

RICE UNIVERSITY

**A Sequence-Based, Population Genetic Model of
Regulatory Pathway Evolution**

by

Troy Ruths

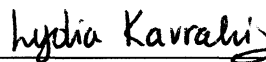
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

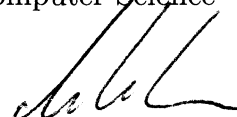
APPROVED, THESIS COMMITTEE:



Dr. Luay K. Nakhleh
Associate Professor, Chair,
Computer Science



Dr. Lydia E. Kavradi
Professor,
Computer Science



Dr. Michael Kohn
Assistant Professor,
Ecology and Evolutionary Biology

HOUSTON, TEXAS

DECEMBER 2010

Abstract

A Sequence-Based, Population Genetic Model of Regulatory Pathway Evolution

by

Troy Ruths

Complex phenotypes with genetic cause are understood through many processes, including regulatory pathways, but our evolutionary understanding of these critical structures is undermined by poor models which fail to preserve the underlying sequence structure and to incorporate population genetics. In response, this thesis builds a pathway model of evolution from its underlying sequence structure and validates it against a pertinent problem in genome evolution which uniquely leverage the developed model. Specically, my model preserves sequence characteristics through a novel data structure and pathway-level mutation and recombination rates which are functions of sequence properties. The utility of the model is validated with a study quantifying the advantages and disadvantages of expansive non-coding DNA regions on the establishment of optimal pathways. Because the model presented in this thesis rectifies many fundamental problems in previous models, it may serve as a critical tool for future work in pathway evolution.

Acknowledgements

This thesis was a collaborative effort and is the result of much hard work and dedication of several individuals. I would like to thank my committee members, especially my advisor Dr. Luay Nakhleh, whose guidance, expertise, and interest were invaluable to the success of this research.

I would also like to thank my wife Ivy for her soups, sandwiches, sanity, support, solicitude, and of course, for being my best friend. I want to thank my brothers for expediting my growth by pointing out my flaws, but also for serving as strong role models on this academic path. To my parents - thank you for always, always supporting my ambitions and raising me into the person I am today, because I know it wasn't easy.

Contents

1	Introduction	1
1.1	Pathways: sequence to function	2
1.2	The effects of non-coding DNA and population size on pathway evolution	5
1.3	Contributions of the thesis	8
2	Model	10
2.1	Population life cycle	10
2.2	The pathway genotype	12
2.3	Phenotype	15
2.4	Non-adaptive processes	20
2.5	Adaptive processes	28
2.6	Summary	34
3	Case study: The establishment of a novel regulatory pathway	36
3.1	Method	38
3.2	Results	42
3.3	Discussion	52
3.4	Conclusions	56
4	Discussion and conclusions	58

List of Figures

1.1	The genotype-phenotype space is a useful abstraction for understanding the phenotypic effects of mutations. In this diagram, circles are distinct genotypes connected by edges with denote mutation from one genotype to the next. Colors represent the phenotype. Robustness is the proportion of neighbors that share the same phenotype, and innovation is the proportion of m-step mutations that yield different phenotypes.	4
1.2	The strong correlation between non-coding sequence (ncDNA) per gene and the effective population size. Pearson correlation coefficient is -0.984. Estimates of population size for bacteria, unicellular eukaryotes, invertebrates, vertebrates, and land plants come from Lynch [1]. Uncertainty in both the estimates in population size and ncDNA are depicted by the gray band.	7
2.1	Life cycle model of the population simulator. The evolutionary processes are annotated for their order in each generation.	11
2.2	A cartoon that illustrates a matrix for a genetic pathway comprising of genes, transcription factors, and binding sites. The matrix representation denotes the binding sites and their affinity. This discretization of binding sites represents where a binding site may arise.	12
2.3	The protein trajectories as computed using the ODE equations generated by RENCO. The <i>continuous</i> phenotype is taken as the last time point once the trajectories have converged.	17
2.4	The time evolution of the example pathway using the <i>discrete</i> method. White signifies up-regulation, black down-regulation, and gray no regulation. The equilibrium concentrations for the gene-products can be found by averaging the cyclic behavior, which yields no regulation (the zero vector on top of the figure).	19

2.5	A recombination event occurs in the 2^{nd} critical binding region. The result of recombination in this binding region is detailed in Figure 2.6. Gene 2 is considered to be downstream and so is effected by the event. The arrows in the recombinants box denotes the crossover that occurred.	24
2.6	All possible results of a recombination event in a critical binding region (one of g in each promoter). By Assumption 1, at most one TFBS may be in each one of these regions. Note that one of the recombinants in scenario C violates Assumption 1.	25
2.7	The probability of each possible recombinant scenario if only one parent has a binding site (left) or both parents have a binding site (right). The binding site length n is equal to 10, so the leftmost x-value is $L/g = n$. The letters in the figure on the right refer to the scenarios mentioned in Figure 2.6.	27
3.1	Contour plots of the median establishment time (yellow is faster) for each population and genomic architecture parameter are shown for four of the nine examined phenotype/fitness functions. Each contour plot was generated using natural neighbor interpolation of the simulation results. On all plots, a noticeable diagonal band indicates the fastest establishment times (least squares fit with dashed line), which parallel the scaling of population and genomic architecture parameter found in nature (scatter points fit with solid black line). The comparison of the fit of the dashed line (fastest establishment times) and the scaling found in nature is shown in Table 3.1.	41
3.2	Median adaptation time (t_a) as a function of $2\mu_l N(1 + \alpha^{-1})$ for all phenotype/fitness functions. Triangles denote results from random fitness functions. The Pearson correlation of the fit is -.98 and is highly significant. The vertical dotted line denotes the point at which median establishment time has “bottomed out” relative to population mutation rate. Median adaptation time can be reduced either by increasing N or increasing the amount of non-coding DNA (decreasing α). . . .	45
3.3	The fixation model which incorporates drift, mutation, and selection. Mutations occur at a rate u , with a probability q of producing an optimal allele. The fitness difference between optimal and sub-optimal alleles is given by s	49

- 3.4 The results of the fixation model are shown on top. This plot explains the joint effect of Nu and the robustness of the optimal-neutral space, q , on the median fixation time. At $Nu \approx 1$, there is a significant increase in the time to fixation. This threshold can be shifted right by increases in s , the selection coefficient, or q . Solid, bold curves depict median fixation time for different phenotype/fitness functions. In this case, $2Nu = 2N(\mu_g + \mu_l) = 2N\mu_l(1 + \alpha^{-1})$. It is evident that results for these phenotype functions follow the pattern explained by the fixation model. Simulations for the fixation model are truncated at 10^8 50
- 3.5 The trough formed by median adaptation and fixation times across values of the population mutation rate ($2N\mu_l[1 + \alpha^{-1}]$). Minimal establishment time occurs at the intersection of these curves, which turns into a diagonal band when examined across values for N and α . The thick line denotes the median establishment time for *discrete-viable*. The several lines for fixation illustrate the strong effect of phenotype space properties on median fixation time, in comparison to the weak effect these properties have on adaptation. 53

List of Tables

3.1	For each phenotype/fitness function, a log-linear fit was calculated for the minimal median establishment times with respect to α . The log-linear fit follows $A\alpha^I$, where A is the amplitude and I is the index. The log of the amplitude and the index are given in the table, along with the Pearson correlation its corresponding p-value. The same log-linear fit for the known organisms is shown in the first row labeled ‘organisms,’ calculated using the estimates for genomic architecture parameter and population size.	40
3.2	Distribution statistics of D , the number of mutations required for a population to adapt. Average, standard deviation, minimum and maximum values were calculated for each $D_{N_e, \alpha}$, and then averaged to report a single value for each phenotype/fitness function.	46
3.3	The robustness, or probability that subsequent mutations are also optimal, during the fixation phase of establishment time for the different phenotype/fitness functions. For each fitness function, Q is the distribution of robustness across all simulated population sizes and genomic architecture parameter values.	51

Chapter 1

Introduction

The path from genotypes to complex phenotypes with genetic cause is mediated by many processes, including regulatory pathways, but our evolutionary understanding of these critical structures falls short. Current models of pathway evolution fail to preserve sequence structure and include population dynamics, two critical issues that have preliminarily been shown to determine several unexplained trends in empirical pathways [2]. Despite being the substrate for regulatory gene interactions, the effect of non-coding regions on the evolution of pathways has been overlooked. Furthermore, evolutionary dynamics, as determined by mutation, recombination, drift, and selection, can only be determined within a finite population, but population-genetic studies are effectively non-existent in the field of pathway evolution with the exception of a few recent works [2, 3]. This thesis presents a solution to both these issues by way of a sequence-based, population-genetic model of pathway evolution.

1.1 Pathways: sequence to function

Over a decade into the sequencing era, science is still at a loss for explaining how differences in the genome manifest as the entire gamut of observable characteristics that distinguish individuals and organisms [4]. These observable traits, or phenotypes, were once thought to be predominantly functions of protein structure, and, consequently, changes between organisms or individuals were largely caused by structural modifications to genes and the proteins they produced [4]. However, with advancements in sequencing and gene expression analysis, a growing body of work has identified the strong effect of mutations in the non-coding regions of the genome.

A major class of these studies, and the center-point of the investigation in this thesis, is the evolution of genetic regulatory pathways, which describe the transcriptional and translational control of cooperating genes. Regulatory pathways are useful in describing complex relationships and therefore elucidate the effect of mutations in non-coding regions on phenotype [4, 5, 6, 7, 8, 9, 10]. Pathways, then, are useful for explaining how mutations in non-coding regions can cause phenotypic diversity within organisms and between individuals. For example, recent work in the comparative analysis of dog genomes indicated that much of the animal's great phenotypic diversity arises in regulatory regions [8].

These regulatory regions harbor important sequences called *transcription factor binding sites* that serve as amenable locations for regulatory proteins, called *transcription factors*, to bind and, by a variety of mechanisms, either increase or decrease the transcription of neighboring genes. The gain and loss of these regulatory connec-

tions turn on, off, or modulate gene expression. Hence, holding the coding sequence constant, the rewiring of genetic regulatory pathways corresponds to the spontaneous gain, loss and modification of binding sites [11, 12, 13]. Consequently, a comparative understanding of regulatory relationships amounts to an evolutionary model of binding site gain, loss, and modification.

However, much of the evolutionary analysis of regulatory pathways revolve around the calculation of their structural and dynamic properties, which although is integral to invoke adaptation, elides the powerful non-adaptive forces of evolution [14, 15, 16, 17, 18, 19, 20, 21, 22]. Concomitant with phenotypic calculation is the estimation of mutational robustness and innovation. Robustness is defined as the proportion of mutations on the pathway level that are neutral with respect to the phenotype, while innovation is defined as the proportion of n -step mutations that yield novel phenotypes (see Figure 1.1) [22]. Taken together, these measures comprise the “evolvability” of a pathway by quantifying the effect of binding site mutations on the phenotype level. Ciliberti and Wagner calculated the inverse correlation between these two measures for the regulatory pathway genotype [22]. The inherent trade-off between robustness and innovation transcends the pathway genotype, a debate which has stimulated much discussion and analysis (see [23] for a good review). But the reconciliation to the debate over robustness and innovation is also, in fact, the same omission by the adaptation-focused models of pathway evolution: populations, not individuals, evolve.

This simple fact is also a complicating factor. Mutation, robustness, and genetic drift complicate both theoretical and simulation studies, but also serve as an

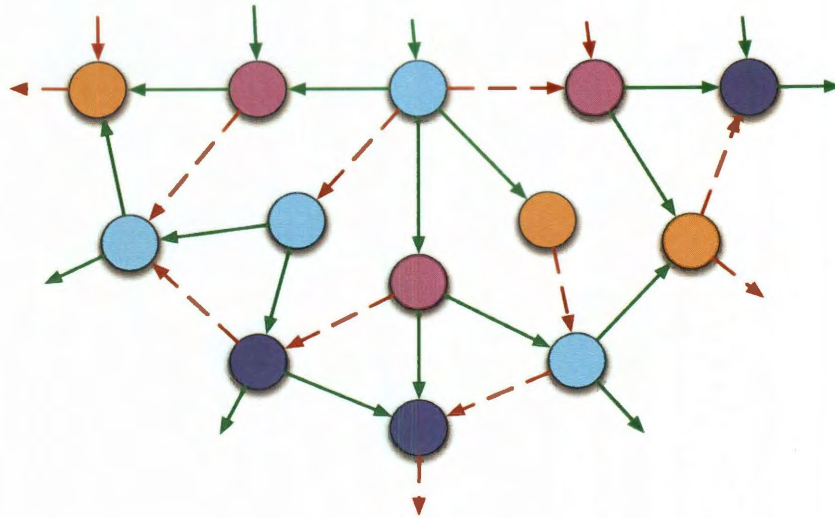


Figure 1.1: The genotype-phenotype space is a useful abstraction for understanding the phenotypic effects of mutations. In this diagram, circles are distinct genotypes connected by edges with denote mutation from one genotype to the next. Colors represent the phenotype. Robustness is the proportion of neighbors that share the same phenotype, and innovation is the proportion of m -step mutations that yield different phenotypes.

important null model of evolution. For instance, Draghi *et al.* quantified the non-monotonic relationship of robustness and innovation in finite populations [24]. By invoking only mutation and genetic drift, their results explained that high robustness supplies substrate for subsequent adaptive mutations and thereby reduces mean time to find a beneficial allele. In other words, a population may be both robust and evolvable through the robustness and innovation of each individual. As for the pathway genotype, Martin *et al.* quantified the effect of recombination within a population on robustness and also measured population-genetic properties like genetic load and diversity [3]. However, the evolutionary model used by Martin *et al.* allowed for

only one mutation rate to govern the gain, loss, and modification of binding sites and one recombination rate for crossover events outside of genes and their regulatory regions. As shown by Lynch, a single mutation rate and recombination rate are not appropriate to evolutionary modeling of regulatory pathways [2]. Consequently, it is difficult to reflect pathway-level insights to the sequence. In fact, only recently did Lynch propose a model of pathway evolution that incorporated features of non-coding DNA, the critical substrate of transcriptional regulation [2].

1.2 The effects of non-coding DNA and population size on pathway evolution

The increased number of completed genomic sequences has revealed many highly significant and interesting patterns across the tree of life [25, 26, 27, 1, 28, 29, 30]. One such pattern, even labeled a paradox by some [26, 27], involves the positive correlation between the expansion of non-coding DNA and total genome size. Given that the number of genes and amount of coding sequence remain relatively constant in comparison [27], this expansion of non-coding nucleotides is valid on a per-gene basis. In addition, along with an increase in genome size is a decrease in the population mutation rate ($2N_e u$) [29]. Since the per base pair mutation rate (u) is actually increasing with genome size in eukaryotes [25], this results in a significant decrease in the effective population size (N_e) to accommodate for the overall decrease in population mutation rate. Combined together, these results suggest a strong negative correlation between the amount of non-coding sequence per gene in a genome and the

effective population size (see Figure 1.2). This correlation is strongly log-linear, with a Pearson correlation coefficient of -0.984.

Lynch and Conery hypothesized that expanding, maladaptive non-coding regions resulted from the weakening of selective pressures by significant decreases in effective population size [29]. Recently, however, this has been challenged on two fronts: first, the phylogenetic implausibility that genetic drift alone caused the accumulation of genome complexity [31], and second, the surprising non-correlation of population size and genome size in seed plants [30]. In either case, both studies argued that population size and genetic drift are not sufficient to explain the expansion of non-coding DNA and proposed other factors like different mating systems. Understanding these regions on a more functional level is clearly needed. While these large intergenic regions are recognized as the hallmark of multicellularity and complexity, explanations for the development of specific functionality in intergenic sequences, and consequently their effect on the evolution of an organism, have thus far been established on a case-by-case basis [32, 12, 33, 1, 34].

In this thesis, I confine my investigation to the aforementioned transcriptional regulatory elements of transcription factors and transcription factor binding sites (TFBS) known to exist between coding DNA. Lynch parameterized the ratio of regulatory interaction loss versus gain using an estimate of the number of base pairs per gene that may harbor TFB sites, or promoter and enhancer regions [2]. Since this ratio is representative of the relative amount of coding and non-coding base pairs in a genome, I will refer to it in this thesis as the *genomic architecture parameter*. Because the genomic architecture parameter scales with the size of intergenic regions per gene

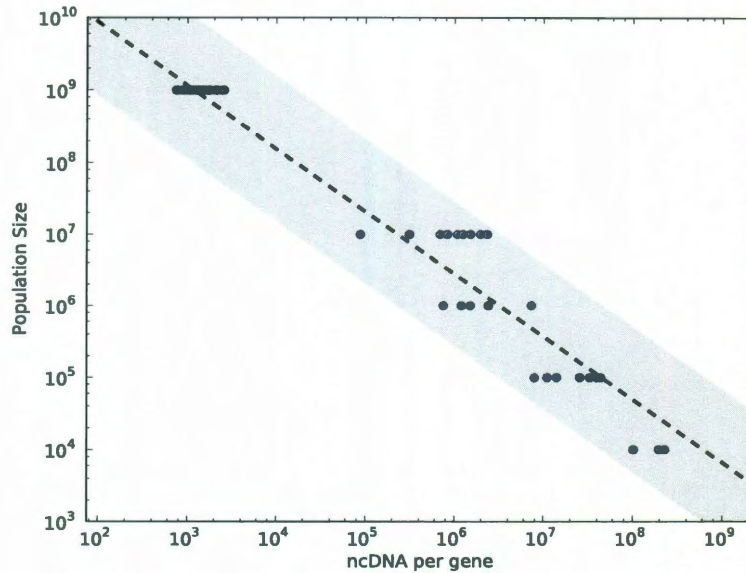


Figure 1.2: The strong correlation between non-coding sequence (ncDNA) per gene and the effective population size. Pearson correlation coefficient is -0.984. Estimates of population size for bacteria, unicellular eukaryotes, invertebrates, vertebrates, and land plants come from Lynch [1]. Uncertainty in both the estimates in population size and ncDNA are depicted by the gray band.

and ascribes functionality to the amount of non-coding DNA surrounding each gene, it is useful in determining the effect of promoter region size on the evolution of regulatory pathways. For example, Lynch applied this parameter to show the possible neutral origins of regulatory pathway complexity in eukaryotes [2].

1.3 Contributions of the thesis

In this thesis, I propose a novel representation of the pathway genotype which captures the sequential underpinnings of the regulatory pathway so as to better represent the effect of mutation and recombination within critical binding regions. Previous work on the effects of mutation and recombination represented a regulatory pathway with an adjacency matrix [22, 3, 2]. An adjacency matrix is a square matrix where each element (i, j) indicates the pairwise interaction strength between gene- i and gene- j in the pathway. This formulation hides important effects of mutation: by random chance, binding sites may co-occur within promoter sequences, especially when those regions are around 100 kilobases in length. Allowing for duplicate binding sites might have a major impact on robustness and plasticity of regulatory pathways, but this has not been investigated. Furthermore, it is impossible to model the effect of recombination (crossover event) within regulatory sequences using an adjacency matrix because crossovers occur on the sequence, but there is no sequential information in an adjacency formulation. Consequently, Lynch's implementation on the effect of recombination within regulatory regions on a three gene pathway are inaccurate (see Appendix for a detailed explanation).

My work leverages the genomic architecture parameter introduced by Lynch [2] in order to understand the effect of promoter and enhancer region size on the evolution of regulatory pathways in a finite population. By further refining a sequence-based model of pathway evolution, results from understanding the evolution on the pathway level will apply back to the sequence. Hence the model of this study uses pathways to

understand the functional importance of the variable length of non-coding sequences within the genome and thereby provide quantitative measures for the advantages, disadvantages, and trends created by non-coding DNA on the evolution of novel pathways.

This thesis, then, bridges several research areas under a pathway evolution perspective: first, understanding the evolutionary advantages and disadvantages of non-coding DNA on regulatory pathways; second, the determination and examination of neutral forces on the evolutionary trajectory of pathways; and finally, a sequence-realistic refinement of the regulatory pathway model.

Chapter 2

Model

In this chapter, I describe a novel sequence-based regulatory pathway model which evolves within a finite population. The description of the model is decomposed into four sections: the population life cycle, the refinement of the genotype and phenotype of the pathway allele, the modeling of non-adaptive evolutionary forces, and finally, the modeling of adaptation. Throughout this chapter special itemizations keep track of important parameters and assumptions made in constructing the model.

2.1 Population life cycle

The model I employ follows a Wright-Fisher life cycle wherein mutation, recombination, drift, and selection take place in non-overlapping generations, as shown in Figure 2.1.

The cycle begins with a population of zygotes (with size $2N$) that produce an

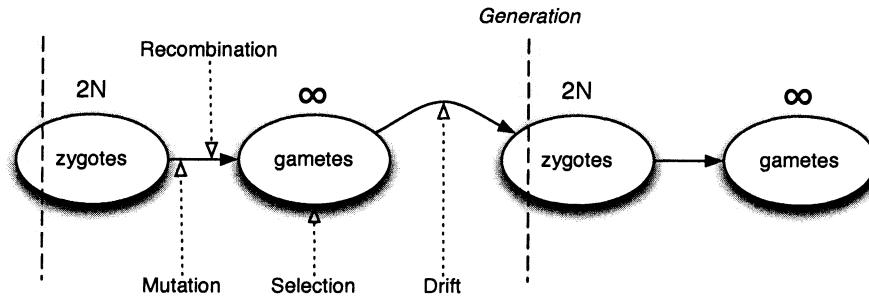


Figure 2.1: Life cycle model of the population simulator. The evolutionary processes are annotated for their order in each generation.

infinite number of gametes. Individuals are considered to be haploids, but diploids can be constructed by pairing the population. Mating is assumed to be random throughout; however, intricacies of mating systems, especially with regard to plant and animal distinctions, can be considered in future studies. During the creation of the gametes (or, depending on the organism, throughout the life of a zygote), mutations occur in the germ line, and recombination occurs between chromatids in a diploid organism or through some other recombination mechanism (transformation, conjugation, or transduction). The proportion of gametes in this population will be skewed from random mating by soft selection since fitter individuals will contribute more gametes, or genetic material, in comparison to individuals with lower fitness. Lastly, random genetic drift occurs in the finite sampling of haploid individuals from the infinite gamete pool to determine the allelic frequency for the next generation.

Parameter 1 (N_e) *The effective population size.*

Parameter 2 (G) *The number of generations for which to model evolution.*

2.2 The pathway genotype

Now that the assumed life cycle is given, the effect of mutation and recombination on the pathway genotype and selection on the phenotype must be determined. A requisite for either of these attributions is the definition of the pathway genotype and phenotype.

2.2.1 Genotype

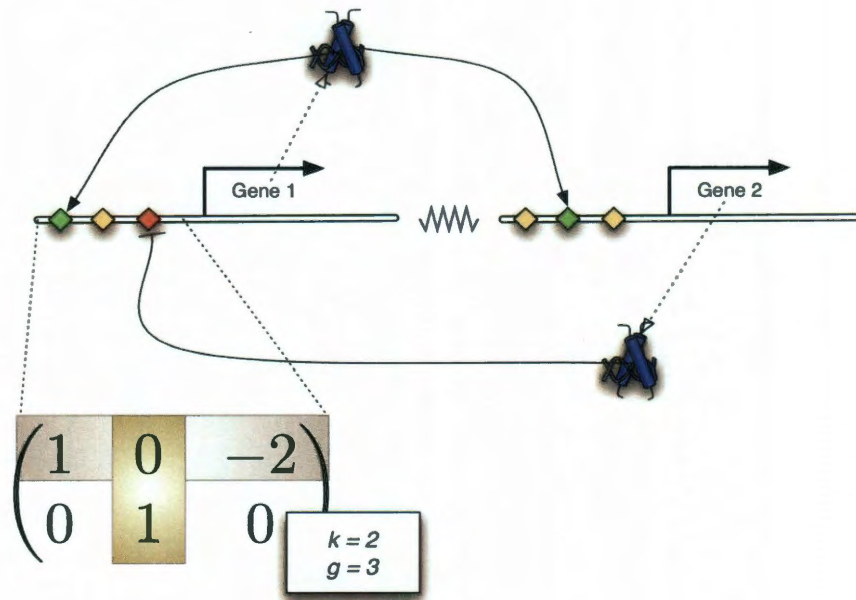


Figure 2.2: A cartoon that illustrates a matrix for a genetic pathway comprising of genes, transcription factors, and binding sites. The matrix representation denotes the binding sites and their affinity. This discretization of binding sites represents where a binding site may arise.

Regulatory pathways consist of genes that encode transcription factors that either

activate or inhibit the transcription of their own or other genes. The binding sites that mediate this regulation exist along the sequence, both upstream, within, and downstream of the gene. To preserve the sequential ordering of binding sites, the promoter and enhancer region for a gene is discretized into g regions, where only one binding site may occur in each of these g regions.

Assumption 1 *Only one binding site may occur in each of the g regions. For a given promoter, at most g binding sites may exist.*¹

The pathway genotype is encoded as a $k \times g$ matrix, where k is the number of genes in the pathway, as shown in Figure 2.2. An allele is an instantiation of the $k \times g$ matrix:

$$M = \begin{pmatrix} b_{i,j} \end{pmatrix},$$

where $b_{i,j}$ encodes the status of the j^{th} binding region for the i^{th} gene. The regulatory region for gene i is encoded as row i in the matrix, and the columns preserve the ordering of binding sites along the sequence. The binding site status, or $b_{i,j}$, is either 0, for no binding site present, or a value $[1, k]$ for activation and $[-1, -k]$ for inhibition, representing the index of the transcription factor $|b_{i,j}|$ that binds to the given site. Formally,

$$b_{i,j} \in \{-k, \dots, k\}$$

¹The quality of this assumption increases with g . Obviously, $g \leq L/n$.

$$b_{i,j} = \begin{cases} d > 0 & i \text{ activated by } d \\ 0 & \text{no regulation on } i \\ d < 0 & i \text{ inhibited by } d \end{cases}$$

Parameter 3 (k) *The number of genes under investigation.*

Parameter 4 (g) *The number of discretized regions of a promoter sequence, where each of the g regions may only contain one binding site.*

It is important to stress that this data structure is not an adjacency matrix. Previous approaches used adjacency matrices to encode genetic networks, but this leads to a departure from the sequential underpinnings of regulatory pathways [2, 22, 3]. An adjacency matrix is a square matrix $A_{i,j}$ ($k \times k$) that encodes the weight, or presence, of an edge between two vertices i and j . For a given pathway topology (or adjacency matrix) there are several configurations of binding sites along the genomic sequence. Hence, mutations and recombinations that could greatly change the binding site data structure may appear neutral on the pathway topology.

To calculate the adjacency matrix from a regulatory pathway allele:

$$A_{i,j} = \sum_{0 < h \leq g} \delta(|b_{j,h}| - i) \text{sign}(b_{j,h}), \quad (2.1)$$

where

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{sign}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}.$$

Here, the edge weight is positive for activation and negative for inhibition. If there is an activating and inhibiting binding site for the same gene in a given promoter, the adjacency matrix will report a value of $1 + -1 = 0$, which hides the regulatory effect of the binding sites. More complicated methods for determining the regulatory influence of competing binding affinities can be investigated in future work.

Assumption 2 *A promoter region can harbor regulatory relationships of different types for the same transcription factor.*

2.3 Phenotype

Previous work on mapping regulatory pathways to phenotypes used a variety of techniques, and in this thesis work I employ three of them:

1. the use of ODEs as a model of equilibrium gene product concentrations (*continuous*) [35],
2. the iterative up and down regulation using a regulatory matrix (*discrete*) [22, 3],
and,
3. a simple viability constraint requiring that all genes are regulated (*discrete-viable, continuous-viable*) [2].

Additional phenotype functions may easily be added in future extensions, since it is only a matter of implementation.

2.3.1 Continuous

The continuous approach approximates regulatory pathway dynamics using ODEs, a common technique in assessing the function of a pathway [14, 35, 15, 19]. Given a pathway allele matrix M , it is relatively straightforward to calculate the differential equations of the gene products. The dynamics associated with each gene amount to transcription, translation, and degradation. Transcription is modeled using the Hill formula for activation. For multiple possible TF bindings, for instance if a given gene is regulated by more than one TF, it is necessary to sum the transcription rates for each combination of TFs. Translation and degradation are modeled as linear fluxes. This process is automated by RENCO, which takes as input genes, gene products, and regulatory interactions to produce a set of ODEs (for an example output see Figure 2.3) [35]. For simplicity, uniform rates for activation, repression, translation, and degradation are used in this thesis, but the opportunity exists for incorporating specific dynamics using user-supplied rates.

Assumption 3 (Uniform rates) *The rate parameters for calculating the ODE regulatory pathway dynamics are assumed to be uniform across all genes. The absolute value of these rates are not as important as their relative relationships. For instance, the ratio of mRNA synthesis over degradation explains the average number of proteins synthesized per molecule of mRNA.*

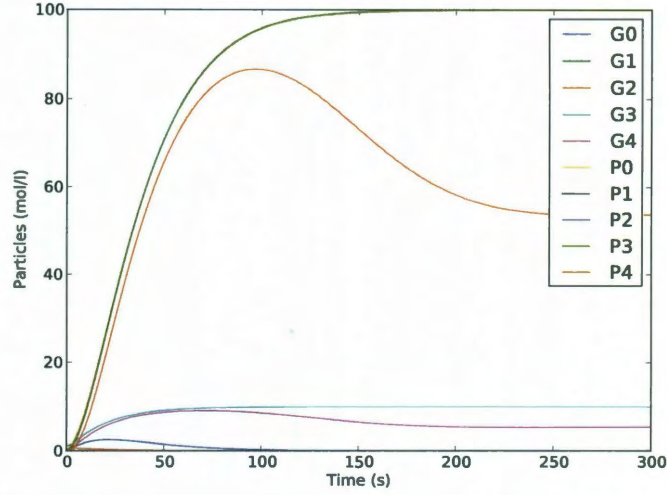


Figure 2.3: The protein trajectories as computed using the ODE equations generated by RENCO. The *continuous* phenotype is taken as the last time point once the trajectories have converged.

2.3.2 Discrete

Like the ODE approach, the work by Wagner determines an expression pattern of k genes as a time series $S(t) = [S_1(t), S_2(t), \dots, S_k(t)]$, but calculates $S_i(t + \tau) = \sigma[\sum_{j=1}^k w_{ij} S_j(t)]$, where τ is some constant time step, σ is the sign function, and $w_{i,j}$ is the regulatory influence between gene- i and gene- j [22, 3]. A product state at any time is either -1 for down regulation, 0 for no regulation, or 1 for up-regulation. In the update equation $S_i(t + \tau)$, influence outside of these values are scaled using the sign function (σ). The influence between genes can either be activating ($w_{i,j} > 0$), inhibiting ($w_{i,j} < 0$) or absent ($w_{i,j} = 0$). The matrix $w = (w_{i,j})$ is the adjacency matrix of the regulatory pathway.

The equilibrium regulatory influence on each gene is given by S_∞ , which evolves from a starting state $S(0)$. This equilibrium state is viable if $S(t)$ is either convergent or cyclic within k^2 timesteps, at which point the equilibrium concentrations are given by $f(S(k^2))$ if $S(t)$ converges, or $f(\text{avg}(S(p : k^2)))$, where p is the period of the cycle and f maps regulatory influence to gene-product concentration. The simplest mapping corresponds to $f(S) = 1 + S$ so that down-regulation corresponds to a concentration of 0, no regulation corresponds to basal transcription of 1, and up-regulation corresponds to 2. The scalar quantities for each case can be parameterized within the f mapping.

To determine the *discrete* phenotype given a an allele M , first the adjacency matrix w using Equation 2.1 is computed. Then, S_i is computed for k^2 time steps from the start state $S(0) = [1, -1, -1, \dots, -1]$ (see Figure 2.4). If S converges or is cyclic within k^2 steps, then S_∞ is the equilibrium regulatory influences and the equilibrium protein concentrations are calculated as $1 + S_\infty$.

2.3.3 Viability

Lynch employed a simple viability constraint in his evolutionary analysis of pathways that required all genes to be regulated [2]. Intermediate transcription factors could be initiator signals for other pathways downstream, and so the loss of regulation of an intermediate gene which encodes a transcription factor would lead to the loss of the downstream pathway as well. Such a scenario would be considered fatal to the cell.

This viability criterion is imposed on the aforementioned *discrete* and *continuous*

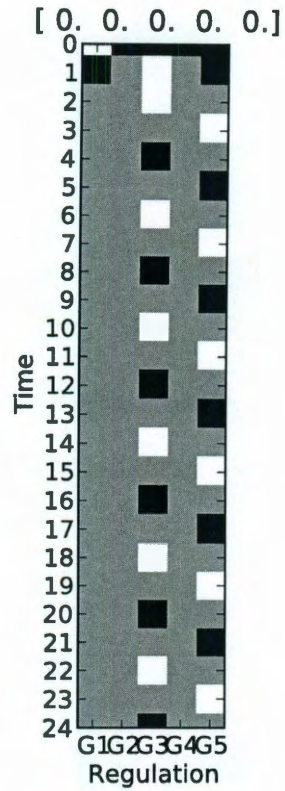


Figure 2.4: The time evolution of the example pathway using the *discrete* method. White signifies up-regulation, black down-regulation, and gray no regulation. The equilibrium concentrations for the gene-products can be found by averaging the cyclic behavior, which yields no regulation (the zero vector on top of the figure).

phenotype functions. The calculation of viability is simple: the pathway is nonviable if any row in the pathway genotype is all zeros (no binding sites exist for the gene), and viable otherwise. If a pathway is not viable, then it is considered invalid.

2.4 Non-adaptive processes

With genotype defined as a matrix of binding sites, it is now possible to understand how mutations to the DNA sequence and crossover events between two sequences manifest on the pathway level. This section explains the effect and modeling of mutation and recombination on the pathway allele.

2.4.1 Mutation

Since my model deals only with the non-coding mutations that effect pathway structure, pertinent mutations occur in the regulatory regions of the gene. Base pair mutations within regulatory regions either have no effect, remove a binding site, or result in the gain of a binding site. Lynch proposed formulas for the loss and gain rates of binding sites (μ_l and μ_g respectively) [2]. The loss rate is the per base pair mutation rate u times the length of a binding site n . The gain rate is given by $Lnu/4^n$, where L is the length of DNA that may harbor binding sites. The ratio of these rates $\alpha = \mu_l/\mu_g$, or genomic architecture parameter, scales with the size of regulatory regions. Lynch showed that the genomic architecture parameter is a function of the length of a binding site n and length of regulatory sequence L : $\alpha = 4^n/L$.

Parameter 5 (n) *The average length of a binding site.*

Parameter 6 (L) *The length of the critical regulatory sequence where binding sites may arise for a particular gene.*

Parameter 7 (μ_l) *The loss rate of a transcription factor binding site in a promoter*

region of length L/g . $\mu_l = nu$, where u is the per site mutation rate and n is the length of a binding site.

Parameter 8 (μ_g) *The gain rate of a transcription factor binding site for a specific TF in a promoter region of length L . $\mu_g = Lnu/4^n$, where u is the per site mutation rate. Therefore, the rate of gain of a binding site for any TF in the pathway is $k\mu_g$.*

While previous models allowed mutations to also change the interaction strength of a regulatory relationship, estimating such strength is not entirely clear [3, 22]. Therefore, this model only supports the presence and absence of binding sites without any affiliated weight.

Lynch recently calculated estimates of the per base pair mutation rate for organisms across all major phyla [25]. Most mutation rates range between 10^{-8} and 10^{-10} , although viruses have extremely accelerated rates around 10^{-4} . Therefore, if the binding site length is 10 bp, the loss rate of binding sites ranges from 10^{-7} to 10^{-9} . The genomic architecture parameter ranges from 10^{-3} in mammals to 10^4 in bacteria, and so the gain rate has a wide spread of 10^{-4} to 10^{-13} .

Modeling a mutation is straightforward. For a pathway allele M , the number of potential gain sites, $gain(M)$, is the count of zero values in the matrix M and the number of loss sites, $loss(M)$, is the count of the number of non-zero entries. The total number of random loss mutations for the allele can be calculated using the binomial distribution, where $freq(M)$ is the frequency of allele M in the population:

$$\#_{losses} = binomial(N freq(M), loss(M)\mu_l)$$

And similarly the number of gain mutations:

$$\#_{gains} = \text{binomial}(Nfreq(M), gain(M)\mu_g)$$

Once the number of losses has been determined for a given network allele, $\#_{losses}$ mutant variants of M are constructed by randomly (uniformly) selecting a binding site and removing it (setting $b_{i,j}$ to 0). This process may introduce new alleles into the population; however, not all mutant variants may be unique.

Similarly, once the number of gains has been determined, $\#_{gains}$ mutant variants of M are constructed by randomly selecting an empty binding site and setting it equal to a random gene index in the pathway. The determination of an inhibition versus activation is determined by ρ , the probability of creating an activation regulatory effect. In this thesis, I use $\rho = 1/2$.

Parameter 9 (ρ) *The probability of creating an activation binding site. $(1 - \rho)$ is the probability of creating an inhibitory site.*

To update the affinity of a gain mutation at binding site $b_{i,j}$:

$$b_{i,j} = s_\rho(\text{uniform}(0, 1)) \times \text{uniform}_{int}(1, k)$$

$$s_\rho(x) = \begin{cases} 1 & : x < \rho \\ -1 & : x \geq \rho \end{cases}$$

2.4.2 Recombination

Recombination is the breaking and joining of DNA. There are various ways for recombination to occur, but the fundamental result is the same: genetic sequence from one allele is recombined with the genetic sequence from a different allele via a crossover event. Other more complicated forms of recombination exist can be investigated as extensions to the model. In this model, the interesting recombination events occur in the g discretized regulatory regions of each gene since these regions harbor binding sites. Shuffling binding sites through recombination can drastically alter the topology of the pathway, leading to a unique expression pattern or even loss of viability [2]. The rate of recombination events in the discretized regulatory regions, r_b , can be calculated knowing the per site recombination rate and the length of the sequence in each discretized regulatory region (L/g).

Parameter 10 (r_b) *The rate of recombination events in the discretized regulatory region. Since recombination rate over a distance d on the genome is calculated by $r(d, c) = 0.5(1 - e^{-2dc})$ and $r(d, c) \approx dc$ when d is small, then $r_b = r(L/g, c)$, where c is the per site recombination rate.*

Recombination events that occur between gene promoter regions, either in the gene itself or in the non-critical intergenic material, occurs at a different rate than in the binding site regions. This is the recombination that has most recently been studied in [3]. Since no assumptions are made about the base pair distance between genes in the pathway, this recombination rate can range anywhere from 0 for genes in high linkage disequilibrium to 1/2 for genes on different chromosomes. A crossover

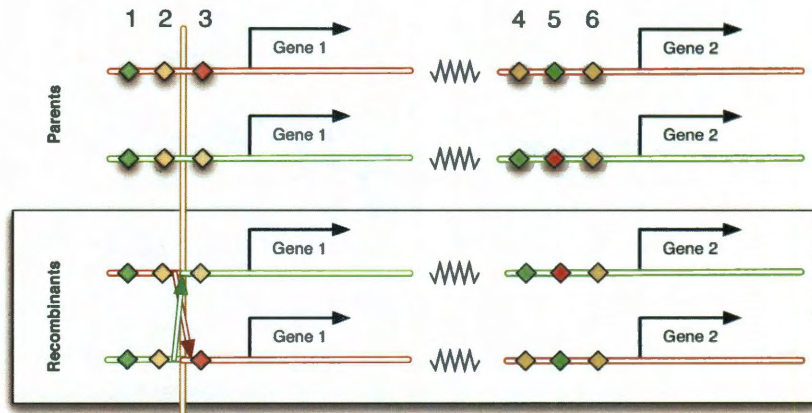


Figure 2.5: A recombination event occurs in the 2nd critical binding region. The result of recombination in this binding region is detailed in Figure 2.6. Gene 2 is considered to be downstream and so is effected by the event. The arrows in the recombinants box denotes the crossover that occurred.

event between genes is simply a switching of the rows in the allele matrix M . Not surprisingly, according to Martin and Wagner, point mutations are more costly than crossover events between genes in genetic networks [3]. However, if the crossover event occurs in the non-coding regions that harbor binding sites for genes in the pathway, recombination would shuffle the binding patterns between alleles. The effects of such an event could be drastic on the pathway topology (see Figure 2.6).

Parameter 11 (r_g) *The rate of recombination events that occur between promoter regions of different genes. For completely unlinked genes (genes on different chromosomes or significantly distant on the same chromosome), r_g would be 0.5.*

The order of the genes on the genome become relevant when recombination occurs. This is also the case with the order of the binding sites. Figure 2.5 illustrates this

point.

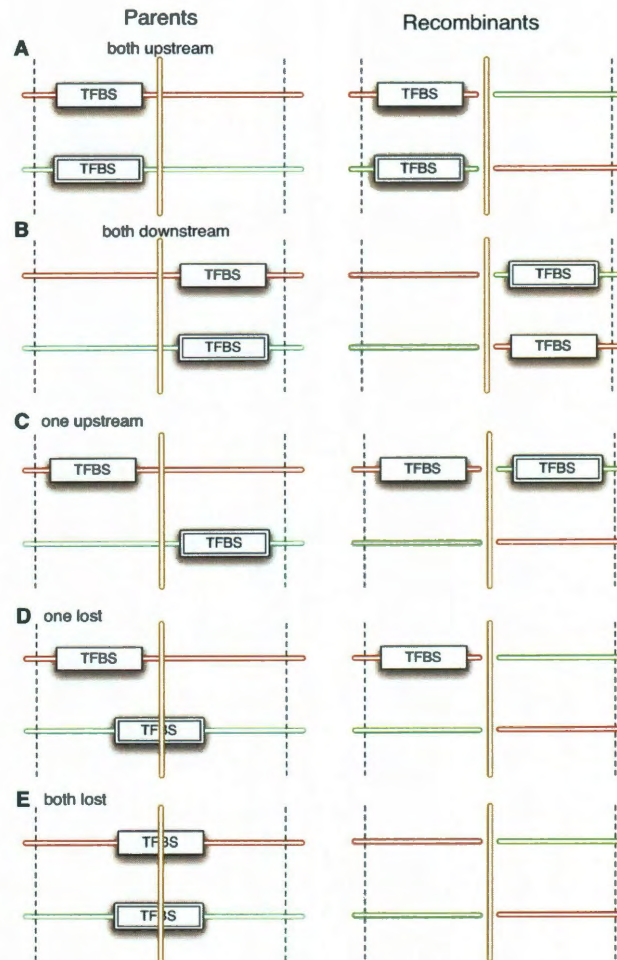


Figure 2.6: All possible results of a recombination event in a critical binding region (one of g in each promoter). By Assumption 1, at most one TFBS may be in each one of these regions. Note that one of the recombinants in scenario C violates Assumption 1.

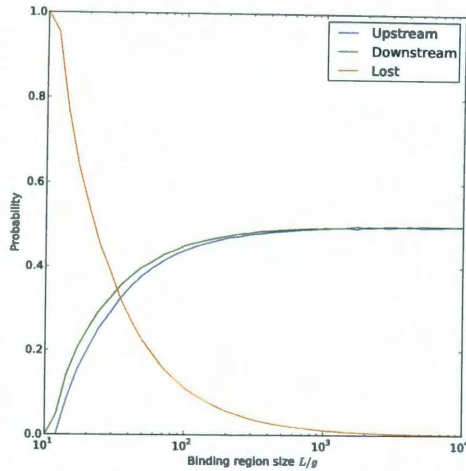
When recombination occurs in a specific discretized section of the regulatory region, there may be at most one binding site for each parent (by Assumption 1). There

are many outcomes for such a recombination. For the case where both parents have a binding site, see these outcomes in Figure 2.6.

First, consider what happens if only one parent has a binding site. Based on where the crossover occurs relative to the binding site, the site may be unchanged if it is upstream, moved if it is downstream, or lost if it is coincident. The probability of each of these events based on the size of this binding site region L/g is shown in Figure 2.7. The probability that recombination destroys a binding site decreases exponentially with respect to the size of the binding region. As one would expect, it is equally probable for the binding site to be upstream or downstream of the recombination event.

In the case where both parents have a binding site in a certain region, there are several more scenarios to consider. These scenarios are illustrated in Figure 2.6. For scenarios **C** and **D** there are symmetric cases where the recombinant TFBS are mirrored. For example, **A** and **B** show the symmetric case for binding sites co-occurring upstream or downstream of the crossover. Of interest are the scenarios that cause redundancy (**C**), loss of one binding site (**D**), and loss of both binding sites (**E**, **C**). Figure 2.7 shows the impact of each scenario with respect to the length of a binding region L/g . When $L/g = n$, the crossover is guaranteed to eliminate both TFBS. However, this probability drops rapidly. When $L/g = 2n$, the probability of losing both decreases by more than half its value, and it is equal to the probability of losing only one (**D**). When $L/g = 4n$, the most likely scenario is to keep both TFBS. Losing both binding sites only becomes a factor with the likelihood of scenario **C**, but this is still less likely than both binding sites being either upstream or downstream

Only one parent has a binding site



Both parents have binding sites

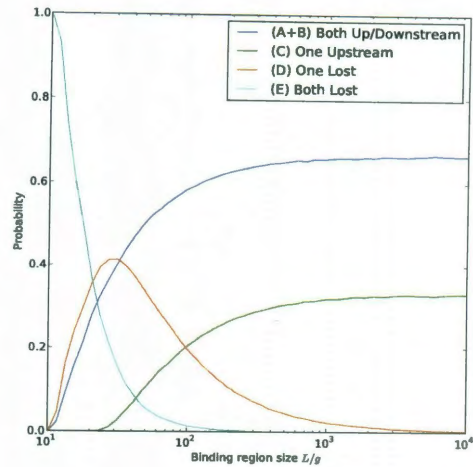


Figure 2.7: The probability of each possible recombinant scenario if only one parent has a binding site (left) or both parents have a binding site (right). The binding site length n is equal to 10, so the leftmost x-value is $L/g = n$. The letters in the figure on the right refer to the scenarios mentioned in Figure 2.6.

of the crossover event. If $g < L/10^2$, then Assumption 1 is a reasonable one, since scenario **C** is relatively small (the failure of the assumption occurs 50% of the time in the case of **C**). The probability of each scenario is a function of the binding region size and length of a binding site (L/g). Interestingly, for bacteria, L/g could be less than 100 bp, and so there is a significant probability that recombination would eliminate one or both binding sites. However, for eukaryotes, L/g is much larger, and a greater chance exists of swapping binding sites or causing redundancy (**A**, **B** and **C**).

To perform a crossover on the pathway genotype, let M_1 and M_2 be parent alleles with a crossover event occurring in the $b_{i,j}$ region. Let m be the $k \times g$ pathway matrix

unraveled into a kg length vector such that $M_{i,j} = m_{ig+j}$. A recombinant is given as $m = m_1[1 : ig + j - 1] + m_2[ig + j : kg]$, with $m[ig + j]$ sampled from the relevant distributions in Figure 2.7. For the scenario that violates Assumption 1, the upstream binding site is chosen. Disregarding redundancy in the discretized region does not preclude redundantly regulated recombinants to form, it only precludes redundancy within a position on the allele matrix. To avoid this issue altogether, it is possible to minimize L/g by more finely discretizing the regulatory region (increasing g).

To perform a crossover between gene i and gene $i + 1$, steps are taken as before to construct the parental kg length vectors. Then, a recombinant is given by $m = m_1[1 : ig + g - 1] + m_2[(i + 1)g : kg]$.

2.5 Adaptive processes

Selection refers to the process by which heritable traits confer a reproductive advantage to an individual. The nature of a fitness function which maps phenotype to selective advantage is highly dependent on the study at hand. Consequently, it is easy to over-constrain a general model, like the one presented in this thesis, with a specific fitness function. However, it is still important to design a realistic fitness function so as to better understand realistic fitness landscapes for the pathway genotype. As a reconciliation, this model employs both a fitness function framework that reflects the benefits and costs of a pathway and a random, parameterized fitness function that provides an important null model. Unlike other studies which use only one or the other, studies on this model can assess the similarity and differences between

parameterized fitness functions and ones motivated by biological principles.

2.5.1 Designing a fitness function

A fitness function calculates the ‘goodness’ of a pathway in context with the environmental conditions and other pathway alleles in a population. For the time being, fitness is considered independently of allelic frequencies, but frequency dependent fitness is an interesting extension of the model which allows for more sophisticated community evolution. Two classifications of fitness functions exist: *soft*, where the function assigns marginal fitness gains, and *hard*, where the function designates the allele as fit or unfit.

Most work on pathway evolution deals with hard, or binary, fitness functions. Lynch imposed only viability [2], which is a binary fitness function that is true only if each gene in the pathway is regulated. Other work in regulatory pathway evolution classified pathways by their phenotype [22, 3]. If a pathway produced different gene expression levels at equilibrium, then it would belong to a distinct phenotype group. However, such a distinction cannot explain relative fitness and reproductive success of the individuals in different phenotype groups, so it is only useful as a calculation of hard fitness.

However, the implications of marginal fitness are well studied in population genetics, and have dramatic effect on adaptation, fixation, and establishment time of a population [36, 37]. Furthermore, empirical studies reported that pathways fine tune to specific environmental conditions, which is a strong argument for the importance of marginal fitness in evolutionary simulations [19, 14]. Consequently, my model em-

employs a soft fitness function, however, a hard fitness function can easily be designed and supplied to the model.

According to Dekel *et al.* , fitness can be decomposed into the benefits over the costs of the pathway [19]:

$$fitness = \frac{benefit}{cost}.$$

This equation states that cost must be greater than zero and the benefit must be positive. Intuitively, a negative fitness is not biologically plausible (a negative difference in fitness is allowed). Rather, fitness must be a positive value where any value less than 1 means the costs outweigh the benefits, and any value greater than 1 means the benefits outweigh the costs.

For example, for the *lac* Operon pathway, the growth benefit is proportional to the amount of *LacZ* bound to lactose [19]. However other forms of benefit may include time delay, oscillation frequency, or depression of a signal. My model of pathway evolution borrows from the *lac* Operon pathway. Let G_k be a target gene in the pathway that performs some important function in the cell, like *LacZ*. Therefore the benefit function of a pathway M is proportional to the expression of G_k :

$$B(M) = \sigma[G_k]. \tag{2.2}$$

This equation elides any function of the environment, since, for the time being, a constant environment is assumed where the production of G_k is beneficial.

Assumption 4 (Fitness benefit) *The benefit of a regulatory network is propor-*

tional to the expression of the target gene G_k that performs some important function in the cell.

Parameter 12 (σ) *The per molecule growth rate benefit conveyed by the target gene.*

Now that a benefit function is defined, where do costs come from? Stoebel *et al.* summarizes [16]:

- Transcription could be costly because it uses nucleotides that could be incorporated into other RNAs.
- Transcription occupies RNA polymerases [20] that might be better used to transcribe genes whose products increase fitness.
- Translation wastes charged tRNAs and occupies free ribosomes.
- The proteins produced by translation tie up amino acids that might be better incorporated into other beneficial proteins.
- Costly activities of the proteins, e.g., insertion of the permease in the membrane, might allow protons to leak into the cytoplasm, thereby partially dissipating the proton motive force. In addition, insertion might affect membrane fluidity and/or occupy space needed for other membrane proteins.

Stoebel *et al.* created a general fitness model for the *lac* Operon circuit (composed of three proteins) that computes relative fitness between two strains [16].

$$\text{Fitness difference between two strains} = \beta_1(\text{difference in LacZ amount}) + \beta_2(\text{difference in LacY amount}) + \beta_3(\text{difference in LacA amount}).$$

In addition, Stoebel *et al.* showed that the major cost of regulatory pathways (in particular the *lac* circuit) is dominated by the process of transcription and translation. The costs of excess proteins or costly activities of the proteins are minor in comparison. So, the proposed fitness model incorporates only the costs of creating the protein product. Therefore, for a general pathway M with k genes, where $[G_i]$ is the concentration of the i^{th} gene at equilibrium and β_i is proportional to the gene product size, the cost is:

$$\eta(M) = \sum_{i=1}^k \beta_i [G_i]. \quad (2.3)$$

For any pathway M , $\eta(M) > 0$, so it will serve as a valid denominator in the calculation of fitness.

Assumption 5 (Fitness cost) *The major cost of a regulatory network is in the process of creating the protein products.*

Futhermore, Equation 2.3 follows the model from [16] in that a difference between the equilibrium concentrations of two pathways is equivalent to the cost of their differences:

$$\begin{aligned} \eta(M) - \eta(M') &= \sum_{i=1}^k \beta_i [G_i] - \sum_{i=1}^k \beta_i [G_i]' \\ &= \sum_{i=1}^k \beta_i ([G_i] - [G_i]') \end{aligned}$$

It is also possible to determine the optimal fitness in the case where only the target gene is produced and all other genes are not present in equilibrium. For an optimal

genotype M^* ,

$$fitness(M^*) = \frac{\sigma[G_k]}{\beta_k[G_k]} = \sigma/\beta_k.$$

Therefore if $\beta_i = 1/k$ and $\sigma = 1$, the optimal fitness is equal to k . A normalized fitness, with optimality at 1 and minimality at 0, is:

$$fitness_{norm}(M) = \frac{\beta_k fitness(M)}{\sigma}.$$

Then, in the normalized case, σ is not needed.

As presented, fitness is a function of the pathway genotype, although all calculations are done using the phenotype, or equilibrium concentrations. Therefore an implementation of fitness must first convert the genotype to equilibrium protein concentrations (using the *continuous*, *discrete*, or other methods) and then impose the benefit and cost functions. Alternatively, fitness could operate on other properties and phenotypes. In this case, depending on the study, the fitness function could use the benefit and cost ratio as guidelines for the development of a relevant soft fitness function. For situations where the benefits and costs are not known or results cannot be constrained by a particular fitness function, it is possible to supply parameterized random fitness functions that describe marginal selective advantage.

2.5.2 Random fitness

Random fitness landscapes are useful parameterizations of empirically unknown distributions by providing the effect of mutation on a genotype’s fitness [38]. I implemented a discretized “stairway to heaven” landscape where the distribution of

selection coefficients is the same for all genotypes. As further work, more fitness landscapes can be tested.

The implemented random fitness landscape is defined as *random-Q*, with

$$P(\text{fitness}(M) = kQ^{-1}) = Q^{-1}, k \in \mathbb{Z}_{[0,Q]}.$$

Therefore, the distribution is independent and identical for every allele. The Q parameter is meaningful both before and after the population has discovered an optimal genotype. Prior to its discovery, Q^{-1} is the probability that an optimal pathway is found in the next mutation and so can be considered as the innovation parameter. Once an optimal genotype has been found, and the population is in the process of fixing on a single or a set of optimal genotypes, Q^{-1} is the probability that a subsequent mutation is also optimal, and so represents robustness.

For example, for a pathway allele M under random-10, the fitness is calculated by uniformly selecting an integer z from 0 to 10 and returning $1/z$. Since the same procedure is done for any arbitrary allele, the distributions are independent and identical.

2.6 Summary

The developed model provides a sequence-based realization of the pathway genotype within a population genetic framework. The fundamental data structure for a pathway preserves sequential order of the binding sites and genes by the order

of columns and rows of the pathway matrix. This novel data structure allows for more accurate representations of how mutation and recombination effect pathways. Furthermore, realistic rates for mutation and recombination on the pathway can be calculated using their well-known respective base pair rates and other estimable quantities (length of regulatory regions, binding site length, and pathway size). Finally, populations of genotypes evolve on both biologically motivated and random fitness landscapes.

This model is implemented in Python/C with the core routines and data structures written in C with Python wrappers.

Chapter 3

Case study: The establishment of a novel regulatory pathway

In this chapter I investigate the effect of non-coding DNA on the time to establishment of novel regulatory pathways within a finite population. Applying the developed sequence-based model of pathway evolution, I find that the length of regulatory promoter and enhancer regions is a major driving force of the establishment time of a novel pathway. The minimal establishment time for large genomes occurs in small populations and for small genomes it occurs in large populations, a pattern that closely matches the empirical scaling between population size and amount of intergenic DNA in nature. Furthermore, I ran simulations under the various fitness landscapes described in Chapter 2 and discovered that all landscapes, including random, yielded the same observations with regard to establishment times. These results provide new insight and theory on the functional role of non-coding DNA from

a pathway evolution perspective.

Establishment time of an allele measures the mutational origin plus fixation [36]. Since I am interested in optimal pathways, the time to mutational origin is the adaptation time, or time until an adaptive allele is first discovered in the population. Since the non-adaptive processes of evolution can have a significant and opposing influence on adaptation and fixation time, establishment time can capture evolutionary pitfalls that are missed when examined by only adaptation or only fixation. Therefore this analysis calculates establishment time as the sum of adaptation time and fixation time of optimal pathways. Since the developed model is sequence-based, results from understanding the evolution on the pathway level apply back to the sequence. Hence this analysis uses pathways to understand the functional importance of the variable length of non-coding sequences within the genome and thereby provide quantitative measures for the advantage and disadvantage imposed by non-coding DNA on the creation of novel pathways.

While theoretical work on the establishment time for abstract complex adaptations provides important insights that may be applied to regulatory pathways, pathway evolution is governed by several parameters and processes, such as the genomic architecture parameter, and thus analytical results may be overly simplistic [36, 37]. For example, prior calculations of establishment time solve for specific evolutionary scenarios like neutral intermediates, deleterious intermediates and diploidy [36]. For this reason, actual simulations, which take into account the full range of complexities, are needed to elucidate information about the establishment time.

3.1 Method

I designed an experiment to measure the establishment time of an optimal pathway from an initial population of pathways of solely self-regulating genes. The experiment begins with a monomorphic population of self-regulating genes and simulates the spontaneous gain and loss of enhancer and silencer transcription factor binding sites in the pathway. The time until the occurrence of the first optimal genotype is the population's *adaptation time*. There are potentially several genotypes that have optimal fitness, which is a reasonable occurrence in nature [14, 13]. *Establishment time* has been defined as the time elapsed until all pathways in a population have optimal fitness [36].

3.1.1 Genomic Architecture Parameter

Depending on the size of the regulatory substrate surrounding a gene, it may be more likely for a binding site to lose its affinity by a one-off mutation or for the promoter to spontaneously gain a new binding site. The ratio of the loss and gain rate of binding sites (μ_l/μ_g), or genomic architecture parameter (α), has been shown to be $4^n/L$, where n is the length of a binding site and L is the size of promoter and enhancer regions for the gene [2]. When the binding site length is held constant, low values ($\alpha \ll 1$) correspond to large regulatory regions whereas high values ($\alpha \gg 1$) correspond to small regulatory regions. A realistic range for known organisms is from 10^{-3} for mammals to 10^3 for prokaryotes.

For prokaryotic gene structures, L is determined by the average length of intergenic

regions upstream of genes and can therefore be approximated per gene by the number of non-coding bases divided by the number of genes in the genome. These numbers are readily available at NCBI-Entrez. For eukaryotic gene structures, promoters and enhancers exist upstream, within introns, and downstream of the gene; however, not all base pairs of this surrounding non-coding DNA provide adequate substrate for novel binding interactions. Recent work estimated the fraction of the human genome under selection somewhere between 2.4-11.8% [39]. Since the coding portion of the human genome comprises 1.2%, roughly 1-10% of non-coding DNA is under selection. For humans, then, the amount of regulatory substrate is an order of magnitude less (10%) than the total number of non-coding base pairs. Consequently, to calculate the number of nucleotides that may harbor a binding site in eukaryotes, the total length of non-coding DNA per gene was multiplied by 10%. Furthermore, while the length of a binding site varies greatly, this analysis will focus on the typical range of around 10 base pairs in length and hold the length of the binding site (b) constant for the calculation of the genomic architecture parameter. Therefore, changes in the genomic architecture parameter (α) arise from changes in the gain rate (μ_g), not the loss rate (μ_l), which is held constant with the length of a typical binding site.

3.1.2 Simulations

I simulated evolution under the given scenario for twenty populations ranging in size from 10^2 to 10^9 and twenty genomic architecture parameter values ranging from 10^{-3} to 10^4 , with an average of 100 samples for each combination of population size and α . Results for each phenotype/fitness function yielded a distribution of

Table 3.1: For each phenotype/fitness function, a log-linear fit was calculated for the minimal median establishment times with respect to α . The log-linear fit follows $A\alpha^I$, where A is the amplitude and I is the index. The log of the amplitude and the index are given in the table, along with the Pearson correlation its corresponding p-value. The same log-linear fit for the known organisms is shown in the first row labeled ‘organisms,’ calculated using the estimates for genomic architecture parameter and population size.

Fitness Function	index	log-amplitude	Pearson-r	p-value
organisms	0.87	6.4	0.98	0
discrete	0.70	7.5	0.88	3.8e-04
discrete-viable	0.79	7.5	0.94	1.2e-05
ode	0.73	7.7	0.82	2.0e-03
ode-viable	0.56	7.1	0.96	3.3e-06
random10	1.00	7.0	0.98	9.7e-08
random20	0.94	6.9	0.95	1.0e-05
random50	0.72	6.9	0.90	1.5e-04
random100	0.83	7.3	0.98	1.5e-07

establishment times $F_{N_e, \alpha}$ for each population size (N_e) and genomic architecture parameter (α). Due to the exponential spread of establishment times, averages are severely right-skewed. Because of this, I examine the median of establishment times (for a further discussion on this decision see Section 3.3.1). To quantify the optimal population size for a given genomic architecture parameter, I will use the minimal median establishment time with respect to α , which is calculated by:

$$\text{minimal-median}(\alpha) = \underset{N_e}{\operatorname{argmin}} \operatorname{median}(F_{N_e, \alpha}).$$

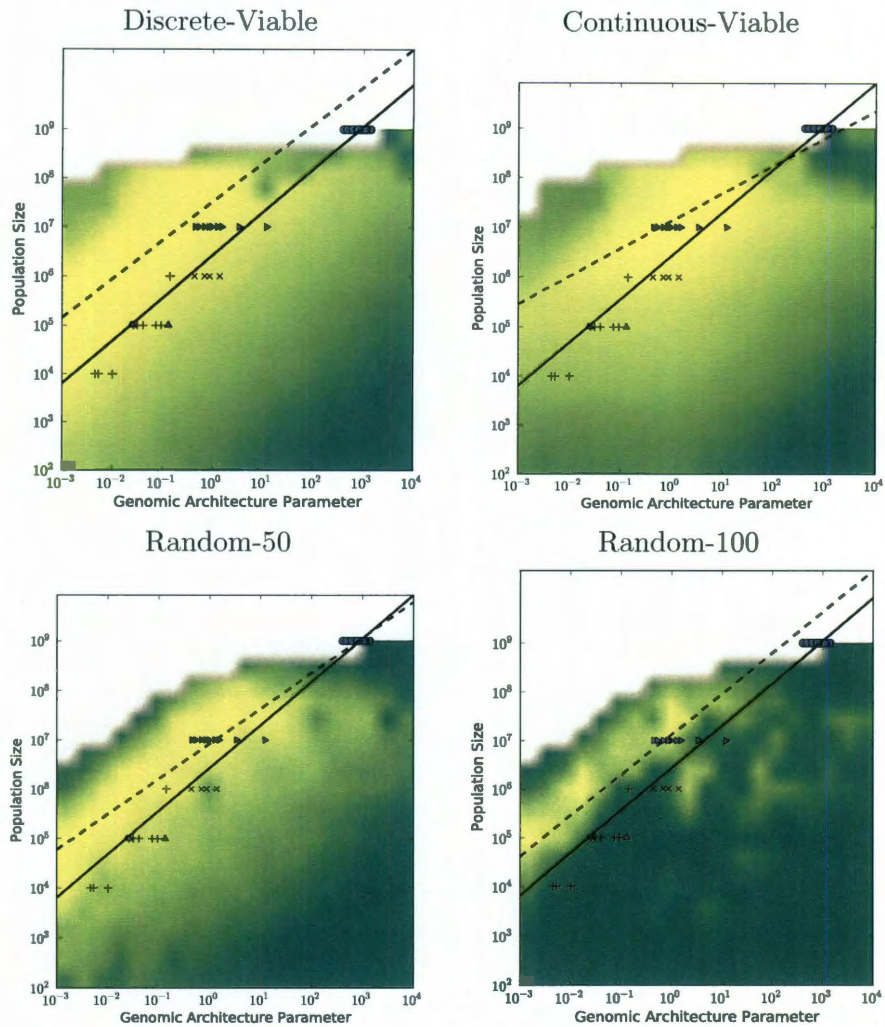


Figure 3.1: Contour plots of the median establishment time (yellow is faster) for each population and genomic architecture parameter are shown for four of the nine examined phenotype/fitness functions. Each contour plot was generated using natural neighbor interpolation of the simulation results. On all plots, a noticeable diagonal band indicates the fastest establishment times (least squares fit with dashed line), which parallel the scaling of population and genomic architecture parameter found in nature (scatter points fit with solid black line). The comparison of the fit of the dashed line (fastest establishment times) and the scaling found in nature is shown in Table 3.1.

3.2 Results

3.2.1 Minimal median establishment time coincides with natural scaling of population and non-coding DNA

Establishment of a novel pathway involves two steps: first the discovery of the optimal pathway, and second, the movement of the unfit portion of the population to optimal alleles. This study is ultimately interested in the time it takes to complete both steps as neither are trivial depending on the population size and genomic architecture parameter. Figure 3.1 depicts the results for each fitness function as a contour plot of the median establishment time for each population and genomic architecture parameter. The optimal population size for minimizing the median establishment time can be quantified as a log-linear function of the genomic architecture parameter (see Table 3.1). The close fit of fast median establishment times with the scaling of population and non-coding DNA can be visually understood by the strong diagonal band evident in all phenotype/fitness functions (see Figure 3.1). Indeed, the optimality of establishment time for novel regulatory pathways coincides with the scaling of population size and genomic architectures of the natural system.

Two major properties of this diagonal pattern are interesting: first, small populations with large regulatory regions are at a significant advantage to larger populations in terms of establishment time; and second, this pattern persists across all tested fitness functions. In order to understand the emergence of this pattern, adaptation and fixation, the two components of establishment, must be examined.

Establishment time decomposes into the summation of time to adaptation, the

discovery of an optimal pathway, and time to fixation, where we relax fixation to mean that every individual in the population has an optimal pathway, but not necessarily all of them have the same optimal pathway allele. Once a population has adapted, it is in the process of fixation until it completes establishment. An allele (in our case an optimal pathway) is fixed once every individual shares that same allele or an allele with equivalent fitness. We can therefore examine the causes of long establishment times as a function of adaptation and fixation.

3.2.2 Median adaptation time scales with genomic architecture parameter

In order to understand adaptation time on the pathway genotype, let d be the number of mutations (either loss or gain) required to find an optimal pathway. Because the number of mutations required to find an optimal allele changes depending on the evolutionary trajectory of the population, d can take on a variety of values based on the population size and other factors. Define a random variable D as the number of mutations required for the population to adapt, and let $\mathbb{E}(D) = d$. Distribution statistics for D , given in Table 3.2, identify d to be on average around 3-4 mutations, which is in the range studied by other works on complex adaptations [36, 37]. However, in very few cases, D can reach to around 60 mutations, which undoubtedly would have a strong effect on the adaptation time, but not the overall median adaptation time.

For simplicity, let the d intermediate pathways be neutral until a final optimal

pathway is discovered. Lynch provided intuition for the adaptation time in this scenario: a population’s ability to adapt is proportional to the arrival rate of mutants [36]. Because d varies negligibly with phenotype/fitness function, population size, and genomic architecture parameter (see Table 3.2), its variation must not be a major factor in the determination of adaptation time for these simulations. Rather, we expect the arrival rate of these d mutations, which is governed by the population size and genomic architecture parameter, to have a significant effect on adaptation time.

This study uses two mutation rates, one for gain and one for loss, and so the population mutation rate is $2N(\mu_l + \mu_g)$. Rearranging terms to tease out the genomic architecture parameter, the population mutation rate is $2N\mu_l(1 + \alpha^{-1})$. Since $\mu_l = nu$, where b is the length of the binding site and u is the mutation rate, is constant in this experiment, we can understand the effect of population mutation rate on adaptation time solely as a function of the population size and genomic architecture parameter.

As shown in Figure 3.2, adaptation time scales positively with $(1 + \alpha^{-1})$ and negatively with the population size. This relationship is strongly log-linear with respect to $N(1 + \alpha^{-1})$ (Pearson R is -0.97, p-value is 10^{-24}). The calculation of this relationship is truncated when the population mutation rate exceeds 10, wherein the population mutation rate has limited effect on the adaptation time. When the population is producing on average 10 mutants per generation, the properties of the phenotype space constrain the discovery of beneficial alleles, through fitness valleys, plains, or inclines. At this high rate of mutants, the entire neighborhood of one-off mutations is explored, and so a fitness valley, wherein no genotypes are optimal,

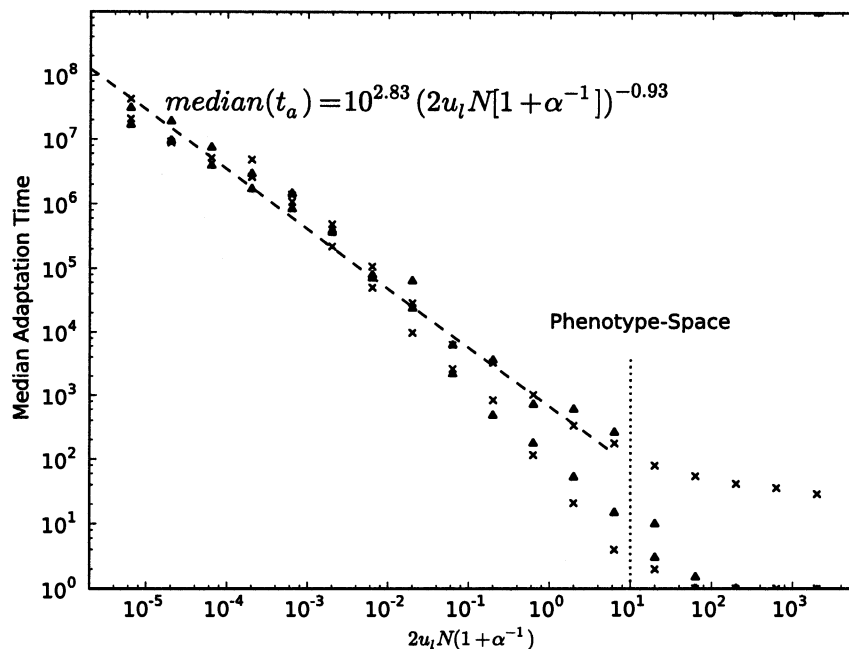


Figure 3.2: Median adaptation time (t_a) as a function of $2\mu_l N(1 + \alpha^{-1})$ for all phenotype/fitness functions. Triangles denote results from random fitness functions. The Pearson correlation of the fit is -0.98 and is highly significant. The vertical dotted line denotes the point at which median establishment time has “bottomed out” relative to population mutation rate. Median adaptation time can be reduced either by increasing N or increasing the amount of non-coding DNA (decreasing α).

limits the discovery of optimal alleles. However, in this interval where the population mutation rate is greater than 10, median adaptation times are minimal (< 10) in comparison.

When $\alpha > 1$, the population size largely determines the population mutation rate (holding μ_l constant), and so the effect of promoter region length on adaptation time occurs when promoters are large, or $\alpha < 1$. While other variables factor into the variation of median adaptation time, such as properties of the phenotype/fitness

Table 3.2: Distribution statistics of D , the number of mutations required for a population to adapt. Average, standard deviation, minimum and maximum values were calculated for each $D_{N_e, \alpha}$, and then averaged to report a single value for each phenotype/fitness function.

Fitness Function	$avg(D) \pm std(D)$	min(D)	max(D)
discrete	3.12 ± 1.83	2	58
discrete-viable	2.91 ± 2.36	2	58
ode	3.01 ± 1.45	2	44
ode-viable	3.98 ± 1.18	3	29
random10	2.48 ± 1.03	1	10
random20	2.91 ± 1.16	1	10
random50	3.40 ± 1.34	1	10
random100	3.68 ± 1.46	1	14

function, they are relatively small in comparison to the joint effect of population size and genomic architecture parameter.

While this follows the intuition in [36], it is surprising when compared to the results investigating the effect of robustness and innovation on adaptation time by Draghi *et al.* [24]. Draghi *et al.* reported the strong effect of robustness on the mean time to the first beneficial allele; however, in our study there is little effect that phenotype/fitness function properties, such as robustness, have on median adaptation time in comparison to the effect of population size or amount of regulatory substrate. Other factors not considered in [24], like a soft fitness landscape during adaptation, could account for these differences. With that said, properties like robustness affect the adaptation time within an order of magnitude, which may be quite large, especially when the range is $10^7 - 10^8$. Unlike population size or amount of non-coding DNA, robustness is a static property of an allele and its associated genotype/phenotype space. Across

a generation, the population size and amount of non-coding DNA may change, but the genotype/phenotype space does not. Therefore, evolution more likely operates within expanding and shrinking of population and promoter size.

The stronger effect of population size and non-coding DNA over phenotype/fitness properties has interesting biological implications: regardless of the pathway function, holding the amount of non-coding DNA constant, adaptation time decreases with an increase in population size. On the other hand, holding the population size constant, expansion of non-coding DNA regions that serve as substrate for novel regulatory interactions decreases adaptation time. Based solely on these two observations, the optimal population for median adaptation time would be both large in population and genome size. However, adaptation time does not consider the further accumulation of mutations following the discovery of an optimal allele, which would be rampant in such a population. Intuitively, the forces that drive rapid adaptation also undermine fixation. Hence, to understand this tradeoff we must examine fixation time.

3.2.3 Fixation time increases with promoter and population size

A population is established when all pathways in the population are optimal, and so prolonged fixation time occurs when there are deleterious pathways in the population. In a given generation prior to fixation, these deleterious pathways either already existed from the previous generation or resulted from mutation because the arrival of mutants is faster than the population can purge them through selection or

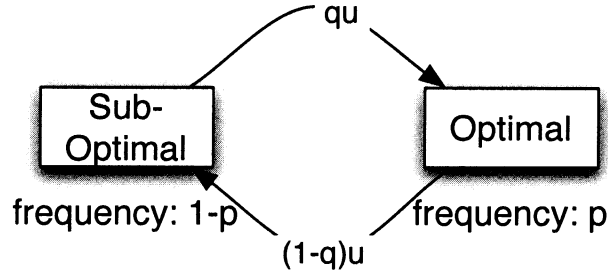
drift. To examine this phenomenon, I constructed a simple model of the fixation of an allele (depicted in Figure 3.3):

- representing optimal pathways with frequency p ,
- mutation rate u ,
- N as population size,
- q as the probability of producing an optimal mutant (or the robustness of the optimal alleles), and
- s as the selective advantage of the optimal pathway

$$\text{fitness}(\textit{optimal}) = \text{fitness}(\textit{sub.optimal}) + s.$$

Mutation between optimal and suboptimal alleles occur with rate $(1 - q)u$.

Results from computational simulations of this model are shown in Figure 3.4. This model is useful because it explains the effect of $N, u, q,$ and s on fixation time. First, as Nu increases, on average more mutations occur each generation which have a deleterious effect on fixation time, especially when $N(1 - q)u > 1$. Second, as q decreases (or $1 - q$ increases), the increased arrival rate of deleterious mutations also increases fixation time. Finally, as s increases, selection becomes a stronger force and decreases fixation time. In terms of phenotype/fitness function properties, fixation time decreases as either the robustness increases or the difference in reproductive advantage widens between the optimal and sub-optimal alleles. For any given robustness



$$\text{fitness}(\text{Optimal}) = \text{fitness}(\text{Sub-Optimal}) + s$$

Figure 3.3: The fixation model which incorporates drift, mutation, and selection. Mutations occur at a rate u , with a probability q of producing an optimal allele. The fitness difference between optimal and sub-optimal alleles is given by s .

and phenotype function, fixation time monotonically increases with the population mutation rate, and there is a significant increase in fixation time around $Nu > 1$.

We can compare this model to the patterns of fixation time discovered by the populations of this study (see Figure 3.4). Fixation time of the simulated populations follows the insights of the model. As before, the population mutation rate under the genomic architecture parameter is $N\mu_i(1 + \alpha^{-1})$, which is a function of N and α because μ_i is constant. Take, for example, the median fixation times of *ode-viable* and *discrete-viable*. The robustness values as measured by Q during fixation for the various landscapes are shown in Table 3.3. The robustness during fixation over all simulations for *discrete-viable* was found to be 0.48 and *ode-viable* to be 0.37. According to this model, all else equal, the minor difference in robustness will have a dramatic effect on fixation time when $Nu > 1$. Indeed, when $Nu \approx 10$, the difference between the two phenotype functions becomes around 10^4 , where previously their separation was

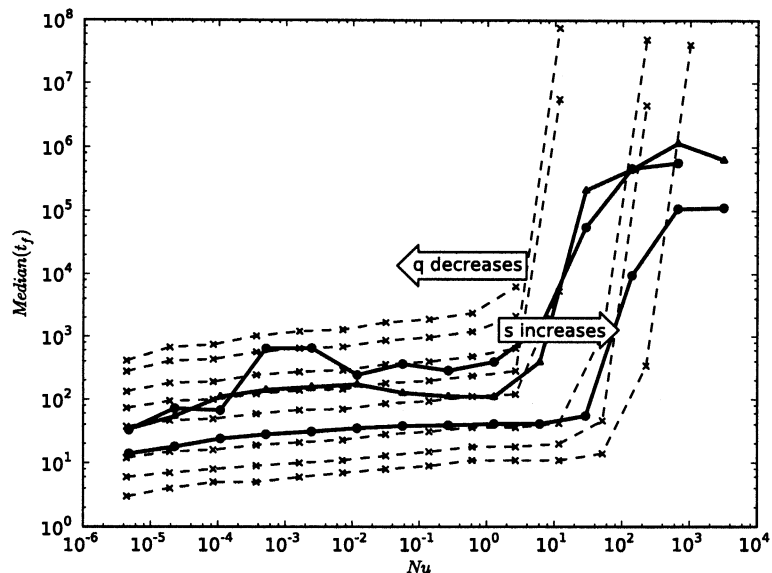


Figure 3.4: The results of the fixation model are shown on top. This plot explains the joint effect of Nu and the robustness of the optimal-neutral space, q , on the median fixation time. At $Nu \approx 1$, there is a significant increase in the time to fixation. This threshold can be shifted right by increases in s , the selection coefficient, or q . Solid, bold curves depict median fixation time for different phenotype/fitness functions. In this case, $2Nu = 2N(\mu_g + \mu_l) = 2N\mu_l(1 + \alpha^{-1})$. It is evident that results for these phenotype functions follow the pattern explained by the fixation model. Simulations for the fixation model are truncated at 10^8 .

roughly 10^2 . In addition, according to the model, subsequent mutations in the *ode-viable* landscape on average have a minor change to the section coefficient ($s = 0.1$), whereas mutations in the *discrete-viable* landscape on average are more consequential ($s = 0.5$). Because of this, selection is a stronger force in the fixation of populations under the *discrete-viable* function, and so fixation time is diminished.

Table 3.3: The robustness, or probability that subsequent mutations are also optimal, during the fixation phase of establishment time for the different phenotype/fitness functions. For each fitness function, Q is the distribution of robustness across all simulated population sizes and genomic architecture parameter values.

Fitness Function	$avg(Q) \pm std(Q)$
discrete	0.48 ± 0.09
discrete-viable	0.43 ± 0.08
ode	0.41 ± 0.07
ode-viable	0.37 ± 0.1
random10	0.10 ± 0.05
random20	0.05 ± 0.03
random50	0.02 ± 0.03
random100	0.01 ± 0.02

One major difference between the fixation model and the simulation results is the plateauing of fixation time in high mutation environments. Where the model predicts further increases in fixation time, the simulated results plateau depending on the phenotype landscape. This is an interesting contrast to adaptation time, where the phenotype function played little role in the variation of median adaptation times. We can attribute this plateau to the saturation of the pathway genotype. High mutation environments ($Nu \gg 1$) correspond with a high binding site gain rate. As the number of mutations accumulate over time, the pathway ultimately saturates with binding sites. In this study, discretizing the pathway genotype imposes this limit, but other studies have imposed this same saturation criteria using hard limits on the number of occupied binding sites [22, 23].

While there exists clear variation between the fixation model used here and the simulation results, the overarching trends in median fixation time can be explained

by the population size, genomic architecture parameter, robustness, and selection coefficient.

3.3 Discussion

When the results for adaptation and fixation time are superimposed (see Figure 3.5), the scaling of minimal median establishment times rests around a population mutation rate of 1 to 100. Because the pathway is a unique genotype that is affected by non-coding DNA in a predictable manner, it is possible to understand this scaling as a function of intergenic DNA using the genomic architecture parameter. When this trough is expanded across valid combinations of N and α , the diagonal scaling emerges.

By decomposing the contribution of adaptation and fixation to the overall establishment time, I elucidated the components under study that gave rise to variations on the overall diagonal-band pattern. First, adaptation time is largely governed by the population mutation rate, which can be expressed as a function of population size and the genomic architecture parameter ($2Nu$ where $u = \mu_l + \mu_g = \mu_l(1 + \alpha^{-1})$). Differences between phenotype spaces accounted for very little variation between median adaptation times. However, this is a surprising result provided the amount of attention that innovation and robustness are given in the adaptive process [24, 22, 23, 40, 41, 42, 43, 44]. Our results indicate that commensurate changes to population size or mutation rate would have a more powerful effect than changes to robustness or innovation. The pathway, as an allele, is unique because it is possible

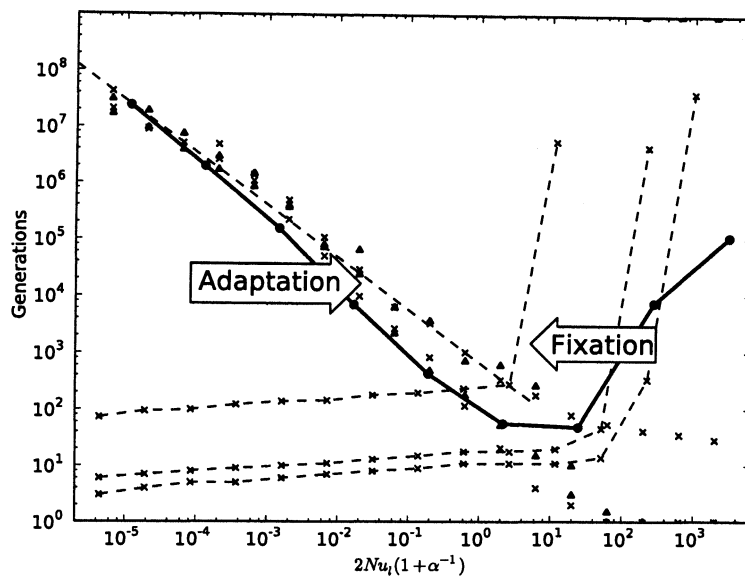


Figure 3.5: The trough formed by median adaptation and fixation times across values of the population mutation rate ($2N\mu_l[1 + \alpha^{-1}]$). Minimal establishment time occurs at the intersection of these curves, which turns into a diagonal band when examined across values for N and α . The thick line denotes the median establishment time for *discrete-viable*. The several lines for fixation illustrate the strong effect of phenotype space properties on median fixation time, in comparison to the weak effect these properties have on adaptation.

to increase or decrease the mutation rate of the pathway by increasing or decreasing the amount of regulatory substrate. Therefore, as an evolutionary strategy, it may be easier for a population to grow or shrink intergenic regions to accommodate for low or high mutational robustness. Furthermore, in this study, the effect of robustness manifests in fixation time, not in adaptation. Though this observation is specific to the pathway genotype model presented in this thesis, further studies should investigate the effect of robustness and innovation on fixation for other allele-types or pathway

models, and moreover for the entire establishment process.

A distinguishing feature of this study, in comparison to others that investigate establishment, is that mutations could continue to accumulate after a target allele was discovered by the population [36, 37]. The results in [36, 37] do not allow for continued mutation during the time to fixation; rather, the authors assume that the population reaches a fitness peak or evolutionary “dead end.” Our results indicate that this is not the case, and mutation will continue to drive the population to a mutation-selection-drift balance, which is the determining factor in fixation.

Since we implemented two pathway phenotype models reported in the literature, it is worth comparing them under the perspective of establishment time. First, the viability constraint decreases robustness of both *discrete* and *continuous* by 10% each. Because viability reduces the number of possible optimal pathways, this decrease is expected. However, the reduced number of optimal pathways imposed by viability weakly affected the average number of mutations required to find an optimal allele. Viability, then, is a weak constraint that has a minor effect on establishment time.

The differences between *discrete* and *continuous*, however, were much stronger. In terms of adaptation time, the closest optimal allele for the continuous function required three subsequent mutations from the genotype of the initial population, whereas the closest discrete allele only required two. As for fixation, the continuous function had lower robustness and the unfit neighbors had a minimal difference in fitness, reducing the effect of selection and increasing the fixation time. For the discrete function, higher robustness and higher differences in fitness between the optimal and unfit alleles resulted in decreased fixation time.

While intuitively a realistic pathway model, like *discrete* or *continuous*, should fit better to the empirical data of population size and non-coding DNA per gene, the best fit resulted from the random-100 function (see Table 3.1). Because the empirical data averages the amount of non-coding DNA across the entire genome, it is an aggregate statistic across all the pathways present in an individual. All these pathways do not operate under the optimality criteria of maximizing the target gene production, and, as such, should not necessarily fall into the optimal establishment times explained by such a criterion. Instead, as we found in the simulations, an aggregate statistic would most likely behave independent of the pathway function but dependent on the pathway structure and genotype. Hence, random-100 provides a best fit to the empirical data.

3.3.1 Median vs. mean

Central to the findings of this study is the use of the median statistic. Almost all other works estimate establishment time, adaptation time and fixation time using their mean [36, 37, 24]. This is not because the median has “lesser” biological meaning, rather because the median is difficult to determine theoretically since it requires a complete understanding of the cumulative distribution function. However, simulations approximate the cumulative distribution function, and, in doing so, allows for the estimation of non-moment based statistics like the median.

Furthermore, the same analysis performed with the mean, rather than the median, yields highly divergent results. In fact, the mean shows no scaling with the empirical population size and genomic architecture. Instead, the mean establishment

time decreases, more or less, with the population mutation rate. However, the same analysis performed with the third quartile (75th percentile) still displays the scaling with empirical data. Therefore, in this study, averages represent the slowest 25% of population ensembles.

3.4 Conclusions

As shown in many previous studies, organism complexity scales not with number of genes or genome size but rather with the amount of non-coding DNA [1]. This phenomenon revolves around the eukaryotic gene structure, which allows coding regions to be couched within even larger amounts of non-coding nucleotides by way of introns and long 5' UTRs. In [1], Lynch proposed a null theory for the origin of eukaryotic gene structure that targets the emergence of this non-coding padding to the shrinking effective population sizes of complex organisms, resulting in evolution dominated by random drift rather than selection. Since smaller populations provide an amenable environment to the fixation of deleterious mutations, these smaller populations were unable to purge nearly neutral expansions of non-coding regions and the eukaryotic gene as we know it today took shape.

It is not clear whether this drift-based solution to complexity is sufficient to explain the success of the eukaryotic genome [31]. To understand the benefits and costs of non-coding DNA to a population, one must assign function to mutations and variation within the expanding non-coding regions of genomes. In order to understand the effects of mutations in these enigmatic regions, I approached the problem with a

pathway-based perspective. By attributing the ratio of loss and gain of transcription factor binding sites to the length of potential regulatory-harboring sequence per gene, I quantified the relative effects of drift, mutation, and selection in forming novel pathways. In addition, by using parameterized random fitness landscapes, results can be generalized beyond the specific optimality criteria enforced by the *discrete* and *continuous* phenotype functions. In the end, we are left with a model that predicts the correlation of population size and length of non-coding promoter regions as defined by minimal establishment times of novel pathways. These results explain that as the effective population size shrinks, there is an indirect selection on larger promoter regions for the development of novel pathways.

This evolutionary pathway perspective provides a reasonable quantification for the known covariation of population size and expansion of non-coding genomic regions. In doing so, I have quantified the evolutionary advantages and disadvantages of non-coding DNA on pathways using establishment time, determined and examined the important parameters of establishment time of pathways, and verified a sequence-based regulatory pathway model.

Chapter 4

Discussion and conclusions

The sequence-based pathway model developed in this thesis fills an important gap in current pathway models, and, in doing so, provides insight to open problems and also the ability to reexamine problems from a population-genetic, pathway-based perspective. I verified the usefulness of this model by revisiting the issue of strong correlation between non-coding DNA and population size. By attributing significance to non-coding DNA using minimal median establishment time, I showed that optimal population sizes and lengths of non-coding DNA for establishing novel pathways coincided with the known population sizes and non-coding expansions for organisms across the tree of life. These results underline three critical and distinguishing features of my pathway model: first, the preservation of sequence information in the pathway genotype; second, the modeling of evolution within a population; and finally, the use of simulation rather than analytical formulations.

Preserving sequence within the pathway model is reflected in the choice of data

structure, the use of two, yet related, mutation rates that are a function of sequence mutation rates, and two recombination rates that, too, are a function of sequence recombination rates. The data structure forms the basis for the genotype and defines how mutation and recombination operate. A poor choice of data structure may encode too little information to reliably model mutation and recombination. For instance, previous models of pathway data structures used the adjacency matrix of the pathway, but this choice divorces the pathway from its sequential underpinnings [2, 3, 22, 14]. Mutation on the adjacency matrix amounts to the gain and loss of gene interactions - not binding sites. If each gene interaction maps to a distinct binding site, gene interactions and binding sites are synonymous; otherwise, they are not. However, repeated binding motifs and pleiotropy of binding factors are observed in nature, and therefore are missing in current pathway models [45]. Furthermore, the strong results in Chapter 3 center around mutation rates that can be calculated from empirical data. Without empirical rates, comparing to observed data would be unfruitful, and in fact many evolutionary studies cannot make informative comparisons to empirical data [3, 24, 22, 36, 37]. By carefully developing a sequenced-based pathway model such that pathway level mutations are a function of sequence characteristics, simulation results can be compared back to these observable quantities.

As for recombination, it is impossible to properly model recombination between binding sites using an adjacency matrix. However, in studying recombination on the robustness of a pathway, Martin *et al.* argued that a recombination event is unlikely to occur between binding sites because they are closely linked, so only free recombination between unlinked genes was included in their model [3]. Martin *et al.* did not justify

this assumption. In fact, according to results from the HapMap project, transcription factor binding sites show decreased levels of linkage disequilibrium in comparison to coding sequences [46]. In addition, given that the per base pair recombination rate c ranges from 10^{-6} to 10^{-10} and the amount of regulatory substrate surrounding eukaryotic genes ranges from several to hundreds of kilobases, the rate of recombination within regulatory regions is between 10^{-3} and 10^5 (using $r = 0.5[1 - e^{2dc}]$). From this, in each generation there may be multiple recombination events between binding sites within the population. Consequently, the analysis of recombination on robustness done by Martin *et al.* is missing a major piece of the puzzle. In order to correctly incorporate recombination within binding regions, though, the fundamental data structure must change from the adjacency matrix to one that preserves sequence order. Unfortunately, in [2], Lynch recognized the importance of recombination between binding sites but still used the same adjacency matrix formulation of a pathway, resulting in an incorrect implementation of recombination within regulatory regions. Since the study was simulation based, the effect of recombination on evolving redundancy for a three gene pathway is undoubtedly incorrect as well. Ultimately, despite the usefulness of the graph abstraction, evolution operates on DNA, and so any model of pathway evolution must be based on genome sequence evolution. Therefore, the representation of the pathway should reflect the underlying genomic sequence.

A population genetics approach to evolution can reveal surprising insights. For example, previous to the results of [24], the intrinsic dichotomy between robustness and innovation created an unmanageable tradeoff: how could evolution operate on a

genotype that is robust, and how could an innovative genotype protect against the vagaries of evolution? Draghi *et al.* reconciled this issue using population genetics to show that populations can be both robust and evolvable at the same time [24]. When it comes to pathways, the literature is ripe with adaptive arguments for patterns of redundancy, robustness, motifs, and modularity [14, 19, 16, 17, 18, 19, 20, 21, 22]. However, only a few models of pathway evolution incorporate neutral forces, and none incorporates them completely correctly [3, 2]. Consequently, the majority of insights on the formation of pathway structure are devoid of non-adaptive influence [2]. The success of applying population genetics and pathways can be seen in the results of the establishment time study. This study leveraged both adaptive and non-adaptive forces in the calculation of establishment time for novel pathways. Because the pathway model developed in this thesis supports mutation, selection, and drift, it is possible to determine the balance of these three forces combined within a population.

This thesis provides a pathway model built from the sequence up and designed for simulation studies. While much work in population genetics uses simulation as a validation tool, the major contributions are presented as an analytical solution [36, 37, 24, 38]. However, due to the randomness introduced by drift, recombination, and mutation, a deterministic solution must be a statistic of the underlying probability distribution. For ease of developing analytical solutions, these analyses almost always measure means or other moments. Other useful statistics, like the median used in the establishment study, are avoided because estimating quantiles requires knowledge of the cumulative distribution function, or the complete description of the probability distribution. Thus much of population theory revolves around the

mean, which despite being a very informative measure, poorly represents skewed distributions. On the contrary, simulations approximate the entire distribution, thereby allowing for any meaningful statistical analysis. The establishment study in this thesis exemplifies the benefits of a good simulated model. As determined by the simulations, establishment time is a highly skewed distribution. Consequently, the mean lies above the 90th percentile and so poorly represents a biologically meaningful quantity. Reasonable quantiles (from 10% to 75%), however, are corroborated by the empirical evidence of non-coding DNA and population size and represent more reasonable biological quantities. Furthermore, complex alleles like pathways introduce configuration spaces that are intractable from an analytical perspective [2]. Solving various population genetic measures for each configuration, or possible pathway topology, is not a viable approach to understanding pathways. Simulation studies are not hindered by large configuration spaces, and so future work in the field of pathway evolution will rest on good digital models. In addition, since the model presented in this thesis rectifies many fundamental problems in previous approaches, it will serve as a critical base for future work in pathway evolution.

In conclusion, this thesis presents a model of pathway evolution that preserves sequence structure, incorporates population-genetics, and can handle the complexities introduced by the pathway allele and population genetics using simulations. In combination, these three features distinguish the model from previous work and fill a critical need which is evidenced by the ill designed models of pathway evolution in the literature. As a validation step, I successfully leveraged this model in the investigation of non-coding DNA and its effect on the establishment of optimal pathways, which

resulted in strong agreement with known amounts of non-coding DNA and population sizes. Hence, the developed model was central in quantifying an important advantage provided by non-coding DNA. Therefore, this thesis builds a pathway model of evolution from its underlying genomic context and validates the model against a pertinent and open problem in genome evolution.

Bibliography

- [1] M. Lynch, “The origins of eukaryotic gene structure,” *Mol Biol Evol*, vol. 23, pp. 450–68, Feb 2006.
- [2] M. Lynch, “The evolution of genetic networks by non-adaptive processes,” *Nature Reviews Genetics*, vol. 8, pp. 803–13, Oct 2007.
- [3] O. C. Martin and A. Wagner, “Effects of recombination on complex regulatory circuits,” *Genetics*, vol. 183, pp. 673–84, Oct 2009.
- [4] P. N. Benfey and T. Mitchell-Olds, “From genotype to phenotype: systems biology meets natural variation,” *Science*, vol. 320, pp. 495–7, Apr 2008.
- [5] J. R. True and E. S. Haag, “Developmental system drift and flexibility in evolutionary trajectories,” *Evol Dev*, vol. 3, pp. 109–19, Jan 2001.
- [6] J. P. Balhoff and G. A. Wray, “Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites,” *Proc Natl Acad Sci USA*, vol. 102, pp. 8591–6, Jun 2005.
- [7] S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews, and C. Boone, “Systematic mapping of genetic interaction networks,” *Annu Rev Genet*, vol. 43, pp. 601–25, Jan 2009.
- [8] J. M. Akey, A. L. Ruhe, D. T. Akey, A. K. Wong, C. F. Connelly, J. Madeoy, T. J. Nicholas, and M. W. Neff, “Tracking footprints of artificial selection in the dog genome,” *Proc Natl Acad Sci USA*, vol. 107, pp. 1160–5, Jan 2010.
- [9] O. R. Homann, J. Dea, S. M. Noble, and A. D. Johnson, “A phenotypic profile of the candida albicans regulatory network,” *PLoS Genet*, vol. 5, p. e1000783, Dec 2009.

- [10] S. B. Carroll, "Evolution at two levels: on genes and form," *PLoS Biol*, vol. 3, p. e245, Jul 2005.
- [11] J. R. Stone and G. A. Wray, "Rapid evolution of cis-regulatory sequences via local point mutations," *Molecular Biology and Evolution*, vol. 18, pp. 1764–70, Sep 2001.
- [12] M. W. Hahn, J. E. Stajich, and G. A. Wray, "The effects of selection against spurious transcription factor binding sites," *Molecular Biology and Evolution*, vol. 20, pp. 901–6, Jun 2003.
- [13] A. E. Tsong, B. B. Tuch, H. Li, and A. D. Johnson, "Evolution of alternative transcriptional circuits with identical logic," *Nature*, vol. 443, pp. 415–20, Sep 2006.
- [14] W. Ma, A. Trusina, H. El-Samad, W. A. Lim, and C. Tang, "Defining network topologies that can achieve biochemical adaptation," *Cell*, vol. 138, pp. 760–73, Aug 2009.
- [15] E. Dekel, S. Mangan, and U. Alon, "Environmental selection of the feed-forward loop circuit in gene-regulation networks," *Physical Biology*, vol. 2, pp. 81–8, Jun 2005.
- [16] D. M. Stoebel, A. M. Dean, and D. E. Dykhuizen, "The cost of expression of escherichia coli lac operon proteins is in the process, not in the products," *Genetics*, vol. 178, pp. 1653–60, Mar 2008.
- [17] R. Losick and C. Desplan, "Stochasticity and cell fate," *Science*, vol. 320, pp. 65–8, Apr 2008.
- [18] M. Acar, J. Mettetal, and A. van Oudenaarden, "Stochastic switching as a survival strategy in fluctuating environments," *Nature Genetics*, vol. 40, no. 4, pp. 471–475, 2008.
- [19] E. Dekel and U. Alon, "Optimality and evolutionary tuning of the expression level of a protein," *Nature*, vol. 436, pp. 588–92, Jul 2005.
- [20] T. Ferenci, "Maintaining a healthy sparc balance through regulatory and mutational adaptation," *Molecular Microbiology*, vol. 57, pp. 1–8, Jul 2005.

- [21] R. U. Ibarra, J. S. Edwards, and B. O. Palsson, "Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth," *Nature*, vol. 420, pp. 186–9, Nov 2002.
- [22] S. Ciliberti, O. C. Martin, and A. Wagner, "Innovation and robustness in complex regulatory gene networks," *Proc Natl Acad Sci USA*, vol. 104, pp. 13591–6, Aug 2007.
- [23] A. Wagner, "Neutralism and selectionism: a network-based reconciliation," *Nature Reviews Genetics*, vol. 9, pp. 965–74, Dec 2008.
- [24] J. A. Draghi, T. L. Parsons, G. P. Wagner, and J. B. Plotkin, "Mutational robustness can facilitate adaptation," *Nature*, vol. 463, pp. 353–5, Jan 2010.
- [25] M. Lynch, "Evolution of the mutation rate," *Trends Genet*, vol. 26, pp. 345–52, Aug 2010.
- [26] M. W. Hahn and G. A. Wray, "The g-value paradox," *Evol Dev*, vol. 4, pp. 73–5, Jan 2002.
- [27] R. J. Taft, M. Pheasant, and J. S. Mattick, "The relationship between non-protein-coding dna and eukaryotic complexity," *Bioessays*, vol. 29, pp. 288–99, Mar 2007.
- [28] M. Lynch, B. Koskella, and S. Schaack, "Mutation pressure and the evolution of organelle genomic architecture," *Science*, vol. 311, pp. 1727–30, Mar 2006.
- [29] M. Lynch and J. S. Conery, "The origins of genome complexity," *Science*, vol. 302, pp. 1401–4, Nov 2003.
- [30] K. D. Whitney, E. J. Baack, J. L. Hamrick, M. J. W. Godt, B. C. Barringer, M. D. Bennett, C. G. Eckert, C. Goodwillie, S. Kalisz, I. J. Leitch, and J. Ross-Ibarra, "A role for nonadaptive processes in plant genome size evolution?," *Evolution*, vol. 64, pp. 2097–109, Jul 2010.
- [31] K. D. Whitney and T. Garland, "Did genetic drift drive increases in genome complexity?," *PLoS genetics*, vol. 6, Jan 2010.
- [32] M. Irimia, S. W. Roy, D. E. Neafsey, J. F. Abril, J. Garcia-Fernandez, and E. V. Koonin, "Complex selection on 5' splice sites in intron-rich organisms," *Genome Res*, vol. 19, pp. 2021–7, Nov 2009.

- [33] B. B. Tuch, H. Li, and A. D. Johnson, “Evolution of eukaryotic transcription circuits,” *Science*, vol. 319, pp. 1797–9, Mar 2008.
- [34] M. Lynch, “Streamlining and simplification of microbial genome architecture,” *Annu Rev Microbiol*, vol. 60, pp. 327–49, Jan 2006.
- [35] S. Roy, M. Werner-Washburne, and T. Lane, “A system for generating transcription regulatory networks with combinatorial control of transcription,” *Bioinformatics*, vol. 24, pp. 1318–20, May 2008.
- [36] M. Lynch and A. Abegg, “The rate of establishment of complex adaptations,” *Molecular Biology and Evolution*, vol. 27, pp. 1404–14, Jun 2010.
- [37] M. Lynch, “Scaling expectations for the time to establishment of complex adaptations,” *Proc Natl Acad Sci USA*, vol. 107, pp. 16577–82, Sep 2010.
- [38] S. Kryazhimskiy, G. Tkacik, and J. B. Plotkin, “The dynamics of adaptation on correlated fitness landscapes,” *Proc Natl Acad Sci USA*, vol. 106, pp. 18638–43, Nov 2009.
- [39] E. H. Margulies, G. M. Cooper, G. Asimenos, D. J. Thomas, C. N. Dewey, A. Siepel, E. Birney, D. Keefe, A. S. Schwartz, M. Hou, J. Taylor, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, J. B. Brown, P. Bickel, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, E. A. Stone, K. R. Rosenbloom, W. J. Kent, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, A. Hinrichs, H. Trumbower, H. Clawson, A. Zweig, R. M. Kuhn, G. Barber, R. Harte, D. Karolchik, M. A. Field, R. A. Moore, C. A. Matthewson, J. E. Schein, M. A. Marra, S. E. Antonarakis, S. Batzoglou, N. Goldman, R. Hardison, D. Haussler, W. Miller, L. Pachter, E. D. Green, and A. Sidow, “Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome,” *Genome Res*, vol. 17, pp. 760–74, Jun 2007.
- [40] C. B. Ogbunugafor, J. B. Pease, and P. E. Turner, “On the possible role of robustness in the evolution of infectious diseases,” *Chaos*, vol. 20, p. 026108, Jun 2010.

- [41] P. A. Lind, O. G. Berg, and D. I. Andersson, “Mutational robustness of ribosomal protein genes,” *Science*, vol. 330, pp. 825–7, Nov 2010.
- [42] R. E. Lenski, J. E. Barrick, and C. Ofria, “Balancing robustness and evolvability,” *PLoS Biol*, vol. 4, p. e428, Dec 2006.
- [43] C. Gokhale, Y. Iwasa, M. Nowak, and A. Traulsen, “The pace of evolution across fitness valleys,” *J Theor Biol*, vol. 259, no. 3, pp. 613–620, 2009.
- [44] E. Bornberg-Bauer and L. Kramer, “Robustness versus evolvability: a paradigm revisited,” *HFSP Journal*, vol. 4, pp. 105–108, May 2010.
- [45] G. P. Wagner and V. J. Lynch, “The gene regulatory logic of transcription factor evolution,” *Trends Ecol Evol (Amst)*, vol. 23, pp. 377–85, Jul 2008.
- [46] A. V. Smith, D. J. Thomas, H. M. Munro, and G. R. Abecasis, “Sequence features in regions of weak and strong linkage disequilibrium,” *Genome Res*, vol. 15, pp. 1519–34, Nov 2005.

Appendix

Error in binding site recombination on an adjacency matrix

In [2], Lynch investigates the effect of crossover events in the promoter region of a gene. In his implementation, he uses an array to store the frequencies of the different genotypes.

Allele designations:

```
Set[1...6][1...6][0...1]
```

First two numbers denotes the status of A and B:

s implies self-regulating;

^ implies it drives C;

> implies it drives the other transcription factor:

1. X
2. sX
3. sX>
4. X>
5. ^X
6. ^sX

7. $\sim sX$

8. $\sim X$

Third number denotes whether C is self-regulated (1) or not (0).

If we define a matrix $w_{i,j}$ as the regulatory relationship gene i regulates gene j , let's construct the row for gene $B = 1$ ($C = 0$ and $A = 2$) based on the above enumeration.

1. $(0 \ 0 \ 0)$

2. $(0 \ 1 \ 0)$

3. $(0 \ 1 \ 1)$

4. $(0 \ 0 \ 1)$

5. $(1 \ 0 \ 0)$

6. $(1 \ 1 \ 0)$

7. $(1 \ 1 \ 1)$

8. $(1 \ 0 \ 1)$

We can see that the implementation encodes the row type with a number and therefore the array called 'Set' keeps track of the frequency for each possible combination of rows. It is also important to note that the row does not represent a physical promoter region; rather, the column represents the physical promoter region in front of a gene, since the relationship 'A regulates B' implies a binding site in front of gene B for gene A. This means that each element in this row represents a binding site in a different promoter region.

In the code for recombination, each genotype pair will create recombinants at rates r_0 , r_1 , and r_2 , for 0, 1, and 2 crossover events, respectively. In this code below *pgen2c* array is the previously mentioned 'Set'.

```

/* impose recombination */

for (ig=1; ig<=8; ++ig) {
for (jg=1; jg<=8; ++jg) {
for (kg=0; kg<=1; ++kg) {
pgen2c[ig][jg][kg] = 0.0;
}}}

for (ig=1; ig<=8; ++ig) {
for (jg=1; jg<=8; ++jg) {
for (kg=0; kg<=1; ++kg) {
for (ig1=1; ig1<=8; ++ig1) {
for (jg1=1; jg1<=8; ++jg1) {
for (kg1=0; kg1<=1; ++kg1) {

pgen2c[ig][jg][kg] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec0);
pgen2c[ig1][jg1][kg1] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec0);

pgen2c[ig][jg][kg1] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec1);
pgen2c[ig1][jg1][kg] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec1);

pgen2c[ig][jg1][kg1] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec1);
pgen2c[ig1][jg][kg] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec1);

pgen2c[ig][jg1][kg] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec2);
pgen2c[ig1][jg][kg1] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec2);

}}}}}}

```

Basically, recombinations are implemented as switching of rows of the $w_{i,j}$ matrix. For example, in the calculation for r_2 , the nested indexing variables ig, jg, kg and $ig1, jg1, kg1$ show the switching of rows twice ($ig, jg1, kg$ and $ig1, jg, kg1$).

```

pgen2c[ig][jg1][kg] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec2);

```

```
pgen2c[ig1][jg][kg1] += ((pgen2b[ig][jg][kg] * pgen2b[ig1][jg1][kg1]) * rec2);
```

From a physical perspective, this does not make sense. Since rows are being switched, physical contiguity is thrown out the window as promoter regions are arbitrarily broken. Thus, this implementation is incorrect for any network incorporating 3 or more genes. It is necessary to keep track of the sequence of each binding site in the promoter, as explained in Section 2.2.1.