EVOLUTION OF TRANSCRIPTIONAL REGULATORY SEQUENCES

by

Murat Tuğrul

May, 2016

A Thesis
Presented to the Faculty of the Graduate School of the
Institute of Science and Technology Austria, Klosterneuburg, Austria
in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy

Supervisor: Nicholas H. Barton, IST Austria, Klosterneuburg, Austria

Co-supervisor: Gašper Tkačik, IST Austria, Klosterneuburg, Austria

Committee Member: Calin Guet, IST Austria, Klosterneuburg, Austria

Committee Member: Johannes Jaeger, Konrad Lorenz Institute, Klosterneuburg, Austria

Program Chair: Gašper Tkačik, IST Austria, Klosterneuburg, Austria

[redacted]

I hereby declare that this dissertation is my own work, and it does not contain other peoples work without this being so stated; and this thesis does not contain my previous work without this being stated, and that the bibliography contains all the literature that I used in writing the dissertation, and that all references refer to this bibliography.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Signature: ____________________________

Murat Tuğrul
July, 2016

Biographical Sketch

M.S., Physics, September 2009, The Institute for Cross-Disciplinary Physics and Complex Systems (University of Balearic Islands-CSIC), Palma de Mallorca, Spain

- Thesis Title: Simple Branching Models for Macroevolution
- Advisor: Emilio Hernández García
- Area of Study: Macroevolution (Speciation), Branching Process

M.S., Computational Sciences & Engineering, December 2007, Koç University, Sarıyer, Istanbul, Turkey

- Thesis Title: The Structure and Dynamics of Gene Regulation Networks
- Advisor: Alkan Kabakçıoğlu
- Area of Study: Gene Regulatory Networks, Boolean Dynamics

B.S., Physics, June 2005, Middle East Technical University, Ankara, Turkey

- With Honors in Advanced Group
- Mathematical Physics specialisation
- Minor degree in Philosophy and History of Science (at Philosophy Department)

List of Publications

*** = an outcome of this PhD thesis.

1. *** **Murat Tuğrul**, Tiago Paixão, Nicholas H. Barton, Gašper Tkačik (2015). "*Dynamics of Transcriptional Factor Binding Site Evolution*", PLoS Genetics 11(11): e1005639.

2. Stephanie Keller-Schmidt, **Murat Tuğrul**, Víctor M. Eguíluz, Emilio Hernández-García, and Konstantin Klemm. (2015) "*Anomalous scaling in an age-dependent branching model*," Physical Review E 91 (2), 022803.

3. **Murat Tuğrul** & Alkan Kabakçıoğlu. (2010) "*Anomalies in the transcriptional regulatory network of Saccharomyces Cerevisiae*", J. Theoretical Biology 263 (3), 328-336.

4. Emilio Hernández-García, **Murat Tuğrul**, E. Alejendro Herrada, Víctor M. Eguíluz and Konstantin Klemm. (2010) "*Simple Models for Phylogenetic Trees*", Int. J. Bifurcation and Chaos 20 (03), 805-811.

5. **Murat Tuğrul** & Alkan Kabakçıoğlu. (2010) "*Robustness of Transcriptional Regulation in Yeast-like Model Boolean Networks*", Int. J. Bifurcation and Chaos 20 (03), 929-935.

6. **Murat Tuğrul** (2009), Thesis for M.S. in Physics: "*Simple Branching Models for Macroevolution*", available at http://ifisc.uib-csic.es/publications

7. **Murat Tuğrul** (2007), Thesis for M.S. in Computational Sciences & Engineerings: "*The Structure and Dynamics of Gene Regulation Networks*", available at http://arxiv.org/abs/0802.1989

Acknowledgments

This PhD thesis may not have been completed without the help and care I received from some people during my PhD life. I am especially grateful to Tiago Paixão, Gašper Tkačik, Nick Barton, not only for their scientific advices but also for their patience and support. I thank Calin Guet and Jonathan Bollback for allowing me to "play around" in their labs and get some experience on experimental evolution. I thank Magdalena Steinrueck and Fabienne Jesse for collaborating and sharing their experimental data with me. I thank Johannes Jaeger for reviewing my thesis. I thank all members of Barton group (aka bartonians) for their feedback, and all workers of IST Austria for making the best working conditions. Lastly, I thank two special women, Nejla Sağlam and Setenay Doğan, for their continuous support and encouragement. I truly had a great chance of having right people around me.

The work described in this thesis partly has appeared in the following publication:

Chapter 2: **Murat Tuğrul**, Tiago Paixão, Nicholas H. Barton, Gašper Tkačik (2015). "*Dynamics of Transcriptional Factor Binding Site Evolution*", PLoS Genetics 11(11): e1005639.

Abstract

Evolution of gene regulation is important for phenotypic evolution and diversity. Sequence-specific binding of regulatory proteins is one of the key regulatory mechanisms determining gene expression. Although there has been intense interest in evolution of regulatory binding sites in the last decades, a theoretical understanding is far from being complete. In this thesis, I aim at a better understanding of the evolution of transcriptional regulatory binding sequences by using biophysical and population genetic models.

In the first part of the thesis, I discuss how to formulate the evolutionary dynamics of binding sequences in a single isolated binding site and in promoter/enhancer regions. I develop a theoretical framework bridging between a thermodynamical model for transcription and a mutation-selection-drift model for monomorphic populations. I mainly address the typical evolutionary rates, and how they depend on biophysical parameters (e.g. binding length and specificity) and population genetic parameters (e.g. population size and selection strength).

In the second part of the thesis, I analyse empirical data for a better evolutionary and biophysical understanding of sequence-specific binding of bacterial RNA polymerase. First, I infer selection on regulatory and non-regulatory binding sites of RNA polymerase in the *E. coli* K12 genome. Second, I infer the chemical potential of RNA polymerase, an important but unknown physical parameter defining the threshold energy for strong binding. Furthermore, I try to understand the relation between the lac promoter sequence diversity and the LacZ activity variation among 20 bacterial isolates by constructing a simple but biophysically motivated gene expression model. Lastly, I lay out a statistical framework to predict adaptive point mutations in *de novo* promoter evolution in a selection experiment.

Table of Contents

List of Figures

"The texture of this world is made up of necessity and chance."

Johann Wolfgang von Goethe

1

General Introduction

3.9 billion years of organic evolution has created an immense diversity on the earth. There are millions of different species, each of which typically consists of characteristically differing populations, which are likely to have individuals differing in phenotypic traits. Phenotypic diversity is not restricted to overall body shapes, but ranges from molecular levels to behavioural patterns, and for humans, includes medically important matters such as antibiotic resistance or susceptibility to cancer. One of the main subjects of evolutionary biology is to elucidate the sources, mechanisms and conditions that drive the evolutionary dynamics creating these variations.

An important observation has been that phenotypic traits typically show high heritability. In other words, they are systemically passed over from parents to offspring (Barton and Keightley, 2002). The importance of hereditary mechanisms was clear even for Darwin and Wallace's evolutionary theory (Darwin, 1859), however, they lacked molecular insights for the mechanistic basis. The rediscovery of Mendel's work in the beginning of the twentieth century, followed by the discovery that chromosomes are the carriers of heritable materials provided the first understanding of hereditary mechanisms and established the science of genetics. As a concrete body of mathematical theory synthesising the Darwinian evolutionary theory and the Mendelian genetics, population genetics was founded mostly by Fisher, Haldane, Wright around $1920 - 1930$. It has been extremely successful but constrained to describe evolutionary dynamics of alleles in a population, without dealing much with molecular details explaining how new alleles or genetic loci arise *de novo* and why they are beneficial or deleterious. Clearly, a better understanding of the molecular details of the heritable material was required.

It took another half century to discover the structure of DNA, and that RNAs and proteins, the main functional molecules in cells, are coded at certain loci of DNA called genes (hereafter referred to as coding-DNA). Prokaryotic molecular biology, pioneered by Jacob and Monod around 1960's, developed the first methods to show that gene expression (i.e. RNA or protein levels) is regulated in the cell in order to affect metabolic processes (for a historical review, see Beckwith (2011)). Developmental biology studies followed to show the importance of gene regulation in eukaryotes. Different cell types in complex multicellular organisms show different metabolic and structural properties, e.g. a human eye cell versus a muscle cell, despite the fact that they share the same DNA content. The differences between the cells of a single individual are coded in gene regulatory dynamics driving differentiation during development. But what is the role of regulated gene expression for the phenotypic variation that we observe across individuals, populations and species?

Gene expression, similar to other phenotypic traits, shows heritable variation, due mostly to non-coding DNA regions (hereafter referred to as regulatory DNA) (Fay and Wittkopp, 2007; Zheng et al., 2011; Romero et al., 2012). Yet, the contribution to larger scale phenotypic diversity from variation in gene expression was largely ignored, mostly due to the difficulty in detecting precise regulatory DNA loci and function. With the development of cheap sequencing techniques around the 1980's, examples of coding sequence similarity between especially closely related species have been shown. In

2000's, the human genome project established that we have $\sim 25,000$ genes, surprisingly much less than expected (Venter et al., 2001). This, together with the other genome projects, suggests that complexities are not necessarily due to number of genes (coding DNA). Furthermore, regulatory changes are thought to be less pleiotropic than coding sequence changes, allowing evolution to create diversity easily (Prud'homme et al., 2006). Consequently, the hypothesis that most phenotypic diversity is due to regulatory rather than structural variation was brought forward (Britten and Davidson, 1971) (for a review, see Wittkopp (2013)). Especially in the last decades, there have been intensive descriptive studies trying to lay out the nature of gene expression in terms of variation, and its relation to large scale phenotypes. There have been clear examples showing the effect of losing/gaining regulatory DNA on gene expression affecting large scale phenotypes, such as the pelvis adaptation of the stickleback fish, caused by losing enhancer regions (Chan et al., 2010). Yet, the examples are not numerous and the bridge between molecular variation to larger scale phenotypic variation is not at all clear. Structural versus regulatory origins of phenotypic diversity are still debated in the literature (Hoekstra and Coyne, 2007). Clarification will only be possible with a better understanding of function and evolution of regulatory DNA, both from theoretical and empirical points of view (for a review, see Romero et al. (2012)).

Regulation of gene expression in the cell is performed at transcriptional, translational, and degradation levels. Complex relations between these different regulatory levels exist, and a complete mapping from DNA to RNA and protein levels remains elusive (for a review, see Vogel and Marcotte (2012)). The main focus of this thesis is on transcriptional regulation; to be more precise, on the initiation of transcription, since it has been better studied from a biophysical and genetic point of view, and thereby is more amenable to thorough evolutionary investigation. Although details differ, the core mechanisms for transcription initiation are conserved across all living organisms. RNA polymerase (RNAP) binds specifically to sequences upstream of the coding region, in order to transcribe the DNA sequence into RNA molecules. Different RNAP binding sequences influence binding, and therefore, transcription initiation rates (Gross et al., 1998; Paget and Helmann, 2003; Murakami, 2015). There also exist *transcription factor* (TF) proteins that bind to DNA in a sequence specific manner, either in *promoter regions* (i.e. near transcription start sites), or *enhancer regions* (i.e. far away from transcription start sites; only exist in eukaryotes), modulating transcription (Dowell, 2010; Wittkopp and Kalay, 2012). A better understanding of the evolution of sequence specific binding of RNAP and TFs is needed to understand the evolution of gene regulation.

Especially in the last decades, there have been many comparative genomics studies in eukaryotes that describe the evolution of regulatory binding (see Villar et al. (2014) for a recent review). It is considered that most of the binding variation is due to *cis*-acting genetic elements (i.e. promoter and enhancer regions) rather than *trans*-acting genetic elements (i.e. coding sequences of transcription factors) (Dowell, 2010; Wittkopp and Kalay, 2012). In particular, by using mice carrying human chromosome, Wilson et al. (2008) and Schmidt et al. (2010) have shown that characteristics of regulatory binding profiles are primarily due to sequence rather than cellular environment or regulatory proteins. Moreover, fast evolution of binding in comparison to speciation time scales has been observed, even under strong developmental constraints on gene expression (Ludwig et al., 1998). There is also evidence from prokaryotes suggesting a similar flexible and fast evolution of regulatory DNA (Kim et al., 2012). However, most of these studies were qualitative and descriptive, and do not explain the evolutionary and molecular (biophysical) mechanisms driving the evolution of transcriptional regulatory sequences.

This thesis aims at a quantitative and predictive understanding of transcriptional regulatory binding sequence evolution, as a complement to the aforementioned qualitative and descriptive approaches. This necessarily requires a realistic but mathematically tractable description of how genotypes harbouring binding sequences are connected with mutations, and mapped to gene expression that is the phenotype under selection. Throughout this thesis, I will address the basic mathematical principles of the evolutionary dynamics of transcriptional binding sequences; how selective signatures on transcriptional regulatory DNA can be inferred quantitatively; how simple genotype-phenotype mapping models can be used to understand the mechanistic and evolutionary relation between regulatory DNA and gene expression. These will be carried out by bridging the theoretical frameworks of biophysics and population genetics, as briefly described below, before giving a more detailed overview of the exact questions addressed in each chapter of this thesis.

In order to construct realistic genotype-phenotype mapping from DNA binding regions to gene expression, biophysical models of sequence-specific protein-DNA interaction are considered throughout this thesis. In particular, I implement thermodynamical models for the binding probability (or the fraction of bound time) of regulatory factors on DNA, which determine gene expression. Thermodynamical models assume that the dynamics of binding and unbinding of RNAP and TF to DNA sequences quickly reaches an equilibrium, relative to typical time scales of gene expression. By only knowing the binding energies and the chemical potential (concentration) of the regulatory factors, binding probabilities are expressed by considering the corresponding Boltzmann weights (i.e. stationary probabilities of thermodynamical systems) of all possible molecular configurations on regulatory DNA. Statistical thermodynamics goes back to Ludwig Boltzmann around 1870's in Vienna, but the first applications of thermodynamic equilibrium to protein-DNA interaction to model transcriptional gene regulation is, to my knowledge, due to Shea and Ackers (1984). This was expanded by von Hippel and Berg (1986) to lay out the essential understanding of regulatory factors's specificity for DNA sequence. Since then, different *in vivo* and *in vitro* molecular techniques and bioinformatic methods have been developed to provide specificity of regulatory factors (for reviews, see Stormo and Fields (1998) and Stormo and Zhao (2010)). In particular, chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-chip) and by sequencing (ChIP-Seq) enhanced the comparative genomic studies for TF binding (Dowell, 2010). Specificity of a regulatory factor is generally reported with a position weight or an energy matrix where, respectively, the observed frequency or free energy contribution of each nucleotide at each position is given. This assumes the independency of positions, which is considered valid at least for sequences a few mutations away the preferred (consensus) sequence (Maerkl and Quake, 2007). I follow this additivity assumption in the entire thesis, and make use of the energy matrices of RNAP and cAMP receptor protein (CRP) (Kinney et al., 2010). Overall, a quantitative framework of a biophysically realistic genotype-phenotype mapping has already been initiated in the literature (for reviews, see Segal and Widom (2009), or the textbooks by Bialek (2012) and Phillips et al. (2012).

In this thesis, the population genetics of regulatory DNA sequences is modelled under evolutionary forces of mutation, selection and random genetic drift. Other evolutionary forces such as recombination and migration are neglected. The first mathematical foundations of the population genetics of mutation, selection, and genetic drift were provided by Fisher, Wright and Haldane in the early times of population genetics. Kimura later provided rigorous derivations of the stationary and dynamical properties in the diffusion theory formalism (see a review by Kimura (1964)). There are also recent studies for better understanding of the complexity of the mutation-selection-drift evolutionary dynamics; for example, Desai and Fisher (2007) studied the question in a more general perspective. All together, a general treatment for the population genetics of mutation, selection, and genetic drift is already mature enough to be used for a theoretical understanding of regulatory DNA evolution (see the textbooks by Rice (2004), Crow and Kimura (2009), Gillespie (2010), Charlesworth (2010) and Ewens (2012)).

There have been a number of key studies which paved the way for building a sound bridge between biophysics and evolutionary theory to understand the evolution of regulatory binding sequences. Following their seminal paper (von Hippel and Berg, 1986) on the thermodynamic equilibrium of protein-DNA interactions, Berg and von Hippel (1987) discussed the interplay between the sequence specificity and selection to understand the statistical deviations from the neutral distribution of the binding sequences in real genomes. Gerland and Hwa (2002); Berg et al. (2004); Sella and Hirsh (2005), and Stewart and Plotkin (2012) contributed along this line, elaborating the mutation, selection and genetic drift population genetic model for the evolution of transcriptional regulatory sequences. Mustonen and Lässig (2005); Mustonen et al. (2008), and Haldane et al. (2014) used the steady state distribution of mutation, selection and genetic drift model to infer selection on the binding energies of several transcription factors. However, the dynamical properties of evolution of transcriptional regulatory sequences have not been addressed in detail which is the major focus of my thesis, as outlined below.

The first part of the thesis is devoted to a theoretical investigation of transcriptional regulatory sequences. **Chapter 2** entitled *"Dynamics of Transcription Factor Binding Site Evolution"* addresses the expected rates of binding sequence evolution under mutation, selection and genetic drift by considering a thermodynamic model for fitnesses of binding sequences. By using exact mathematical expressions which are confirmed by computational simulations, the chapter shows how evolutionary rates at a single binding site scale with biophysical parameters such as binding length and specificity, as well as with population genetic parameters such as population size and selection strength. Furthermore, it discusses how fast the evolutionary dynamics at single site approach a stationary state. This chapter also deals with the evolution of binding sites at larger DNA sequences, i.e. promoters and enhancers, by using computational simulations and approximate mathematical expressions.

The second part of the thesis is reserved for my applied investigation of transcriptional regulatory sequences. **Chapter 3** entitled *"Binding Site Evolution of Bacterial RNA Polymerase"* deals with the evolution of binding sequences of bacterial RNAP by using the reported biophysical characteristics of sequence-specific interactions of RNAP with DNA. First, it analyses the whole genome, the promoter regions and the experimentally verified transcription start sites of *E.coli* K12 to infer selection for RNAP binding by using a population genetic theory. Secondly, it deals with inference of the chemical potential of RNAP, an unknown but crucial thermodynamic parameter, from evolved sequences. Furthermore, the chapter aims at understanding the coevolution of the LacZ protein activity and lac promoters among 20 bacterial species, diverged over millions of years. Lastly, the chapter lays out a statistical framework for predictive understanding of the adaptive mutations in *de novo* promoter evolution in *E.coli* K12.

This thesis is organised cumulatively, i.e. each research chapter is written in the style of a research article, which can be read without the knowledge of the rest of the thesis, at the expense of some inevitable repetition. The reader will find a summary and a general discussion of the thesis together with future research directions in **Chapter 4** entitled *"Conclusions"*.

2

Dynamics of Transcriptional Factor Binding Site Evolution

The work presented in this chapter has been done in collaboration with Tiago Paixão, Nicholas H. Barton and Gašper Tkačik. This chapter was published as Tuğrul *et al.* (2015) "*Dynamics of Transcriptional Factor Binding Site Evolution*", PLoS Genetics 11(11): e1005639.

2.1 Introduction

Evolution produces heritable phenotypic variation within and between populations and species on relatively short timescales. Part of this variation is due to differences in gene regulation, which determines how much of each gene product exists in every cell. These gene expression levels are heritable quantitative traits subject to natural selection (Fay and Wittkopp, 2007; Zheng et al., 2011; Romero et al., 2012). While the importance of their variability for the observed phenotypic variation is still debated (Hoekstra and Coyne, 2007), it is believed to be crucial within closely related species or in populations whose proteins are functionally or structurally similar (Wittkopp, 2013). The genetic basis for gene expression differences is thought to be non-coding regulatory DNA, but our understanding of its evolution is still immature; this is due, in part, to the lack of precise knowledge about the mapping between the regulatory sequence and the resulting expression levels.

Transcriptional regulation is the most extensively studied mechanism of gene regulation. Transcription factor proteins (TFs) recognize and bind specific DNA sequences called binding sites, thereby affecting the expression of target genes. Eukaryotic regulatory sequences, i.e., enhancers and promoters, are typically between a hundred and several thousand base pairs (bp) in length (Yao et al., 2015), and can harbor many transcription factor binding sites (TFBSs), each typically consisting of $6-12$ bp. The situation is different in prokaryotes: they lack enhancer regions and have one or a few TFBSs which are typically longer, between 10 to 20 bp in length (Wunderlich and Mirny, 2009; Stewart and Plotkin, 2012). Differences in TF binding are thought to arise primarily due to changes in the regulatory sequence at the TF binding sites rather than changes in the cellular environment or the TF proteins themselves (Schmidt et al., 2010). Nevertheless, a theoretical understanding of the relationship between the evolution of the regulatory sequence and the evolution of gene expression levels remains elusive, mostly because of the complex interaction of evolutionary forces and biophysical processes (Stefflova et al., 2013).

From the evolutionary perspective, the crucial question is whether and when these regulatory sequences can evolve rapidly enough so that new phenotypic variants can arise and fix in the population over typical speciation timescales. Comparative genomic studies in eukaryotes provide evidence for the evolutionary dynamics of TF binding, highlighting the possibility for rapid and flexible TFBS gain and loss between closely related species on timescales of as little as a few million years (Dowell, 2010;

Villar et al., 2014). Examples include quick gain and loss events that cause divergent gene expression (Doniger and Fay, 2007), or the compensation of such events by turn-over at other genome locations (Moses et al., 2006); gain and loss events sometimes occur even in the presence of strong constraints on expression levels (Ludwig et al., 1998; Paris et al., 2013). Furthermore, such events enabled new binding sites on sex chromosomes that arose as recently as $1-2$ million years ago (Ellison and Bachtrog, 2013; Alekseyenko et al., 2013). There are examples of rapid regulatory DNA evolution across and within populations requiring shorter timescales, i.e. $10.000-100.000$ years (Contente et al., 2002; Kasowski et al., 2010; Zheng et al., 2011; Chan et al., 2010). On the other hand, strict conservation has also been observed at orthologous regulatory locations even in distant species (e.g., (Vierstra et al., 2014)). Taken together, these facts suggest that the rates of TFBS evolution can extend over many orders of magnitude and differ greatly from the point mutation rate at a neutral site. To study the evolutionary dynamics of regulatory sequences and understand the relevant timescales, we set up a theoretical framework with a special focus on the interplay of both population genetic and biophysical factors, briefly outlined below.

Sequence innovations originate from diverse mutational mechanisms in the genome. While tandem repeats (Gemayel et al., 2010) or transposable elements (Feschotte, 2008) may be important in evolution, the better studied and more widespread mutation types still need to be better understood in the context of TFBS evolution. Specifically, we ask how the evolutionary dynamics are affected by single nucleotide (point) mutations, as well as by insertions and deletions (indels). New mutations in the population are selected or eliminated by the combined effects of selection and random genetic drift. Although the importance of selection (Hahn et al., 2003; He et al., 2011; Arnold et al., 2014) and mutational closeness of the initial sequences (MacArthur and Brookfield, 2004; Nourmohammad and Lässig, 2011) for TF binding site evolution has already been reported, the belief in fast evolution via point mutations without selection (i.e., neutral evolution) persists in the literature (e.g.,(Wittkopp, 2013; Villar et al., 2014)), mainly due to Stone and Wray (2001)'s misinterpretation of their own simulation results (see MacArthur and Brookfield (2004)). This likely reflects the current lack of theoretical understanding of TFBS evolution in the literature, even under the simplest case of directional selection. Basic population genetics shows that directional selection is expected to cause a change, e.g., yield a functional binding site, over times on the order of $1/(NsU_b)$, where N is the population size, s is the selection advantage of a binding site, and U_b is the beneficial mutation rate (Berg et al., 2004). This process can be extremely slow, especially under neutrality, if several mutational steps are needed to reach a sequence with sufficient binding energy to confer a selective advantage. As already pointed out by Berg et al. (2004), this places strong constraints on the length of the binding sites, if they were to evolve from random sequences.

Several biophysical factors, such as TF concentration and the energetics of TF-DNA and TF-TF interactions, might play an important role in TFBS evolution. Quantitative models for TF sequence specificity (von Hippel and Berg, 1986; Berg and von Hippel, 1987; Stormo and Fields, 1998; Stormo and Hartzell, 1989; Stormo and Zhao, 2010; Zhao et al., 2009) and for thermodynamic (TD) equilibrium of TF occupancy on DNA (Shea and Ackers, 1984; Berg and von Hippel, 1987; Bintu et al., 2005a,b; Hermsen et al., 2006, 2010) were developed in recent decades and, in parallel with developments in sequencing, have contributed to our understanding of TF-DNA interaction biophysics. These biophysical factors can shape the characteristics of the TFBS fitness landscape over genotype space in evolutionary models (Gerland et al., 2002; Gerland and Hwa, 2002; MacArthur and Brookfield, 2004; Berg et al., 2004; Stewart and Plotkin, 2012, 2013; Payne and Wagner, 2014). There are also intensive efforts to understand the mapping from promoter/enhancer sequences to gene expression (Segal et al., 2008; Hermsen et al., 2006; Samee and Sinha, 2014; He et al., 2010). Despite this recent attention, there have been relatively few attempts to understand the evolutionary dynamics of TFBS in full promoter/enhancer regions (MacArthur and Brookfield, 2004; He et al., 2012; Hermsen et al., 2010; Duque et al., 2013; Duque and Sinha, 2015), especially using biophysically realistic but still mathematically tractable models. Such models are necessary to gain a thorough theoretical understanding of binding site evolution.

Our aim in this study is to investigate the dynamics of TFBS evolution by focusing on the typical evolutionary rates for individual TFBS gain and loss events. We consider both a single binding site at an isolated DNA region and a full enhancer/promoter region, able to harbor multiple binding sites. In the following section, we lay out our modeling framework, which covers both population genetic and biophysical considerations, as outlined above. Using this framework, we try to understand **i)** what typical gain and loss rates are for a single TFBS site; **ii)** how quickly populations converge to a stationary distribution for a single TFBS; **iii)** how multiple TFBS evolve in enhancers and promoters; **iv)** how early history of the evolving sequences can change the evolutionary rates of TFBS; and **v)** how cooperativity between TFs affects the evolution of gene expression. We find that, under realistic parameter ranges, both gain and loss of a single binding site is slow, slower than the typical divergence time between species. Importantly, fast emergence of an isolated TFBS requires strong selection and favorable initial sequences in the mutational neighborhood of a strong TFBS. The evolutionary process approaches the equilibrium distribution very slowly, raising concerns about the use of equilibrium assumptions in theoretical work. We proceed to show that the dynamics of TFBS evolution in larger sequences can be understood approximately from the dynamics of single binding sites; the TFBS gain times are again slow if evolution starts from random sequence in the absence of strong selection or large regulatory sequence “real estate.” Finally, we identify two factors that can speed up the emergence of TFBS: the existence of an initial sequence distribution biased towards the mutational neighborhood of strongly binding sequences, which suggests that ancient evolutionary history can play a major role in the emergence of “novelties” (Villar et al., 2015); and the biophysical cooperativity between transcription factors, which can partially account for the lack of observed correlation between identifiable binding sequences and transcriptional activity (Stefflova et al., 2013).

2.2 Models & Methods

2.2.1 Population genetics

We consider a finite population of N diploid individuals whose genetic content consists of an evolvable L base pair (bp) contiguous regulatory sequence $\boldsymbol{\sigma}$ to which TFs can bind. Given that $\sigma_i \in \{A, C, G, T\}$ where $i = 1, 2, ..., L$ indexes the position in regulatory sequence, there are 4^L different regulatory sequences in the genotype space. Each TF is assumed to bind to a contiguous sequence of n bp within our focal region of L bp (Fig. 2.1A,B). Regulatory sequences evolve under mutation, selection, and sampling drift. The rest of the genome is assumed to be identical for all individuals and is kept constant. In the first part of our study we consider the regulatory sequence comprised of a single TFBS (i.e. $L = n$). Later, we consider the evolution of a longer sequence (i.e. $L \gg n$) in which more than one TFBS can evolve. For simulations, we use a Wright-Fisher model where N diploid individuals are sampled from the previous generation after mutation and selection. Our analytical treatment is general and corresponds to setups where a diffusion approximation to allele frequency evolution is valid. We neglect recombination since typical regulatory sequences are short, $L \leq 1000$. To be consistent with most of the population genetics literature we assume diploidy, but since we do not consider any dominance effects, our results also hold for a haploid population with $2N$ individuals.

Evolutionary dynamics simplify in the low mutation limit where the population consists of a single genotype during most of its evolutionary history (the fixed state population model). Desai and Fisher (2007) have shown that the condition $\frac{\log 4N\Delta f}{\Delta f} \ll \frac{1}{4NU_b\Delta f}$ needs to hold for a fixed state population assumption to be accurate. The term on the left is the establishment time of a mutant allele with a selective advantage Δf relative to the wild type; the term on the right-hand side is the waiting time for such an allele to appear, where U_b is the beneficial mutation rate per individual per generation. Note that, in binding site context, U_b refers to the rate of mutations which increase the fitness, for instance, by increasing binding strength. Its exact value depends on the current state of the genotype; nevertheless, typical value estimates help model the evolutionary dynamics. In multicellular eukaryotes, where most evidence for the evolution of TFBSs has been collected and which provide the motivation for this manuscript, the number of mutations per nucleotide site is typically low, e.g. $4Nu \sim 0.01$ in *Drosophila* and $4Nu \sim 0.001$ in humans (Lynch and Conery, 2003), where u is the point mutation rate per generation per base pair. For a single binding site of typical length $n \sim 5-15$, one therefore expects the fixed state population model to be accurate. For longer regulatory sequences, one expects that beneficial mutations are rare among all possible mutations, so that the fixed state population model can be assumed to hold as well.

Evolution under the fixed state assumption can be treated as a simple Markovian jump process. The transition rate from a regulatory sequence $\boldsymbol{\sigma}$ to another regulatory sequence $\boldsymbol{\sigma}'$ in a diploid population

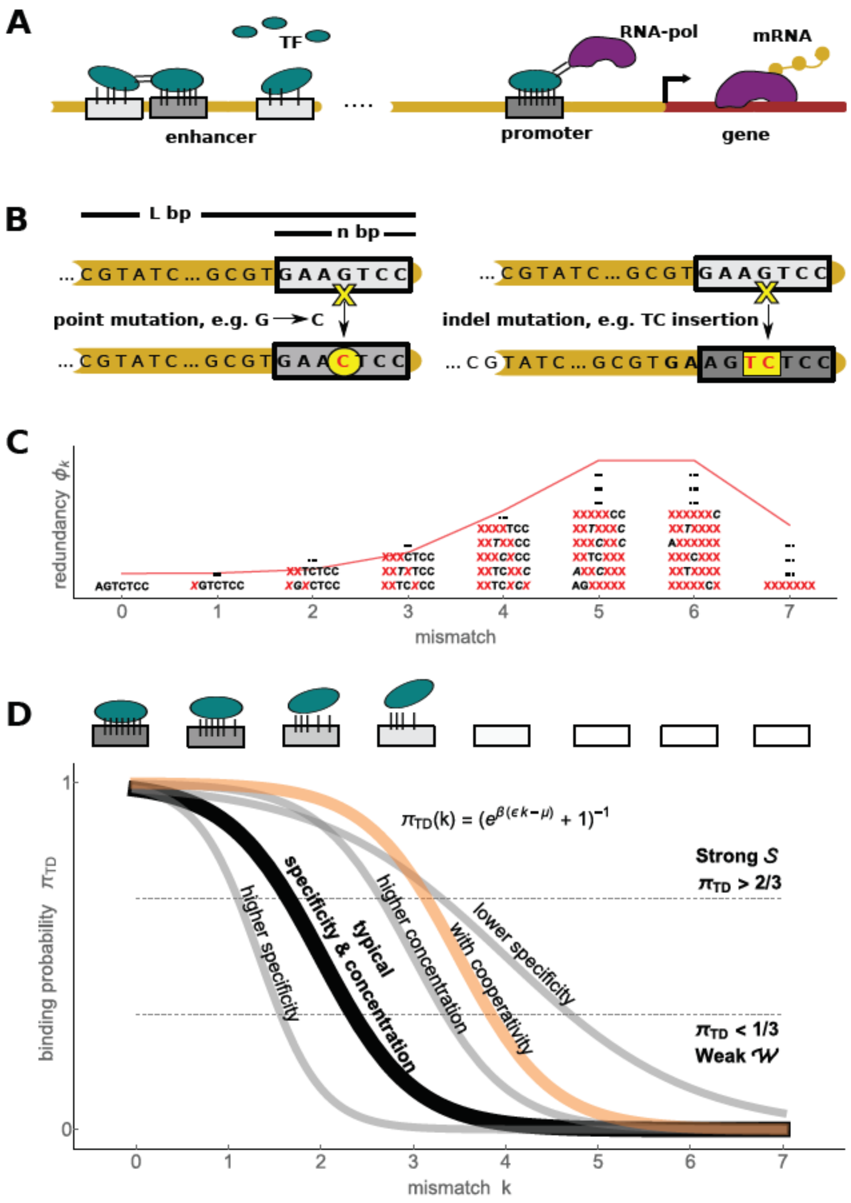


Figure 2.1: **Biophysics of transcription regulation.** **A)** TFs bind to regulatory DNA regions (promoters and enhancers) in a sequence-specific manner to regulate transcriptional gene expression (mRNA production) level via different mechanisms, such as recruiting RNA polymerase (RNA-pol). **B)** A schematic of two types of mutational processes that we model: point mutations (left) and indel mutations (right). **C)** The mismatch binding model results in redundancy of genotype classes, with a binomial distribution (red) of genotypes in each mismatch class (some examples of degenerate sequences shown) **D)** The mapping from the TFBS regulatory sequence to gene expression level is determined by the thermodynamic occupancy (binding probability) of the binding site. If each of the k mismatches from the consensus sequence decreases the binding energy by ϵ, the occupancy of the binding site is $\pi_{\mathrm{TD}}(k) = (1+e^{\beta(\epsilon k-\mu)})^{-1}$, where μ is the chemical potential (related to free TF concentration). A typical occupancy curve is shown in black ($\epsilon = 2\,k_BT$ and $\mu = 4\,k_BT$); the gray curves show the effect of perturbation to these parameters ($\epsilon = 1\,k_BT$, $\epsilon = 3\,k_BT$ and $\mu = 6\,k_BT$); the orange curve illustrates the case of two cooperatively binding TFs ($k_c = 0$ and $E_c = -3\,k_BT$, see text for details). We pick two thresholds, shown in dashed lines, to define discrete binding classes: strong $\mathcal{S}$ ($\pi_{\mathrm{TD}} > 2/3$) and weak $\mathcal{W}$ ($\pi_{\mathrm{TD}} < 1/3$).

is

$$R_{\sigma',\sigma} = 2N\, U_{\sigma',\sigma}\, P_{\text{fix}}(N,\, \Delta f_{\sigma',\sigma}) \tag{2.1}$$

where $\Delta f_{\sigma',\sigma} = f(\boldsymbol{\sigma}') - f(\boldsymbol{\sigma})$ is the fitness difference and $U_{\sigma',\sigma}$ is the mutation rate from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma}'$. The fixation probability P_{fix} of a mutation with fitness difference Δf in a diploid population of N individuals is

$$P_{\text{fix}}(N,\, \Delta f) = \frac{1 - e^{-2\Delta f}}{1 - e^{-4N\Delta f}} \approx \frac{2\Delta f}{1 - e^{-4N\Delta f}}, \tag{2.2}$$

which is based on the diffusion approximation (Kimura, 1962). Note that the fixation probability scaled with $1/N$ approximates to $2N\Delta f$ when $N\Delta f \gg 1$. Evolutionary dynamics therefore depend essentially on how regulatory sequences are mutationally connected in genotype space, and how fitnesses differ between neighboring genotypes, i.e., on the fitness landscape.

2.2.2 Directional selection on biophysically motivated fitness landscapes

In this study, we focus on directional selection by assuming that fitness f is proportional to gene expression level g which depends on regulatory sequence, i.e.

$$f(\boldsymbol{\sigma}) = s\, g(\boldsymbol{\sigma}) \tag{2.3}$$

where s is the selection strength. It is important to note that this choice does not imply that directional selection is the only natural selection mechanism. It simply aims at obtaining the theoretical upper limits for the rates of gaining and losing binding sites.

To analyze a realistic but tractable mapping from the regulatory sequence to fitness, we primarily assume that the proxy for gene expression is the binding occupancy (binding probability) π at a single TF binding site, or the sum of the binding occupancies within an enhancer/promoter region (based on limited experimental support (Giorgetti et al., 2010)). This corresponds to

$$f(\boldsymbol{\sigma}) = s \sum_i \pi^{(i)}(\boldsymbol{\sigma}) \tag{2.4}$$

where $\pi^{(i)}$ is the binding occupancy of a site starting at the nucleotide i in sequence $\boldsymbol{\sigma}$, and s can be interpreted as the selective advantage of a strongest binding to a weakest binding at a site. We assume all binding sites have equal strength and direction in their contribution towards total gene activation. Sites acting as repressors in our simple model would enter into Eq. (2.4) with a negative selection strength, s. Future studies developing mathematically tractable models should consider more realistic case of unequal contribution with combined activator and repressor sites responding differentially to various regulatory inputs (Duque and Sinha, 2015). Although one can postulate different scenarios that map TF occupancies in a long ($L \gg n$) promoter to gene expression, we chose the simplest case which allows us to make analytical calculations. Later we relax our assumption on noninteracting binding sites and consider the effects of several kinds of interactions on gene expression and thus on evolutionary dynamics.

The occupancy of the TF on its binding site is assumed to be in thermodynamic (TD) equilibrium (Shea and Ackers, 1984; Berg and von Hippel, 1987; Bintu et al., 2005a,b; Hermsen et al., 2006, 2010). While this might not always be realistic (Hammar et al., 2014; Cepeda-Humerez et al., 2015), there is empirical support for this assumption (particularly in prokaryotes) (Segal et al., 2008; Brewster et al., 2012; Razo-Mejia et al., 2014), and more importantly, it is sufficient to capture the essential nonlinearity in this genotype-phenotype-fitness mapping (Haldane et al., 2014). In thermodynamic equilibrium, the binding occupancy at the site starting with the i-th position in regulatory sequence is given by

$$\pi^{(i)}_{\mathrm{TD}}(E_i) = \left(1 + e^{\beta(E_i - \mu)}\right)^{-1}. \tag{2.5}$$

Here, μ is the chemical potential of the TF (related to its free concentration) (Gerland et al., 2002; Weinert et al., 2014); E_i is the sequence specific binding energy, where lower energy corresponds to tighter binding, and $\beta = (k_B T)^{-1}$. We compute the binding energy E_i by adopting an additive energy model which is considered to be valid at least up to a few mismatches from the consensus sequence (Maerkl and Quake, 2007; Zhao et al., 2009; Stormo and Zhao, 2010; Kinney et al., 2010), i.e.

$$E_i(\boldsymbol{\sigma}) = \sum_{j=i}^{i+n-1} \xi_{\sigma_j, j} \tag{2.6}$$

where ξ stands for the energy matrix whose $\xi_{\sigma_j,j}$ element gives the energetic contribution of the nucleotide σ_j appearing at the j-th position within TFBS. With this, Eq. (2.4) can be rewritten more formally as

$$f(\boldsymbol{\sigma}) = s \sum_i \pi^{(i)}_{\mathrm{TD}}(E_i(\boldsymbol{\sigma})) \tag{2.7}$$

To allow analytical progress, we make the “mismatch assumption,” i.e., the energy matrices contain identical $\epsilon > 0$ entries for every non-consensus (mismatch) base pair; the consensus entries are set to zero by convention. A single binding sequence with k mismatches therefore has the binding energy $E = k\epsilon$. We will refer to ϵ as “specificity.” Specificity is provided by diverse interactions between DNA and TF, including specific hydrogen bonds, van der Waals forces, steric exclusions, unpaired polar atoms, etc. (McKeown et al., 2014). ϵ is expected to be in the range $1 - 3\ k_B T$, which is consistent with theoretical arguments (Gerland et al., 2002) as well as direct measurements (Fields et al., 1997; Kinney et al., 2010; Maerkl and Quake, 2007). Note that we explicitly check the validity of the analytical results based on the mismatch assumption by comparing them against simulations using realistic energy matrices. The redundancy (i.e., normalized number of distinct sequences) of a mismatch class k at a single site in a random genome can be described by a binomial distribution ϕ (Fig. 2.1C) where the probability of encountering a mismatch class k is

$$\boldsymbol{\phi_k}(n, \alpha) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \tag{2.8}$$

where $\alpha = 3/4$ in the case of equiprobable distribution over the four nucleotides.

We focus on selection in a single environment, which in this framework corresponds to a single choice for the TF concentration. We therefore fix the chemical potential to a baseline value of $\mu = 4\, k_B T$, which maps changes in the sequence (mismatch class k) to a full range of gene expression levels, as shown in Fig. 2.1D. We subsequently vary μ systematically and report how its value affects the results.

After these preliminaries, the equilibrium binding probability of Eq. (2.5) reduces to

$$\pi_{\mathrm{TD}}(k) = \left(1 + e^{\beta(\epsilon\, k - \mu)}\right)^{-1}. \tag{2.9}$$

This function has a sigmoid shape whose steepness depends on specificity ϵ and whose midpoint depends on the ratio of chemical potential to specificity, μ/ϵ (Fig. 2.1D). To simplify discussion, we introduce two classes of sequences: genotypes are associated with "strong binding" $\mathcal{S}$ and "weak binding" $\mathcal{W}$ if $\pi_{\mathrm{TD}} > 2/3$ and $\pi_{\mathrm{TD}} < 1/3$, respectively. The thresholds that we pick are arbitrary, while still placing the midpoint of the sigmoid between the two classes; our results do not change qualitatively for other choices of thresholds. In the mismatch approximation, the genotype classes $k = \{0,\, 1,\, ...,\, k_{\mathcal{S}}\} \in \mathcal{S}$ and $k = \{k_{\mathcal{W}},\, k_{\mathcal{W}} + 1,\, ...,\, n\} \in \mathcal{W}$ correspond to strong and weak binding, respectively. $k_{\mathcal{S}}$ and $k_{\mathcal{W}}$ are defined as the closest integers to the thresholds defined above; these values depend on ϵ and μ. We also define a "presite" as the mismatch class that is 1 bp away from the threshold for strong binding, i.e., a class with $k_{\mathcal{S}} + 1$ mismatches. Note that binding length n extends the tail of the fitness landscape for a single site and shifts the center of redundancy rich mismatch classes (Fig. 2.1C).

The formulation in Eq. (2.7) reduces to

$$f(k) = s\, \pi_{\mathrm{TD}}(k) \tag{2.10}$$

in a mismatch approximation at a single site, which we will investigate extensively for Ns scaling of TFBS gain and loss rates. We consider a wide range of Ns values: $Ns < 0$ for negative selection, $Ns = 0$ for neutral evolution, $Ns \sim 1$ for weak positive selection, $Ns \gg n \log(2)/2$ for strong positive selection (see below for this particular choice of the threshold).

In order to study the effects of interacting TFBSs in large regulatory sequences, we relax our assumption of non-interacting TFBS in Eq. (2.7) and study three simple models. In the main text, we report the cooperative physical interaction between two TF molecules binding two nearby sites where the binding probability at a site is modified as

$$\pi_{\mathrm{coop}}(k,\, k_c) = \frac{e^{-\beta(\epsilon k - \mu)} + e^{-\beta(\epsilon(k + k_c) - 2\mu - E_c)}}{1 + e^{-\beta(\epsilon k - \mu)} + e^{-\beta(\epsilon k_c - \mu)} + e^{-\beta(\epsilon(k + k_c) - 2\mu - E_c)}}, \tag{2.11}$$

where k_c stands for the mismatch class at the co-binding site and E_c for cooperativity. In this study we consider that cooperative energy ranges from an intermediate strength ($E_c = -2\, k_B T$) to a high strength ($E_c = -4\, k_B T$) (Hermsen et al., 2006). Fig. 2.1D shows an example of the binding probability when a strong co-binding site exists. As a function of k alone, at fixed k_c, this formulation of cooperativity is consistent with the zero-cooperativity ($E_c = 0$) case but with a changed effective chemical potential. We take cooperative interactions into account if the two TFs are binding within 3 bp of each other, and we only consider the strongest binding of the cooperative partner (i.e., the proximal location with the lowest k_c).

In Supporting Information, we discuss the other two models of interacting TFBS. In one model, gene expression is determined only by the binding probability of the strongest site in the regulatory sequence. In the other model, gene expression is determined by the probability of the joint occupancy of 2 strongest binding sites, anywhere in the regulatory sequence; this model is a toy version of synergistic "non-physical" interaction of TFs which compete with nucleosomal binding for the occupancy of regulatory regions in eukaryotes (see Mirny (2010) for a detailed model).

2.2.3 Point and indel mutations

Point mutations and indels are the only mutational processes in our framework. Point mutations with a rate u convert the nucleotide at one position into one of the 3 other nucleotide types. For a single binding site, the probability that a point mutation changes the mismatch class from k to k' is

$$\boldsymbol{P}^{(\mathrm{point})}_{k',k} = \big(1 - k/n\big)\,\delta_{k',k+1} + \big(k/3n\big)\,\delta_{k',k-1} + \big(2k/3n\big)\,\delta_{k',k} \tag{2.12}$$

where $\delta_{a,b} = 1$ if $a = b$ and 0 otherwise.

We define the indel mutation rate per base pair such that it occurs with rate $\theta\, u$ at a position where a random nucleotide sequence is either inserted, or an existing nucleotide sequence is deleted. For mathematical simplicity, we assume that insertions and deletions are equally likely; in fact, a slight bias towards deletions is reported in the literature with a ratio of deletion to insertion $\sim 1.1 - 3.0$ (Taylor et al., 2004; Brandström and Ellegren, 2007; Park, 2015). Parameter θ is the ratio of indel mutation rate to point mutation rate, and is reported to be in the range $0.1 - 0.2$ (Cartwright, 2009; Chen et al., 2009; Lee et al., 2012). We consider two cases: the baseline of $\theta = 0$ for no indel mutations, and $\theta = 0.15$ for the combined effect of indel and point mutations. Since we fix the length of the regulatory sequence, indels shift existing positions away from or inwards to some reference position (e.g., transcription start site). For consistency, we fix the regulatory sequence at its final position and assume that sequences before the initial position are random. Indel lengths vary, with reports suggesting a sharply decreasing but fat-tail frequency distribution (Keightley and Johnson, 2004). For simulations we consider only very short indels of size $1 - 2$ bp, occurring proportional with their reported frequencies of 0.45 and 0.18, respectively. We do not need to assume any particular indel length for analytical calculations (below). While sufficient for our purposes, this setup would need to be modified when working with real sequence alignments of orthologous regions.

For a single binding site (i.e. $L = n$) one can exactly calculate the probability of an indel mutation changing the mismatch class from k to k' as

$$\boldsymbol{P}^{(\text{indel})}_{k',k} = \sum_{i=1}^{n} (1/n) \sum_{x=0}^{k'} p(X_i = x \mid k)\, p(Y_i = k' - x). \tag{2.13}$$

Here, i is the index for the position of an indel mutation within the binding site. The distribution over possible positions is uniform (hence $1/n$). The indel mutation defines two distinct parts in the binding site in terms of mismatches: nucleotides behind the indel mutation preserve their mismatch information, yet the nucleotides within and after indel mutation completely lose it. The new mismatches at these distinct parts X_i and Y_i are binomial random variables,

$$\begin{aligned} p(X_i = x \mid k) &= \boldsymbol{\phi_x}(i-1,\, \alpha = k/n) \\ p(Y_i = y) &= \boldsymbol{\phi_y}(n-i+1,\, \alpha = 3/4) \end{aligned} \tag{2.14}$$

where $\boldsymbol{\phi_k}(n,\, \alpha)$ is defined in Eq. (2.8). Fig. 2.6 shows that Monte Carlo sampling of indel mutations at a single binding site matches the analytical expression in Eq. (2.13).

The two types of mutations can be combined into the mutation rate matrix as follows:

$$\boldsymbol{U}_{k',k} = \begin{cases} n\, u \left(\boldsymbol{P}^{(\text{point})}_{k',k} + \theta\, \boldsymbol{P}^{(\text{indel})}_{k',k} \right) & k' \neq k \\ -\sum_{k' \neq k} \boldsymbol{U}_{k',k} & k' = k \end{cases}. \tag{2.15}$$

2.2.4 Evolutionary dynamics of single TF binding sites

For a sequence that consists of an isolated TFBS (i.e., $L = n$), analytical treatment is possible under the fixed state assumption. Let $\boldsymbol{\psi}(t)$ be a distribution over an ensemble of populations, whose k-th component, $\psi_k(t)$, denotes the probability of detecting a genotype with k mismatches at time t. In the continuous time limit, the evolution of $\boldsymbol{\psi}(t)$ is described by

$$\frac{d}{dt}\boldsymbol{\psi}(t) = \boldsymbol{R} \cdot \boldsymbol{\psi} \tag{2.16}$$

which accepts the following solution:

$$\boldsymbol{\psi}(t) = e^{\boldsymbol{R}\,t} \cdot \boldsymbol{\psi}(0). \tag{2.17}$$

Here, $\boldsymbol{R}$ is the transition rate matrix defined as

$$\boldsymbol{R}_{k',k} = \begin{cases} 2N\, \boldsymbol{U}_{k',k}\, P_{\text{fix}}(N,\, \Delta f_{k',k}) & k' \neq k \\ -\sum_{k' \neq k} \boldsymbol{R}_{k',k} & k' = k \end{cases}. \tag{2.18}$$

This dynamical system is a continous-time Markov chain and there exists a unique stationary distribution $\hat{\psi}$ corresponding the genotype distribution over an ensemble of populations at large time points. It can be calculated by decomposing the transition rate matrix $\boldsymbol{R}$ into its eigenvalues and eigenvectors.

The normalised left eigenvector with zero eigenvalue corresponds to the stationary distribution. This can also be expressed analytically as

$$\hat{\psi}_k \propto e^{F(k,N)+H(k \mid n,\alpha)}, \tag{2.19}$$

where $F(k, N) = 4Nf(k)$ captures the relative importance of selection to genetic drift, and $H(k \mid n, \alpha)$ is the mutational entropy, describing how a particular mismatch class k is favored due to redundancy and connectivity of the genotype space. For point mutations alone ($\theta = 0$), $H = \log \phi_k(n, \alpha)$, with the binomial distribution $\phi_k(n,\ \alpha)$ as defined in Eq. (2.8). Obtaining a closed form expression for H is difficult when considering indel mutations ($\theta > 0$), yet the eigenvalue method solution suggests a similar shape for θ in the range of interest. The form of the stationary distribution was known for a long time in population genetics literature for a single locus or many loci with linkage equilibrium (Wright, 1931). It has recently been generalised to arbitrary sequence space under the fixed state assumption (Berg et al., 2004; Sella and Hirsh, 2005), resulting in the form of Eq. (2.19) with a close analogy in the energy-entropy balance of statistical physics (Barton and Coe, 2009), and become a subject of theoretical interest (Mustonen and Lässig, 2005; Mustonen et al., 2008; Manhart et al., 2012; Haldane et al., 2014).

Under weak directional selection for high expression (and thus high binding site occupancy), the stationary distribution shows a bimodal shape, with one peak located around the fittest class, $k \sim 0$, and another at the core of mutational entropy, $k \sim \alpha n$ (recall that $\alpha = 3/4$ for a completely random genome). This bimodal shape collapses to a unimodal one, either at no selection or at strong selection. The threshold value for Ns distinguishing strong and weak selection regimes primarily depends on the TFBS binding length, n. In a sigmoidal fitness landscape and approximating the binomial distribution by a normal distribution as appropriate, the sizes of these two peaks are roughly proportional to $\exp(4Ns - n\log 4)$ and $\sqrt{2\pi\alpha(1-\alpha)n}$, respectively. Therefore, we expect the threshold Ns to scale as $\frac{1}{4}\Big(n\ \log 4 - \frac{1}{2}\log 2\pi\alpha(1-\alpha)n\Big)$. For typical n, the linear term is dominant, suggesting that

$$Ns \sim n\log(2)/2 \tag{2.20}$$

corresponds to the threshold for strong selection in TFBS evolution (cf. Fig. 2.7). Note that this n scaling differs from the $\log(n)$ scaling which is expected in simple fitness landscapes (Paixao et al., 2015). Our argument assumes that the system is at evolutionary equilibrium, which, as we will see, is not necessarily the case even under strong selection, providing further motivation for focusing on dynamical aspects of evolution.

We define the time needed to gain (or lose) a TFBS as the time it takes for a strong binding site to emerge from a weak one (and vice versa), as schematized in Fig. 2.1D. For an isolated TFBS, these times can be computed from the Markovian properties of the evolutionary dynamics, by calculating the average first hitting times (Otto and Day, 2007). We will use the notations $\langle t \rangle_{\mathcal{S}\leftarrow k}$ and $\langle t \rangle_{\mathcal{W}\leftarrow k}$, respectively, for average gain and loss times when evolution starts from mismatch class k. Obviously, $\langle t \rangle_{\mathcal{S}\leftarrow k} = 0$ if k is among the strong binding classes ($k \in \mathcal{S}$) and $\langle t \rangle_{\mathcal{W}\leftarrow k} = 0$ if k is among the weak binding classes ($k \in \mathcal{W}$). The average gain times from other mismatch classes can be found by considering the relation $\langle t \rangle_{\mathcal{S}\leftarrow k} = 1 + \sum_{k' \notin \mathcal{S}} \boldsymbol{P}_{k,k'} \langle t \rangle_{\mathcal{S}\leftarrow k'}$, where $\boldsymbol{P}_{k,k'}$ is the probability of transition from k' to k in one generation. One can compute the average gain times by writing it in terms of linear algebraic equation:

$$\boldsymbol{T}_{\mathcal{S}\leftarrow} = \left(\mathbf{R}_{\notin \mathcal{S}}\right)^{-\mathrm{T}} \cdot (-\mathbf{1}) \tag{2.21}$$

where $\boldsymbol{T}_{\mathcal{S}\leftarrow}$ is a column vector listing non-trivial gain times, i.e. $\{\langle t \rangle_{\mathcal{S}\leftarrow k}\}$ for $k = k_{\mathcal{S}} + 1,\ ...,\ n$. $\mathbf{R}_{\notin \mathcal{S}}$ is the $\mathbf{R}$ matrix with all rows and columns corresponding to $k \in \mathcal{S}$ deleted and $-\mathrm{T}$ is the matrix operator for the transpose after an inverse operation. $\mathbf{1}$ is a vector of ones. Similarly one can find the loss times,

$$\boldsymbol{T}_{\mathcal{W}\leftarrow} = \left(\mathbf{R}_{\notin \mathcal{W}}\right)^{-\mathrm{T}} \cdot (-\mathbf{1}) \tag{2.22}$$

where $\boldsymbol{T}_{\mathcal{W}\leftarrow}$ is a column vector listing non-trivial loss times, i.e. $\{\langle t \rangle_{\mathcal{W}\leftarrow k}\}$ for $k = 1,\ 2,\ ...\ k_W - 1$. $\mathbf{R}_{\notin \mathcal{W}}$ is the $\mathbf{R}$ matrix with all rows and columns corresponding to $k \in \mathcal{W}$ deleted.

In the case of point mutations alone ($\theta = 0$), the $\mathbf{R}$ matrix is tri-diagonal and one can deduce simpler formulae for gain and loss times:

$$\begin{aligned} \langle t \rangle^{(\mathrm{point})}_{\mathcal{S}\leftarrow k} &= \textstyle\sum_{i=k_{\mathcal{S}}+1}^{k} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \frac{1-\hat{\boldsymbol{\Psi}}_{i-1}}{\hat{\boldsymbol{\psi}}_{i}} \\ \langle t \rangle^{(\mathrm{point})}_{\mathcal{W}\leftarrow k} &= \textstyle\sum_{i=k+1}^{k_{\mathcal{W}}} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \frac{\hat{\boldsymbol{\Psi}}_{i-1}}{\hat{\boldsymbol{\psi}}_{i}} \end{aligned} \tag{2.23}$$

where we use $\hat{\boldsymbol{\Psi}}_i = \sum_{j=0}^{i} \hat{\boldsymbol{\psi}}_j$ to denote the cumulative stationary distribution. For very strong selection, the second term in the sums approaches unity, resulting in even simpler formulae (Berg et al., 2004), called the "shortest path" (sp) solution:

$$\begin{aligned} \langle t \rangle^{(\mathrm{sp})}_{\mathcal{S}\leftarrow k} &= \textstyle\sum_{i=k_{\mathcal{S}}+1}^{k} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \\ \langle t \rangle^{(\mathrm{sp})}_{\mathcal{W}\leftarrow k} &= \textstyle\sum_{i=k+1}^{k_{\mathcal{W}}} \frac{1}{\boldsymbol{R}_{i-1,\,i}} \end{aligned} \quad . \tag{2.24}$$

These equations can be used to quickly estimate gain and loss rates of interest. For example, the gain rate from presites under strong selection is approximately $2\,N s\,u \frac{k_{\mathcal{S}}+1}{3}(f(k_{\mathcal{S}}) - f(k_{\mathcal{S}}+1))$. Although the exact value depends on the binding specificity and chemical potential, one can see that it is about $N s\,u$ for the parameter range of interest. Similarly, one can see that the rate of loss from strong sites is about $2n\,|Ns|\,u$ when there is strong negative selection.

2.3 Results

2.3.1 Single TF binding site gain and loss rates under mutation-selection-drift are typically slow

We first studied the evolutionary rates for a single TF binding site at an isolated DNA sequence of the same length under mutation, genetic drift, and directional selection for high gene expression level (i.e., tighter binding). As detailed in the Models & Methods section, we combined a thermodynamically motivated fitness landscape with the mismatch approximation, and assumed that the mutation rate is low enough for the fixed state population approximation to be valid. Under these assumptions, we could calculate the inverse of the average TFBS gain and loss times as a function of the starting genotype, using either an exact method or Wright-Fisher simulations. We considered point mutations alone, or point mutations combined with short indel mutations, in order to understand under which conditions the rates of gaining and losing binding sites can reach or exceed the rates $2-3$ orders of magnitude greater than point mutation rate, and thus to become comparable to rates observed in comparative genomic studies.

Fig. 2.2A shows the dependence of the TFBS gain rate on the selection strength (with respect to genetic drift), Ns. For parameters typical of eukaryotic binding sites (length $n = 7$ bp, specificity $\epsilon = 2\,k_BT$), the TFBS gain rates are extremely slow (practically no evolution) when there is negligible selection pressure ($Ns \sim 0$), indicating the importance of selection for TFBS emergence. Indeed, the effective selection needs to be very strong, e.g., $Ns > 100$, for TFBS evolution to exceed the per-nucleotide mutation rate by orders of magnitude and become comparable to speciation rates.

Even if strong selection were present, the gain rate depends crucially on the initial genotype. While gain rates from presites, i.e., genotypes one mutation away from the threshold for strong binding, are roughly $Ns\,u$ for the strong Ns regime (as estimated by Berg et al. (2004)), they decrease dramatically if more mutational steps are needed to evolve a functionally strong binding site. This is illustrated in the inset to Fig. 2.2A, showing an exponential-like decay in the gain rates as a function of the number of mismatches, even for a TFBS of a modest length of 7 bp. As argued in the Models & Methods section (see Eq. (2.20)), we confirmed that the threshold for the strong Ns regime scales as $n\log(2)/2$ and not as $\log(n)$ which is the case for simple fitness landscapes (Paixao et al., 2015).

The availability of a realistic fraction of indel mutations (here, $\theta = 0.15$) can speed up evolution when starting from distant genotypes (cf. solid and dashed red line in Fig. 2.2A). This is because indels connect the genotype space such that paths from many to few mismatches are possible within a single mutational step. Nevertheless, the improvement due to indel mutations does not alleviate the need for very strong selective pressure and the proximity of the initial to strongly-binding sequence, in order to evolve a functional site.

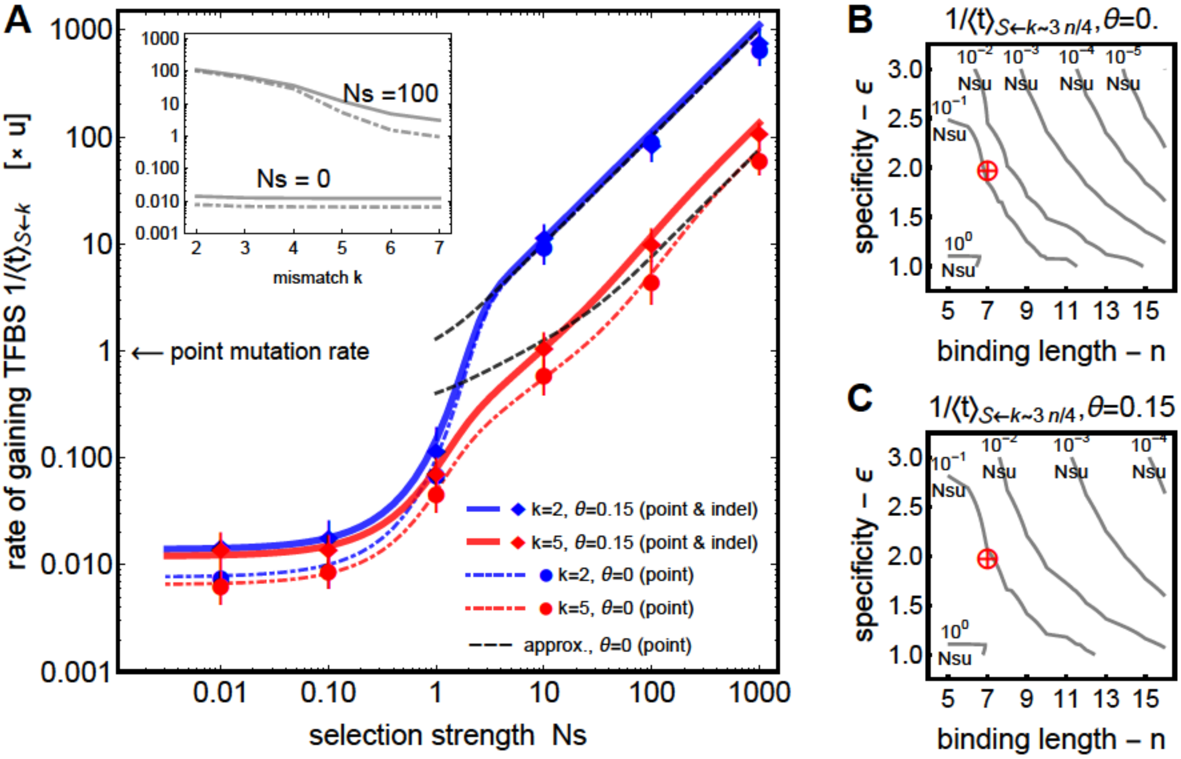


Figure 2.2: **Single TF binding site gain rates at an isolated DNA region. A)** The dependence of the gain rate, $1/\langle t \rangle_{S \leftarrow k}$ shown in units of point mutation rate, from sequences in different initial mismatch classes k (blue: $k = 2$, red: $k = 5$), as a function of selection strength. Results with point mutations only ($\theta = 0$) are shown by dashed line; with admixture of indel mutations ($\theta = 0.15$) by a solid line. For strong selection, $Ns \gg n \log(2)/2$, the rates scale with Ns, which is captured well by the "shortest path" approximation (black dashed lines in the main figure) of Eq. (2.24). The biophysical parameters are: site length $n = 7$ bp; binding specificity $\epsilon = 2\,k_B T$; chemical potential $\mu = 4\,k_B T$. Points correspond to Wright-Fisher simulations with $Nu = 0.01$ where error bars cover ± 2 SEM (standard error of mean). Inset shows the behavior of the gain rates as a function of the initial mismatch class k for $Ns = 0$ and $Ns = 100$. **B, C)** Gain rates from redundancy rich classes ($k \sim 3n/4$, typical of evolution from random "virgin" sequence) under strong selection, without (B) and with (C) indel mutations supplementing the point mutations. Red crosshairs denote the cases depicted in panel A. Contour lines show constant gain rates in units of Nsu as a function of biophysical parameters n and ϵ. Wiggles in the contour lines are not a numerical artefact but a consequence of discrete mismatch classes.

Biophysical parameters—the binding site length n, the chemical potential μ, and the specificity ϵ—influence the shape of the fitness landscape and thus the TFBS gain rates. This is especially evident when we consider *de novo* evolution starting from random sequence. As shown in Figures 2.2B, C, increases in specificity or length cause a sharp drop in the gain rates from initial sequences in the most redundancy rich class, which can be only partially mitigated by the availability of indel mutations. This especially suggests that adaptation of TFBS from random sequences for TF with very large binding lengths and very strong specificities is unlikely with point and indel mutations which can constrain the evolution of TF lengths and TF specificity, which is consistent with Berg et al. (2004)'s earlier numeric observation. Importantly, the binding specificity and length show an inverse relation with the logarithm of the gain rates. This is due to the fact that a decrease in specificity allows more genotypes to generate appreciable binding and therefore fitness (see Fig. 2.1D), which partially compensates the increase in mutational entropy at larger binding site lengths. Variation of the chemical potential μ corresponding to an order-of-magnitude change in the free TF concentration does not qualitatively affect the results.

Typically slow TFBS evolution is a consequence of the sigmoidal shape of the thermodynamically motivated fitness landscape, where adaptive evolution in the redundant but weakly binding classes $\mathcal{W}$ must proceed very slowly due to the absence of a selection gradient. To illustrate this point, we generated alternative fitness landscapes that agree exactly with the thermodynamically motivated one from the fittest class to the threshold class for strong binding, $k_{\mathcal{S}}$, but after that decay as power laws, π_{pl}, with a tunable exponent (see SI text). As seen in Fig. 2.8, this exponent is a major determinant of the gain rates, suggesting that a biophysically realistic fitness landscape is crucial for the quantitative understanding of TFBS evolution.

To check that the assumption of the fixed state population is valid at $Nu = 0.01$, the value used here that is also relevant for multicellular eukaryotes (Lynch and Conery, 2003), we performed Wright-Fisher simulations as described in the Models & Methods section. Fig. 2.2A shows excellent agreement between the analytical results and the simulation. We further increased the mutation rate to $Nu = 0.1$, a regime more relevant for prokaryotes where polymorphisms in the population are no longer negligible, to find that the analytical fixed state assumption systematically overestimates the gain rates, as shown in Fig. 2.9. In the presence of polymorphism, therefore, evolution at best proceeds as quickly as in monomorphic populations, and generally proceeds slower, so that our results provide a theoretical bound on the speed of adaptive evolution under directional selection. This is expected since the effects of clonal interference kick in after a certain Nu, where two different beneficial mutants start competing with each other, and eventually decrease the fixation probability in comparison to one beneficial mutant sweeping to fixation as in the monomorphic population case.

To check that the mismatch assumption does not strongly affect the reported results, we analyzed evolutionary dynamics with more realistic models of TF-DNA interaction. Different positions within the binding site can have different specificities, and one could suspect that this can significantly lower the evolutionary times. First, some positions within the TFBS may show almost no specificity for any nucleotide, most likely due to the geometry of TF-DNA interactions (e.g, when the TF can contact the nucleic acid residues only in the major groove); we have not simulated such cases explicitly, but simply take the binding site length n to be the effective sequence length where TF does make specific contacts with the DNA. Second, the positions that do exhibit specificity might do so in a manner that is more inhomogeneous than our mismatch assumption, which assigns zero energy to the consensus and a constant ϵ to any possible mismatch. We thus generated energy matrices where ϵ was drawn from a Gaussian distribution with the same mean $\langle\epsilon\rangle = 2\,k_BT$ as in our baseline case of Fig. 2.2A, but with a standard deviation $0.5\,k_BT$. Fig. 2.10 shows that both equal and unequal energy contributions produce statistically similar behaviors, indicating that inhomogeneous binding interactions cannot substantially enhance the evolutionary rates.

We further investigated the rate of TFBS loss (Fig. 2.11). Here too strong (negative) selection is needed to lose a site on reasonable timescales, and it is highly unlikely that a site would be lost in the presence of positive selection. In contrast to the TFBS gain case, however, negative selection and mutational entropy act in the same direction for TFBS loss, reducing the importance of the initial genotype and making selection more effective at larger n and ϵ.

Taken together, these results suggest that the emergence of an isolated TFBS under weak or no selection is typically slow relative to the species' divergence times, and gets rapidly slower for sites that are either longer or whose TFs are more specific than the baseline case considered here. This suggests that biophysical parameters themselves may be under evolutionary constraints; in particular, if point mutations and indels were the only mutational mechanisms, the evolution of long sites, e.g. $n \gg 10 - 12$, would seem extremely unlikely, as has been pointed out previously (Berg et al., 2004). Absent any mechanisms that could lead to faster evolution and which we consider below, isolated TFBS are generally only likely to emerge in the presence of strong directional selection and a favorable distribution of initial sequences that is enriched in presites.

2.3.2 Convergence to the stationary distribution is slow and depends strongly on initial conditions

A number of previous studies (e.g., (Mustonen and Lässig, 2005; Mustonen et al., 2008; Haldane et al., 2014)) assumed that a stationary distribution of mismatch classes is reached in the evolution of isolated TFBS and thus an equilibrium solution, Eq. (2.19), is informative for binding sequence distributions. In contrast, our results for average gain and loss times suggest that the evolution of an isolated TFBS is typically slow. To analyze this problem in a way that does not depend on arbitrary thresholds defining "strong" and "weak" binding classes $\mathcal{S}$ and $\mathcal{W}$, we first examined the evolution of the distribution $\psi(k)$ over the mismatch classes as a function of time in Fig. 2.3A. For typical parameter values it takes on

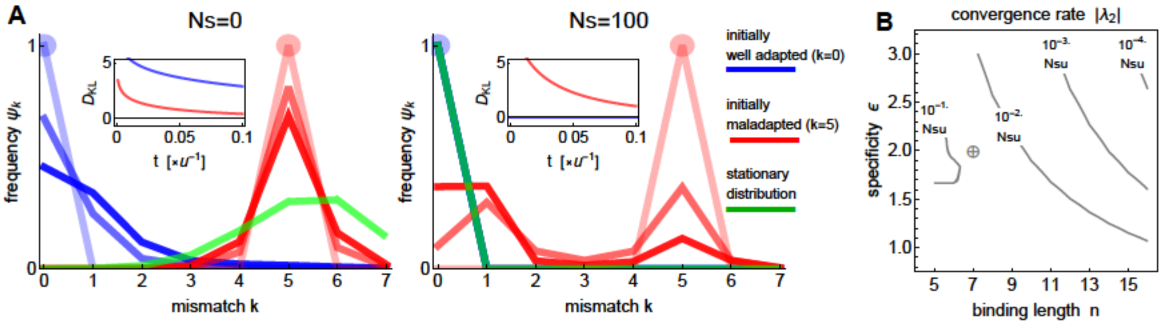


Figure 2.3: **Convergence to the stationary distribution of TFBS sequences.** **A)** Evolutionary dynamics of the mismatch classes distribution $\psi(k)$ for an isolated TFBS under point and indel mutations ($\theta = 0.15$), directional selection for stronger binding, and genetic drift is shown for initially well ($k = 0$, blue) and badly ($k = 5$, red) adapted populations. At left, no selection ($Ns = 0$); at right, strong selection ($Ns = 100$). Different curves show the distribution of genotype classes at different time points ($t = 0u^{-1}$, $0.05u^{-1}$, $0.1u^{-1}$ as decreasing opacity); stationary distribution is shown in green. Insets show the time evolution to convergence for initially well ($k = 0$, blue) and badly ($k = 5$, red) adapted populations, measured by the Kullback-Leibler divergence $D_{KL}[\psi(t)\,||\,\psi(t=\infty)]$. The biophysical parameters are: $n = 7$ bp, $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. **B)** Rate of convergence to the stationary distribution for different ϵ and n values under strong selection ($Ns \gg n\log(2)/2$; here specifically $Ns = 100$) and for $\theta = 0.15$. Crosshairs represent the parameters used in a).

the order of the inverse point mutation rate to reach the stationary distribution for populations that start off far away from it, even with strong selection.

A systematic study of the convergence rates can be performed by computing the (absolute value of the) second eigenvalue, $|\lambda_2|$, of the transition rate matrix $\mathbf{R}$ from Eq. (2.18), and exploring how this depends on the biophysical parameters n and ϵ. Consistent with previous results, we observe large increases in convergence times as n and ϵ increase. For example, an increase in the binding site length from $n = 7$ to $n = 11$ at baseline specificity of $\epsilon = 2k_BT$ would result in a ten-fold increase in the convergence time.

The intuitive reason behind the slow convergence rates is in the bimodal nature of the distribution $\psi(k)$ on the thermodynamically motivated fitness landscape, similar to that reported by Lynch and Hagner (2015). One "attractor" is located around the fittest class ($k \sim 0$, due to directional selection), while the other is located around the redundancy-rich mismatch classes ($k \sim 3/4n$). These two attractors are separated by a typically sharp fitness landscape, and the redundancy-rich attractor lacks selection gradients needed to support fast adaptation. The temporal evolution of the distribution $\psi(k)$ from, e.g., a maladapted state, can thus be best understood as the probability weight "switching" from resting approximately within one attractor to the other one, while maintaining the bimodal shape throughout, rather than a gradual shift of a unimodal distribution from a maladapted initial value of k to the value favored by selection. This is especially true when n gets larger: although adaptation within the functional sites can still happen, adaptation from the most random mismatch classes becomes extremely slow, even under strong selection (see Fig. 2.15).

These results suggest that stationary distributions of isolated TFBS sequences may not be realizable on the timescales of speciation, which should be a cause of concern when stationarity is assumed without prior critical assessment. For example, applications assuming the stationary distribution might wrongly infer selection on regulatory DNA.

2.3.3 Evolution of TF binding sites in longer sequences

So far we have shown that the evolution of isolated TFBS is typically slow. How do the results change if we consider TFBS evolution in a stretch of sequence L bp in length, where $L \gg n$, e.g., within a promoter or enhancer? Here we focus on *de novo* evolution under strong directional selection for high gene expression, by simulating the process in the fixed state population framework. Compared to the isolated TFBS case, we need to make one further assumption: that the expression level of the selected gene is proportional to the summed TF occupancy on all sites within the regulatory region of length L (see Models & Methods for details). While this is the simplest choice, it is neither unique nor perhaps the most biologically plausible one, although limited experimental support exists for such additivity (Giorgetti et al., 2010); it does, however, represent a tractable starting point when the interactions between individual TF binding sites are not strong and the contribution of each site is equal and of the same sign. To address the interactions, we look at the cooperative binding case in the following section. In Supporting Information, we also discuss the competition of TFBSs for the strongest binding, and the “nonphysical” synergetic interaction by two strongest TFBSs.

We propose a simple analytical model for the time evolution of the number of strongly binding sites, $z(t)$, in the promoter, derived from isolated TFBS gain and loss rates, λ_{gain} and λ_{loss}. Assuming constant rates, one can write

$$\frac{d}{dt}z(t) = \lambda_{\mathrm{gain}}\Big(z_{\mathrm{max}} - z(t)\Big) - \lambda_{\mathrm{loss}}z(t) \tag{2.25}$$

where z_{max} is the maximum number of TFBS that can fit into the regulatory sequence of length L bp. If the sites can overlap, $z_{\mathrm{max}} = L - n + 1$, otherwise $z_{\mathrm{max}} \approx L/n$. The solution for Eq. (2.25) is

$$z(t) = \Big(z_{\mathrm{o}} - \frac{B}{A}\Big)e^{-At} + \frac{B}{A} \tag{2.26}$$

where $A = \big(\lambda_{\mathrm{gain}} + \lambda_{\mathrm{loss}}\big)$, $B = z_{\mathrm{max}}\lambda_{\mathrm{gain}}$ and $z_{\mathrm{o}} = z(t = 0)$. Under strong positive selection, i.e. $Ns \gg n\log(2)/2$, the loss rate λ_{loss} can be ignored. If the distribution of the initial mismatch classes in the promoter is ψ_k, one can approximate $z_{\mathrm{max}} - z_{\mathrm{o}} = z_{\mathrm{max}} \sum_{k=k_{\mathcal{S}}+1}^{n} \psi_k$ to obtain:

$$z(t) - z_{\mathrm{o}} = \big(1 - e^{-\lambda_{\mathrm{gain}}t}\big)\, z_{\mathrm{max}} \sum_{k=k_{\mathcal{S}}+1}^{n} \psi_k. \tag{2.27}$$

There are two limiting regimes in which we can examine the behavior of Eq. (2.27). Over a short timescale, evolutionary dynamics will search over all possible positions, $z_{\max} = L - n + 1$, to pull out the presites, since they are fastest to evolve into the strong binding class $\mathcal{S}$, i.e.:

$$\lambda_{\text{gain}} \approx \lambda_{\text{gain}}^{\text{presite}} = \Big(\sum_{k \notin \mathcal{S}} \psi_k\Big)^{-1} \psi_{k_{\mathcal{S}}+1} / \langle t \rangle_{\mathcal{S} \leftarrow k_{\mathcal{S}}+1} \tag{2.28}$$

As the process unfolds and new sites are established, new TFBS will only be able to emerge at a smaller set of positions due to possible overlaps, so that $z_{\max} \approx L/n$. On the other hand, evolution from higher mismatch classes will also start to contribute towards new sites:

$$\lambda_{\text{gain}} \approx \lambda_{\text{gain}}^{\text{all}} = \Big(\sum_{k \notin \mathcal{S}} \psi_k\Big)^{-1} \sum_{k \notin \mathcal{S}} \psi_k / \langle t \rangle_{\mathcal{S} \leftarrow k} \tag{2.29}$$

Fig. 2.4 shows how new TFBSs with length $n = 7$ bp emerge over time in a promoter of $L = 30$ bp in length. Consistent with the predictions of our simplified model, we can distinguish the early, intermediate, and late epochs. In the early epoch, $t < 0.01u^{-1}$, presites are localized among all possible locations and are established as binding sites. During this period, the growth in the expected number of new TFBSs is linear with time. The importance and predictive power of presites at early epoch remain even under different models of gene expression, including interaction between TFBSs (see Fig. 2.14). In the intermediate epoch, new binding sites accumulate at the rate that is slightly above that expected by establishment from presites alone, as the mutational neighborhood is explored further. In the late epoch, $t > 0.1u^{-1}$, initial sites in the immediate mutational vicinity have been exhausted, and established sites have constrained the number of positions where new sites can evolve from more distant initial sequences, leading to the saturation in the number of evolved TFBS.

Using the simple analytical model, we explored in Fig. 2.4B,C how the binding length n and specificity ϵ affect the number of newly evolved TFBS. Increasing n leads to a steep decrease in the number of expected sites, with a somewhat weaker dependence on ϵ, especially at early times. Simulations at other values of biophysical and evolutionary parameters confirm the qualitative agreement between the analytical model and the simulation (Fig. 2.12); given that the model is a simple heuristic, it cannot be expected to match the simulations in detail, yet it nevertheless seems to capture the gross features of evolutionary dynamics. Together, these results show that at early times under strong selection, the number of newly evolved sites will grow linearly with time and proportional to L, before evolution from higher mismatch classes can contribute and ultimately before the sites start interacting, with a consequent slowdown in their evolution. Thus, evolution in longer regulatory regions ($L = 10^2 - 10^3$ bp) could feasibly give rise to tens of binding sites at $Ns = 10^2 - 10^3$ within a realistic time frame $t \sim 0.001u^{-1}$, if the sites are sufficiently short ($n \sim 7$ bp). Explaining the evolution of longer sites, e.g., $n > 10 - 12$ bp, especially within short promoters found in prokaryotes, would likely necessitate invoking new mechanisms.

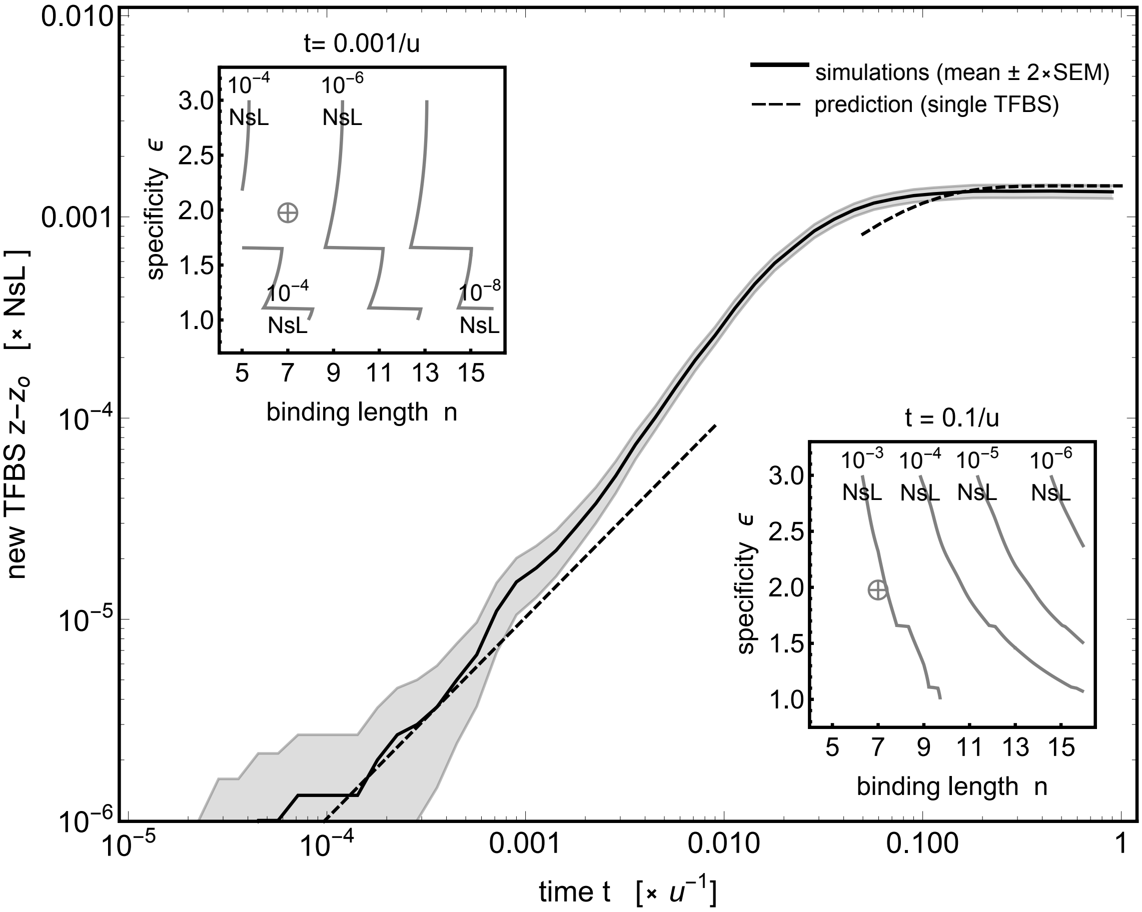


Figure 2.4: **TF binding site evolution in a longer sequence of $L = 30$ base pairs.** The expected number of newly evolved TF binding sites with length $n = 7$ bp, under strong directional selection ($Ns = 100$) and both point and indel mutations ($\theta = 0.15$). Time is measured in inverse mutation rates; the number of newly evolved sites is scaled to the selection strength and the sequence length. 1000 replicate simulations were performed with different initial sequences. Average number of sites shown by a solid black line; the gray band shows ± 2 SEM (standard error of the mean) envelope. Dashed curves are analytical predictions based on single TFBS gain rates at an isolated DNA region, given by Equations (2.27,2.28,2.29). Biophysical parameters used: $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. **Insets:** Expected number of newly evolved sites from a random sequence of length L at $t = 0.001u^{-1}$ (left) and $t = 0.1u^{-1}$ (right) for different binding length and specificity values, computed using the analytical predictions. Crosshairs denote the values used in the main panel.

2.3.4 Ancient sites and cooperativity between TFs can accelerate binding site emergence

Finally, we briefly examine two mechanisms that can further speed up the evolution of TF binding sites in longer sequences.

The first possibility is that the sequence from which new TFBS evolve is not truly random; as discussed previously, presites have a strong influence on the early accumulation of new binding sites. There are a number of mechanisms that could bias the initial sequence distribution towards presites: examples include transposable elements, DNA repeats, or CG content bias. Here we consider an alternative mechanism that we refer to as the “ancient TFBS scenario,” in which a strong TFBS existed in the sequence in the ancient past, after which it decayed into a weak binding site, possibly due to the relaxation of selection (i.e., $Ns \sim 0$).

As we demonstrated in the context of isolated sites, TFBS loss rates are slow and the remains of the binding site will linger in the sequence for a long time before decaying into the most redundancy rich mismatch classes. This biased initial distribution of mismatches $\boldsymbol{\Psi}$ in a sequence of length L with a single ancient site can be captured by writing:

$$\boldsymbol{\Psi} = \frac{1}{L-n+1}\,\boldsymbol{\psi}(t') + \frac{L-n}{L-n+1}\,\boldsymbol{\phi} \tag{2.30}$$

where $\boldsymbol{\phi}$ is the binomial distribution, Eq. (2.8), characteristic of the random background, and $\boldsymbol{\psi}(t')$ is the distribution of mismatches due to the presence of the ancient site. Time t' refers to the interval in which the isolated ancient TFBS has been decaying under relaxed selection, and the corresponding $\boldsymbol{\psi}(t')$ can be solved for using Eq. (2.17).

Fig. 2.5A shows that the ancient site scenario can enhance the number of newly evolved sites by resurrecting the ancient site, even after it has decayed for $t' = 0.1u^{-1}$. Simulation results agree well with the simple analytical model using the biased initial sequence distribution of Eq. (2.30). Importantly, such a mechanism is particularly effective for longer binding sites of high specificity, indicating that regulatory sequence reuse could be evolutionarily beneficial in this biophysical regime (see Fig. 2.13).

Fig. 2.5A and Fig. 2.13 also show the emergence of new sites when the ancient site was not a full consensus (preferred) sequence but differed from it by a certain number of mismatches. The results qualitatively agree with the case of perfect consensus. Importantly, this shows that the applicability of the ancient site scenario extends to cases where the ancient site belonged to a different TF (albeit with a preferred sequence similar to the studied TF), which has recently been reported to be a frequent phenomenon by Payne and Wagner (2014), possibly due to evolution of TFs by duplication and divergence (Weirauch et al., 2014).

The second mechanism that we consider is the physical cooperativity between TFs: when one site is occupied, it is favorable for the nearby site to be occupied as well. We extended the thermodynamic

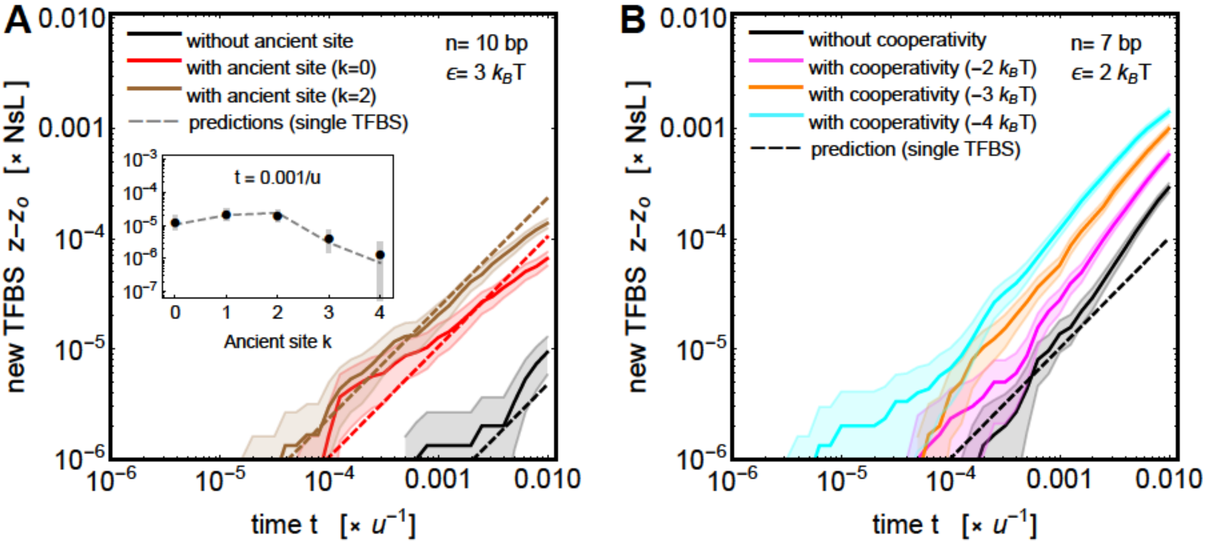


Figure 2.5: **Ancient sites and cooperativity can accelerate the emergence of TF binding sites in longer regulatory sequences.** **A)** The expected number of newly evolved TFBS in the presence (red and brown) or absence (black) of an ancient site, for binding site length $n = 10$ bp, and specificity, $\epsilon = 3\,k_BT$. In this example, the ancient site was a consensus site ($k = 0$) or two mismatches away from it ($k = 2$) that evolved under neutrality for $t' = 0.1/u$ prior to starting this simulation. Dashed lines show the predictions of a simple analytical model, Eq. (2.30). The inset shows how the number of newly evolved TFBS at $t = 0.001/u$ scales with the mismatch of the ancient site k (plot markers: simulation means; error bars: two standard errors of the mean; dashed curve: prediction). **B)** The expected number of newly evolved TFBS without (black) and with cooperative interactions (for different cooperativity strengths, magenta: $E_c = -2\,k_BT$, yellow: $E_c = -3\,k_BT$, cyan: $E_c = -4\,k_BT$, see Eq. (2.11) in the Models & Methods and text) for binding site length $n = 7$ bp, and specificity, $\epsilon = 2\,k_BT$. Both panels use $\mu = 4\,k_BT$, strong selection ($Ns = 100$) and a combination of point and indel mutations ($\theta = 0.15$), acting on a regulatory sequence of length $L = 30$ bp. Thick solid lines show an average over 1000 simulation replicates, shading denotes ± 2 SEM.

model to incorporate cooperativity (see the Models & Methods, Eq. (2.11) and Fig. 2.1D). The genotype of a nearby site will then influence whether a given site acts as a strongly or weakly binding site. The presence of a cooperative site acts as a local shift in the chemical potential, which changes the weak/strong threshold, so that an individually weak site can become a strongly binding site. Simulations using cooperative binding presented in Fig. 2.5B illustrate how cooperativity can increase the speed of evolution. This is specifically effective for short binding sites of intermediate or low specificity, where a cooperative energy contribution can strongly influence the number of sites in the strong binding class (see Fig. 2.13).

2.4 Discussion

In this study, we aimed at a better theoretical understanding of which biophysical and population genetic factors influence the fast evolution of TFBSs in gene regulatory DNA, making sequence specific TF binding a plausible mechanism for the evolution of gene regulation and for generating phenotypic diversity. Following Berg et al. (2004), we combined a biophysical model for TF binding with a simple population genetic model for the rate of sequence evolution. The key assumptions are that binding probability is determined by a thermodynamic equilibrium; that fitness depends linearly on binding probability; and that populations are typically homogeneous in genotype, and so evolve by substitution of single point and short insertion/deletion (indel) mutations. Remarkably, the biophysical and the evolutionary models take the same mathematical form: in the biophysical model, binding probability depends on the binding energy, relative to thermal fluctuations, βE, whilst in the evolutionary model, the chance that a mutation fixes depends on its selective advantage, relative to random sampling drift, Ns.

For single TFBS evolution, we calculated the average transition time between genotypes, the inverse being a measure for the speed of the evolution. Our results indicate that TFBS evolution is typically slow unless selection is very strong. It is important to emphasize that gaining a TFBS by point mutations under neutral evolution is very unlikely, contrasting with the belief in the current literature (e.g., (Wittkopp, 2013; Villar et al., 2014)). This is mainly due to Stone and Wray (2001)'s argument that functional sites could readily be found by a random walk; however, their argument assumed that individuals follow independent random walks, which grossly overestimates the rate of evolution (see MacArthur and Brookfield (2004)). Indeed, fast rates of gaining a single TFBS require not only strong selection but also initial sequences in the mutational neighborhood of the functional sites. Especially, “presites,” i.e. sequences 1 bp away from threshold sequences, can be crucial since they can evolve to functional sites by single mutations. Indel mutations can increase the rate of gaining a single TFBS from distant sequences, since they connect the genotype space extensively, but their effect is limited under realistic indel mutation rates (Cartwright, 2009; Chen et al., 2009). Future studies should consider the updates in estimates of indel mutation rates, since they are currently not as precise as point mutation rates, although we do not expect big qualitative departures from our results.

Considering the evolution of a single TFBS from random sequence, we showed that biophysical parameters, binding length and specificity, are constrained for realistic evolutionary gain rates from the most redundant mismatch classes. The rates drop exponentially with binding length, making TF whose binding length exceeds $10 - 12$ bp difficult to evolve from random sites, at least under the point and indel mutation mechanisms considered here. As a consequence of the biophysical fitness landscape, binding specificity and length show an inverse relation for the same magnitude of the gain rate from the most redundant mismatch class. Such an inverse relation is observed in position weight matrices of TFs collected from different databases for both eukaryotic and prokaryotic organisms, by Stewart and Plotkin (2012). In the same study, they reproduce this observation using a simple model which assumes that a trade-off between the selective advantage of binding to target sites, versus the selective disadvantage of binding to non-target sequence. Their model assumes a stationary distribution, and that sites are functional if they are mismatched at no more than one base. It would be interesting to explore a broader range of models that account for the dynamical coevolution between transcription factor binding specificity, its length, and its binding sites (Lynch and Hagner, 2015). One idea can be to combine the evolutionary dynamical constraints (against large binding length and high specificity, which we show here) with simple physical constraints of TF dilution in non-target DNA (against short binding length and low specificity, again in an inverse relation (Gerland et al., 2002)).

For a single TF binding site, the stationary distribution for the mismatch with the consensus binding sequence depends on the binding energy, but also on the sequence entropy – that is, the number of sequences at different distances from the consensus. Typically, the distribution is bimodal: either the site is functional, and is maintained by selection, or it is non-functional, and evolves almost neutrally. We show that it may take an extremely long time for the stationary distribution to be reached. Functional sites are unlikely to be lost if selection is strong (i.e., $Ns \gg 1$), whilst function is unlikely to evolve from a random sequence by neutral evolution, even if predicted under stationarity assumption. Therefore, typical rapid convergence to stationary distribution should be considered with caution in theoretical studies.

We showed that the dynamics of TFBS evolution in longer DNA sequences can be understood from the dynamics of single TFBS. The rate of evolution of new binding sites will be accelerated in proportion to the length of the promoter/enhancer sequence in which that can be functional; however, because this increase is linear in promoter/enhancer length, it will have a weaker influence than the exponential effect changes in specificity or length of binding site. Especially the earlier dynamics (relevant for speciation timescales) are determined by the availability of presite biased sequences. Any process that allowed selection to pick up more distant sequences or that increased presite ratio among non-functional sites would accelerate adaptation from “virgin” sequences.

A key factor for an enrichment in presite ratio may arise through variation in GC content or through simple sequence repeats (especially if the preferred sequence has some repetitive or palindromic structure). In this study, we showed that it may also arise from ancient sites, i.e. sites that were functional in earlier evolutionary history and decayed into nonfunctional classes in evolution. Since loss of function is slow (comparable to the neutral mutation rate once selection becomes ineffective), this is plausible for sites that are under intermittent selection, or where there is a shift to binding by a new TF with similar preferred sequence (Payne and Wagner, 2014; Weirauch et al., 2014). This effect of the earlier evolution can be especially important for long binding TFs as convergence to a truly randomized sequence distribution requires much longer times. MacArthur and Brookfield (2004) showed that real promoter sequences may acquire functional sites more quickly than random sequence, but it is not clear whether that is due to a different general composition, or to the ghosts of previous selection. New studies are required to test our enriched presite-biased sequence hypothesis, especially for orthologous regions where functional TFBS is observed in sister populations or species. In a recent study, Villar et al. (2015) provide evidence that enhancer DNA sequence structure is older than other DNA portions, suggesting the reuse of such regions in evolution, plausibly by gaining and losing TFBSs in repetitive manner. Nourmohammad and Lässig (2011) showed evidence suggesting that local duplication of sequences followed by point mutations played important role in binding site evolution in Drosophila species (but surprisingly, not in yeast species). Another interesting option would be the existence of "mobile" presites or their fragments, e.g., as sequences embedded into transposable elements that could be inserted before the gene under selection for high expression (Feschotte, 2008). Presites can be considered as concrete examples of cryptic sequences (Rajon and Masel, 2013), potential source of future diversity and evolvability. We believe that understanding the effects of presites would contribute to the predictability of genetic adaptations regarding gene regulation, especially in important medical applications such as antibiotic resistance or virus evolution.

We also showed that the evolution of a functional binding site in longer DNA can be accelerated by cooperativity between adjacent transcription factors. When a TF occupies a co-binding site, sufficient transcriptional activity can be achieved from sequences of larger mismatch classes, an effect similar to a local increase in TF concentration. This mechanism permits faster evolution towards strongly binding sequences, and seems most effective for short TFBS where it creates a selection gradient already in the redundancy rich mismatch classes. Cooperative physical interactions might allow the evolution of binding occupancy and thus expression without large underlying sequence changes, which might be a reason for the observed weak correlation between sequence and binding evolution at certain regulatory regions. Importantly, TFBS clustering in eukaryotic enhancers can be a consequence of the fast evolution with cooperativity, as also supported by a recent empirical study (Stefflova et al., 2013).

Our theoretical framework is relevant more broadly for understanding the evolution of gene regulatory architecture. Since the speed of TFBS evolution from random sequences is proportional to NsL, our results suggest that population size N and the length of regulatory sequences L can compensate for each other in terms of the rate of adaptation. This is exactly what is observed: eukaryotes typically have longer regulatory DNA regions but small population sizes, while prokaryotes evolve TFBS within shorter regulatory sequence fragments but have large population sizes. Similarly, prokaryotes might have achieved longer TF binding lengths n, as large population size allowed them to overcome the exponential decrease in the gain rates with increasing n. If relevant, these observations would suggest that an important innovation in eukaryotic gene regulation must have been the ability of the transcriptional machinery to integrate the simultaneous occupancy of many low-specificity transcription factors bound over hundreds of basepairs of regulatory sequence, a process for which we currently have no good biophysical model.

2.5 Supporting Information

2.5.1 Other fitness models for comparison & for interacting TFBSs

Power-law decaying fitness models for comparison:

In order to understand the importance of the thermodynamically-motivated sigmoid shape for the binding probability, we compare our results to those obtained with power-law functions that decay with exponent γ (note that $\gamma = \infty$ corresponds to a step-like fitness landscape), formally defined as

$$\pi_{\mathrm{pl}}(k) = \begin{cases} \pi_{\mathrm{TD}}(k) & k \leq k_{\mathcal{S}} \\ \left(k_{\mathcal{S}}/k\right)^{\gamma}\, \pi_{\mathrm{TD}}(k_{\mathcal{S}}) & k > k_{\mathcal{S}} \end{cases}. \tag{2.31}$$

Fig. 2.8 shows that the power-law exponent is a major determinant of the gain rates, suggesting that a biophysically realistic fitness landscape is crucial for the quantitative understanding of TFBS evolution.

Fitness models of interacting TFBSs in larger regulatory sequence:

In addition to physical cooperativity between nearby TFs on promoter/enhancers (see the Models & Methods, Fig. 2.5 and Fig. 2.13), here we also consider two other models. The first additional model assumes that the binding occupancy of the strongest binding site in the regulatory sequence is the proxy for the gene expression level and the fitness, i.e.

$$f(\boldsymbol{\sigma}) = s\,\mathrm{MAX}\{\pi^{(\mathrm{i})}(\boldsymbol{\sigma})\}. \tag{2.32}$$

Note that different TFBSs interact with each other to compete for the strongest binding within a promoter or an enhancer.

The second additional model addresses synergistic interaction between the two strongest-binding TFBS, located anywhere in the regulatory sequence. This example is a simplified version of a biophysical model where TFs, binding anywhere in a regulatory region, compete for the occupancy of that region with a nucleosome (for a more elaborative modeling framework, see Mirny (2010)). We call this type of interaction between two TFs "non-physical" because TFs don't interact directly; their interaction is effectively mediated by some other biophysical process. The probability of the joint occupancy of the two TFs at promoter or enhancer can be used as the proxy for gene expression level and the fitness, i.e.

$$f(\boldsymbol{\sigma}) = s\,\frac{e^{-\beta(\epsilon(k_1+k_2)-2\mu)}}{1+e^{-\beta(\epsilon k_1-\mu)}+e^{-\beta(\epsilon k_2-\mu)}+e^{-\beta(\epsilon(k_1+k_2)-2\mu)}}, \tag{2.33}$$

where k_1 and k_2 correspond to the genotypes of two TFBSs with the smallest mismatches in the regulatory sequence.

Do these models yield different result for the emergence of strong binding sites from random sequences at early evolutionary times ($\sim$ speciation time scales), in comparison to our main model, where the sum of binding occupancies is used as a proxy for gene expression level [Eq.(2.7) in the main text]? For typical biophysical parameters (binding lenght: $n = 7$ bp, binding specificity: $\epsilon = 2\ k_BT$ and chemical potential: $\mu = 4\ k_BT$), we show in Fig. 2.14 that these modified models do not differ extensively from results of our main model.

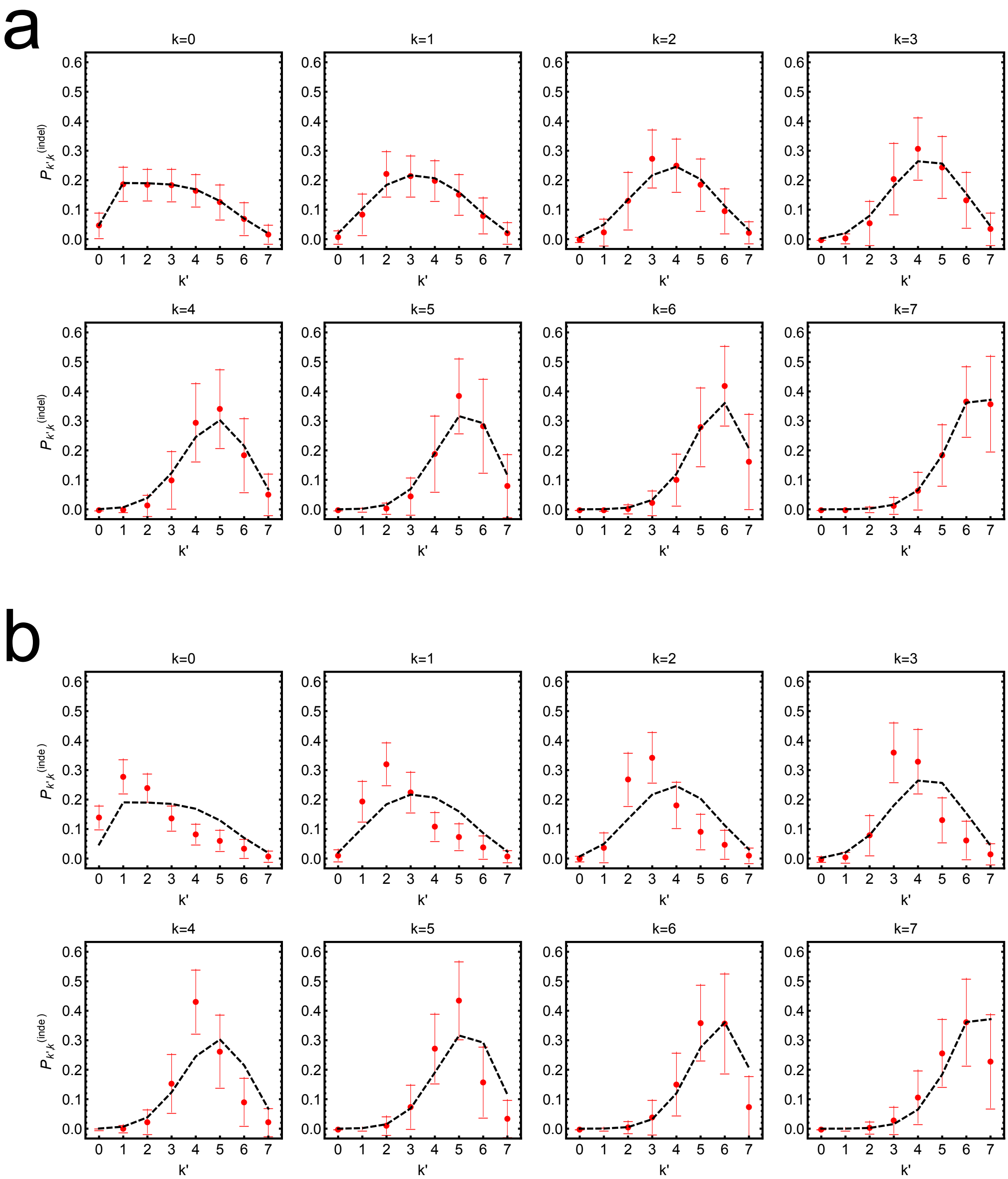


Figure 2.6: **Indel mutations connect the mismatch genotype space differently from point mutations.** **a)** Probability that a binding site with k mismatches mutates to k' mismatches, for a single binding site of length $n = 7$ bp, according to our indel mutation model in a fixed genomic window (see the Models & Methods section). Dashed curve = analytical prediction according to Eq. (2.13). Red points = mean ± 1 std of 10^3 replicate realizations of the frequency distribution (for each replicate, 1 consensus suence is created and 10^4 mutations are simulated for each k). **b)** The same analysis as in a), but allowing for a flexible genomic window for alignment after insertion mutations. We pick the minimal mismatch case to asses the quality of our approximation. As expected, this creates a bias towards smaller mismatch classes, but suggests that our approximation is still reasonable.

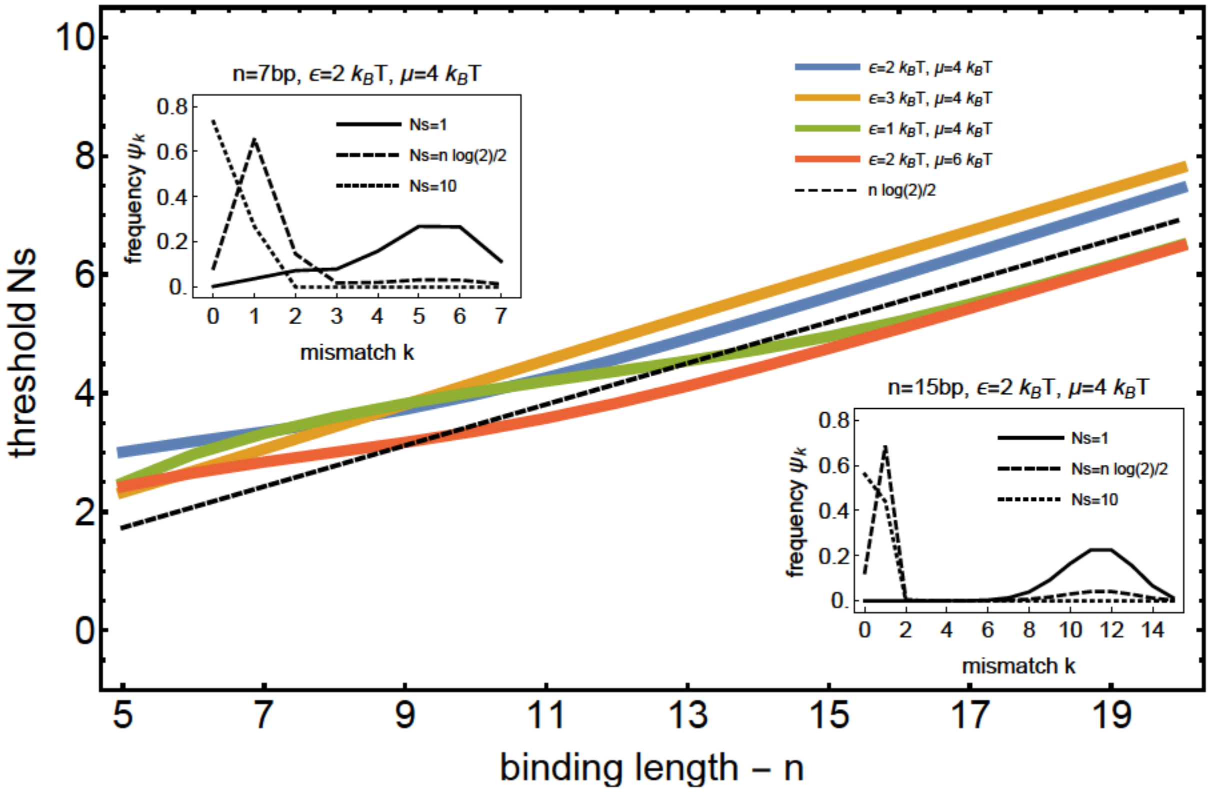


Figure 2.7: **Threshold value of Ns for bimodality (i.e., threshold between strong and weak selection regimes).** The value of Ns at which 5% of the probability weight in the stationary distribution is in non-strong mismatch classes, i.e. $k > k_S$. For selection stronger than this threshold, the stationary distribution is concentrated at low k (high fitness) classes and is practically unimodal. Different colors correspond to different biophysical parameters (see legend), analytical prediction $n\log(2)/2$ is in black (see the Models & Methods section and Eq. (2.20)). Insets show examples of stationary distributions for different Ns values for short and long binding sites.

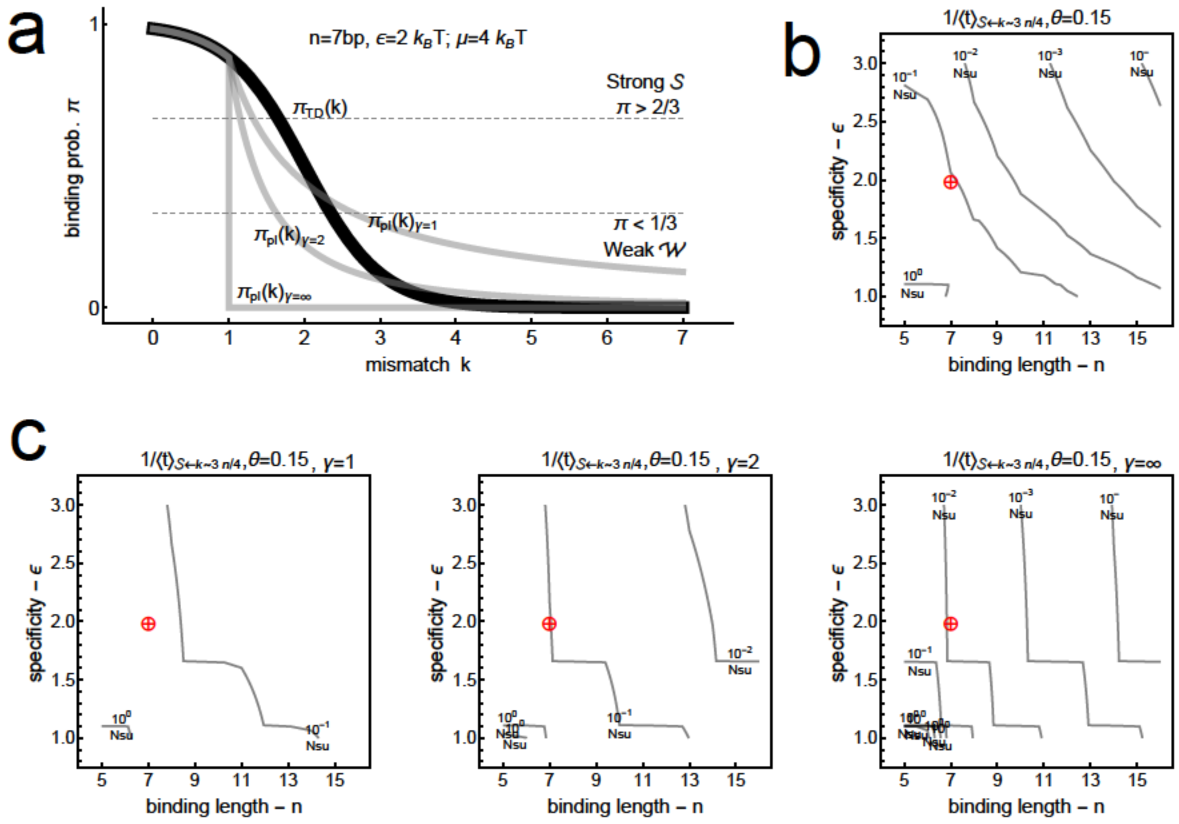


Figure 2.8: **Single TFBS gain rates in modified fitness landscapes with a power-law tail.** The thermodynamic fitness landscape has been modified to have a power-law decaying tail of exponent γ for $k > k_{\mathcal{S}}$, as in Eq. (2.31) in SI text. We tested $\gamma = 1, 2$ and ∞ corresponding to smooth, intermediate and step-like decay. Plot conventions are the same as in Fig. 2.2C. **b)** Isolated TFBS gain rate from the most redundant mismatch class for the thermodynamic model, replotted from Fig. 2.2C for reference. **c)** Plots analogous to b) using modified fitness landscapes defined by the power-law exponent γ. Gain rates are higher for small $\gamma = 1$ and lower for the step landscape ($\gamma = \infty$), relative to the reference.

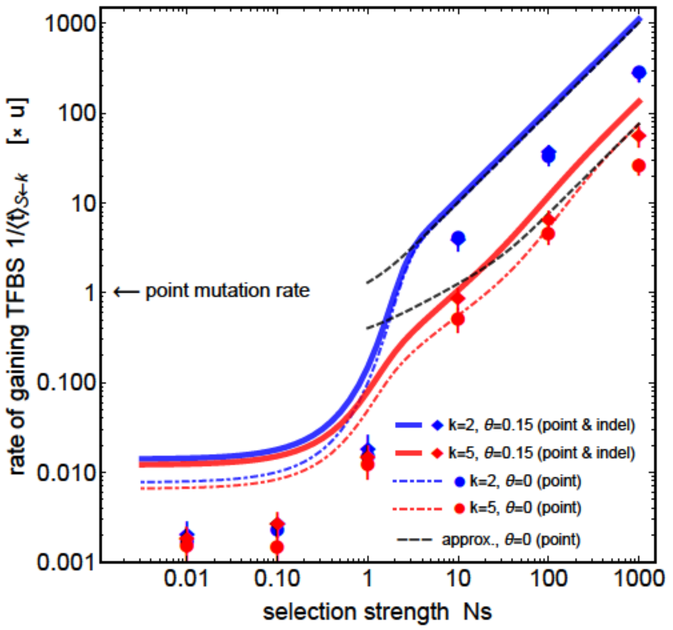


Figure 2.9: **The effect of polymorphisms on the single TFBS gain rate at higher mutation rates.** Wright-Fisher simulation results (point markers, error bars = 2 standard errors of the mean) at $4Nu = 0.1$, in comparison to the fixed state model (continuous curves). Plot conventions are the same as in Fig. 2.2. Biophysical parameters used: $n = 7$, $\epsilon = 2$ k_BT, $\mu = 4$ k_BT. Polymorphisms generally decrease TFBS gain rates.

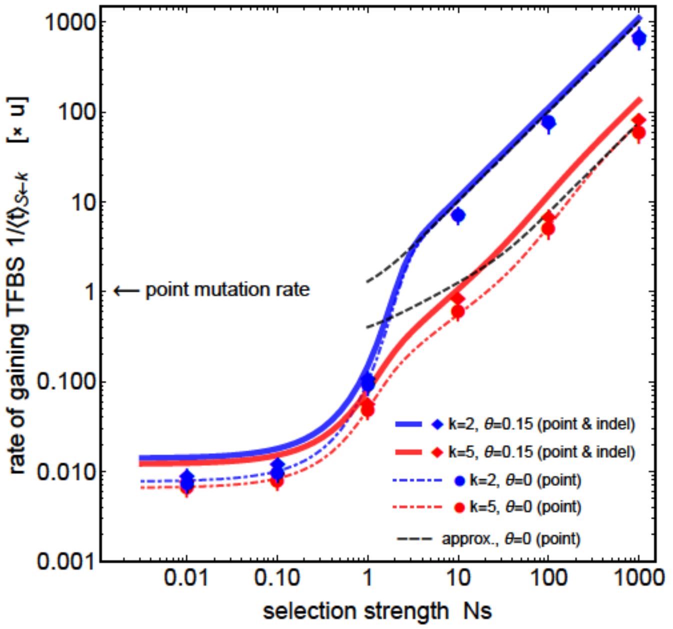


Figure 2.10: **Relaxing the mismatch assumption.** Fig. 2.2, but using energy matrices whose nonzero entries are gaussian random variables ε_i, such that $\langle \varepsilon_i \rangle = \epsilon = 2k_BT$ and $\sigma_\varepsilon = 0.5k_BT$; $n = 7$, $\mu = 4k_BT$. The analytical results under the equal mismatch assumption are shown in continuous lines.

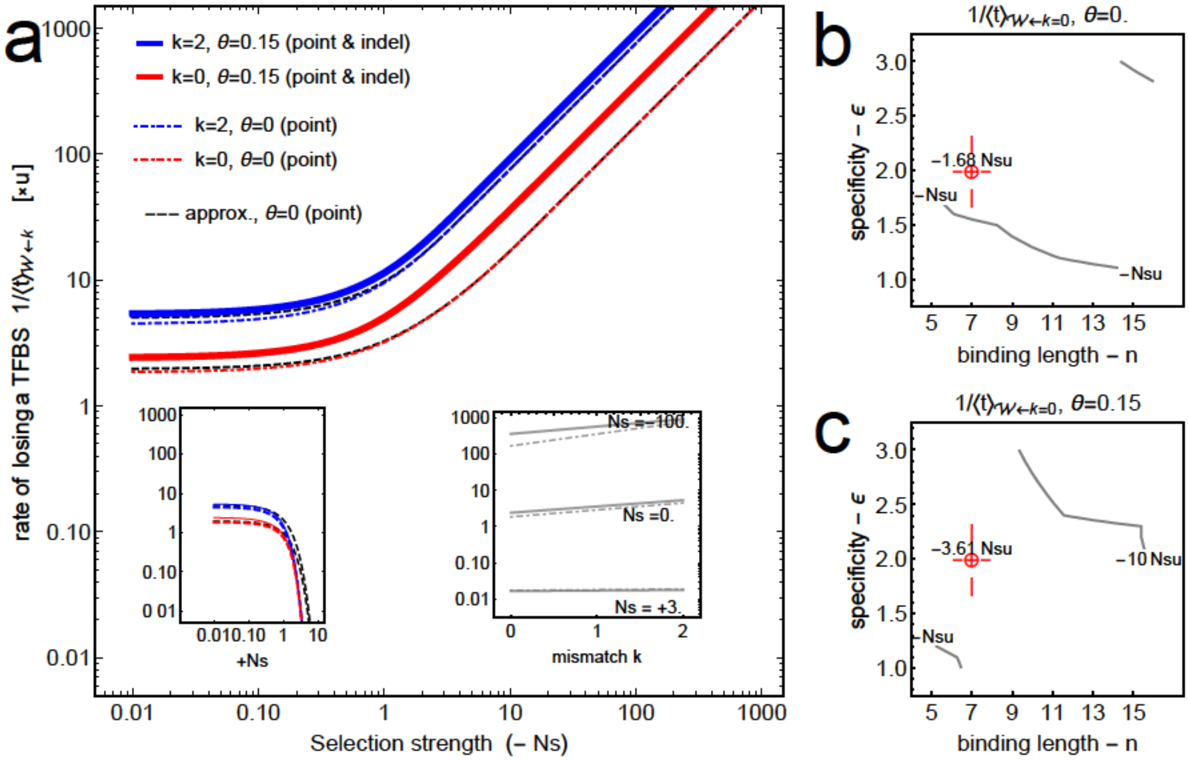


Figure 2.11: **Single TF binding site loss rates at an isolated DNA region.** The dependence of the loss rate, $1/\langle t\rangle_{\mathcal{W}\leftarrow k}$ shown in units of point mutation rate, from sequences in different initial mismatch classes k (blue: $k = 2$, red: $k = 0$), as a function of negative selection strength. Results with point mutations only ($\theta = 0$) are shown by dashed line; with admixture of indel mutations ($\theta = 0.15$) by a solid line. For strong selection, $|Ns| \gg 1$, the rates scale with $2|Ns|nu$, which is captured well by the "shortest path" approximation (black dashed lines in the main figure) of Eq. (2.24). The biophysical parameters are: site length $n = 7$ bp; binding specificity $\epsilon = 2\ k_B T$; chemical potential $\mu = 4\ k_B T$. Left inset: Ns-scaling with positive selection. Right inset: gain rates as a function of the initial mismatch class k for different Ns. **b, c)** Loss rates from the consensus sequence ($k = 0$) under strong negative selection, without (b) and with (c) indel mutations supplementing point mutations. Red crosshairs denote the cases depicted in panel a). Contour lines show constant loss rates in units of $N s\, u$ as a function of biophysical parameters n and ϵ.

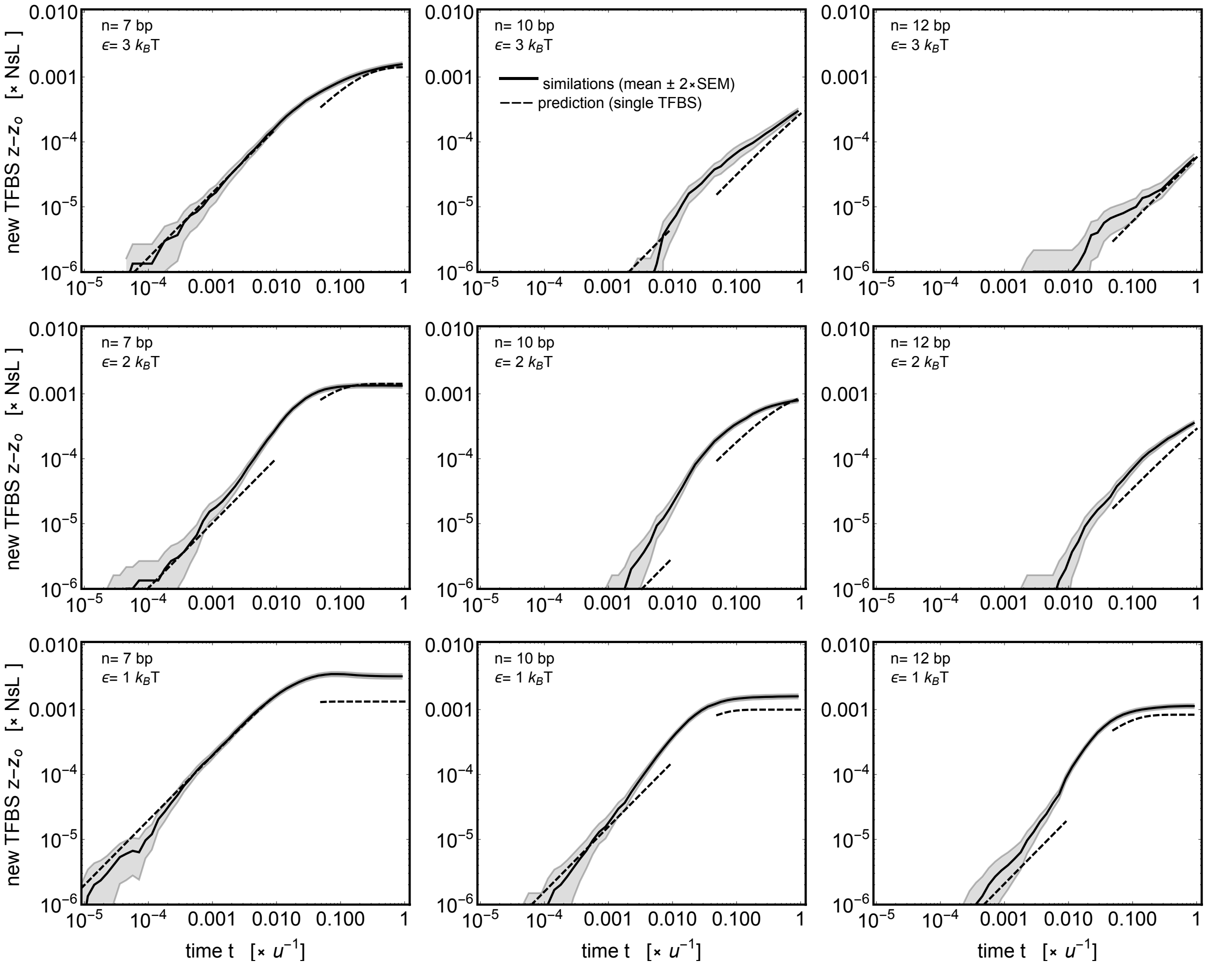


Figure 2.12: **TFBS evolution in longer sequences.** Example simulations (black solid line) and analytic predictions based on single TFBS gain/loss rates (black dashed line), for different binding length n and specificity ϵ. Details are identical to Fig. 2.4.

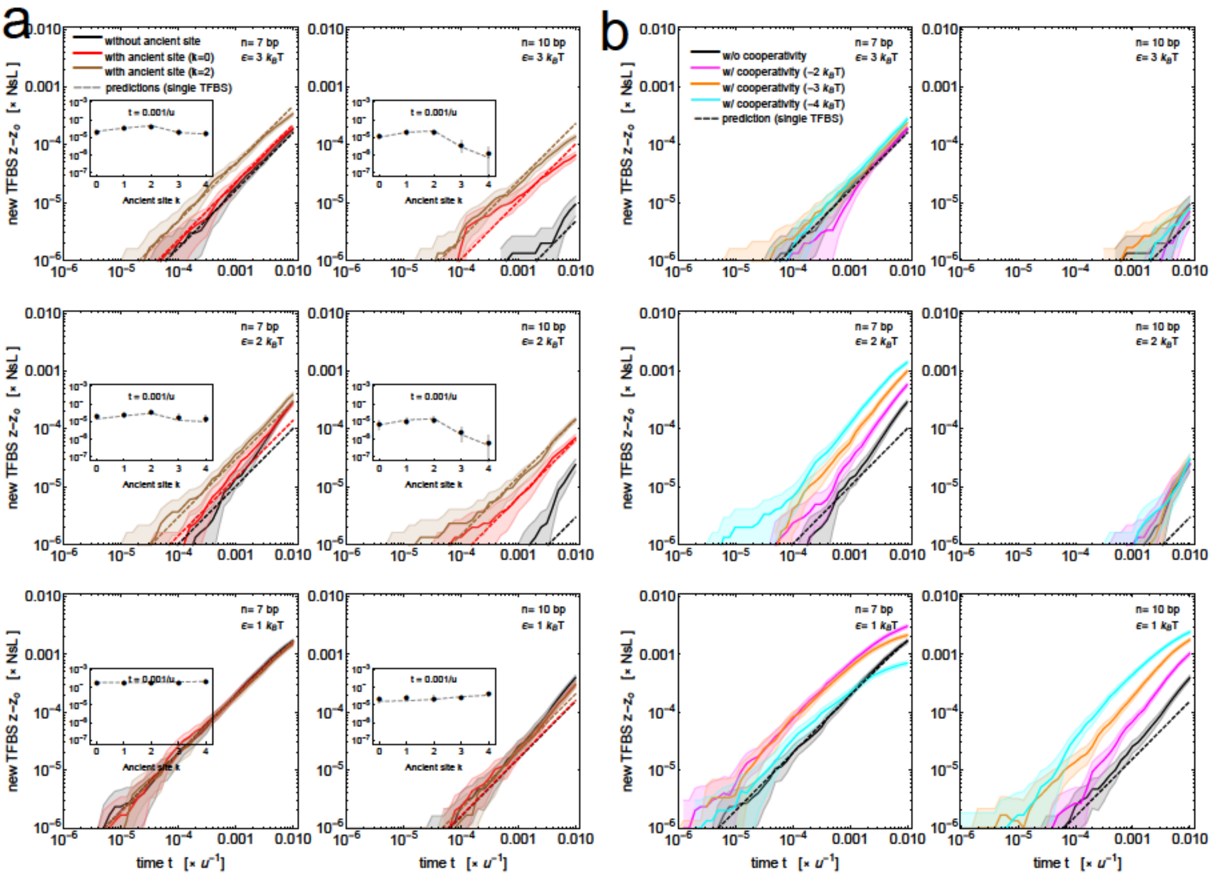


Figure 2.13: **The effect of ancient sites (a) and cooperativity (b) for different binding lengths and specificities.** Simulations of TFBS evolution in longer sequences (colored lines) and analytic predictions based on single TFBS gain and loss rates (dashed black lines), analogous to Fig. 2.5. Different panels show different choices of TFBS binding length n and specificity ϵ. Ancient sites specifically facilitate the emergence of longer sites of high specificity, whereas cooperativity specifically facilitates the emergence of shorter sites of intermediate or low specificity.

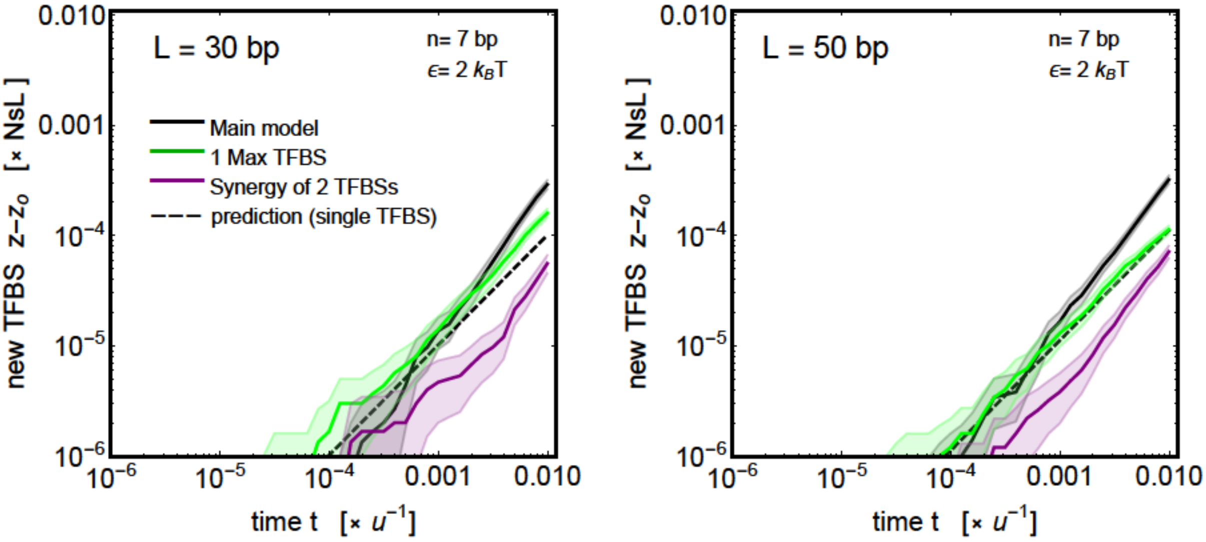


Figure 2.14: **Fitness models of interacting TFBSs.** The expected number of newly evolved TFBS for binding site length $n = 7$ bp, specificity $\epsilon = 2\,k_BT$, and chemical potential $\mu = 4\,k_BT$ are shown for different fitness models. The solid black curve is the non-interacting model used in the main text (dashed curve: theoretical prediction). The green curve stands for the model of Eq. (2.32) in SI text, where only the strongest binding site in the regulatory sequence determines gene expression. The purple curve stands for the model of Eq. (2.33) in SI text, where two strongest TFBS synergistically determine the gene expression level. Shading denotes ± 2 SEM. The simulations use regulatory sequences of length $L = 30$ bp (left) and $L = 50$ bp (right).

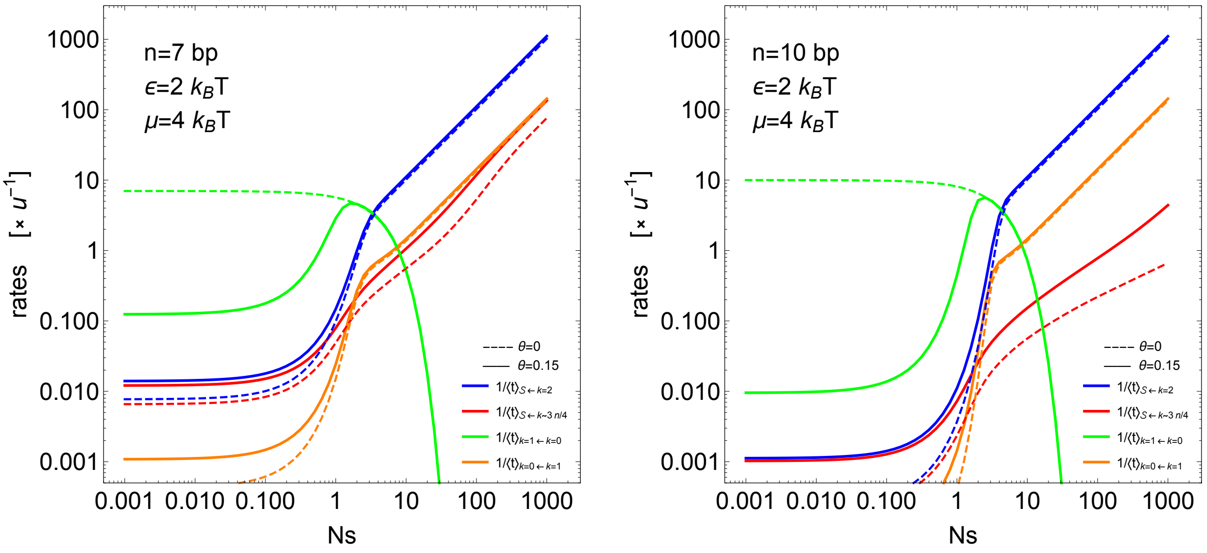


Figure 2.15: **Comparison rates of TFBS gain rates and sequence turnover rates within functional TFBSs** Average first hitting times to particular mismatch k_j state can be calculated with a minor modification to Eq. (2.21) by replacing $\mathcal{S}$ with k_j. The figures compare the rates of evolution of TFBS within the functional sites (i.e. $1/\langle t\rangle_{k=0\leftarrow k=1}$ and $1/\langle t\rangle_{k=1\leftarrow k=0}$). Plot conventions are the same as in Fig. 2.2-A. Biophysical parameters used: $n = 7$ bp (left), $n = 10$ bp (right) $\epsilon = 2\ k_BT$, $\mu = 4\ k_BT$. It shows that for weak selection, the rates to evolve from $k = 0$ to $k = 1$ can be relatively faster. Also, although adaptation from random sites slows down with increasing n, we see that the adaptation rate to evolve from $k = 1$ to $k = 0$ can stay high.

3

Binding Site Evolution of Bacterial RNA Polymerase

The work presented in this chapter has been done in collaboration with Fabienne Jesse, Magdalena Steinrück, Jonathan P. Bollback, Calin C. Guet, Tiago Paixão, Gašper Tkačik and Nicholas H. Barton. The experimental data were produced and kindly shared by F. Jesse and M. Steinrück. The data are made available in the repository IST DataRep (DOI: `10.15479/AT:ISTA:43`).

3.1 Introduction

Regulation of gene expression is a heritable process contributing to overall phenotypic diversity (Wittkopp, 2013). Yet, little is known about the principles governing the evolutionary genetics of gene regulation. Bacterial transcriptional regulatory mechanisms and elements have been relatively well understood, particularly due to the studies on the *lac* operon in *E.coli* (Beckwith, 2011), and are amenable to an evolutionary investigation. Bacterial regulation of gene expression is due partly to sequence specific binding of RNA polymerase (RNAP), which has been shown to be an important genetic locus of adaptation (Blank et al., 2014). Better understanding of the physical and evolutionary characteristics of bacterial RNAP binding would contribute to the functional and evolutionary understanding of gene regulation.

Bacterial RNAP's core complex, consisting of five subunits, is sufficient to elongate the DNA, and to terminate transcription. However, it requires a sixth subunit, sigma factor, to form a holoenzyme that binds to DNA in a sequence-specific manner. After forming a closed complex with DNA, the RNAP holoenzyme isomerases double stranded DNA into single strands, forming an open complex to initiate transcription. Having transcribed approximately ten nucleotides, the RNAP holoenzyme disassociates the sigma factor, and the core RNAP continues elongation and transcription. Sigma70 is the principal (housekeeping) sigma factor in *E.coli*, and has been studied well (for reviews, see Gross et al. (1998); Paget and Helmann (2003); Murakami (2015)). The binding specificity of sigma70 dependent RNAP is primarily due to the two well conserved domains contacting two distinct hexamers on the DNA (Paget and Helmann, 2003). They are named, respectively, the -35 and -10 boxes referring to their bp positions (typically, from -36 to -31 and from -12 to -7) measured from transcription start site $(+1)$. The strongest binding (consensus) sequences of the -35 and -10 boxes are "TTGACA" and "TATAAT." The spacer length between two hexamers is known to be flexible in a short range of $15-21$ bp, with the most typical value of 17 bp, based on bioinformatic studies (Harley and Reynolds, 1987) and more elaborate experiments (Mulligan et al., 1985; Dombroski et al., 1996). Although the spacer length can impose an energetic cost on the RNAP binding, on which there is a limited direct study (Weindl et al., 2007), the spacer region is considered to have no specificity to DNA, i.e. no binding difference for different nucleotides. The exceptions are the -15 and -14 contacts where in some promoters a specificity to "TG" is observed, which now defines the extended -10 box. The general view is that binding deficiency at the -35 box can be compensated by the extended -10 box (Paget and Helmann, 2003), but there is also evidence exhibiting the complex interaction of the extended -10 and -35 boxes (Hook-Barnard et al., 2006). Overall, it is difficult to model the sequence specific binding of sigma70 dependent RNAP holoenzyme (hereafter simply referred to as RNAP) in every detail, and therefore simple mathematical approaches are favoured.

Thermodynamic equilibrium models have been widely used to understand sequence-specific protein-DNA interactions. With an additivity assumption, an energy matrix is generally used to represent the specificity, where the free energy contribution of each nucleotide across the binding interface is reported. Recently, Kinney et al. (2010) provided an energy matrix for RNAP, largely consistent with the known consensus sequences and with earlier models. Their experiment was based on high-throughput mutagenesis of the *E.coli lac* promoter sequence around the principal RNAP binding site (which is known to have a spacer length of 18 bp), and on followed-up measurements of gene expression. The specificity values were inferred by an information theoretic method combined with a thermodynamical model of transcription. However, it is not only the binding energy but also the chemical potential of a regulatory factor which determines the binding occupancy (probability) (Stormo and Zhao, 2010). Although Kinney et al. (2010) successfully provided the binding energy matrix model, their approach was uninformative about the chemical potential of the RNAP, which defines the threshold energy for their energy matrix. The absolute numbers of core RNAP ($2.600-13.000$ per cell) and sigma70 ($4.700-17.000$ per cell) have already been reported (Grigorova et al., 2006), but it is difficult to estimate the number of free RNAP determining the chemical potential.

Sequence-specific binding of RNAP is expected to leave selective signatures in genomes by shifting the sequence distribution from a neutral expectation. Hahn et al. (2003) have analysed the underrepresentation of the number of consensus sequences of the -10 and -35 boxes in 41 different eubacteria. They have inferred the average values for Ns across genomes per site as -0.09 and -0.15 for the -10 and -35 boxes, respectively, where N is the (effective) population size and s is the selection strength. They have concluded that a weak negative selection acts to reduce the number of RNAP bindings at off-targets, which is expected if spurious bindings are costly for the cell, e.g. by lowering the RNAP concentrations at target sites, and by producing non-functional transcripts. These results need to be inspected with different methods taking into account all sequence information, not only the consensus one. Because the consensus sequences are known to be stably bound by RNAP in some cases, limiting the successful formation of an open complex and transcription initiation (Knaus and Bujard, 1988). Mustonen and Lässig (2005) have developed such a method to infer selection for binding energies. It has been used for different transcription factors, mostly documenting positive selection (Mustonen and Lässig, 2005; Mustonen et al., 2008; Haldane et al., 2014), but not for RNAP so far, plausibly due to the difficulties in modelling of the binding energy of RNAP. By using an energetic approach, Weindl et al. (2007) have analysed the promoters of *E.coli* and showed a decrease of RNAP binding energies around the transcription start sites, yet they did not attempt to analyse them from an evolutionary perspective.

Multiple RNAP binding sites exist in some promoters (Huerta and Collado-Vides, 2003; Mendoza-Vargas et al., 2009). They lead to transcription of the same gene differing in the 5' noncoding sides of the mRNAs. This fact should be accounted for the understanding of the regulatory code, i.e. the mapping from a promoter sequence to the gene expression. This will likely help unravel the coevolution of a promoter region and the corresponding gene expression. So far studies have mostly focused on the relation between a specific site and gene expression. Early studies showed that mutations towards consensus-like sequences of a RNAP binding site increase gene expression (Hawley and McClure, 1983). Using the energy matrix of Kinney et al. (2010), Brewster et al. (2012) have synthesised several *lac* promoter constructs of varying RNAP binding strengths at the principal site, and have consistently shown a good correlation with their predictions from a thermodynamic model of transcription. Understanding the diversity in homologous promoter regions and the corresponding gene expression in natural organisms has proved to be a more difficult task. Recently, Razo-Mejia et al. (2014) have studied the variation in *lac* promoter sequence and LacZ protein (i.e. β-galactosidase) among bacterial isolates, but they did not find a convincing correlation with the predictions from a thermodynamic model of transcription. One reason might be that their model did not take into account the multiple RNAP binding sites within *lac* promoter, which is also widely ignored in textbook descriptions (Reznikoff, 1992). Understanding the multiple and flexible locations of RNAP binding for the regulatory code is also important to understand *de novo* promoter evolution, which is expected to primarily depend on RNAP binding evolution in an initially non-functional promoter sequence. There are theoretical studies on *de novo* promoter evolution (e.g. Tuğrul et al. (2015), i.e. Chapter 2 in this thesis). However, empirical knowledge about bacterial *de novo* promoter evolution is currently limited, mostly confined to genome rearrangement of the functional or semi-functional DNA with duplications and insertion elements (M. Steinrueck & C.C. Guet, *personal communication*), or studied in different contexts, e.g. evolution of expression noise (Wolf et al., 2015).

In this study, we aim to understand the binding sequence evolution of RNAP by analysing the data from an existing database and two collaborative projects in the light of the theoretical tools of biophysics and population genetics. We try to understand **i)** what selection signatures are inferred for regulatory and non-regulatory DNA sequences; **ii)** whether we can infer the chemical potential of RNAP from evolved sequences; **iii)** whether the multiple and alternative RNAP binding sites can explain the observed diversity of the *lac* promoter and LacZ protein activity among bacterial isolates; and **iv)** whether biophysics and population genetics theory can bring a predictive understanding of *de novo* promoter evolution in an experimental setup. We find that a moderate positive selection on regulatory regions, and a weak negative selection on non-regulatory regions (i.e. off-targets) have likely been acting on the genome of *E.coli* K12 throughout its evolutionary history. We inferred the chemical potential of RNAP corresponding to a binding energy of $\sim 4{\pm}1$ mismatches from the consensus sequences of the -35 and -10 boxes of RNAP. Moreover, we show that an alternative RNAP site is the main locus of *lac* promoter region causing the variation in the LacZ activity levels. Finally, we show that part of *de novo* evolution can be understood and predicted, at least better than by a neutral model, by considering biophysics and population genetics of RNAP binding.

3.2 Models & Methods

3.2.1 Thermodynamic modelling of RNA polymerase binding

We assume that a thermodynamical equilibrium model estimates the binding occupancy (probability) π_i at a particular position i in a sequence $\boldsymbol{\sigma}$ (Shea and Ackers, 1984; Berg and von Hippel, 1987; Bintu et al., 2005a,b), i.e.

$$\pi_i(\boldsymbol{\sigma}) = \pi_i(\mathrm{E_i}, \mu) = \left(1 + \mathrm{e}^{\beta(\mathrm{E_i} - \mu)}\right)^{-1}. \tag{3.1}$$

Here, E_i and μ are the binding energy and the chemical potential, respectively, and $\beta = (k_B T)^{-1}$. The chemical potential is related to the free concentration (Gerland et al., 2002; Weinert et al., 2014) and defines the threshold energy determining which sequences are strongly/weakly bound (Fig. 3.1C). Its value for RNAP is not known, therefore we consider it as the free parameter in our model, and check our results over a wide parameter range.

We assume that the DNA binding sequence determines the binding energy with an additive contribution model across the contact interface, i.e. an energy matrix model. Kinney et al. (2010) inferred the matrices of RNAP and CRP with physical energy units, and kindly provided them for our use[1]. To allow a more general binding energy model with a flexible spacer length $\boldsymbol{l}$ in a range of 15-21 bp (Mulligan et al., 1985; Harley and Reynolds, 1987; Dombroski et al., 1996), we only use the -35 and -10 boxes which carry the most of the sequence specificity (Fig. 3.1B). We denote these separate matrices as $\xi^{(-10)}$ and $\xi^{(-35)}$ where the element $\xi_{\sigma_j,j}$ gives the energetic contribution of the nucleotide σ_j appearing at the j-th position at the binding sequence. The average value of the non-consensus entries of the $\xi^{(-10)}$ and $\xi^{(-35)}$ is $\epsilon = 2.83\ k_B T$. In representation of our results, we show the binding energies in the scale of ϵ, in order to get an insight into the approximate number of mismatches from the consensus sequence. We will also consider a mismatch energy model (i.e. an energy matrix using a simple energetic value ϵ for each non-consensus entry) to evaluate the importance of the homogenous energy matrix elements.

We extend the RNAP binding energy model to include an energetic cost of spacer length $\mathrm{c}(\boldsymbol{l})$. Weindl et al. (2007) estimated a cost function for a spacer length range of $15-19$ bp, together with energy

[1] We downloaded the matrices from `https://github.com/jbkinney/09_sortseq` on December 6, 2015 and use the ones from the experiment labelled as ”full-wt”.

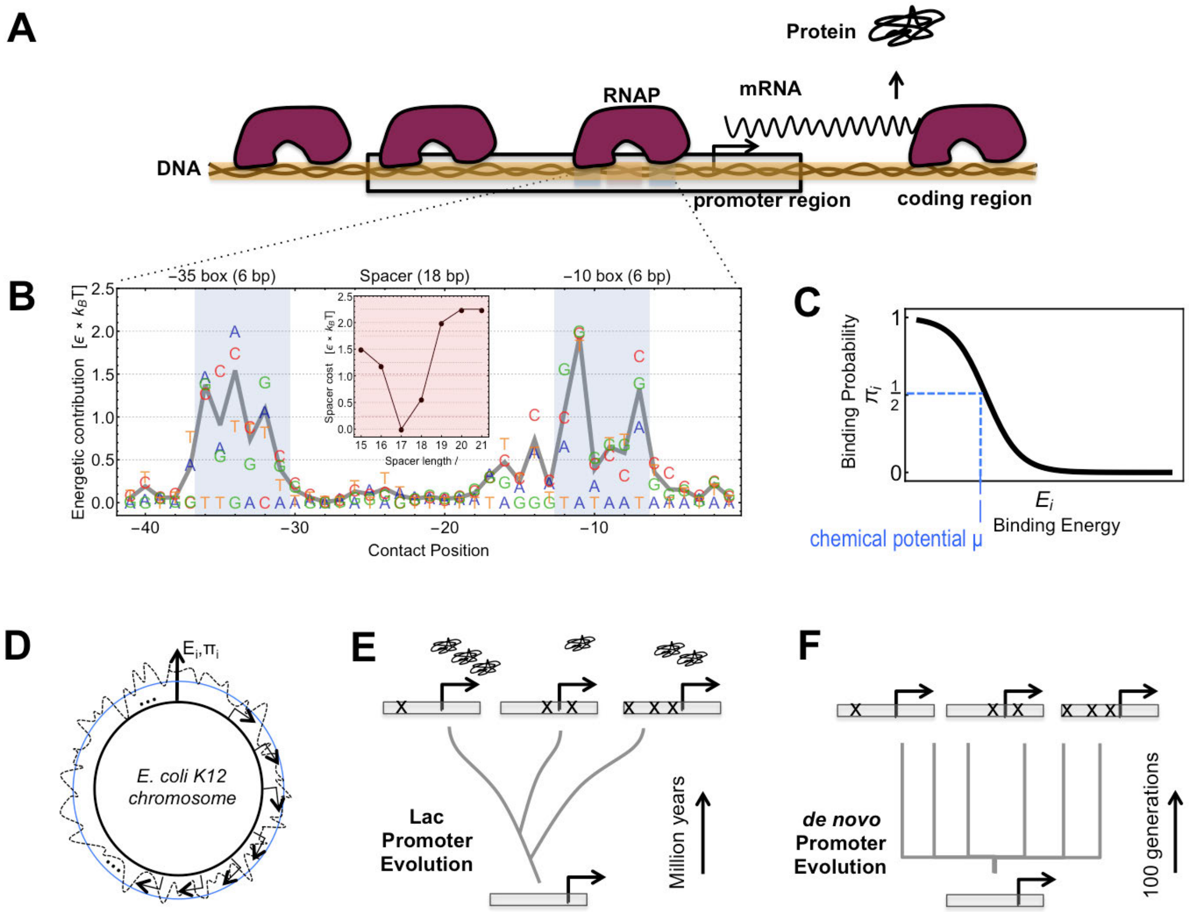


Figure 3.1: **Biophysics and evolution of bacterial RNA polymerase binding.** **A)** We study the binding sequence evolution of bacterial RNAP by using a biophysical model of sequence-specificity in a population genetics framework. We consider that multiple RNAP binding sites in a promoter sequence determine gene expression. **B)** We utilise Kinney et al. (2010)'s energy matrix for RNAP. The figure shows the energetic contribution of each nucleotide (A,C,G,T) at each contact position. Black curve shows the mean of the non-consensus entries at each contact position. **inset:** we consider and extended Weindl et al. (2007)'s inference for the energetic cost of the spacer length. **C)** We consider the thermodynamic model for the binding probability which depends on the binding energy and the chemical potential. In this study, we aim **D)** to infer selection on the binding energies of RNAP, and a realistic range of the chemical potential of RNAP from evolved sequences of the *E.coli* K12; **E)** to study how *lac* promoter evolution gives rise to lacZ protein activity variation among 20 bacterial isolates; **F)** to study *de novo* promoter evolution in *E.coli* K12 by focusing on a selection experiment for higher expression of an antibiotic resistance gene, whose original promoter sequence has been replaced by a random promoter sequence.

matrices of the -35 and -10 boxes with an arbitrary unit (au) from a microarray binding assay. Although their energy matrix shows discrepancy with the known consensus sequence, their estimate of the spacer energy cost, to our knowledge, is the only one in the literature, and therefore we utilise it here. We gleaned their estimate of the spacer cost function from their Figure 1. In order to convert it to a physical unit, we make use of the fact that both their cost function and their energy matrices take values in a range of ~ 0.57 au. By equating this to the physical energy range of Kinney et al. (2010)'s energy matrix, i.e. $\sim 2.00\,\epsilon$ (Fig. 3.1B), we obtained $c(l) = 1.51\epsilon,\ 1.19\epsilon,\ 0,\ 0.56\epsilon,\ 2.00\epsilon\ \mathrm{k_BT}$, respectively for $\boldsymbol{l} = 15,\ 16,\ ...,19$ bp. We also extended it to include 20 and 21 bps by consulting the related studies (Mulligan et al., 1985; Harley and Reynolds, 1987; Dombroski et al., 1996), and chose an *ad hoc* value of $c(l) = 2.25\epsilon$ for both $\boldsymbol{l} = 20$ and 21 (Fig. 3.1B). We checked and confirmed that our results do not change drastically with slight perturbations on the values of $c(l)$ we assigned. Although we do not expect a large deviation from our conclusions, future studies should aim at obtaining more accurate values of the energetic cost of the spacer length. Furthermore, we explicitly assume that the conformation of RNAP determines the spacer length such that the sum of the binding energy at the -35 box and the cost energy of the spacer length is minimised. Therefore, the RNAP binding energy E_i when the -10 box's rightmost (3') side contacts at the $(i-6)^{th}$ position in the sequence $\boldsymbol{\sigma}$ is expressed as

$$\mathrm{E_i}(\boldsymbol{\sigma}) = \sum_{\mathrm{j}=1}^{6} \xi^{(-10)}_{\sigma_{\mathrm{i}-12+\mathrm{j}},\mathrm{j}} + \mathrm{MIN}\Big\{\mathrm{c}(\boldsymbol{l}) + \sum_{\mathrm{j}=1}^{6} \xi^{(-35)}_{\sigma_{\mathrm{i}-18+\mathrm{j}+\boldsymbol{l}},\mathrm{j}}\Big\}_{\boldsymbol{l}=15,\dots,21} \tag{3.2}$$

Apart from checking our results with a simple mismatch model, we will also consider $c(\boldsymbol{l}) = 0$ to understand the importance of the spacer cost model.

3.2.2 Inferring selection

The distribution of different states of binding sequences under an equilibrium of the mutation-selection-genetic drift within a monomorphic population follows a Boltzmann-like distribution (Wright, 1931; Berg et al., 2004; Sella and Hirsh, 2005; Barton and Coe, 2009; Manhart et al., 2012). Mustonen and Lässig (2005) showed that this is also valid if the states of the binding sequences are taken to be continuous values of binding energy E as molecular phenotype i.e.

$$\psi(E) \propto \phi(E)\, e^{2Nf(E)} \tag{3.3}$$

where N is the population size, and $f(E)$ is the fitness of the binding sequences with the binding energy E. The $\phi(E)$ is the expected distribution under no selection, i.e. neutral evolution. This representation not only allows for a continuous biophysical (instead of a discrete bioinformatic) description of binding sequences, but also provides a method to infer selection (Mustonen and Lässig, 2005). Under the observed distributions of binding energies $\psi(E)$ and hypothesising a realistic $\phi(E)$, one can estimate the product of the population size and fitness up to a constant, i.e.

$$2Nf(E) = \log[\psi(E)/\phi(E)] - constant \tag{3.4}$$

In order to eliminate the constant term, we focus on the differences between the inferred fitness. We define the selection strength s as the difference between the maximum and minimum fitness $\Delta f(E)$, so that we can report Ns values for the inference of selection.

In this study, we obtain the $\phi(E)$, i.e. neutral distribution, by randomising the corresponding sequences so that the A,C,G,T content is kept constant on average. This serves as the simplest model for a neutral evolution, and future studies should check our results with more realistic models.

Note that this method to infer selection should be used with caution. We showed in our earlier work (Tuğrul et al., 2015) (i.e. Chapter 2 in this thesis) that the convergence to the stationary-state of binding sequences is typically slow, taking on the order of inverse point mutation rate. Here, we use this method to infer selection along the *E.coli* K12's evolutionary history over a larger time scales than the inverse point mutation rate. A modification to the method is needed if the divergence times are much less than the inverse point mutation rate, e.g. homologous loci among different strains or species, which we do not deal in this study.

3.2.3 A simple mapping from promoter sequence to gene expression

In this study, we want to understand how multiple binding sites in a promoter region can jointly affect gene expression. We construct a genotype-phenotype mapping from promoter sequence to gene expression by assuming that each RNAP binding in a promoter can transcribe a mRNA with equal weight, and that the transcription rate at each binding position is proportional to the binding probability at this position. Therefore, the dynamics of the amount of mRNA whose 5' end corresponds $(i+1)^{th}$ position in the promoter sequence can be expressed by a simple differential equation, i.e.,

$$\frac{d}{dt}m_i(\boldsymbol{\sigma}) = \Big(\gamma_\pi \, \pi_i(\boldsymbol{\sigma}) - \gamma_{deg} \, m_i(\boldsymbol{\sigma})\Big) \tag{3.5}$$

where γ_π and γ_{deg} represent the rate constants of binding and degradation. The steady state of these dynamics dictate a direct proportionality between m_i and π_i. Furthermore, if mRNAs with different length of the 5' end do not change protein expression dynamics, the sum of the binding probabilities across the promoter can be used as a proxy for the level of gene expression, i.e.

$$g(\boldsymbol{\sigma}) = \sum_i m_i(\boldsymbol{\sigma}) \; = \gamma_\pi/\gamma_{deg} \sum_i \pi_i(\boldsymbol{\sigma}). \tag{3.6}$$

We neglect the proportionality term (i.e. $\frac{\gamma_\pi}{\gamma_{deg}} = 1$), and use this gene expression model to understand the correlation between variation of promoter sequence and gene expression. We also check our results when a weight for each binding site position is introduced, i.e.

$$g(\boldsymbol{\sigma}) = \sum_i \alpha_i \, \pi_i(\boldsymbol{\sigma}). \tag{3.7}$$

For this study, we infer the positional weights α_i from the observed distributions of the functional RNAP binding sites (Huerta and Collado-Vides, 2003; Mendoza-Vargas et al., 2009). Most of the TSSs fall into a typical promoter length of 200 bps. The distribution of the distances between the TSSs and the first codon (AUG) sites exhibits a scattered and long-tailed exponential-like distribution with the optimal distance at 26 bp (see SI Fig. 3.12).

3.2.4 Probability of observing a particular point mutation under directional selection for higher gene expression

We also interpret our gene expression model in a relevant evolutionary experiment for *de novo evolution*. We consider a mutation-selection-drift population genetic model. In case of directional selection for higher gene expression, which will be an appropriate selection scheme for our experimental evolution set-up, the Malthusian fitness of a promoter sequence $\boldsymbol{\sigma}$ can be modelled as

$$f(\boldsymbol{\sigma}) = s\, g(\boldsymbol{\sigma}) \tag{3.8}$$

where s is the selection strength.

Transversion and transition mutation rates differ which needs to be accounted for calculating substitution rates. We consider a recent report of the spontaneous point mutation rates in *E.coli K12* by Lee et al. (2012), i.e. A:T>C:G $= 0.65$; A:T>G:C $= 0.80$; A:T>T:A $= 0.30$; C:G>A:T $= 0.51$; C:G>G:C $= 0.30$; C:G>T:A $= 1.37$ where the values are $\times 10^{-10}$ per generation, and we use the notations : and $>$ to refer the base pair and the direction of the mutation, respectively. Although mutations rates vary depending on conditions and organisms, we assume that the relative ratios, and the low mutation rate per population (see below) still hold in our experimental setup.

When the mutation rate per population is small and the selection is strong, one can realistically model the evolution of populations as successive sweeps of genotypes through a typically monomorphic population (Desai and Fisher, 2007). In the selection experiment we consider, the population size during the bottle-neck is expected to be in the range of $N = 10^5 - 10^8$ which suffices the low mutation assumption if we consider the point mutation rate in *E. coli* as $u \sim 10^{-10}$ per generation (Lee et al., 2012). We explicitly assume that typical time to sweep is shorter than the selection experiment, which is confirmed by observations in the experiment. The evolutionary dynamics in the monomorphic population model can be treated as a Markovian jump process. The substitution rate from a regulatory sequence $\boldsymbol{\sigma}$ to another regulatory sequence $\boldsymbol{\sigma}'$ in a haploid population can be expressed as

$$R_{\sigma',\sigma} = N\, U_{\sigma',\sigma}\, P_{\text{fix}}(N,\, \Delta f_{\sigma',\sigma}) \tag{3.9}$$

where $\Delta f_{\sigma',\sigma} = f(\boldsymbol{\sigma}') - f(\boldsymbol{\sigma})$ is the fitness difference and $U_{\sigma',\sigma}$ is the mutation rate from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma}'$. The fixation probability P_{fix} of a mutation with fitness difference Δf in a haploid population of N individuals is

$$P_{\text{fix}}(N,\, \Delta f) = \frac{1 - e^{-2\Delta f}}{1 - e^{-2N\Delta f}} \approx \frac{2\Delta f}{1 - e^{-2N\Delta f}}, \tag{3.10}$$

which is based on the diffusion approximation (Kimura, 1962). Note that $N\,P_{\text{fix}}(N,\,\Delta f)$ approximates to $2N\Delta f$ when $N\Delta f \gg 1$ which is valid in our selection experiment.

In our modelling framework, we consider single point mutations in the evolution of the promoter sequence. Using the above substitution rates for all possible point mutation rates, one can simply express the probability of observing a particular point mutation as a function of the relative selection strength Ns and the chemical potential μ, i.e.

$$P_{\sigma',\sigma}(Ns,\mu) = \frac{R_{\sigma',\sigma}}{\sum_{\sigma''} R_{\sigma'',\sigma}}. \tag{3.11}$$

Note that the denominator takes into account all possible point mutations in the promoter sequence. Using this probability of observing a particular point mutation, one can set a statistical framework to predict the outcome of any promoter evolution under directional selection. For example, one can calculate the likelihood of a data set S of the observed single point mutations, i.e.

$$\mathcal{L}(Ns,\mu) = \prod_{\sigma'\in \boldsymbol{S}} P_{\sigma',\sigma}(Ns,\mu). \tag{3.12}$$

This can be also used to infer the evolutionary and biophysical parameters.

3.3 Data

3.3.1 Dataset A: Whole genome and experimentally verified RNAP binding sites of *E.coli* K12

First, we consider the existing sequence data of *E.coli* K12, in order to infer the selective signatures and the chemical potential of RNAP. The whole genome of *E.coli* K12 (MG1655) has been already sequenced (Blattner et al., 1997) and consists of $4.641.652$ bps[2]. Therefore, there are around 9.3 million potential binding sites in the forward and backward directions. The regulatory information for the sigma70 factor dependent RNAP binding was obtained from the RegulonDB database[3] (Salgado et al., 2013). In our study, we consider the data of 788 transcription start sites (TSSs) that have been experimentally verified, in order to avoid any artificial bias in bioinformatic inferences. The distance between the TSS and the position of the -10 box can vary from 4 to 12 bp (the most common value is 7 bp). To obtain the exact location of the RNAP binding sites, we search the sequences in this range and assume that the sequence with the minimum (i.e. the strongest) binding energy corresponds to the experimentally verified binding site. Confirming multiple binding sites in promoters (Huerta and Collado-Vides,

[2] The genome of *E.coli* K12 (MG1655 U00096 .3) was downloaded from Ecogene.org `http://www.ecogene.org` on November 27, 2015

[3] Downloaded from `http://regulondb.ccg.unam.mx` on March 5, 2015

2003; Mendoza-Vargas et al., 2009), these 788 binding sites correspond to 735 different promoter regions, which we arbitrarily define as the 200 bps upstream of the first codon (ATG) following each TSS. We also arbitrarily define the non-regulatory regions as the 500 bps upstream and 200 bps downstream from the first codon following the experimentally verified binding sites. Let us note that other TSSs that are not known are likely placed in the non-regulatory regions in our definition, which can cause the inferences of selection to be underestimated in our analysis.

In short, we explicitly consider four different ensembles of binding sequences in the *E.coli* genome: **i)** 788 experimentally verified binding sites, **ii)** all possible binding sequences in the 735 promoter regions, which we define as the 200 bps upstream from the first codon sites following the experimentally verified binding sites, **iii)** all possible binding sequences in the whole genome except the regulatory regions, which we define as the 500 bps upstream and 200 bps downstream from the first codon sites following the experimentally verified binding sites, **iv)** all possible binding sequences in the whole genome.

3.3.2 Dataset B: *lac* promoter and LacZ activity diversity in 20 bacterial isolates

Secondly, we consider the genetic diversity of *lac* promoters where RNAP binding acts to determine gene expression, and the phenotypic diversity of LacZ activity among 20 natural bacterial isolates from Escherichia genus. The data used here were curated or experimentally generated by Fabienne Jesse[4] (to appear also in F. Jesse and J. P. Bollback). She kindly shared her data, which we made available in a repository[5].

Apart from using K12 for *E.Coli* K12 (MG1655), we use the existing nomenclature for the short names as appeared in the references where the information regarding the strain, species, location of isolation, and ecological origin of these organisms can also be found: M1, M2, M3, M6, M7 (Cravioto et al., 1990), G8, A5, F10, G10, H3 (Ishii et al., 2006), TW10509, TW15838, TW09276, TW09231, TW11588, TW14182,TW15844, TW09308 (Walk et al., 2009). Importantly, some of these species/strains have diverged over millions of years (Walk et al., 2009; Luo et al., 2011), and differ in the environments of their isolation suggesting the difference in the lactose availability as carbon source. Therefore, one can expect that selection for lactose metabolism, or genetic drift in the absence of selection may have caused genetic and phenotypic differences among these organisms, which we study here by focusing on the *lac* promoter sequences and LacZ activity.

Lac operons of these organisms were isolated and inserted into a plasmid in *E.coli* K12 with a deletion of its natural *lac* operon in the chromosome serving as a common genetic background. A Miller assay method, as a quantitative measure of lactose metabolism, was performed for LacZ activity as a proxy for LacZ gene expression (Dodd et al., 2001). 0 mM and 1.00 mM of the inducer IPTG were used,

[4] PhD student in Jon Bollback's group at IST Austria

[5] in the repository IST DataRep, DOI: `10.15479/AT:ISTA:43`

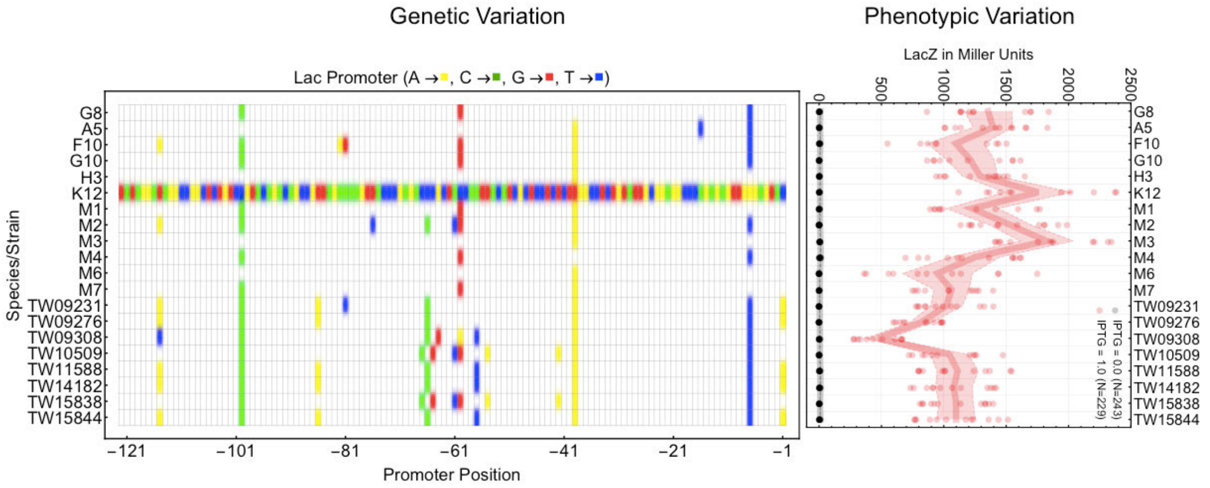


Figure 3.2: **Genetic variation of *lac* promoters and phenotypic variation of LacZ protein activity** **Left:** Nucleotides are shown with a color code (A:yellow, C:green, G:red, T:blue) for *E.coli* K12. For other strains/species, we only show the differentiating nucleotides. **Rigth:** The correlation between the *lac* promoter variation and lacZ activity variation, assayed for 20 strains/species at 0 mM (black) and 1.00 mM (red) of the inducer IPTG levels. Dots show each LacZ activity assay (243 and 229 data points for 0 mM and 1.00 mM, respectively); curves and shades represent mean and 2 SEM over essays.

where the latter is expected to fully suppress the *lac* repressor protein, and make the *lac* promoter a functional regulatory element for RNAP binding. Fig. 3.2 shows the data of the promoter sequences and LacZ activity which show variation among these 20 bacterial isolates. We will consider the average LacZ activity level as a molecular phenotype, ranging from $400 - 2000$ Miller units. Mutations in the coding regions exist (not shown) and might constitute an additional source of variation which needs to be investigated in further studies. Here, we focus on to investigate the sources of the LacZ activity variation due to RNAP binding in lac promoter regions, i.e. the 122 bp between the coding regions of the *lac* repressor protein and the LacZ protein.

3.3.3 Dataset C: Adaptive and random point mutations in an initially non-functional promoter in *E.coli* K12

Lastly, we consider RNAP binding in an initially non-functional promoter region. We consider both evolution of the promoter under a selection experiment, and synthesis of promoter constructs by mutagenising the promoter. Magdalena Steinrück[6] designed and performed all experiments, and kindly shared the data. Evolutionary data are a small subset of her main study (to appear in M. Steinrück and C. C. Guet); the synthetic data were produced for this study. They are made available in a data repository[7].

[6] PhD student in Calin Guet's group at IST Austria

[7] in the repository IST DataRep, DOI: 10.15479/AT:ISTA:43

A 194 bp sequence was randomly generated and followed by a 14 bp sequence including a strong ribosomal binding site[8]. This complete 208 bp sequence is referred here as (*p0*), and serves as a non-functional promoter region that drives a basal level of expression in the *E. coli* K12 chromosome. *p0* is followed by a tetracycline resistance gene (*tetA*). The *p0-tetA* construct is translationally fused to a YFP marker gene for fluorescent measurements. Both adaptive and random point mutations on the promoter sequence *p0* were obtained, referred to as evolutionary and synthetic datasets, respectively.

For evolutionary data, several replicate populations of the construct were subjected to a selection for approximately 100 generations with exponentially increasing antibiotic tetracycline. The antibiotic level was adjusted such that it starts at a subinhibitory concentration and ends at levels above the minimum inhibitory concentration so that the ancestral genotype diminishes in ratio and practically goes extinct, and only the adapted mutations remain. This assumption of the evolutionary rescue by increasing expression of the tetA-YFP was checked and confirmed. From surviving populations, 1 kb of the *tetA*-upstream region was sequenced from single clones. Among other mutations (mobile element insertions, deletions, amplifications), point mutations (single substitutions), both inside and further upstream of *p0*, were repeatedly observed. For this study, we consider 14 independent replicates adapted with single point mutations within *p0*, i.e. -24 T>A (5), -24 T>C (1), -31 C>T (2), -92 G>T (3), -149 T>A (2), -183 C>T (1), where the first number shows the position of the mutation with respect to the translation start site, and the numbers in the parentheses show the number of independent observations.

For synthetic data, 76 mutants with different single point mutations, which largely covers the point mutational neighbourhood of the *p0* between 20 bp and 80 bp *tetA*-upstream, were generated. The mutant library was obtained by a site directed mutagenesis method which uses a pool of primers having a single and random nucleotide change across the target region. As a proxy for transcription of the tetA-YFP gene fusion, the YFP fluorescence of a photographical image of spotted mutant cultures was used (Chait et al., 2010).

3.4 Results

3.4.1 Positive selection on regulatory sites and negative selection on non-regulatory sites (off-targets)

We first studied the inference for the selection of RNAP binding in *E.coli* K12. In an earlier study, Hahn et al. (2003) have shown the underrepresentation of the consensus sequences of RNAP in eubacteria, and inferred a weak negative selection with $Ns \sim -0.1$ to eliminate the consensus sequences (N is the population size and s is the selection strength). Following this study, here we want to obtain the selective

[8] i.e. AGGAGGAATTCACC where the first 6 bp corresponds to the strong ribosomal binding site

signatures more quantitatively on all binding sequences (not only on the consensus sequences) in both regulatory and non-regulatory DNA regions using Dataset A (the Data section 3.3.1). As explained in the Models & Methods, we represent the binding sequences with binding energies calculated by using the models of the energy matrix (Kinney et al., 2010) and the energetic cost of the spacer length (Weindl et al., 2007). In a similar energetic method, Weindl et al. (2007) have shown that the binding energies steadily decrease around transcription start sites (TSSs), which we confirmed to be also valid around translation start sites (see SI Fig. 3.12), suggesting that selection is acting on the regulatory regions. To detect the selective signatures quantitatively, we follow the inference method proposed by Mustonen and Lässig (2005) (see the Models & Methods).

We calculated the binding energy distributions of real and randomised sequences for different categories of binding regions of the *E.coli* K12 genome (Fig. 3.3A). Randomised sequences serve as the model for neutral evolution in our study, and using Eq. (3.4), we computed the inferred selection (fitness) profiles with the binding energy (Fig. 3.3B). Recall that we define the selection strength s as the maximum depth of the fitness landscape, i.e. the difference between the maximum and minimum fitness. First of all, the experimentally verified binding sequences, as expected, exhibit a positive selection towards sequences with lower binding energies (i.e. stronger binding) with a selection strength $Ns \sim 3$. This selection strength and the profile with binding energy are similar to the estimates for several transcription factors (Mustonen and Lässig, 2005; Mustonen et al., 2008; Haldane et al., 2014). Secondly, for the promoter regions (i.e. 200 bp upstream from the first codon sites), we observe a weak selection towards the binding energies corresponding to the values around one mismatch from the consensus sequence. Importantly, there is weak selection to eliminate the consensus sequence in the promoter regions. Lastly, for both the whole genome and the non-regulatory regions (i.e. whole genome except the regulatory regions), we observe a weak negative selection $Ns \sim -0.5$ to eliminate the sequences with the lower binding energies.

Removing the spacer length cost (i.e. $c(\boldsymbol{l}) = 0$), as expected, slightly shifts and squeeze the binding energy distributions (roughly $(0.5 - 1.0)\,\epsilon$) towards lower energy values (not shown). Nevertheless, this did not change the characteristics of the inferred selection, which is also the case when a mismatch assumption (i.e. all non-zero elements in the energy matrix is replaced with the average energy ϵ) is used for binding energy calculations (not shown).

3.4.2 The chemical potential of RNA polymerase likely corresponds to an average energy of ~ 4 mismatches from the consensus sequence

Above we inferred selection on the binding energy of RNAP, as a biophysically more relevant measure of a binding sequence. However, it is not only the binding energy but also the chemical potential which determines the binding probability of RNAP, and as consequence, gene expression (see Eq. (3.1) and

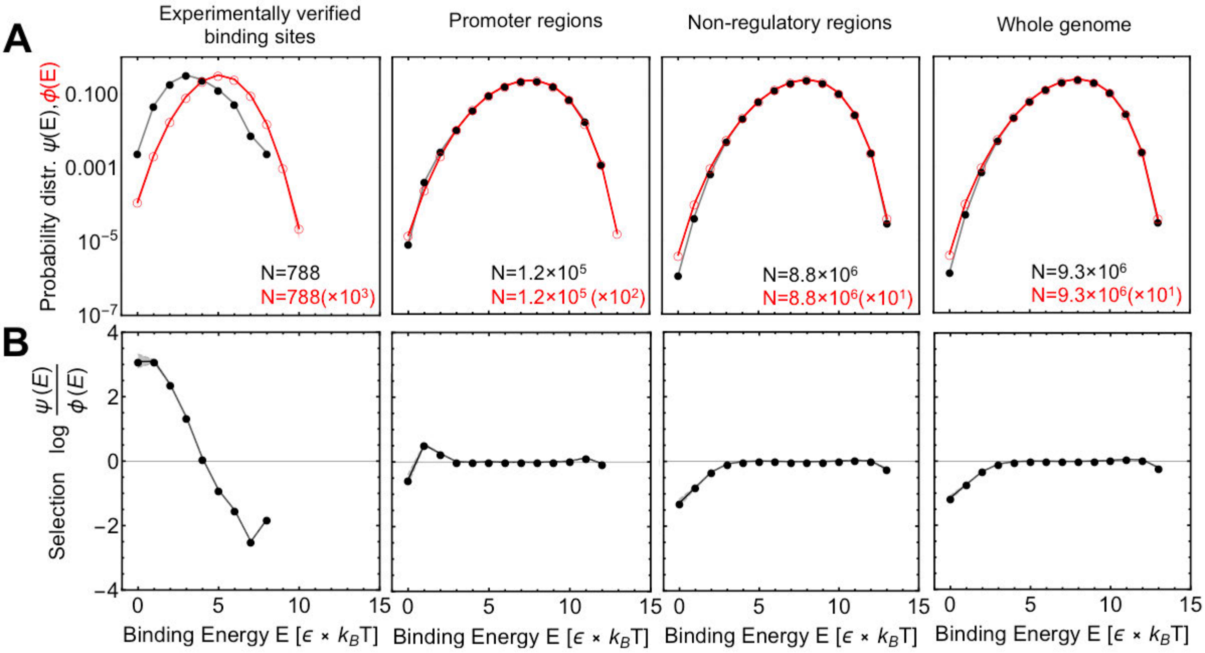


Figure 3.3: **Distribution of the binding energies, and the inference of selection for RNA polymerase binding in E.coli** **A:** Binding energy distributions of evolved (black) and randomised (red) sequences are shown for four different categories of the binding sequences in *E.coli* K12 (see the figure headings). The bin size of ϵ (i.e. the average of the non-zero elements in the -35 and -10 boxes of the RNAP energy matrix) is used to estimate the approximate number of mismatches from the consensus sequence. Red curves also include 2 SEM which is smaller than the marker size. The number of binding sequences in each category is indicated in the right-below corner of the figure where the number of replica for generating random sequences is given in the parentheses. **B:** Using Eq. 3.4, the selection is inferred. Black curves also include 2 SEM, indicating that the deviations from 0 is significant.

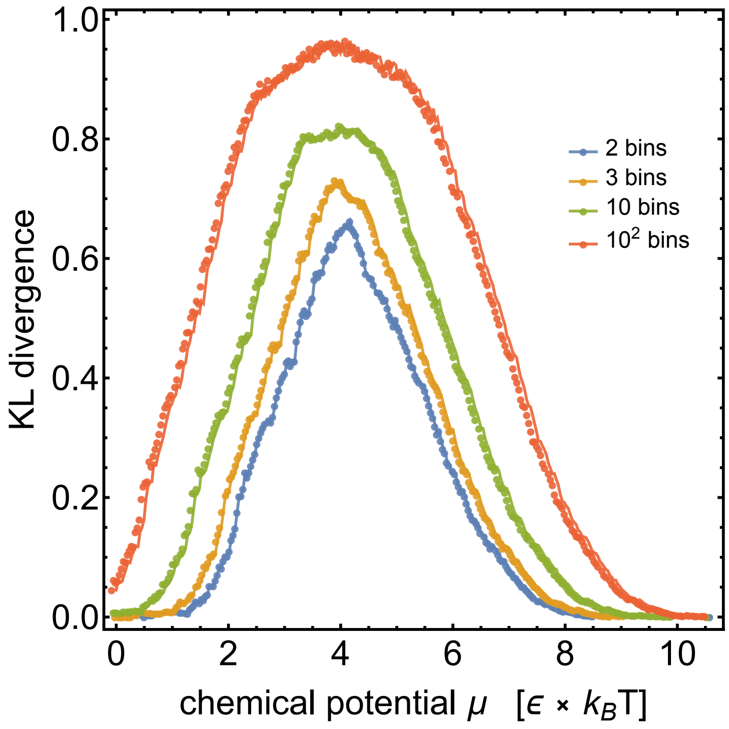


Figure 3.4: **Inferred the Chemical Potential of RNA polymerase in *E.coli*.** To calculate the difference between the distributions of the binding probability for the evolved and random sequences of the experimentally verified RNAP binding sites, we consider the Kullback-Leibler divergence $D_{KL}[\boldsymbol{P_{evolved}}(\pi)\,||\,\boldsymbol{P_{random}}(\pi)]$. The probability distributions are obtained by using different number of bins for the binding probability π. Different bin sizes consistently indicate $\mu \sim 4\,\epsilon\; k_B T$ as maximising the KL divergence.

Fig. 3.1). Yet, the chemical potential μ, i.e. the energetic threshold which defines the strongly and weakly binding sequences, have not been provided for RNAP (Kinney et al., 2010). Here, we estimate this important biophysical parameter from the evolved sequences by considering the experimentally verified RNAP binding sites in *E.coli* K12 using Dataset A (the Data section 3.3.1). For that, we mapped the binding energies of the evolved and neutral (randomised) sequences onto the binding probability values using Eq. (3.1), and searched for the parameter values of μ that best differentiate the evolved and neutral distributions. Fig. 3.4 shows that $\mu \sim 4\,\epsilon\; k_B T$ maximises the Kullback-Leibler (KL) divergence of these two distributions, and suggests that $(4 \pm 1)\epsilon\; k_B T$ can be a realistic range of the chemical potential of RNAP.

Consistent to the earlier statement, we observe that the optimal parameter of the chemical potential maximising the KL divergence shifts slightly ($\mu \sim 3.5\,\epsilon\; k_B T$) when we remove the spacer length cost, i.e. using $c(\boldsymbol{l}) = 0$ (not shown). Surprisingly, adding also a mismatch assumption for binding energy calculation brings back the optimality to $\mu \sim 4\,\epsilon$, although the smoothness of KL divergence scaling with the parameter is lost (not shown).

3.4.3 An alternative binding site in *lac* promoter is responsible for LacZ activity variation

Multiple RNAP binding sites can be functional in a single promoter region. Genotype-phenotype mapping from promoter sequence to gene expression should include this largely ignored fact. In the Models & Methods, we constructed the simplest such model for gene expression which is the sum of the RNAP binding probabilities across a promoter region (Eq. (3.6)). To test this model, we analyse the Dataset B (the Data section 3.3.2) regarding the *lac* promoter sequence and the LacZ protein activity variation among 20 bacterial isolates (Fig. 3.2). We apply our gene expression model to the *lac* promoter sequences across a wide range of the chemical potential parameter μ, and search for a correlation with the average LacZ activity. Fig. 3.5 shows that the correlation is significantly high (Pearson test: $R^2 > 0.5, p < 0.01$; SpearmanRank test $\rho > 0.55, p < 0.01$) in the chemical potential range of $\sim (3 - 4.5)\epsilon\ \mathrm{k_BT}$, consistent with our earlier inference. These results suggests that at least 50% of the variation in the LacZ activity can be explained by sequence-specific RNAP binding. The same figure shows an example plot of the model prediction versus the LacZ activity levels for the optimal value of $\sim 4.3\epsilon\ \mathrm{k_BT}$ (see also SI Fig. 3.13 for supplementary plots for different μ values). Our model is based on a linear relation between RNAP occupancy on promoter and expression. Therefore, Pearson test, a linear correlation test, can be considered more appropriate than Spearman test, a rank ordering test, which misses the optimum in this case.

We further ask which RNAP binding positions influence the variation in our gene expression model. The variance of binding probabilities at each promoter position across 20 bacterial isolates was calculated for different values of μ (Fig. 3.6). The 39th bp upstream promoter position corresponding to the principal (canonical) RNAP binding site does not show any binding difference among the isolates. Surprisingly, we see that a single position, the 54th bp upstream promoter position, is responsible for almost all binding variance. Consistently, this position corresponds to one of the three alternative RNAP binding sites which have been experimentally verified (Xiong et al., 1991; Reznikoff, 1992).

We checked the effect of some of our modelling choices. First of all, we found that a null spacer cost (i.e. $c(\boldsymbol{l}) = 0$) reduces the correlation (the figure not shown; Pearson test: $R^2 < 0.5, p > 0.02$; SpearmanRank test $\rho < 0.43, p > 0.05$), and does not exhibit a clear optimal range of μ. This can be expected since the multiple binding sites in the *lac* promoter are known to have different spacer lengths, and a spacer cost energy is needed to assign the binding sites location. We observed a similar effect of losing the optimality range of μ and reducing the correlation (figure not shown; Pearson test: $R^2 < 0.5, p > 0.024$; SpearmanRank test $\rho < 0.34, p > 0.1$) when we use a mismatch assumption for binding energy, suggesting the importance of the inhomogeneity of the elements in the RNAP energy matrix. Furthermore, we checked the effect of the CRP which binds to DNA and activates the RNAP binding.

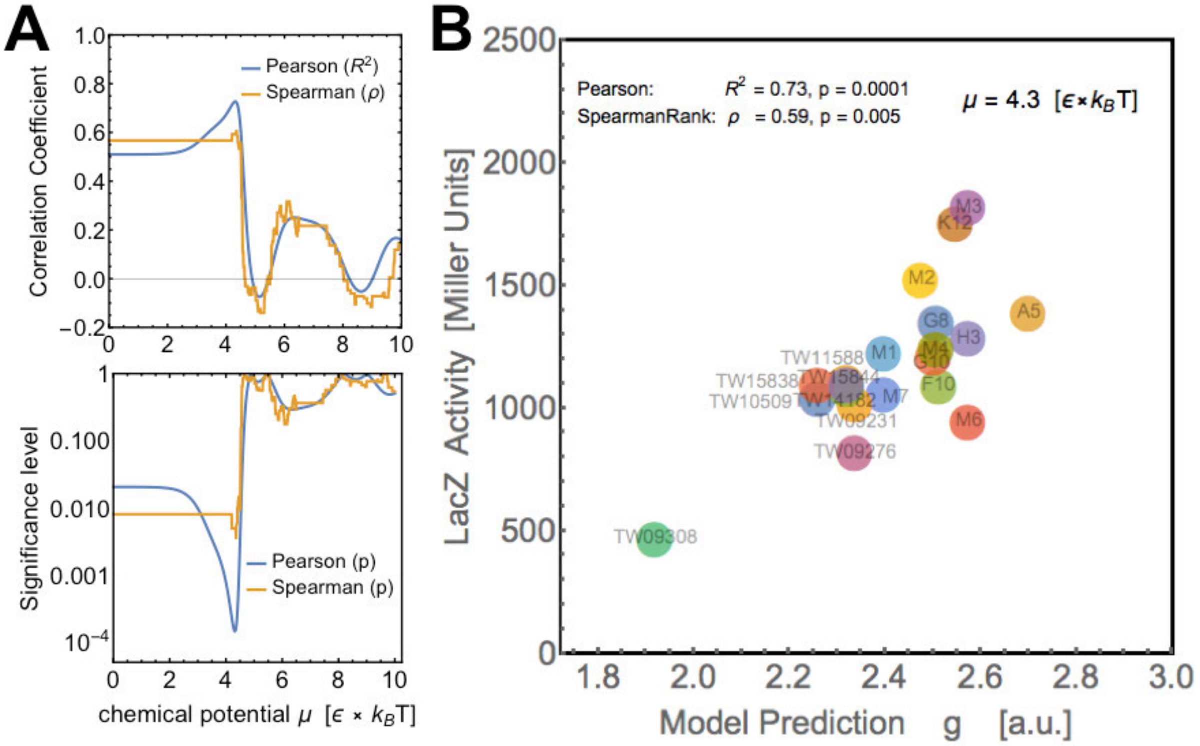


Figure 3.5: **Model prediction versus observed LacZ activity.** **Right:** The chemical potential is the free parameter in our gene expression model. We show how the Pearson and SpearmanRank correlation statistics scale with this parameter (**top:** the correlation coefficients, **bottom:** the significance level). **Left:** An example plot of the model prediction versus the observed lacZ activity for the optimal value of the chemical potential parameter.

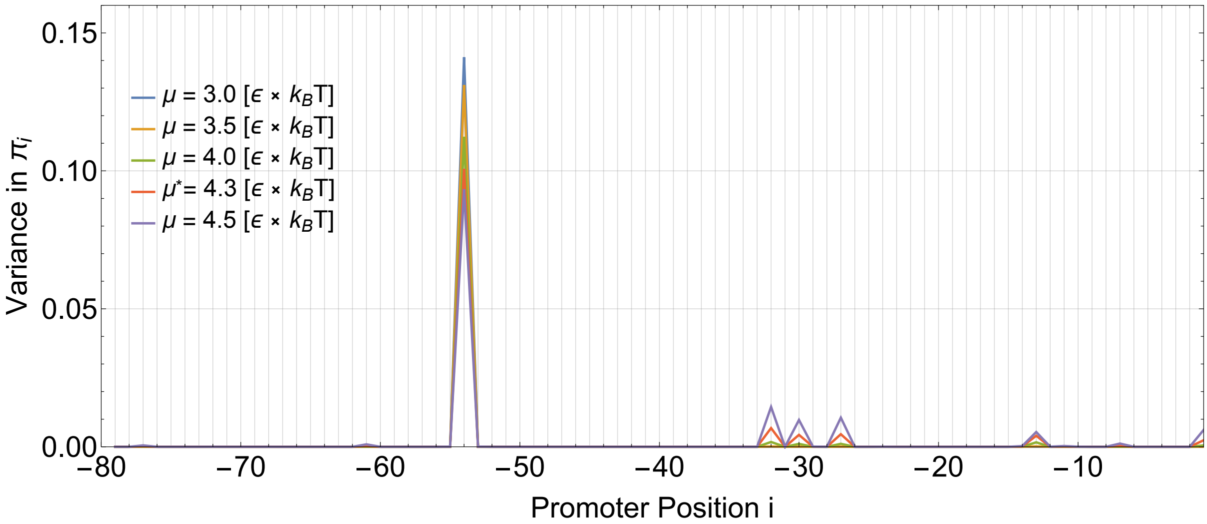


Figure 3.6: **Variance of the RNAP binding probability at each promoter binding position** Our gene expression model is the sum of the binding probability across the *lac* promoter and shows good correlation with the LacZ protein activity level variation. Therefore, the variance of the binding probability could point to the genetic source of the variation. The figure shows that the most of the variance is due to the binding at the 54th upstream position, which is a known alternative binding site that has been experimentally verified. The 39th bp upstream promoter position, which corresponds to the principal (canonical) RNAP binding site, does not show any binding difference.

This cooperation can be captured in a modified thermodynamic model for RNAP binding probability, i.e.

$$\pi_i(\mathrm{E_i}, \mu, \mathrm{E_{crp}}, \mu_{\mathrm{crp}}, \mathrm{E_{coop}}) = \frac{\mathrm{e}^{-\beta(\mathrm{E_i}-\mu)} + \mathrm{e}^{-\beta(\mathrm{E_i}+\mathrm{E_{crp}}-\mu-\mu_{\mathrm{crp}}-\mathrm{E_{coop}})}}{1 + \mathrm{e}^{-\beta(\mathrm{E_i}-\mu)} + \mathrm{e}^{-\beta(\mathrm{E_{crp}}-\mu_{\mathrm{crp}})} + \mathrm{e}^{-\beta(\mathrm{E_i}+\mathrm{E_{crp}}-\mu-\mu_{\mathrm{crp}}-\mathrm{E_{coop}})}}, \tag{3.13}$$

where we utilised Kinney et al. (2010)'s inferences of the CRP energy matrix[9] (to calculate the CRP binding energy E_{crp}), the chemical potential of CRP[10] μ_{crp} and the cooperativity[11] with RNAP E_{coop}. The results (not shown) do not exhibit a visible statistical difference in the correlations. Close inspection shows that the genetic alterations in the *lac* promoters of these bacterial isolates do not significantly affect the CRP binding affinities (i.e. $e^{-\beta(E_{crp}-\mu_{crp})}$), and therefore do not influence the RNAP binding. Furthermore, we also checked the effect of post-transcriptional regulatory mechanisms. First, we used the Salis ribosomal binding site calculator[12] (Salis et al., 2009; Salis, 2011) to estimate the translation rate differences due to the variation in the promoter sequences. The results (not shown) do not yield a correlation between the estimated translation rate and the observed LacZ activity levels (Pearson test:

[9] We downloaded the CRP matrix from `https://github.com/jbkinney/09_sortseq` on December 6, 2015 and use the one from the experiment labelled as "full-wt". As described in Kinney et al. (2010), we added an energy shift of $-11.18\ k_BT$ for each binding energy.

[10] As given in Kinney et al. (2010), we use $\mu_{crp} = -2.76\ k_BT$

[11] As given in Kinney et al. (2010), we use $E_{coop} = -5.28\ k_BT$. We use the interaction distance window as $6-12$ bp and consider the smallest binding energy

[12] `https://salislab.net/`

$R^2 = 0.01,\ p = 1.0$; SpearmanRank test $\rho = 0.08,\ p = 0.7$). Secondly, we implemented an *ad hoc* weight for the contribution of each RNAP binding position to include the transcription or translation efficiency (see the Models & Methods, Eq. (3.7)). However, it did not change the results (not shown), as expected from the existence of a dominant RNAP binding at a single position determining the variation.

3.4.4 Can we understand and predict *de novo* promoter evolution?

Modelling multiple RNAP binding sites in a promoter can be also important to understand *de novo* promoter evolution, i.e. emergence of a transcribing RNAP binding in an initially non-functional promoter. In Models & Methods section, we developed a gene expression model (Eq. (3.6), the one used above), and a statistical framework to predict the adaptive point mutations (Eq. (3.11)). Our framework is based on a biophysical model of sequence-specific binding of RNAP with a free parameter for the chemical potential μ, and on a population genetic model with a free parameter for the selection strength Ns. To test this approach, we used Dataset C (the Data section 3.3.3) from an evolution experiment where a higher expression of an antibiotic resistance gene with an initially non-functional (random) promoter is selected for approximately 100 generations in independent replicates. 14 replicates resulted in single adaptive point mutations, comprising the data set $\boldsymbol{S}$[13]. We first calculated the likelihood of $\boldsymbol{S}$ according to our model (Eq. (3.12)) in a wide parameter range of Ns and μ, and compared it with a neutral model ($Ns = 0$) (Fig. 3.7). As expected, strong selection (i.e. $Ns >> 1$) values are needed to explain the observed data set. However, this is not sufficient, the chemical potential parameter range of $\mu \sim (3.0 - 4.5)\,\epsilon\ k_BT$ is also needed for high likelihood ratios ($\sim 10^7$). This μ range is consistent with our earlier inferences above. The optimal parameter values are $Ns = 66$ and $\mu = 3.8\,\epsilon\ k_BT$, having a likelihood ratio of $\sim 10^{8.46}$, i.e. explaining the observed dataset almost one billion times better than the neutral evolution model. This high likelihood ratio in the optimal range of the chemical potential parameter is due to the model prediction of gene expression increase for all the observed point mutations (SI Fig. 3.14). Although the model predicts the observed set of adaptive point mutations better than the neutral model, the Pearson chi-square test suggests that the observed data set is unlikely to be drawn from the expected distribution produced by the model (Eq. (3.11)) at any parameter value ($p < 0.01$; figure not shown). To understand the reasons, we separate the population genetic model, and continue to evaluate the power of the biophysically motivated genotype-phenotype mapping model in predicting the levels of gene expression.

In order to collect synthetic data for *de novo* promoter activity, 76 mutant promoters with single point mutations were generated, and their YFP florescence were measured as a proxy for the level of gene expression (Dataset C in the Data section 3.3.3)). We checked the correlation statistics of our gene

[13] $\boldsymbol{S} =$ { -24 T>A (5), -24 T>C (1), -31 C>T (2), -92 G>T (3), -149 T>A (2), -183 C>T (1) } , where the first number shows the position of the mutation with respect to the translation start site, and the numbers in the parentheses show the number of such mutations observed in the independent replicas.

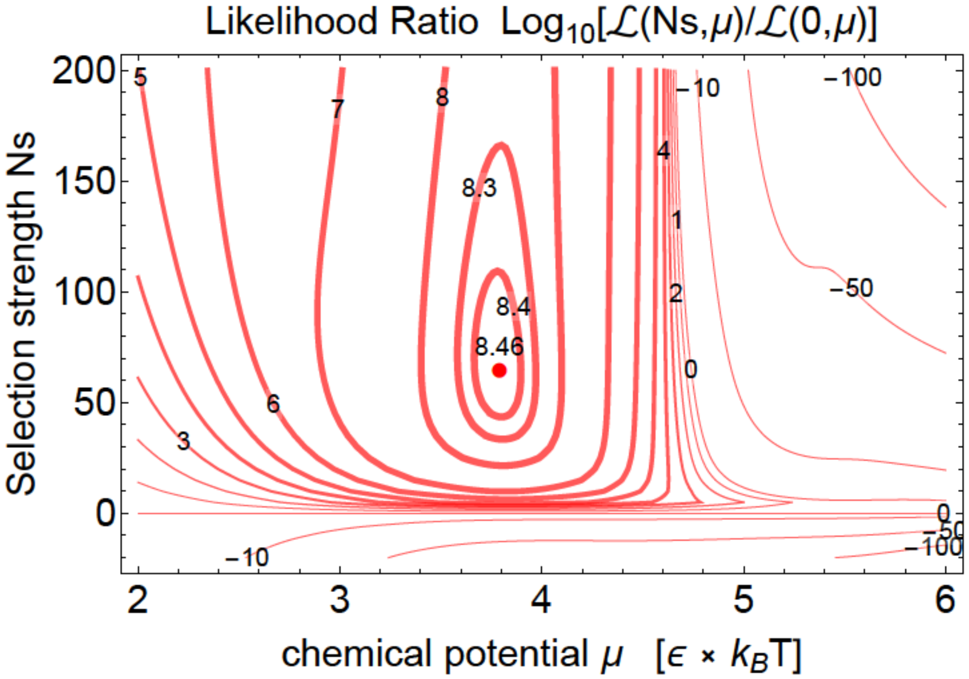


Figure 3.7: **Likelihood ratio of the evolutionary data set.** Using Eq. (3.11) and Eq. (3.12), we calculated the likelihood of the 14 adaptive point mutations at different selection strength Ns and chemical potential μ. We give the likelihoods in a ratio to the neutral evolution case ($Ns = 0$), and in a log scale. Some contour values are indicated.

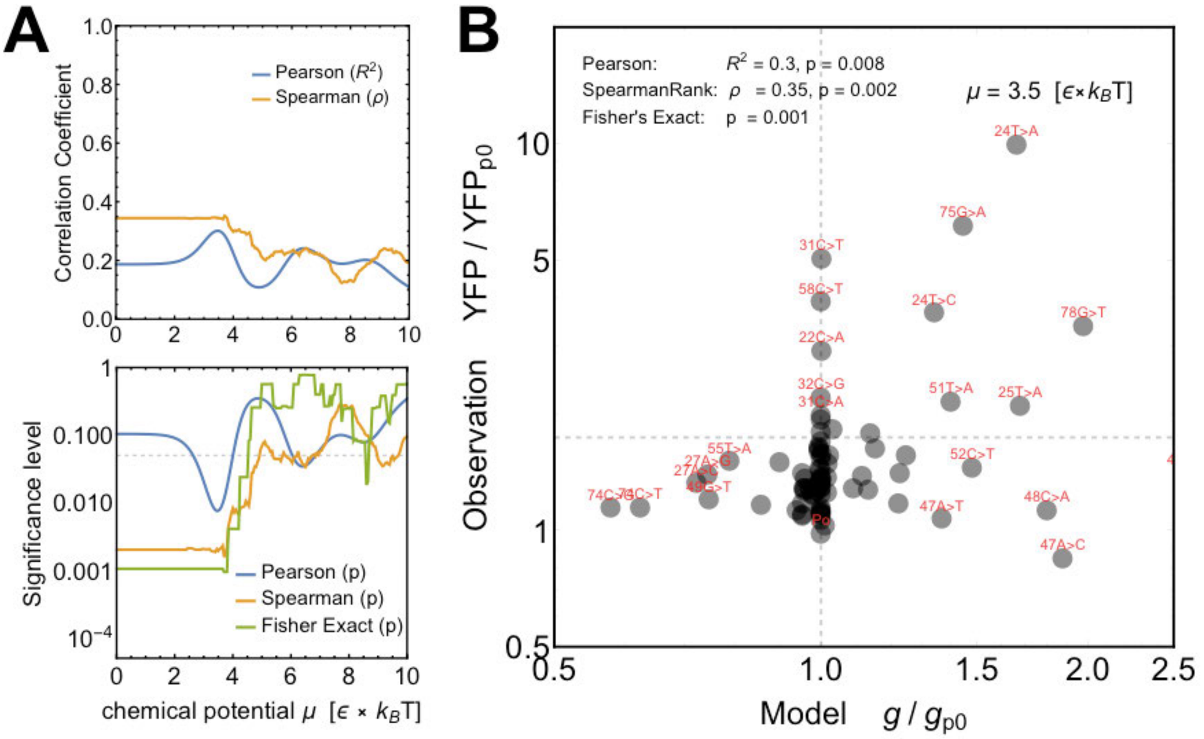


Figure 3.8: **Comparison of our main model's predictions and the YFP measurements.** **A)** We show how the Pearson, SpearmanRank and Fisher's Exact test statistics scale with the chemical potential μ (**top:** the correlation coefficients, **bottom:** the significance level (p-value)). **B)** We show an example plot of the model prediction versus the observed YFP level for $\mu = 3.5\,\epsilon\, k_B T$. For convenience, the model and the observations are scaled so that the values of the promoter sequence *p0* are 1. Those mutations with clear deviations from *p0*'s are labeled above the data points.

expression model (Eq. (3.6) versus the measured YFPs over a wide range of the chemical potential μ (Fig. 3.8). A weak but statistically significant correlation is observed in the parameter range of $\mu \sim (3-4)\,\epsilon\, k_B T$ (Pearson Correlation test: $R^2 \sim 0.2-0.3$, $p < 0.05$, SpearmanRank test: $\rho \sim 0.3$, $p < 0.01$, Fisher's Exact Test: $p < 0.01$). The weak correlation suggests a limited predictive power of the model, but note that the best parameter range is consistent with the earlier results in this study.

Our gene expression model is based on the transcriptional effect of the sequence-specific binding of RNAP. Other regulatory mechanisms are also influenced by promoter mutations, such as post-transcriptional regulation by ribosomal binding. This can be a reason why the correlation between the model predictions and the observed YFP levels is weak. In order to address this, we use the Salis ribosomal binding site calculator[14] to estimate the translation rates of the promoter constructs (Salis et al., 2009; Salis, 2011). Fig. 3.9 shows the comparison of these estimated translation rates with the observed YFP levels. Although the correlation is still weak (and even not significant in a SpearmanRank test or a Fisher's Exact test, i.e. $p > 0.05$), we observe that some mutations which cannot be explained

[14] https://salislab.net/

in our main model (e.g. -31C$>$T) do have high estimates of translation rate. This suggests to us that a better predictive power can be obtained by combining the transcriptional and translational effects. For that, we expand our model by considering a simple differential equation for the cellular dynamics of the protein expression level r as a function of the regulatory sequence $\boldsymbol{\sigma}$, i.e.

$$\frac{d}{dt}r(\boldsymbol{\sigma}) = \Big(\gamma_p\, \tau(\boldsymbol{\sigma})g(\boldsymbol{\sigma}) - \gamma_{pdeg}\, r(\boldsymbol{\sigma})\Big). \tag{3.14}$$

where $\tau(\boldsymbol{\sigma})$ is the sequence dependent translation rate estimated by the Salis's ribosomal binding site calculator (Salis et al., 2009; Salis, 2011), and γ_p and γ_{pdeg} are the rate constants. Recall that g is the mRNA expression level whose steady state expressed in Eq. (3.6). In steady-state, the protein expression level can be expressed as

$$r(\boldsymbol{\sigma}) \propto \Big(\tau(\boldsymbol{\sigma})\, g(\boldsymbol{\sigma})\Big) \tag{3.15}$$

In other words, the multiplication of the transcription rate estimated by our main model with the translation rate estimated by the Salis ribosomal binding site calculator becomes the new model prediction for the protein expression level. Fig. 3.10 shows how the correlation statistics scales with the chemical potential parameter μ, and gives an example plot for the model versus observation. Importantly, the Pearson test ($R^2 \sim 0.4 - 0.6$, $p \sim 10^{-3} - 10^{-7}$) indicates a significant improvement for a linear correlation in a similar parameter range (i.e. $\mu \sim (2-4)\,\epsilon\; k_BT$). However, the SpearmanRank and Fisher's Exact test indicate an opposite trend ($\rho \sim 0.25$, $p > 0.01$ and $p > 0.05$, respectively).

Lastly, we checked some of our modelling assumptions to test for a better correlation (results not shown). We did not observe any improvement by using a null spacer cost (i.e. $c(\boldsymbol{l}) = 0$) or a mismatch assumption (i.e. all non-zero elements in the energy matrix is replaced with the average energy ϵ). Positional weights of the RNAP binding in promoter region (see the Models & Methods, Eq. (3.7)) did not improve the model's prediction, either. We also checked whether the binding of the -10 box alone can improve our predictive power as an alternative mechanism, but did not observe any significant signature.

3.5 Discussion

In this study, we aimed to understand the evolutionary and biophysical characteristics of bacterial RNAP binding by analysing empirical data. We constructed a biophysical and population genetic framework to model a mapping from sequence to binding, gene expression, and fitness. The key assumptions are that a thermodynamic equilibrium determines binding probability; additive energetic contributions at the binding interface (Kinney et al., 2010) and the spacer length of RNAP (Weindl et al., 2007) determine the binding energy of RNAP; gene expression is the sum of RNAP binding probabilities across each promoter binding position; and fitness is proportional to the gene expression.

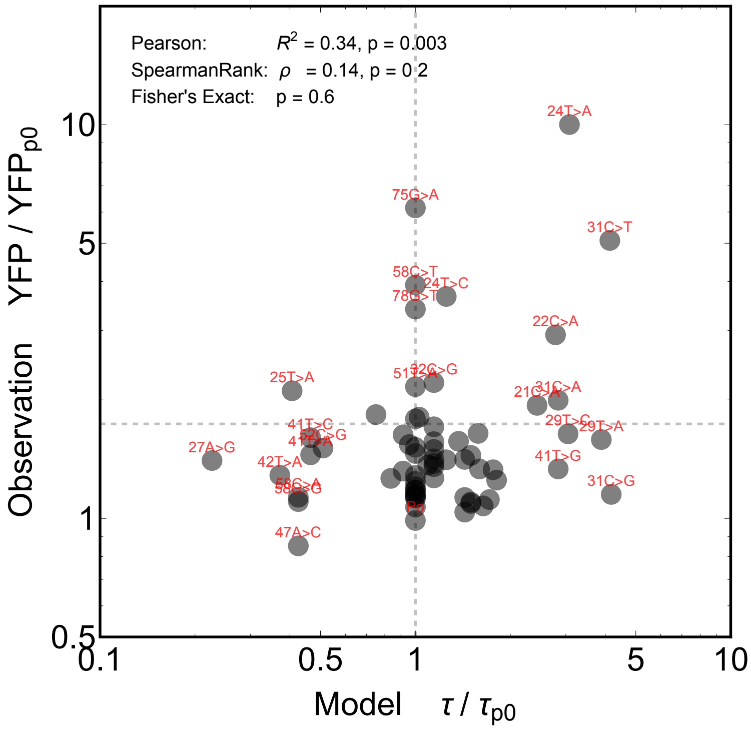


Figure 3.9: **Comparison of Salis's model predictions for translation rate and the YFP measurements.** We use the Salis ribosomal binding site calculator (Salis et al., 2009; Salis, 2011) to estimate the translation rates of the promoter constructs which can be used as a proxy to the protein expression levels. For convenience, the model and the observations are scaled so that the values of the promoter sequence *p0* are 1. Those mutations with clear deviations from *p0*'s are labeled above the data points.

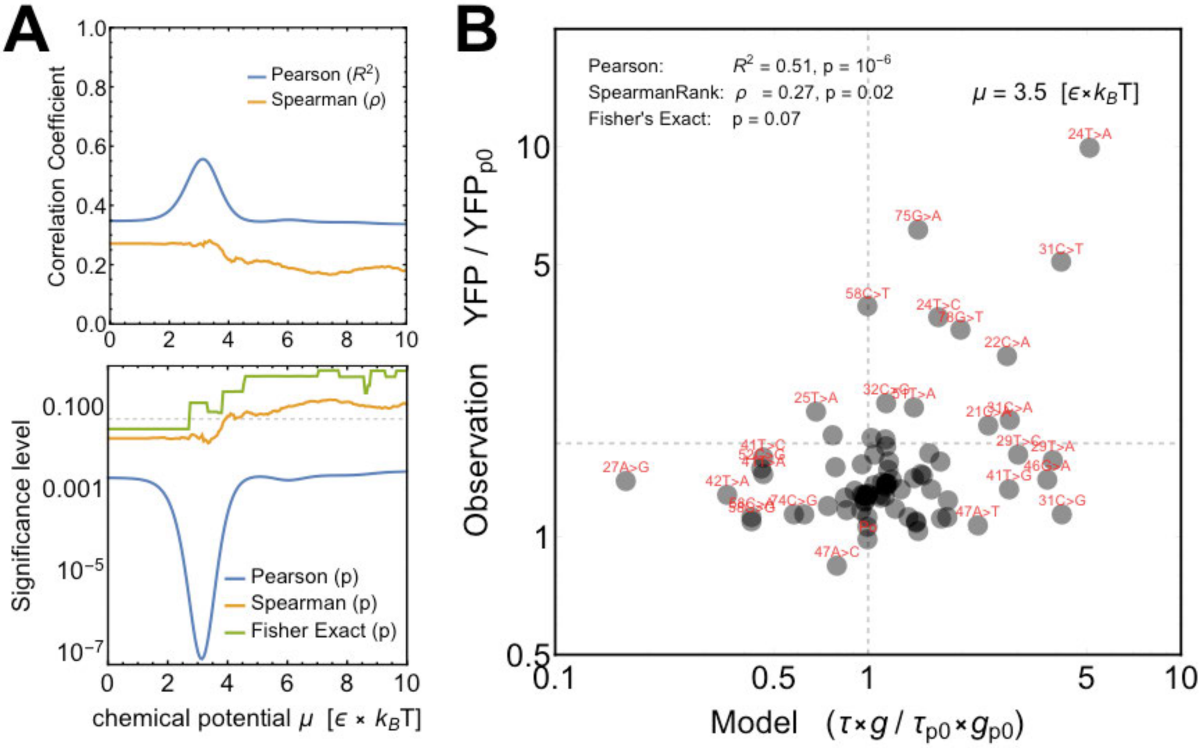


Figure 3.10: **Comparison of the combined model (our main model and Salis's model) predictions and the YFP measurements.** **A)** We show how he Pearson, SpearmanRank and Fisher's Exact test statistics scale with the chemical potential μ (**top:** the correlation coefficients, **bottom:** the significance level (p-value) **B)** We show an example plot of the model prediction versus the observed YFP level for $\mu = 3.5\,\epsilon\;k_BT$. For convenience, the model and the observations are scaled so that the values of the promoter sequence *p0* are 1. Those mutations with clear deviations from *p0*'s are labeled above the data points.

We determined the selective signatures of RNAP binding in *E.coli* K12. We used the inference method proposed by Mustonen and Lässig (2005) which is based on the comparison of binding energy distributions in evolved and neutral sequences. We assumed that randomised sequences (keeping the same nucleotide A, C, G, T content) represent neutral evolution. We estimated Ns values where we define the selection strength s as the maximum depth of the fitness landscape. We inferred a positive selection towards stronger binding ($Ns \sim +3$) for the experimentally verified RNAP binding sites (curated in RegulonDB database (Salgado et al., 2013)). For the whole genome, we inferred a weak negative selection to eliminate strong binding sites ($Ns \sim -0.5$). This is most likely a result of the long time evolution reducing off-target bindings in the genome which can be costly to the organisms. Note that our inference is slightly larger than what Hahn et al. (2003) stated for eubacteria ($Ns \sim -0.1$). This might be due to their inference method which is based on counting the consensus sequences, which might be misleading since extreme levels of binding strength are expected to inhibit transcription initiation in some cases (Knaus and Bujard, 1988). This fact can also explain why we observe a small fitness reduction from a linear fitness scaling at the lowest energy of the regulatory sequences. Although our results are not entirely surprising, the inferred strengths of Ns are interestingly moderate, especially for the regulatory sites. Such a weak selection in bacteria have been also reported in different contexts, e.g. the selection for codon bias (Bulmer, 1991; Brandis and Hughes, 2016). Our analyses are based on the assumptions of an averaging over many sites (Ns might change for different RNAP binding sites); an averaging over long time period (selection might have been strong but changing sign); a simple sequence randomising model of neutral evolution, and future studies should address these concerns.

We inferred the effective chemical potential of the RNAP, an important but unknown physical parameter, by maximising the divergence of the evolved and neutral distributions of the binding probability for experimentally verified sites. We determined that a realistic parameter range corresponds to an energy of $\sim 4 \pm 1$ mismatches from the consensus sequence. One caveat in our analysis is that the existence of activator and repressor transcription factors would change the thermodynamics and distributions of the binding probability of RNAP. Therefore, our inference method should only be seen as a crude estimation of the (effective) chemical potential of the RNAP. However, it is still interesting that an actual physical parameter can be inferred from evolved sequences, which is not a usual approach in biology (e.g. Morcos et al. (2011)). One can expect that, with ever increasing number of genome sequencing projects, population genetics and comparative genomics studies will be more in use of inferring biophysical parameters, especially with small effects which are difficult to measure in the lab in short time scales. Our inference of the chemical potential and selection should be updated in future since new developments in technology (e.g. Vvedenskaya et al. (2015)) will likely document frequent transcription events in precise genome position.

We tried to mechanistically understand the relation between the observed variation of LacZ protein activity and the *lac* promoters among 20 bacterial isolates. Mechanisms of the lac operon driving the regulatory function can be very complex with different factors, e.g. DNA looping (Kuhlman et al., 2007). Here we used a simple approach taking into account multiple RNAP binding sites in a promoter, which are widely ignored even in textbook explanations (Reznikoff, 1992). We constructed a biophysically motivated mapping from promoter sequence to gene expression, i.e. the sum of RNAP binding probabilities across the *lac* promoter. It is simple but still epistatic since the binding probabilities are calculated with a (non-linear) thermodynamical model using the chemical potential and the binding energy. We showed that the model predictions correlate well with the observed protein activity at a range of RNAP chemical potential consistent with our earlier inference in this study. We identified that a single RNAP binding position is responsible for almost all the variation in the binding probabilities. Surprisingly, this position is one of the known and experimentally verified alternative RNAP binding sites (Xiong et al., 1991). Razo-Mejia et al. (2014) aimed at a similar project but they only considered the principal RNAP binding in their modelling, which resulted in poor correlations. We propose that they can recover the correlations by repeating their analysis with including the alternative RNAP sites. Indeed, they even realised that mutations on the principal binding region of CRP affect the gene expression but the concentration of CRP does not. These mutations most likely correspond to the alternative RNAP binding site overlapping with the CRP binding site. Our findings suggest that the *lac* promoter evolution influencing the regulation of the LacZ expression in the wild have been carried out by tuning the strength of this alternative RNAP binding site, instead of the principal (canonical) site. The reason for such evolutionary mode of action is a topic of ongoing research. But it is tempting to speculate that such “shadow” or secondary sites are more prone to evolve since they might have the capacity to tune gene regulation in small amounts rather than big effects caused by the principal sites.

Our quantitative modelling needs to be further tested. One idea is to check the lacZ activity when a SNP in an isolate is genetically engineered to the *E.coli* K12 variant (see the predictions in Fig. 3.11). We also plan to design the constructs where only the *lac* promoter regions show variation, in order to test our modelling approach more directly.

In this study, we also aimed at a predictive and quantitative understanding of *de novo* promoter evolution. We combined our simple genotype-phenotype mapping with a population genetic model to lay out the probability of observing all possible single point mutations. We used this framework to interpret the data from an evolution experiment where an antibiotic gene preceded by an initially random (non-functional) promoter region is selected for higher gene expression. We showed that our model explains the evolutionary data many orders of magnitude better than a neutral evolution model in a range of the chemical potential of RNAP that is consistent with our earlier inference. This suggests that our modelling assumptions based on multiple RNAP binding capture some parts of the actual mechanisms for *de novo* promoter evolution. However, the model does not explain all characteristics of the distribution of these observed evolutionary data (e.g. why we see repetitions of certain mutations). For a better testing our gene expression model, we created random single point mutations on the same initial promoter, and

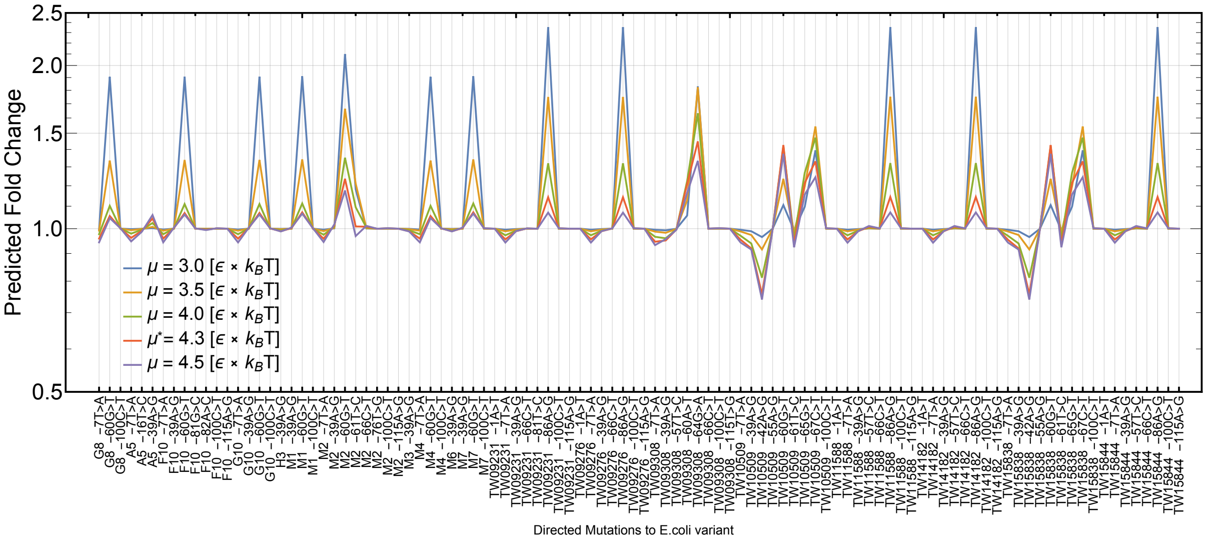


Figure 3.11: **Further predictions for the fold change of the lacZ activity levels under directed mutations.** Our modelling approach allows us to estimate the effect of genetic engineering in the *lac* promoters for the lacZ activity level. We show here the predicted fold-change for further planned directed point mutation experiments.

measured the effect on gene expression. The model possesses some correlation with the observed gene expression in the consistent parameter range, but currently it cannot be taken as a very convincing mechanistic explanation for a mapping from non-functional promoter to gene expression. In future, the sample sizes can be increased for a better evaluation (we had 14 evolutionary data points; 76 synthetic data points for 1 random promoter construct), One concern about our modelling is that we use an energy matrix of RNAP which have been inferred from the mutational effects around a functional promoter (Kinney et al., 2010). Future studies should aim to describe the sequence-specificity of RNAP far away from the consensus sequence. It is also likely that our gene expression model is too simplistic and should take into account other regulatory mechanisms. We obtained some preliminary results by combining our gene expression model with a model to estimate translation rates (Salis et al., 2009; Salis, 2011), indicating a statistically significant improvement, but further work should be followed for this and other regulatory factors.

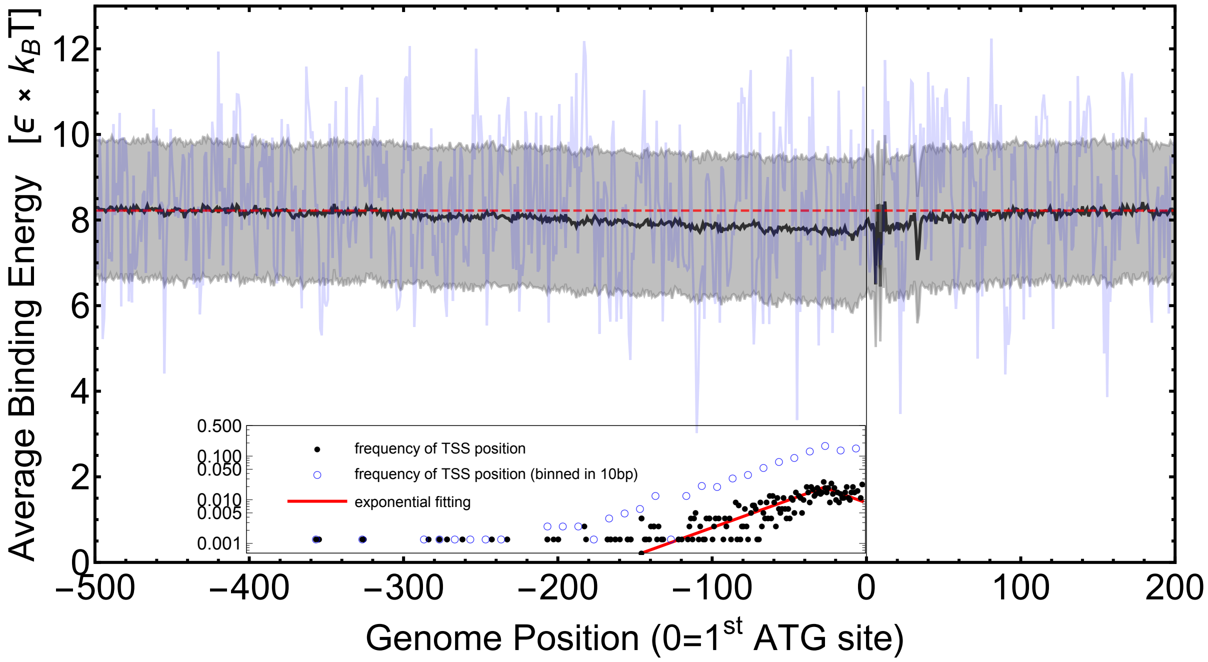


Figure 3.12: **Supplementary Information: The binding energy profiles, and the positions of the experimentally verified transcription start sites for the Sigma70 dependent RNA polymerase in E.coli.** 735 regulatory regions which includes the reported experimentally verified of transcription start sites (TSSs) in the RegulonDB database (Salgado et al., 2013) are aligned with respect to the first codon (AUG) site. The binding energies are assayed by using a sliding window approach (see the Models & Methods). We show the average energy (black curve) and 2 SEM (gray shading) profiles in comparison with the average binding energy expectation for randomised sequence (dashed red line). **Inset:** The frequency distribution of the distance between the experimentally verified TSSs and the first codon (AUG) site is shown in black dots. It exhibits a scattered and long-tailed exponential-like distribution with the optimal distance as 26 bp. Most of the TSSs fall into a typical promoter length of 200 bps, but there are also larger distances with the extremest as 353 bp. An exponential fit ($\propto e^{-0.03\,|x-26|}$) is shown in red. The blue circle shows the binning with 10 bp.

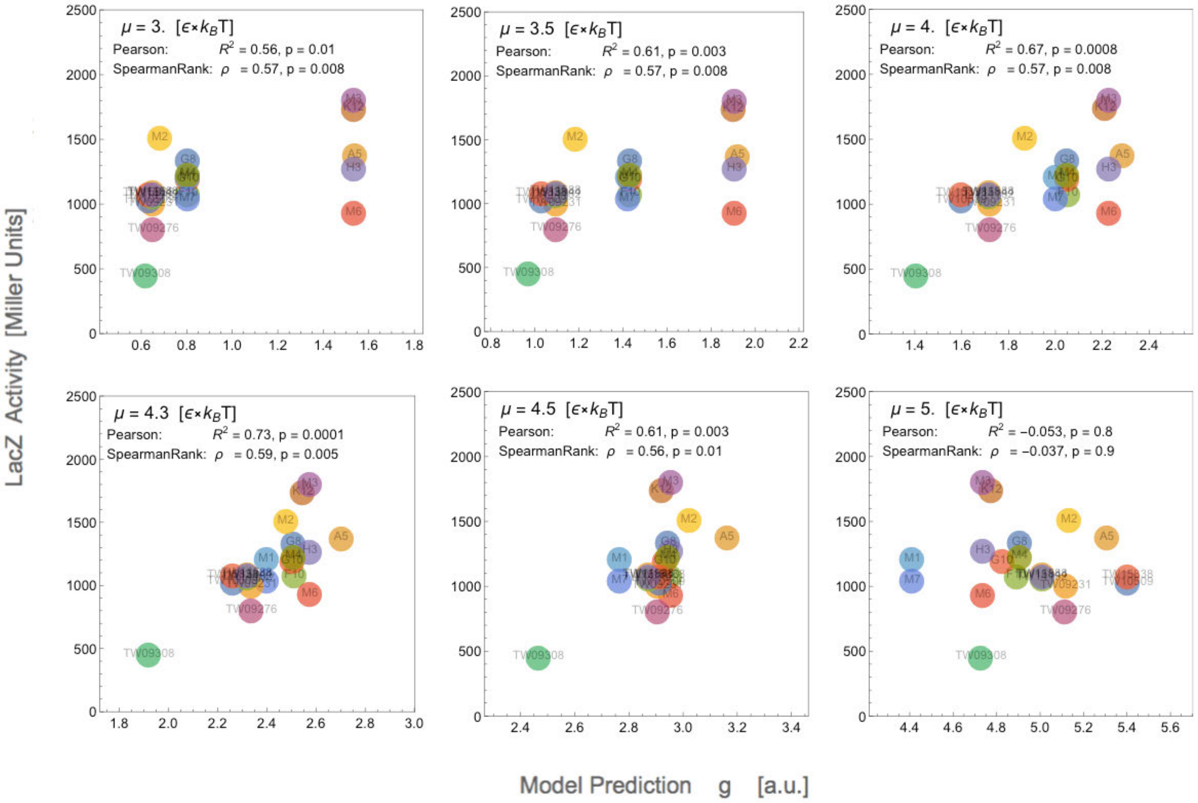
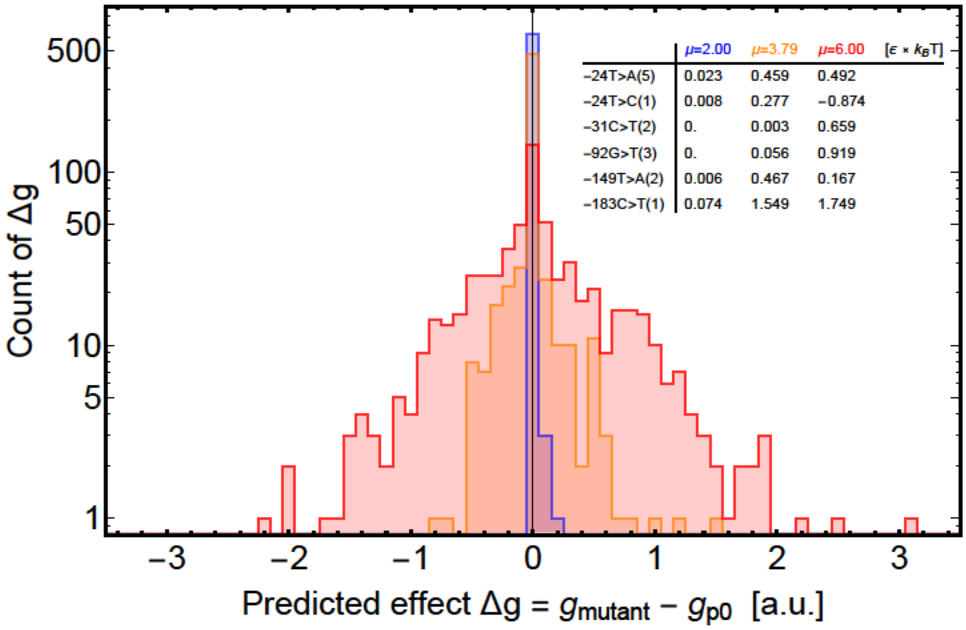


Figure 3.13: **Supplementary Information: the model prediction versus the observed LacZ activity.** As a supplementary to Fig. 3.3, we show the model prediction versus the observed lacZ activity for different chemical potential parameters.



	μ=2.00	μ=3.79	μ=6.00	[$\epsilon \times k_B T$]
−24T>A(5)	0.023	0.459	0.492	
−24T>C(1)	0.008	0.277	−0.874	
−31C>T(2)	0.	0.003	0.659	
−92G>T(3)	0.	0.056	0.919	
−149T>A(2)	0.006	0.467	0.167	
−183C>T(1)	0.074	1.549	1.749	

Figure 3.14: **Supplementary Information: Histogram of the predicted change in the gene expression level.** Using Eq. (3.6), we calculated the gene expression level of each point mutation in the promoter sequence *p0*. We obtained the count distribution of the predicted effect on gene expression using bins of width 0.1. The different coloured distributions refer to different chemical potential μ indicated in the table inset, together with the predicted values for the adaptive point mutations.

4

Conclusions

Evolution of gene regulation is important for phenotypic differences between species, populations and individuals. Sequence-specific binding of regulatory proteins is a key regulatory mechanism determining transcription of gene expression and hence heritable phenotypic variation. In this thesis, I aimed at a better understanding of evolution of the transcriptional regulatory sequences. I used a biophysical (i.e. thermodynamic) model to map from regulatory binding sequence to gene expression and fitness. This was combined with a population genetic (i.e. mutation, selection, and genetic drift) model to comprehend the evolutionary characteristics of the regulatory binding sequence. I obtained a number of conclusions which are summarised below, followed by the anticipated directions of a further research.

The first part of the thesis was devoted to the theoretical understanding of the evolutionary dynamics of transcriptional regulatory sequences. I estimated the rates of gain and loss of regulatory binding sequences in finite populations under both point and insertion/deletion mutations. If selection is not very strong, these rates are typically slow for a single regulatory binding site in an isolated DNA region. Clearly, there are also biophysical constraints on these rates. They decrease with increasing specificity of protein-DNA interactions, or with increasing binding length, making the evolution of sites longer than ~ 10 bp unlikely on typical eukaryotic speciation timescales. Similarly, evolution converges to the stationary distribution of binding sequences very slowly. Therefore, the equilibrium assumption should be used in theory and applications with caution. We identified some factors that can facilitate gain of regulatory binding sites and reconcile theoretical calculations with timescales inferred from comparative genomics. These are the availability of longer regulatory sequences in which multiple binding sites can evolve simultaneously, the presence of “pre-sites” which can be caused by partially decayed ancient (old) sites in the initial sequence, and biophysical cooperativity between transcription factors.

The second part of the thesis is reserved for application to understand the evolutionary and biophysical characteristics of sequence-specific binding of bacterial RNA polymerase (RNAP) from empirical data. First, we inferred selection acting on the binding sequences of RNAP by analysing the genome of *E.coli* K12 and using population genetic theory. As expected, there is an intermediate level of positive selection towards the lower binding energies at the experimentally verified transcription start sites. The selective signatures differ in the non-regulatory parts, suggesting a weak negative selection to remove strong binding sequences which is likely costly for the cell by reducing the RNAP concentration. We also inferred that the chemical potential of RNAP from the evolved sequences, and it corresponds to an energy value of $\sim 4 \pm 1$ mismatches from the consensus sequence. Furthermore, we tried to understand the relation of the observed variations in the lac promoter sequence and in the LacZ activity among 20 bacterial isolates by constructing a simple but biophysically motivated gene expression model including multiple RNAP bindings in promoter regions. Our model correlates well with the observed gene expression in a parameter range of the chemical potential of RNAP which is consistent with our inference. This mechanistic model indicates that the variation in protein activity is mediated by an alternative (non-principal) but experimentally verified RNAP binding in the lac promoters. Lastly, we laid out the statistical framework for a predictive and quantitative approach to *de novo* promoter evolution in *E.coli* K12 in an experimental setup where an initially random and non-functional promoter preceding to an antibiotic resistance gene is selected for higher gene expression. We showed that our model predicts the adapted point mutations many orders of magnitudes better than a neutral model in a parameter range consistent with our inference of the chemical potential and our expectation of a strong selection applied in the experiment. However, the full distribution of the observed mutations does not match our expectation from the model. This is partially also the case in our more direct experiment with random mutagenesis of the promoter and measurement of the gene expression, although a weak correlation still exists at the consistent range of the chemical potential. We conclude that our modelling has limited capacity to understand *de novo* evolution of a promoter from evolution of RNAP binding. This can be due partly to the energy matrix of RNAP not being a good model for sequences far away from the consensus sequence. Certainly, further improvements of the mapping from promoter sequence to gene expression level is necessary.

Future studies can directly follow from the content of this thesis, especially as presented in the first part. First of all, existing data from comparative genomic studies on TF binding sequences in enhancers and promoters can be reanalysed to check whether an enrichment of sequence classes near and at a pre-site is observed, as expected from our theoretical calculations with ancient site hypothesis or from other considerations such as mobile elements carrying near functional sites. Furthermore, the modelling framework in this thesis can be extended to investigate coevolution of binding length, binding specificity and promoter/enhancer sequence, in order to see whether a coevolution of protein and DNA sequence can accelerate the evolutionary rates. Such an extended modelling framework can also help understand evolvability principles of regulatory proteins such as the observed trade-off between binding length and binding specificity; and more profoundly, why we see different strategies of regulatory architecture and mechanisms in eukaryotes and prokaryotes.

A number of issues were not addressed in this thesis, which can be considered in the future. First of all, the modelling throughout this thesis was limited to directional selection towards higher or lower gene expression to simplify the mathematics and to focus on the evolution from non-functional to functional binding sequences (and vice versa). However, gene expression under stabilising selection would also be an interesting scenario from a theoretical perspective, but certainly also a realistic case to model enhancers and promoters of developmental genes. One can think of calculating the turn-over rates in evolution from a functional sequence to another functional sequence with population genetic considerations of crossing fitness valleys. The results should be compared with Chapter 2 to see whether the evolutionary rates are accelerated. Such a mathematical treatment can be applied also to obtain the expected rewiring rates in regulatory interactions defined by binding sequences, as a first step to understand the evolutionary dynamics of gene regulatory networks. Overall, it will be necessary in future research to combine the transcriptional models and analyses of this thesis with other regulatory mechanisms, such as post-transcriptional regulation or epigenetic control, in order to improve our understanding of the evolution of gene regulation.

Bibliography

Alekseyenko, A. A., Ellison, C. E., Gorchakov, A. A., Zhou, Q., Kaiser, V. B., Toda, N., Walton, Z., Peng, S., Park, P. J., Bachtrog, D., Kuroda, M. I., Apr. 2013. Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in Drosophila. Genes & Development 27 (8), 853–858.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C., Stark, A., Jul. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nature Genetics 46 (7), 685–692.

Barton, N., Coe, J., Jul. 2009. On the application of statistical physics to evolutionary biology. Journal of Theoretical Biology 259 (2), 317–324.

Barton, N. H., Keightley, P. D., Jan. 2002. Understanding quantitative genetic variation. Nature Reviews. Genetics 3 (1), 11–21.

Beckwith, J., May 2011. The Operon as Paradigm: Normal Science and the Beginning of Biological Complexity. Journal of Molecular Biology 409 (1), 7–13.

Berg, J., Willmann, S., Lässig, M., Oct. 2004. Adaptive evolution of transcription factor binding sites. BMC Evolutionary Biology 4 (1), 42.

Berg, O. G., von Hippel, P. H., Feb. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of molecular biology 193 (4), 723–750.

Bialek, W. S., 2012. Biophysics: Searching for Principles. Princeton University Press.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Phillips, R., 2005a. Transcriptional regulation by the numbers: applications. Current Opinion in Genetics & Development 15, 125–135.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Phillips, R., 2005b. Transcriptional regulation by the numbers: models. Current Opinion in Genetics & Development 15, 116–124.

Blank, D., Wolf, L., Ackermann, M., Silander, O. K., Feb. 2014. The predictability of molecular evolution during functional innovation. Proceedings of the National Academy of Sciences 111 (8), 3044–3049.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., Shao, Y., Sep. 1997. The Complete Genome Sequence of Escherichia coli K-12. Science 277 (5331), 1453–1462.

Brandis, G., Hughes, D., Mar. 2016. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. PLOS Genet 12 (3), e1005926.

Brandström, M., Ellegren, H., Jul. 2007. The Genomic Landscape of Short Insertion and Deletion Polymorphisms in the Chicken (Gallus gallus) Genome: A High Frequency of Deletions in Tandem Duplicates. Genetics 176 (3), 1691–1701.

Brewster, R. C., Jones, D. L., Phillips, R., Dec. 2012. Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. PLoS Computational Biology 8 (12).

Britten, R. J., Davidson, E. H., Jun. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. The Quarterly Review of Biology 46 (2), 111–138.

Bulmer, M., Nov. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129 (3), 897–907.

Cartwright, R. A., Feb. 2009. Problems and Solutions for Estimating Indel Rates and Length Distributions. Molecular Biology and Evolution 26 (2), 473–480.

Cepeda-Humerez, S. A., Rieckh, G., Tkačik, G., Apr. 2015. Stochastic proofreading mechanism alleviates crosstalk in transcriptional regulation. arXiv:1504.05716 [q-bio]ArXiv: 1504.05716.

Chait, R., Shrestha, S., Shah, A. K., Michel, J.-B., Kishony, R., Dec. 2010. A Differential Drug Screen for Compounds That Select Against Antibiotic Resistance. PLOS ONE 5 (12), e15179.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jonsson, B., Schluter, D., Bell, M. A., Kingsley, D. M., Jan. 2010. Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer. Science 327 (5963), 302–305.

Charlesworth, B., 2010. Elements of Evolutionary Genetics. Roberts and Company Publishers.

Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M., Tian, D., Jul. 2009. Variation in the Ratio of Nucleotide Substitution and Indel Rates across Genomes in Mammals and Bacteria. Molecular Biology and Evolution 26 (7), 1523–1531.

Contente, A., Dittmer, A., Koch, M. C., Roth, J., Dobbelstein, M., Mar. 2002. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nature Genetics 30 (3), 315–320.

Cravioto, A., Reyes, R. E., Trujillo, F., Uribe, F., Navarro, A., Roca, J. M. D. L., Hernandez, J. M., Perez, G., Vazquez, V., May 1990. Risk of Diarrhea During the First Year of Life Associated with Initial and Subsequent Colonization by Specific Enteropathogens. American Journal of Epidemiology 131 (5), 886–904.

Crow, J. F., Kimura, M., Jan. 2009. An Introduction to Population Genetics Theory. Blackburn Press.

Darwin, C., 1859. On the Origin of Species. John Murray.

Desai, M. M., Fisher, D. S., Jul. 2007. Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection. Genetics 176 (3), 1759–1798.

Dodd, I. B., Perkins, A. J., Tsemitsidis, D., Egan, J. B., Nov. 2001. Octamerization of λ CI repressor is needed for effective repression of P RM and efficient switching from lysogeny. Genes & Development 15 (22), 3013–3022.

Dombroski, A. J., Johnson, B. D., Lonetto, M., Gross, C. A., Aug. 1996. The sigma subunit of Escherichia coli RNA polymerase senses promoter spacing. Proceedings of the National Academy of Sciences 93 (17), 8858–8862.

Doniger, S. W., Fay, J. C., May 2007. Frequent Gain and Loss of Functional Transcription Factor Binding Sites. PLoS Comput Biol 3 (5), e99.

Dowell, R. D., Nov. 2010. Transcription factor binding variation in the evolution of gene regulation. Trends in Genetics 26 (11), 468–475.

Duque, T., Samee, M. A. H., Kazemian, M., Pham, H. N., Brodsky, M. H., Sinha, S., Oct. 2013. Simulations of Enhancer Evolution Provide Mechanistic Insights into Gene Regulation. Molecular Biology and Evolution 31 (1), 184–200.

Duque, T., Sinha, S., Jun. 2015. What Does It Take to Evolve an Enhancer? A Simulation-Based Study of Factors Influencing the Emergence of Combinatorial Regulation. Genome Biology and Evolution 7 (6), 1415–1431.

Ellison, C. E., Bachtrog, D., Nov. 2013. Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science 342 (6160), 846–850.

Ewens, W. J., Oct. 2012. Mathematical Population Genetics 1: Theoretical Introduction. Springer Science & Business Media.

Fay, J. C., Wittkopp, P. J., 2007. Evaluating the role of natural selection in the evolution of gene regulation. Heredity 100, 191–199.

Feschotte, C. e., May 2008. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9 (5), 397–405.

Fields, D. S., He, Y.-y., Al-Uzri, A. Y., Stormo, G. D., Aug. 1997. Quantitative specificity of the Mnt repressor 1. Journal of Molecular Biology 271 (2), 178–194.

Gemayel, R., Vinces, M. D., Legendre, M., Verstrepen, K. J., 2010. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. Annual Review of Genetics 44 (1), 445–477.

Gerland, U., Hwa, T., Oct. 2002. On the selection and evolution of regulatory DNA motifs. Journal of Molecular Evolution 55 (4), 386–400.

Gerland, U., Moroz, J. D., Hwa, T., Sep. 2002. Physical constraints and functional characteristics of transcription factor-DNA interaction. Proceedings of the National Academy of Sciences of the United States of America 99 (19), 12015–12020.

Gillespie, J. H., Dec. 2010. Population Genetics: A Concise Guide. JHU Press.

Giorgetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., Pasparakis, M., Milani, P., Bulyk, M. L., Natoli, G., Feb. 2010. Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. Molecular Cell 37 (3), 418–428.

Grigorova, I. L., Phleger, N. J., Mutalik, V. K., Gross, C. A., Apr. 2006. Insights into transcriptional regulation and σ competition from an equilibrium model of RNA polymerase binding to DNA. Proceedings of the National Academy of Sciences 103 (14), 5332–5337.

Gross, C. A., Chan, C., Dombroski, A., Gruber, T., Sharp, M., Tupy, J., Young, B., Jan. 1998. The Functional and Regulatory Roles of Sigma Factors in Transcription. Cold Spring Harbor Symposia on Quantitative Biology 63, 141–156.

Hahn, M. W., Stajich, J. E., Wray, G. A., Jun. 2003. The Effects of Selection Against Spurious Transcription Factor Binding Sites. Molecular Biology and Evolution 20 (6), 901–906.

Haldane, A., Manhart, M., Morozov, A. V., Jul. 2014. Biophysical Fitness Landscapes for Transcription Factor Binding Sites. PLoS Comput Biol 10 (7), e1003683.

Hammar, P., Walldén, M., Fange, D., Persson, F., Baltekin, O., Ullman, G., Leroy, P., Elf, J., Apr. 2014. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nature Genetics 46 (4), 405–408.

Harley, C. B., Reynolds, R. P., Mar. 1987. Analysis of E.Coli Pormoter sequences. Nucleic Acids Research 15 (5), 2343–2361.

Hawley, D. K., McClure, W. R., Apr. 1983. Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Research 11 (8), 2237–2255.

He, B. Z., Holloway, A. K., Maerkl, S. J., Kreitman, M., Apr. 2011. Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genet 7 (4), e1002053.

He, X., Duque, T. S., Sinha, S., Mar. 2012. Evolutionary Origins of Transcription Factor Binding Site Clusters. Molecular Biology and Evolution 29 (3), 1059–1070.

He, X., Samee, A. H., Blatti, C., Sinha, S., 2010. Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. PLOS Computational Biology.

Hermsen, R., Tans, S., ten Wolde, P. R., Dec. 2006. Transcriptional Regulation by Competeing Transcription Factor Modules. PLoS Comput Biol 2 (12), e164.

Hermsen, R., Ursem, B., ten Wolde, P. R., Jun. 2010. Combinatorial Gene Regulation Using Auto-Regulation. PLoS Comput Biol 6 (6), e1000813.

Hoekstra, H. E., Coyne, J. A., May 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution; International Journal of Organic Evolution 61 (5), 995–1016.

Hook-Barnard, I., Johnson, X. B., Hinton, D. M., Dec. 2006. Escherichia coli RNA Polymerase Recognition of a sigma70-Dependent Promoter Requiring a -35 DNA Element and an Extended -10 TGn Motif. Journal of Bacteriology 188 (24), 8352–8359.

Huerta, A. M., Collado-Vides, J., Oct. 2003. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. Journal of Molecular Biology 333 (2), 261–278.

Ishii, S., Ksoll, W. B., Hicks, R. E., Sadowsky, M. J., Jan. 2006. Presence and Growth of Naturalized Escherichia coli in Temperate Soils from Lake Superior Watersheds. Applied and Environmental Microbiology 72 (1), 612–621.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., Snyder, M., Apr. 2010. Variation in Transcription Factor Binding Among Humans. Science 328 (5975), 232–235.

Keightley, P. D., Johnson, T., Mar. 2004. MCALIGN: Stochastic Alignment of Noncoding DNA Sequences Based on an Evolutionary Model of Sequence Evolution. Genome Research 14 (3), 442–450.

Kim, D., Hong, J. S.-J., Qiu, Y., Nagarajan, H., Seo, J.-H., Cho, B.-K., Tsai, S.-F., Palsson, B., Aug. 2012. Comparative Analysis of Regulatory Elements between Escherichia coli and Klebsiella pneumoniae by Genome-Wide Transcription Start Site Profiling. PLoS Genet 8 (8), e1002867.

Kimura, M., Jun. 1962. On the Probability of Fixation of Mutant Genes in a Population. Genetics 47 (6), 713–719.

Kimura, M., Dec. 1964. Diffusion Models in Population Genetics. Journal of Applied Probability 1 (2), 177.

Kinney, J. B., Murugan, A., Callan, C. G., Cox, E. C., May 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences 107 (20), 9158–9163.

Knaus, R., Bujard, H., Sep. 1988. PL of coliphage lambda: an alternative solution for an efficient promoter. The EMBO Journal 7 (9), 2919–2923.

Kuhlman, T., Zhang, Z., Saier, M. H., Hwa, T., Apr. 2007. Combinatorial transcriptional control of the lactose operon of Escherichia coli. Proceedings of the National Academy of Sciences 104 (14), 6043–6048.

Lee, H., Popodi, E., Tang, H., Foster, P. L., Oct. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proceedings of the National Academy of Sciences 109 (41), E2774–E2783.

Ludwig, M. Z., Patel, N. H., Kreitman, M., 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development, 949–958.

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., Konstantinidis, K. T., Apr. 2011. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proceedings of the National Academy of Sciences 108 (17), 7200–7205.

Lynch, M., Conery, J. S., Nov. 2003. The Origins of Genome Complexity. Science 302 (5649), 1401–1404.

Lynch, M., Hagner, K., Jan. 2015. Evolutionary meandering of intermolecular interactions along the drift barrier. Proceedings of the National Academy of Sciences 112 (1), E30–E38.

MacArthur, S., Brookfield, J. F. Y., Jun. 2004. Expected Rates and Modes of Evolution of Enhancer Sequences. Molecular Biology and Evolution 21 (6), 1064–1073.

Maerkl, S. J., Quake, S. R., Jan. 2007. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science 315 (5809), 233–237.

Manhart, M., Haldane, A., Morozov, A. V., Aug. 2012. A universal scaling law determines time reversibility and steady state of substitutions under selection. Theoretical Population Biology 82 (1), 66–76.

McKeown, A. N., Bridgham, J. T., Anderson, D. W., Murphy, M. N., Ortlund, E. A., Thornton, J. W., Sep. 2014. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. Cell 159 (1), 58–68.

Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A. M., Collado-Vides, J., Morett, E., Oct. 2009. Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. PLoS ONE 4 (10), e7526.

Mirny, L. A., Dec. 2010. Nucleosome-mediated cooperativity between transcription factors. Proceedings of the National Academy of Sciences 107 (52), 22534–22539.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., Weigt, M., Dec. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108 (49), E1293–E1301.

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., Eisen, M. B., Oct. 2006. Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Comput Biol 2 (10), e130.

Mulligan, M. E., Brosius, J., McClure, W. R., Mar. 1985. Characterization in vitro of the effect of spacer length on the activity of Escherichia coli RNA polymerase at the TAC promoter. Journal of Biological Chemistry 260 (6), 3529–3538.

Murakami, K. S., 2015. Structural biology of bacterial RNA polymerase. Biomolecules 5 (2), 848–864.

Mustonen, V., Kinney, J., Callan, C. G., Lässig, M., Aug. 2008. Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences of the United States of America 105 (34), 12376–12381.

Mustonen, V., Lässig, M., Nov. 2005. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. Proceedings of the National Academy of Sciences of the United States of America 102 (44), 15936–15941.

Nourmohammad, A., Lässig, M., Oct. 2011. Formation of Regulatory Modules by Local Sequence Duplication. PLoS Comput Biol 7 (10), e1002167.

Otto, S. P., Day, T., 2007. A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press.

Paget, M. S., Helmann, J. D., Jan. 2003. The σ70 family of sigma factors. Genome Biology 4 (1), 203.

Paixao, T., Pérez-Heredia, J., Sudholt, D., Trubenová, B., 2015. First Steps Towards a Runtime Comparison of Natural and Artificial Evolution. In: Genetic and Evolutionary Computation Conference (GECCO 2015). ACM, pp. 1455–1462.

Paris, M., Kaplan, T., Li, X. Y., Villalta, J. E., Lott, S. E., Eisen, M. B., Sep. 2013. Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression. PLoS Genet 9 (9), e1003748.

Park, L., May 2015. Ancestral Alleles in the Human Genome Based on Population Sequencing Data. PLoS ONE 10 (5), e0128186.

Payne, J. L., Wagner, A., Feb. 2014. The Robustness and Evolvability of Transcription Factor Binding Sites. Science 343 (6173), 875–877.

Phillips, R., Kondev, J., Theriot, J., Garcia, H., Oct. 2012. Physical Biology of the Cell, Second Edition. Garland Science.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S.-D., True, J. R., Carroll, S. B., Apr. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440 (7087), 1050–1053.

Rajon, E., Masel, J., Jan. 2013. Compensatory Evolution and the Origins of Innovations. Genetics 193 (4), 1209–1220.

Razo-Mejia, M., Boedicker, J. Q., Jones, D., DeLuna, A., Kinney, J. B., Phillips, R., Apr. 2014. Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. Physical Biology 11 (2), 026005.

Reznikoff, W. S., Sep. 1992. The lactose operon-controlling elements: a complex paradigm. Molecular Microbiology 6 (17), 2419–2422.

Rice, S. H., 2004. Evolutionary Theory: Mathematical and Conceptual Foundations. Sinauer Associates.

Romero, I. G., Ruvinsky, I., Gilad, Y., Jul. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics 13 (7), 505–516.

Salgado, H., et al., Jan. 2013. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic acids research 41 (Database issue), D203–213.

Salis, H. M., 2011. Chapter two - The Ribosome Binding Site Calculator. In: Voigt, C. (Ed.), Methods in Enzymology. Vol. 498 of Synthetic Biology, Part BComputer Aided Design and DNA Assembly. Academic Press, pp. 19–42.

Salis, H. M., Mirsky, E. A., Voigt, C. A., Oct. 2009. Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology 27 (10), 946–950.

Samee, M. A. H., Sinha, S., Mar. 2014. Quantitative Modeling of a Gene's Expression from Its Intergenic Sequence. PLoS Comput Biol 10 (3), e1003467.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., Odom, D. T., May 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science 328 (5981), 1036–1040.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U., Jan. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451 (7178), 535–540.

Segal, E., Widom, J., Jul. 2009. From DNA sequence to transcriptional behaviour: a quantitative approach. Nature Reviews Genetics 10 (7), 443–456.

Sella, G., Hirsh, A. E., 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A 102, 9541–9546.

Shea, M. A., Ackers, G. K., 1984. The OR Control system of bacteriophage lambda: A physical-chemical model for gene regulation. Journal of Molecular Biology, 211–230.

Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D. J., Talianidis, I., Marioni, J. C., Flicek, P., Odom, D. T., Aug. 2013. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell 154 (3), 530–540.

Stewart, A. J., Plotkin, J. B., Nov. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192 (3), 973–985.

Stewart, A. J., Plotkin, J. B., Oct. 2013. The evolution of complex gene regulation by low-specificity binding sites. Proceedings of the Royal Society B: Biological Sciences 280 (1768).

Stone, J. R., Wray, G. A., Sep. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Molecular Biology and Evolution 18 (9), 1764–1770.

Stormo, G. D., Fields, D. S., Mar. 1998. Specificity, free energy and information content in protein-DNA interactions. Trends in biochemical sciences 23 (3), 109–113.

Stormo, G. D., Hartzell, G. W., Feb. 1989. Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences 86 (4), 1183–1187.

Stormo, G. D., Zhao, Y., Nov. 2010. Determining the specificity of protein-DNA interactions. Nature Reviews Genetics 11 (11), 751–760.

Taylor, M. S., Ponting, C. P., Copley, R. R., Apr. 2004. Occurrence and Consequences of Coding Sequence Insertions and Deletions in Mammalian Genomes. Genome Research 14 (4), 555–566.

Tuğrul, M., Paixão, T., Barton, N. H., Tkačik, G., Nov. 2015. Dynamics of Transcription Factor Binding Site Evolution. PLOS Genet 11 (11), e1005639.

Venter, J. C., et al., Feb. 2001. The Sequence of the Human Genome. Science 291 (5507), 1304–1351.

Vierstra, J., et al., Nov. 2014. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346 (6212), 1007–1012.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F., Murchison, E. P., Flicek, P., Odom, D. T., Jan. 2015. Enhancer Evolution across 20 Mammalian Species. Cell 160 (3), 554–566.

Villar, D., Flicek, P., Odom, D. T., Apr. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nature Reviews Genetics 15 (4), 221–233.

Vogel, C., Marcotte, E. M., Apr. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics 13 (4), 227–232.

von Hippel, P. H., Berg, O. G., Mar. 1986. On the specificity of DNA-protein interactions. Proceedings of the National Academy of Sciences of the United States of America 83 (6), 1608–1612.

Vvedenskaya, I., Zhang, Y., Goldman, S., Valenti, A., Visone, V., Taylor, D., Ebright, R., Nickels, B., Dec. 2015. Massively Systematic Transcript End Readout, “MASTER”: Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. Molecular Cell.

Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., Whittam, T. S., Oct. 2009. Cryptic Lineages of the Genus Escherichia. Applied and Environmental Microbiology 75 (20), 6534–6544.

Weindl, J., Hanus, P., Dawy, Z., Zech, J., Hagenauer, J., Mueller, J. C., Nov. 2007. Modeling DNA-binding of Escherichia coli sigma-70 exhibits a characteristic energy landscape around strong promoters. Nucleic Acids Research 35 (20), 7003–7010.

Weinert, F. M., Brewster, R. C., Rydenfelt, M., Phillips, R., Kegel, W. K., Dec. 2014. Scaling of Gene Expression with Transcription-Factor Fugacity. Physical Review Letters 113 (25), 258101.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J. M., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., Hughes, T. R., Sep. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158 (6), 1431–1443.

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., Fisher, E. M. C., Tavaré, S., Odom, D. T., Oct. 2008. Species-Specific Transcription in Mice Carrying Human Chromosome 21. Science 322 (5900), 434–438.

Wittkopp, P. J., 2013. Evolution of Gene Expression. In: The Princeton Guide to Evolution. Princeton University Press, pp. 413–419.

Wittkopp, P. J., Kalay, G., 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics, 59–69.

Wolf, L., Silander, O. K., Nimwegen, E. v., Jun. 2015. Expression noise facilitates the evolution of gene regulation. eLife 4, e05856.

Wright, S., Mar. 1931. Evolution in Mendelian Populations. Genetics 16 (2), 97–159.

Wunderlich, Z., Mirny, L. A., Oct. 2009. Different gene regulation strategies revealed by analysis of binding motifs. Trends in genetics : TIG 25 (10), 434–440.

Xiong, X. F., Cruz, N. d. l., Reznikoff, W. S., Aug. 1991. Downstream deletion analysis of the lac promoter. Journal of Bacteriology 173 (15), 4570–4577.

Yao, P., Lin, P., Gokoolparsadh, A., Assareh, A., Thang, M. W. C., Voineagu, I., Aug. 2015. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nature Neuroscience 18 (8), 1168–1174.

Zhao, Y., Granas, D., Stormo, G. D., Dec. 2009. Inferring Binding Energies from Selected Binding Sites. PLoS Comput Biol 5 (12), e1000590.

Zheng, W., Gianoulis, T. A., Karczewski, K. J., Zhao, H., Snyder, M., 2011. Regulatory Variation Within and Between Species. Annual Review of Genomics and Human Genetics 12 (1), 327–346.