sequence density in the human genome (see §7.3.11). We give evidence that the link between GC content and CDS density is significant but weak.

For chromosome 1, we find that as t varies, the largest correlation between $P^{\text{hs}}(1, 8, x, \varphi_t)$ and $\text{CDS}(1, 8, x)$ is .138. This maximum is attained (uniquely) when $t = 10.308$. Over all the chromosomes, the maximum correlation of $P^{\text{hs}}(i, 8, x, \varphi_t)$ and $\text{CDS}(i, 8, x)$ has a statistically significant ($p < .0005$) mean of 0.121 with a variance of .00359. This maximum is achieved for a mean parameter value of $t = 10.306$ with a variance of 15.780. The outliers were

Figure 7.7. Correlation between pressure and coding sequence distribution



chromosome 18, which achieves maximal correlation at $t = 21.246$, and chromosome 15, which achieves maximal correlation at $t = 0$. Excluding these

two chromosomes gives essentially the same mean ($t = 10.273$), but a much improved variance of 4.98. These results indicate that potentials based on GC content give a weak positive correlation with CDS density. However, the much higher correlation obtained when using the potential φ_{hs} indicates that considering GC content alone is far from optimal in CDS density estimation.

7.4 Application of the potential φ_{hs} to the mouse genome

We further illustrate the biological significance of the potential φ_{hs} by examining the correlation between the coding sequence density of the mouse genome and the topological pressure associated to the potential φ_{hs}. Following the setup of section 7.3.6, we retrieve the mouse genome (build mm9, NCBI build 37) from the UCSC database [38] via Galaxy [42], extract the RefSeq genes, and then define $\text{CDS}^{\text{mm}}(i, n, x)$ and $P^{\text{mm}}(i, n, x)$ for the mouse autosomes. The results of the correlation between the topological pressure $P^{\text{mm}}(i, 8, x, \varphi_{\text{hs}})$ associated to the potential obtained from the human genome and the coding sequence density of the mouse genome $\text{CDS}^{\text{mm}}(i, 8, x)$ are contained in figure 7.8. We see a strong positive correlation. This indicates that codon usage in the mouse is similar to codon usage in humans, and gives further evidence that the potential φ_{hs} genuinely encodes biological information relevant to detecting coding sequence distributions, even across different species.

We follow the maximization procedure outlined in §7.3.6 to obtain potentials $\varphi^i_{\text{mm,max}}$ that maximize the correlation between $P^{\text{mm}}(i, 8, x, \varphi)$ and $\text{CDS}^{\text{mm}}(i, 8, x)$ with respect to φ. Following section §7.3.9, we average over the potentials $\varphi^i_{\text{mm,max}}$ to obtain a 'canonical' potential for the mouse, which we denote by φ_{mm}. In figure 7.9, we include a visualization of the parameter values for φ_{mm} and φ_{hs} to demonstrate the similarities between them. It will be an interesting project to carry out this procedure for a much larger collection of species and investigate the similarities and differences between the potentials that are selected for each species.

Figure 7.8. Plot of Pearson correlation coefficient between the coding sequence density of the *Mus Musculus* genome and the topological pressure associated to the potential φ_{hs}.

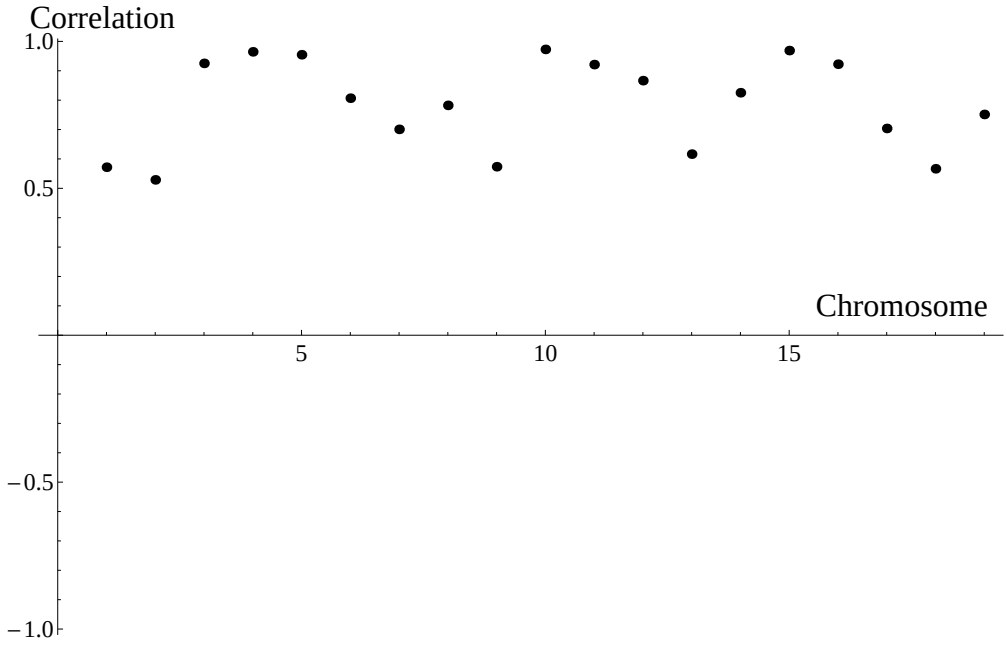
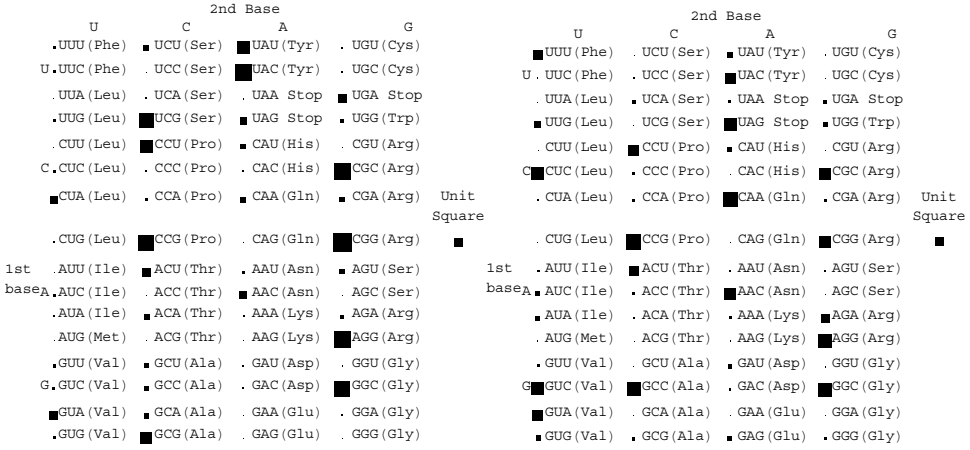


Figure 7.9. Plot of parameter values of median potentials for human and mouse respectively. The area of a square is equal to the corresponding potential value.



50× Parameters of φ_{hs} Overlayed on Genetic Code

50× Parameters of $\varphi_{\text{mm,max}}$ Overlayed on Genetic Code

7.5 Equilibrium measures and DNA

As mentioned in the introduction, an important area of research is to develop single sequence measures that effectively distinguish between short coding se-

quences and short non-coding sequences. Here, we utilize ergodic theory to develop such a measure.

Given a locally constant function ψ on $\mathcal{A}$, the theory of thermodynamic formalism gives us a means of selecting a Markov measure μ_ψ, known as the equilibrium measure for ψ. We adapt this theory to the case of finite sequences. The measure thus obtained reflects the properties of the function ψ. We carry out this procedure for our canonical potential φ_{hs} and obtain a measure that is effective for the analysis of relatively short segments of DNA sequences. We give a brief review of the theory of equilibrium measures in the case of potentials which depend on 3 symbols, and show that the measure thus obtained is meaningful in the finite setting also. We then give numerical results to demonstrate that our measure can distinguish between coding and non-coding sequences with a reasonably high probability of success.

7.5.1 Constructing Markov measures from potentials

The results in this subsection are due, almost in their entirety, to Daniel J. Thompson. We are primarily concerned with functions that depend on 3 symbols, using the alphabet $\mathcal{A} = \{A, C, T, G\}$. That is, we consider potentials $\psi = \log \varphi$, where

$$\varphi = \sum_{w \in \mathcal{A}^3} t_w \mathbf{1}_w, \tag{7.9}$$

so that each $t_w > 0$ is a parameter associated to the word $w \in \mathcal{A}^3$. We review how this this function defines a Markov measure with memory 2 on Σ. The presentation here is a special case of more general expositions given in [3, 9, 64, 77, 78, 107]. Let $\mathcal{B} = \{A, C, G, T\}^2$. Enumerate $\mathcal{B}$ by some natural ordering. For example, let

$$w_1 = AA, w_2 = AC, w_3 = AG, w_4 = AT, w_5 = CA, \ldots, w_{16} = TT$$

Define a $1 - 0$ square matrix S of dimension 16 as follows. Let $S_{ij} = 1$ if and only if the second letter in w_i is the same as the first letter in w_j. Otherwise, set $S_{ij} = 0$.

We now use the potential ψ to define a non-negative matrix M of dimension 16 as follows. Let $g_{ij} = \log t_w$, where if $w_i = IJ$, and $w_j = JK$, then $w = IJK$. Let $g(i,j) = 0$ if the second letter in w_i is not the same as the first letter in w_j. We define M by

$$M_{ij} = S_{ij} e^{g(i,j)}. \tag{7.10}$$

The Perron-Frobenius theorem gives a maximal eigenvalue $\lambda > 0$ and a strictly positive vector r such that

$$Mr = \lambda r.$$

Now define a matrix P of dimension 16 by

$$P_{ij} = \frac{M_{ij} r_j}{\lambda r_i}. \tag{7.11}$$

It is easy to check that P_{ij} is a stochastic matrix and that there is a unique probability vector p so that $pP = p$. More explicitly, p_i is given by normalizing the vector $l_i r_i$, where l is a strictly positive left eigenvector for M. For $a, b, c \in \mathcal{A}$, let $p(ab) = p_i$ when $ab = w_i$, and let $P(ab, bc) = P_{ij}$ when $ab = w_i$ and $bc = w_j$. We use the pair (p, P) to define a measure as follows.

Definition 7.5.2. *We define a probability measure μ_ψ on $\mathcal{A}$, or $\mathcal{A}^n$ for any fixed $n \geq 3$, by the formula*

$$\mu_\psi([x_1 \dots x_k]) = p(x_1 x_2) P(x_1 x_2, x_2 x_3) P(x_2 x_3, x_3 x_4) \dots P(x_{k-2} x_{k-1}, x_{k-1} x_k).$$

We call the measure μ_ψ the equilibrium measure for ψ.

7.5.3 Properties of the equilibrium measure

We recall the classical theory from dynamical systems which explains the importance of μ_ψ. First, we recall the definition of topological pressure for the full shift.

Definition 7.5.4. *The topological pressure of ψ on the full shift Σ over an*

alphabet $\mathcal{A}$ is defined to be:

$$P(\Sigma, \psi) = \lim_{n\to\infty} \frac{1}{n} \log \left(\sum_{u\in\mathcal{A}^n} \exp \sum_{i=0}^{n-1} \psi(\sigma^i u) \right).$$

The following result gives the fundamental relationship between pressure and invariant measures [78, 107].

Theorem 7.5.5 (Variational Principle)**.** *$P(\Sigma, \psi) = \sup_m \left\{ h_m + \int \psi dm \right\}$, where the supremum is taken over all σ-invariant probability measures on Σ, and h_m denotes the measure theoretic entropy, given by*

$$h_m = \lim_{n\to\infty} -\frac{1}{n} \sum_{w\in\mathcal{A}_n} m([w]) \log m([w]).$$

The variational principle illustrates the trade off between structure and complexity which is detected by pressure. Pressure effectively balances the inherent randomness in a sequence while still reflecting the emphasis encoded by the potential φ. Pressure simultaneously maximizes entropy (which is maximized by the uniform measure) and the average value of the potential (the integral itself is maximized by a Dirac measure). A measure achieving the supremum in the variational principle is called an *equilibrium measure* for ψ. The following result, proved in [78, §4], tells us that the measure constructed in the previous section is indeed an equilibrium measure.

Theorem 7.5.6. *The Markov measure $\mu = \mu_\psi$ is the unique equilibrium measure for ψ and*

$$P(\Sigma, \psi) = h_\mu + \int \psi d\mu = \log \lambda,$$

where λ is the Perron-Frobenious eigenvalue of the matrix (7.10)*.*

The relationship between ψ and μ_ψ is captured by the *Gibbs property*, established in [9, 77].

Theorem 7.5.7 (Gibbs property)**.** *For $\psi = \log \varphi$ defined as in* (7.9) *and any $w \in \mathcal{A}^n$,*

$$\mu_\psi([w]) \asymp \exp\{-nP(\Sigma, \psi) + \sum_{i=1}^{n-2} \psi(w_i^{i+2})\},$$

where $a_n \asymp b_n$ means there exists a constant $C > 1$ so that $C^{-1} \le a_n/b_n \le C$ for all n.

Thus, if we normalize ψ so that $P(\Sigma, \psi) = 0$ (which is done by taking a suitable multiple of φ), then

$$\mu_\psi([w]) \asymp \prod_{i=1}^{n-2} \varphi(w_i^{i+2}). \tag{7.12}$$

7.5.8 Relationship between the equilibrium measure and pressure for finite sequences

We apply the theory developed in §7.5.3 to finite sequences. The proof of the following result is similar to that of [107, Theorem 7.30].

Theorem 7.5.9. *When ψ depends on 3 symbols, $P_{max}(\psi, n) = \log \|M^{n-2}\|^{1/n}$, where M is the matrix constructed in (7.10). Since $\|M^{n-2}\|^{1/n}$ converges to λ exponentially fast as $n \to \infty$, then for large n, $P_{max}(\psi, n)$ is very close to $\log \lambda$.*

Proof. For $u \in \mathcal{A}^n$,

$$\exp(\sum_{i=0}^{n-3} \varphi(\sigma^i u)) = \prod_{i=0}^{n-3} e^{\varphi(\sigma^i u)} = \prod_{i=0}^{n-3} e^{g(u_{i+1}u_{i+2}, u_{i+2}u_{i+3})}.$$

Thus,

$$\begin{aligned} \sum_{u \in \mathcal{A}^n} \exp \sum_{i=0}^{n-3} \psi(\sigma^i u) &= \sum_{i_1, \ldots, i_{n-1}=1}^{16} \prod_{j=1}^{n-2} M_{i_j i_{j+1}} \\ &= \sum_{i_1=1}^{16} \sum_{i_{n-1}=1}^{16} (M^{n-2})_{i_1, i_{n-1}} = \|M^{n-2}\|. \end{aligned}$$

□

Thus, the number $\log \lambda$ is still important for finite sequences. The formula (7.12) reveals the meaning of the measure μ_ψ. Sequences which have a relatively high frequency of words $w \in \mathcal{A}^3$ where t_w is large, and a relatively small frequency of words $w \in \mathcal{A}^3$ where t_w is small, will have relatively large

measure. This gives a theoretical underpinning for using μ_ψ to predict coding sequence density.

7.5.10 An equilibrium measure for CDS density estimation

We show that the equilibrium measure has practical applications to distinguishing between coding and non-coding DNA sequences. Recall that in §7.3 we found a potential φ_{hs} for which the pressure of human DNA segments of length $65,536$bp has strong positive correlation with the coding sequence density. We now show that the equilibrium measure associated to $\log \varphi_{\text{hs}}$, which we denote by μ_{hs}, can distinguish between coding and non-coding DNA sequences with a reasonable degree of success.

The advantage of using the measure μ_{hs} rather than pressure associated to φ_{hs} is that the measure is effective in analyzing relatively short DNA sequences (10bp-5000bp). Indeed, when ψ depends on 3 symbols, pressure is only defined for sequences of length at least $4^4 - 4 = 251$, and only becomes an effective tool for much longer sequences where the noise inherent in the calculation of pressure is effectively suppressed. While the equilibrium measure is a cruder tool than the pressure, it is nevertheless effective for analyzing shorter sequences where the pressure is unavailable.

To demonstrate this phenomena, we show that the measure μ_{hs} can partially distinguish between a randomly selected assortment of intron and exon sequences that are more than an order of magnitude shorter than the sequences on which pressure was evaluated. We randomly select $5,000$ intron sequences and $5,000$ exon sequences from Chr(1), each of length $5,000$bp. These sequences are completely un-preprocessed: no information such as ORF's, stop/start codons or repeat masking is utilized. As expected, the measure μ_{hs} reflects the properties of the potential φ_{hs}: exon sequences are typically weighted more heavily than intron sequences. This is demonstrated by figure 7.10, which shows the histogram of $\log(\mu_{\text{hs}})$ evaluated on the test sequences. We also include the ROC curve (true positive rate vs. false positive rate) associated to μ_{hs}. The area under the curve (AUC) is 0.722.

Figure 7.10. Histogram of $\log(\mu_{\text{hs}})$ evaluated on the test set of introns and exons

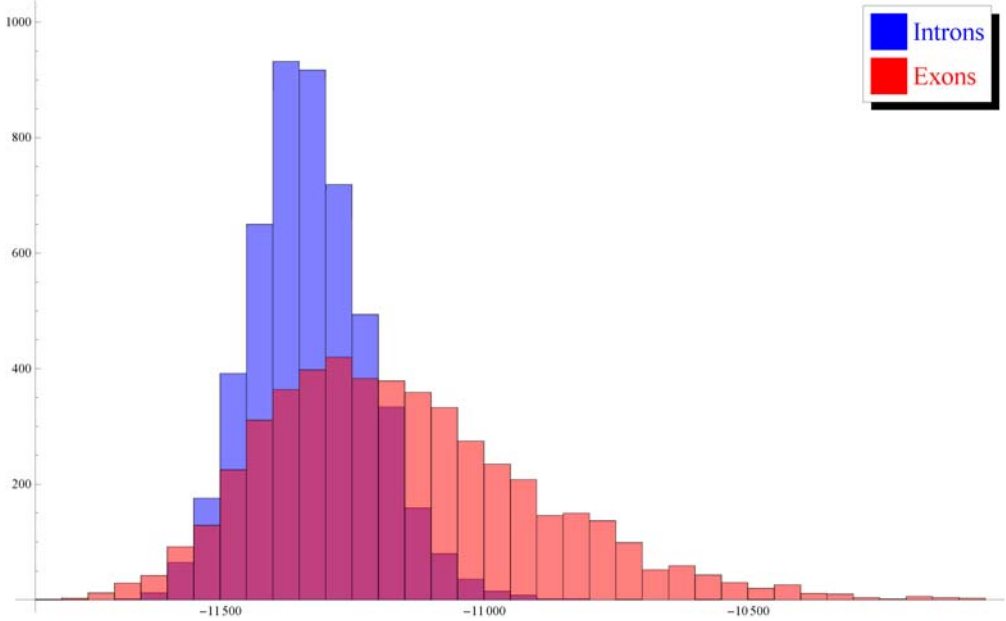
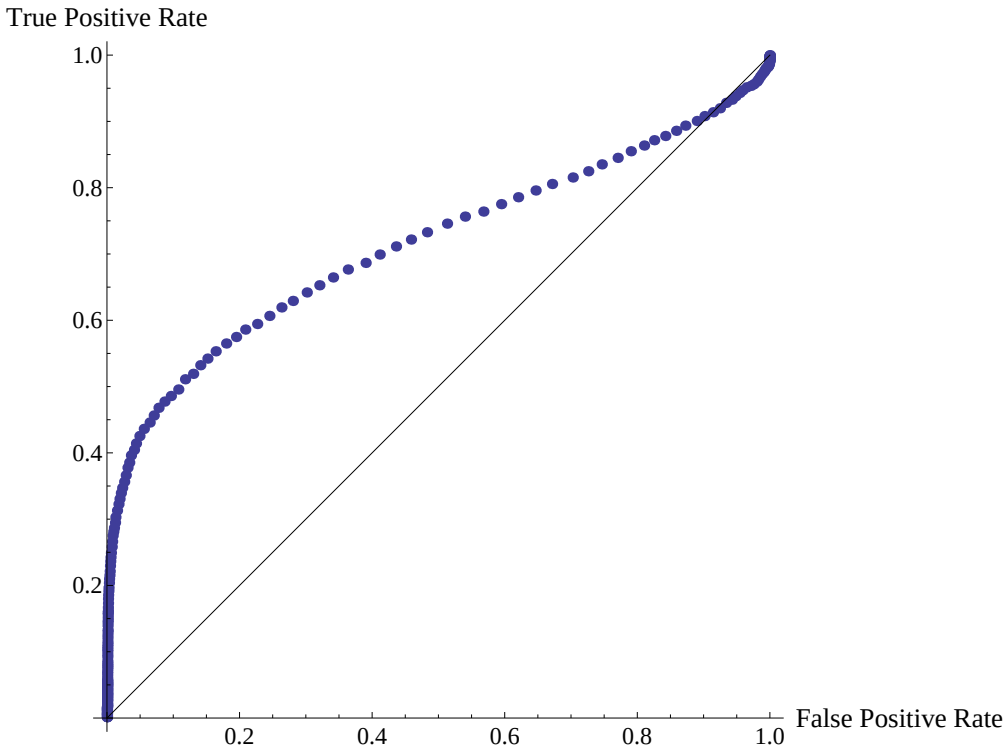


Figure 7.11. ROC curve associated to μ_{hs}.



7.6 Conclusion

We have introduced a definition of topological pressure for finite sequences inspired by, and related to, the classical definition from ergodic theory. We

have applied this definition to DNA sequences in four distinct fashions. First, we obtained a potential that effectively estimates the distribution of coding sequences across the human genome. Second, we gave a detailed analysis of this potential to give new evidence about which codons are most important in coding sequence density estimation. Thus, pressure can be used as a tool for the study of synonymous codon usage. Our analysis effectively measures which codons are most important and not simply most frequently appearing. Third, we used topological pressure to compare coding sequence density in the human and the mouse genome, giving evidence via pressure that codon usage is similar across both species. Lastly, we derived the equilibrium measure associated to a particular potential and showed that this can be used to distinguish between relatively short reads of coding and non-coding sequences.

This study has indicated that topological pressure may help elucidate the nuanced problems of mammalian codon bias. Since topological pressure does not rely on a particular statistical perspective but is motivated by a rigorous implementation of a well developed mathematical theory, we expect that our approach will yield many further applications in genomic analysis in the future. We expect that the inclusion of pressure in comparative studies will contribute to the understanding of the relationship between codon bias and gene expression levels. Furthermore, the ability of equilibrium measures to succinctly encapsulate information obtained on very large scales indicates the usefulness of pressure for the development of measures of coding potential for short DNA sequences.

Bibliography

[1] M. A. Zaks. Multifractal Fourier spectra and power-law decay of correlations in random substitution sequences. *Physical Review E*, 65(1):1–5, 2001.

[2] H. Akashi. Gene expression and molecular evolution. *Current Opinion in Genetics & Development*, 11(6):660–666, 2001.

[3] V. Baladi. *Positive transfer operators and decay of correlations*, volume 16. World Scientific, 2000.

[4] G. Bernardi. The isochore organization of the human genome. *Annual review of genetics*, 23:637–661, Dec. 1989.

[5] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, New York, NY., 1968.

[6] F. Blanchard, A. Maass, and A. Nogueira. *Topics in symbolic dynamics and applications*. Cambridge Univ Pr, Cambridge, 2000.

[7] D. Blankenberg, J. Taylor, I. Schenk, J. He, Y. Zhang, M. Ghent, N. Veeraraghavan, I. Albert, W. Miller, K. Makova, R. Hardison, and A. Nekrutenko. A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly. *Genome research*, 17(6):960–964, 2007.

[8] D. Blankenberg, G. Von Kuster, N. Coraor, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molec. Biol.*, 19(10):1–20, 2010.

[9] R. Bowen. *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, volume 470 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin-New York, 1975.

[10] R. K. Bradley and I. Holmes. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics (Oxford, England)*, 23(23):3258–62, Dec. 2007.

[11] R. Cartwright. Problems and solutions for estimating indel rates and length distributions. *Molecular biology and evolution*, 26(2):473–80, Feb. 2009.

[12] J. V. Chamary and L. D. Hurst. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Molecular biology and evolution*, 21(6):1014–23, June 2004.

[13] J. V. Chamary and L. D. Hurst. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome biology*, 6(9):R75, Jan. 2005.

[14] J. V. Chamary, J. L. Parmley, and L. D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature reviews. Genetics*, 7(2):98–108, Feb. 2006.

[15] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3480–5, Mar. 2004.

[16] A. Colosimo and A. De Luca. Special factors in biological strings. *Journal of theoretical biology*, 204(1):29–46, May 2000.

[17] J. M. Comeron and M. Aguadé. An evaluation of measures of synonymous codon usage bias. *Journal of molecular evolution*, 47(3):268–74, Sept. 1998.

[18] T. M. Creanza, D. S. Horner, A. D'Addabbo, R. Maglietta, F. Mignone, N. Ancona, and G. Pesole. Statistical assessment of discriminative features for protein-coding and non coding cross-species conserved sequence elements. *BMC bioinformatics*, 10 Suppl 6:S2, Jan. 2009.

[19] M. Crochemore and R. Vérin. Zones of low entropy in genomic sequences. *Computers & chemistry*, 23(3-4):275–82, July 1999.

[20] M. Denker and H. Sato. Sierpiski gasket as a Martin boundary. II. The intrinsic metric. *Publications of the Research Institute for Mathematical Sciences*, 35(5):769–794, 1999.

[21] M. Denker and H. Sato. Sierpinski gasket as a Martin boundary I: Martin kernels. *Potential analysis*, 14(3):211–232, 2001.

[22] M. Denker and H. Sato. Reflections on Harmonic Analysis of the Sierpiski Gasket. *Mathematische Nachrichten*, 241(1):32–55, July 2002.

[23] J. Doob. Discrete potential theory and boundaries. *Indiana University mathematics journal*, 8(3):433–458, 1959.

[24] V. Doronina and J. Brown. When nonsense makes sense and vice versa: Noncanonical decoding events at stop codons in eukaryotes. *Molecular Biology*, 40(4):654–663, July 2006.

[25] M. dos Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17):5036–44, Jan. 2004.

[26] F. Durand, B. Host, and C. Skau. Substitutional dynamical systems, Bratteli diagrams and dimension groups. *Ergodic Theory and Dynamical Systems*, 19(4):953–993, Aug. 1999.

[27] E. Dynkin and M. Malyutov. Random walks on groups with a finite number of generators. *Soviet Math. Doklady*, 2:399–402, 1961.

[28] E. B. Dynkin. Boundary theory of Markov processes (the discrete case). *Uspehi Mat. Nauk*, 24(2):1–42, 1969.

[29] K. P. et. al. Ensemble genomes: Extending ensembl across the taxonomic space. *Nucleic Acids Research*, 38(suppl. 1):D563–D569, 2010.

[30] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 48–57. Society for Industrial and Applied Mathematics, 1995.

[31] A. Fedorov, S. Saxonov, and W. Gilbert. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic acids research*, 30(5):1192–7, Mar. 2002.

[32] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[33] S. Ferenczi and C. Mauduit. Substitution dynamical systems: algebraic characterization of eigenvalues. *Ann. Sci. Ecole Norm. Sup.*, 4(29):519–533, 1996.

[34] J. W. Fickett and C. S. Tung. Assessment of protein coding measures. *Nucleic acids research*, 20(24):6441–50, Dec. 1992.

[35] W. M. Fitch and T. F. Smith. Optimal sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America*, 80(5):1382–1386, Mar. 1983.

[36] R. Fleißner, D. Metzler, and A. von Haeseler. Can one estimate distances from pairwise sequence alignments? In E. Bornberg-Bauer, U. Rost, J. Soye, and M. Vingron, editors, *Proceedings of the German conference on bioinformatics*, pages 89–95, Heidelberg, Berlin, 2000. Logos.

[37] N. Fogg. *Substitutions in Dynamics, Arithmetics and Combinatorics*. Springer, Berlin, 2002.

[38] P. Fujita, B. Rhead, A. Zweig, and K. W. Hinrichs, AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D. The UCSC Genome Browser database: update 2011. *Nucleic acids research*, 39(suppl 1):D876–D882, 2011.

[39] A. Gabrielian and A. Bolshowy. Sequence Complexity and DNA Curvature. *Computers & Chemistry*, 23:263–274, 1999.

[40] F. Gao and C.-T. Zhang. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics (Oxford, England)*, 20(5):673–81, Mar. 2004.

[41] I. Gheorghiciuc. The subword complexity of a class of infinite binary words. *Advances in Applied Mathematics*, 39(2):237–259, Aug. 2007.

[42] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 25(11):R86—-R99, 2010.

[43] J. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, 2006.

[44] R. Guigó and J. W. Fickett. Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *Journal of molecular biology*, 253(1):51–60, Oct. 1995.

[45] Y. G. Gursky and R. S. Beabealashvilli. The increase in gene expression induced by introduction of rare codons into the C terminus of the template. *Gene*, 148(1):15–21, Oct. 1994.

[46] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174, 1985.

[47] R. Hershberg and D. A. Petrov. Selection on codon bias. *Annual review of genetics*, 42(iv):287–99, Jan. 2008.

[48] G. Hunt. Markoff chains and Martin boundaries. *Illinois J. Math.*, 4(3):313–340, 1960.

[49] T. Jukes and C. Cantor. Evolution of protein molecules. In H. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, NY., 1969.

[50] V. Kaimanovich. Measure-theoretic boundaries of Markov chains, 0-2 laws and entropy. In *Proceedings of the Conference on Harmonic Analysis and Discrete Potential Theory (Frascati*, pages 0–2. Citeseer, 1991.

[51] K. Karamanos, I. Kotsireas, A. Peratzakis, and K. Eftaxias. Statistical compressibility analysis of DNA sequences by generalized entropy-like quantities: towards algorithmic laws for biology? In *Proceedings of the 6th WSEAS International Conference on Applied Informatics and Communications*, volume 2006, pages 481–491. World Scientific and Engineering Academy and Society (WSEAS), 2006.

[52] S. Karlin, J. Mrázek, and A. Campbell. Codon usages in different gene classes of the Escherichia coli genome. *Molecular microbiology*, 29(6):1341–1355, 1998.

[53] J. Kemeny, J. Snell, and A. Knapp. *Denumerable Markov Chains*. Springer-Verlag, New York, NY., 2nd edition, 1965.

[54] J. Kim and S. Sinha. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics (Oxford, England)*, 23(3):289–97, Feb. 2007.

[55] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 6:111–120, 1980.

[56] J. King. Non-Darwinian evolution. *Science*, 164:788–798, 1969.

[57] O. Kirillova. Entropy concepts and DNA investigations. *Physics Letters A*, 274(5-6):247–253, 2000.

[58] D. Koslicki. Topological entropy of DNA sequences. *Bioinformatics*, 27(8):1061–1067, Feb. 2011.

[59] N. Larsson. *Structures of string matching and data compression*. PhD thesis, 1999.

[60] D. P. Letzring, K. M. Dean, and E. J. Grayhack. Control of translation efficiency in yeast by codon-anticodon interactions. *RNA (New York, N.Y.)*, 16(12):2516–28, Dec. 2010.

[61] W. Li. *Molecular Evolution*. Sunderland, Chicago, 1997.

[62] M. F. Lin, A. N. Deoras, M. D. Rasmussen, and M. Kellis. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS computational biology*, 4(4):e1000067, Apr. 2008.

[63] M. F. Lin, I. Jungreis, and M. Kellis. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics (Oxford, England)*, 27(13):i275–i282, July 2011.

[64] D. Lind and B. Marcus. *An introduction to symbolic dynamics and coding*. Cambridge University Press, 1995.

[65] P. Lio and N. Goldman. Models of molecular evolution and phylogeny. *Genome research*, 8(12):1233–1244, 1998.

[66] G. Lunter. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics (Oxford, England)*, 23(13):i289–296, July 2007.

[67] G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome research*, 18(2):298–309, Feb. 2008.

[68] R. Mantegna, S. Buldyrev, A. Goldberger, S. Havlin, C. Peng, M. Simons, and H. Stanley. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E*, 52(3):2939–2950, 1995.

[69] D. Metzler. Statistical alignment based on fragment insertion and deletion models. *Bioinformatics*, 19(4):490–499, Mar. 2003.

[70] D. Metzler, R. Fleißner, A. Wakolbinger, and A. von Haeseler. Assessing variability by joint sampling of alignments and mutation rates. *Journal of molecular evolution*, 53(6):660–9, Dec. 2001.

[71] I. Miklós, G. Lunter, and I. Holmes. A "long indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529, 2004.

[72] I. Miklós, A. Novák, R. Satija, R. Lyngsø, and J. Hein. Stochastic models of sequence evolution including insertion-deletion events. *Statistical methods in medical research*, 18(5):453–85, Oct. 2009.

[73] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput. J.*, 7:308–313, 1965.

[74] P. Ney and F. Spitzer. The Martin boundary for random walk. *Transactions of the American Mathematical Society*, 121(1):116–132, 1966.

[75] J. Novembre. Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution*, 19(8):1390, 2002.

[76] D. Ornstein and B. Weiss. Entropy is the only finitely observable invariant. *Journal of Modern Dynamics*, 1(1):93–105, 2007.

[77] W. Parry and M. Pollicott. *Zeta functions and the periodic orbit structure of hyperbolic dynamics*. Number 187-188 in Astérisque. Soc. Math. France, 1990.

[78] W. Parry and S. Tuncel. *Classification problems in ergodic theory*, volume 67 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1982. Statistics: Textbooks and Monographs, 41.

[79] J. Peyrière. Random beadsets and birth processes with interaction. *IBM Research Report RC-7417*, pages 1–19, 1978.

[80] J. Peyrière. Process of birth interaction with the neighbors, evolution of charts. *Ann. Inst. Fourier. Grenoble*, 31(4):187–218, 1981.

[81] J. Peyriére. Substituions aléatoires itérées. *Seminaire de Théorie des Nombres de Bordeaux*, 17:1–9, 1981.

[82] J. Peyrière. Frequence des motifs dans les suites doubles invariantes par une subsitution. *Ann. sc. math. Québec*, 11(1):133–138, 1987.

[83] J. B. Plotkin and G. Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1):32–42, Jan. 2011.

[84] M. Queffélec. *Substitution Dynamical Systems - Spectral Analysis, Lecture Notes in Mathematics, 1294*. Springer-Verlag, Berlin, 1980.

[85] A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 547, pages 547–561, 1961.

[86] A. M. Resch, L. Carmel, L. Mariño Ramírez, A. Y. Ogurtsov, S. A. Shabalina, I. B. Rogozin, and E. V. Koonin. Widespread positive selection in synonymous sites of mammalian genes. *Molecular biology and evolution*, 24(8):1821–31, Aug. 2007.

[87] D. Revuz. *Markov Chains*. North-Holland, Amsterdam, 1975.

[88] J. Romiguier, V. Ranwez, E. J. P. Douzery, and N. Galtier. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, 20(8):1001–9, Aug. 2010.

[89] A. H. Rosenberg, E. Goldman, J. J. Dunn, F. W. Studier, and G. Zubay. Effects of consecutive AGG codons on translation in Escherichia coli, demonstrated with a versatile codon test system. *Journal of bacteriology*, 175(3):716–22, Feb. 1993.

[90] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, Oct. 2007.

[91] A. Schmitt and H. Herzel. Estimating the entropy of DNA sequences. *Journal of theoretical biology*, 188(3):369–377, 1997.

[92] W. Seffens and D. Digby. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic acids research*, 27(7):1578–84, Apr. 1999.

[93] E. Senata. *Non-negative matrices and Markov chains*. Springer, Berlin, 2nd edition, 2006.

[94] C. E. Shannon. The mathematical theory of communication. 1963. *M.D. computing : computers in medical practice*, 14(4):306–17, 1948.

[95] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, and Others. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

[96] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[97] T. Stadtman. Selenocysteine. *Annual review of biochemistry*, 65:83–100, 1996.

[98] H. Stanley. Scaling features of noncoding DNA. *Physica A*, 273:1–18, 1999.

[99] J. Sun, M. Chen, J. Xu, and J. Luo. Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *Journal of molecular evolution*, 61(4):437–44, Oct. 2005.

[100] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[101] J. Thorne and H. Kishino. Freeing phylogenies from artifacts of alignment. *Molecular Biology and Evolution*, 9(6):1148–1162, 1992.

[102] J. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33(2):114–124, 1991.

[103] J. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34(1):3–16, 1992.

[104] O. Troyanskaya, O. Arbell, Y. Koren, G. Landau, and A. Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679, 2002.

[105] S. Vinga and J. Almeida. Rényi continuous entropy of DNA sequences. *Journal of theoretical biology*, 231(3):377–388, 2004.

[106] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *Journal of molecular biology*, 235(1):1–12, Jan. 1994.

[107] P. Walters. *An Introduction to Ergodic Theory*, volume 79 of *Graduate Texts in Mathematics*. Springer, New York, 1982.

[108] K. Wargan. *S-adic dynamical systems and Bratelli diagrams*. PhD thesis, George Washington University, 2002.

[109] S. Washietl, S. Findeiss, S. a. Müller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, and N. Goldman. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA (New York, N.Y.)*, 17(4):578–94, Apr. 2011.

[110] S. Whelan, P. Li, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS in Genetics*, 17(5):262–272, 2001.

[111] M. Wilson and K. Makova. Evolution and survival on eutherian sex chromosomes. *PLoS Genet*, 5(7):11, 2009.

[112] M. Wilson and K. Makova. Genomic analyses of sex chromosome evolution. *Annual Review of Genomics and Human Genetics*, 10:333–354, 2009.

[113] W. Woess. *Random Walks on Infinite Graphs and Groups*. Cambridge University Press, Cambridge, 2000.

[114] W. Woess. *Denumerable Markov Chains*. European Math. Soc. Publishing House, Zürich, Switzerland, 2009.

[115] Wolfram. *Mathematica*. Wolfram Research, Inc., Champaign Illinois, 8.0 edition, 2010.

[116] F. Zinoni, A. Birkmann, and W. Leinfelder. Cotranslational insertion of selecocysteine into formate dehydrogenase from Escherichia coli directed by a UGA codon. *Proc. Natl. Acad. Sci.*, 84(May):3156–3160, 1987.

Vita

David Koslicki

David Koslicki was born in southern California in 1986 and grew up in central California and the Pacific Northwest. He was introduced to the pleasures of higher mathematics (via Greta Kocol and the continuum hypothesis) during his time at Skagit Valley College. After beginning his undergraduate degree at Washington State University during Fall 2004, David was fortunate to attend the MASS (Mathematics Advanced Studies Semesters) program at the Pennsylvania State University during Fall 2005. It was during this time that David was thoroughly convinced mathematics was his passion. Here too were the first seeds of this dissertation planted during Sergei Tabachnikov's class on Billiard (via cutting sequences and substitutions). Returning to Washington State University, David was fortunate to work with Judie McDonald, Bill Webb, and David Wollkind in research and teaching appointments. After graduating from WSU with a 4.0 GPA in the Fall of 2006, David began his Ph.D. studies at Penn State during the Fall of 2007.

While completing his thesis research, David has received departmental and university-wide teaching awards, as well had the opportunity to serve on the graduate teaching association committee. After beginning this dissertation with Manfred Denker, David was fortunate to assist with the dynamical systems conference held biennially at Georg-August-Universität in Göttingen, Germany.

David's hobbies include dressage, three-day eventing, and other equestrian sports as well as the occasional jaunt on his motorcycle.

For the time being, David's research interests are in adapting analytic tools from ergodic theory, symbolic dynamics, and thermodynamic formalism to the study of DNA sequences (with an emphasis on alignment-free techniques). The work represented in this dissertation is partially contained in the papers:

- D. Koslicki, Topological Entropy and DNA Sequences, *Bioinformatics*, Apr. 15; 27 (8): 1061-1067
- D. Koslicki, D. J. Thompson, Topological Pressure and Coding Sequence Density Estimation in the Human Genome, arXiv:1109.5999
- D. Koslicki, An Alignment-Free Indel Model of Molecular Evolution, arXiv:1102:1897